



Unravelling Error Patterns in Linear Programming: A Hierarchical Bayesian Approach

Muhammad Ali Gunawan

STIK Kendal, Central Java, Indonesia

guns12380@gmail.com

Abstract

Received :

09/07/2025

Accepted :

13/10/2025

Published :

03/11/2025

Linear programming (LP) is a cornerstone of mathematical optimization, yet students frequently struggle to translate real-world problems into correct LP formulations, often producing systematic error patterns. This study investigates these cognitive errors using a hierarchical Bayesian multinomial logistic model, applied to responses from 12 undergraduate students in a mathematics education program. Errors were categorized into five types: constraint omission, objective misidentification, sign reversal, variable mislabeling, and feasibility misjudgment. Posterior estimates revealed that constraint omission and sign reversal emerged as the most dominant error patterns, while variance decomposition showed that over 40% of the variability stemmed from individual differences. These findings demonstrate the diagnostic power of Bayesian modeling in uncovering persistent misconceptions and offer practical insights for designing targeted instructional interventions in LP learning. The study is limited by its small sample size and single task context, indicating the need for replication with larger and more diverse datasets.

Keywords: Hierarchical Bayesian modeling, linear programming, cognitive errors, small sample, mathematics education

Abstrak

Linear programming (LP) merupakan salah satu pilar utama dalam optimisasi matematika, namun mahasiswa sering mengalami kesulitan dalam menerjemahkan permasalahan nyata ke dalam formulasi LP yang benar sehingga memunculkan pola kesalahan yang sistematis. Penelitian ini mengkaji kesalahan kognitif tersebut menggunakan model regresi logistik multinomial Bayesian hierarkis, yang diterapkan pada jawaban 12 mahasiswa program studi pendidikan matematika. Kesalahan dikategorikan ke dalam lima tipe: penghilangan kendala, kesalahan identifikasi fungsi objektif, pembalikan tanda, salah penamaan variabel, dan salah interpretasi daerah layak. Hasil posterior menunjukkan bahwa penghilangan kendala dan pembalikan tanda merupakan pola kesalahan dominan, sementara analisis variansi mengindikasikan lebih dari 40% variabilitas berasal dari perbedaan individu. Temuan ini menegaskan kekuatan diagnostik pendekatan Bayesian dalam mengungkap miskonsepsi yang persisten serta memberikan wawasan praktis untuk merancang intervensi instruksional yang lebih terarah dalam pembelajaran LP. Penelitian ini terbatas pada ukuran sampel kecil dan satu jenis tugas, sehingga diperlukan replikasi dengan jumlah data lebih besar dan konteks yang lebih beragam.

Kata Kunci: Model Bayesian hierarkis, linear programming, kesalahan kognitif, sampel kecil, pendidikan matematika

To cite this article (in APA style):

Gunawan, M. A. (2025). Unravelling Error Patterns in Linear Programming: A Hierarchical Bayesian Approach with Small-Sample Insights. *Delta: Jurnal Ilmiah Pendidikan Matematika*, 13(2), 101-118. <https://doi.org/10.31941/delta.v13i2.6752>

1. Introduction

Linear programming (LP) occupies a central position in mathematical optimization, with applications spanning operations research, industrial engineering, economics, and education. Despite its broad utility, students frequently encounter persistent difficulties in LP problem-solving, particularly in the early stages of instruction



(He et al., 2025; Sinha & Kapur, 2021; Wakhata et al., 2022). Prior studies have documented recurring challenges in translating real-world scenarios into formal LP models, including misidentification of objective functions (Hartmann et al., 2021; Kenney et al., 2020), misformulation of constraints (Sanjari & Manouchehri, 2024), and inappropriate specification of decision variables (Nargundkar & Kulkarni, 2024; Osaba et al., 2021).

Much of this literature, however, has relied on descriptive statistics or classical psychometric approaches that tend to obscure the latent cognitive structures underlying student performance. These difficulties are not random but manifest as error patterns, which are systematic and predictable cognitive deviations that reveal persistent misconceptions rather than incidental mistakes. In LP learning, these patterns include constraint omission, sign reversal, and variable mislabeling, each of which highlights the need for diagnostic approaches capable of capturing the structure of student reasoning.

Cognitive science research underscores that diagnosing error patterns provides critical insights into underlying conceptual misunderstandings (Adamo et al., 2021; Croskerry et al., 2023; Pelaccia et al., 2025; Rapp & Braasch, 2023). Within LP learning, these errors are not isolated slips but structured manifestations of alternative mental models that deviate from normative mathematical representations. Standard inferential techniques, however, struggle to separate such structured misconceptions from random variation, particularly when sample sizes are constrained, a common condition in classroom-based research (Gehlbach et al., 2023; Robutti, 2020). This raises a methodological dilemma: how can researchers generate valid diagnostic inferences from sparse data without compromising analytical rigor?

Advances in Bayesian educational measurement offer promising solutions. Hierarchical Bayesian multinomial models facilitate the simultaneous estimation of individual- and group-level tendencies, thus capturing variability both within and between learners (Haines et al., 2020; Li et al., 2025). The incorporation of prior knowledge and probabilistic pooling allows these models to enhance analytical robustness under small-sample conditions (Gilholm et al., 2021; Nguyen, 2023; Wu & Stephenson, 2024). Despite their potential, applications of hierarchical Bayesian frameworks in analyzing LP error patterns remain limited, particularly in authentic learning contexts.

This study seeks to address this gap by applying a hierarchical Bayesian multinomial model to examine error patterns in LP formulation among undergraduate mathematics education students. Its contribution is threefold. Theoretically, it refines the conceptualization of LP error taxonomies as structured manifestations of cognitive misconceptions. Methodologically, it demonstrates the viability of hierarchical Bayesian modeling as a diagnostic tool under small-sample constraints. Practically, it offers actionable insights for designing targeted instructional interventions that directly address the most frequent and persistent errors in LP, thereby advancing pedagogy in mathematical optimization.

2. Method

2.1. Research Design

This study employed an explanatory sequential design within a quantitative cognitive diagnostic framework. A hierarchical Bayesian modelling strategy was used to analyze categorical error data from a structured linear programming (LP) task. The study prioritized both individual-level error identification and group-level inference, particularly under the constraint of small sample size. All procedures were approved by the Institutional Review Board of Universitas Pekalongan and adhered to ethical guidelines for human subject research.

2.2. Participants

The study involved 12 undergraduate students enrolled in the Mathematics Education program at Universitas Pekalongan. Participants were purposively sampled based on their enrollment in an advanced operations research course, ensuring prior exposure to LP concepts. The sample consisted of 4 males and 8 females, with an average age of 21.3 years ($SD = 0.7$). Participation was voluntary, and informed consent was obtained from all individuals.

2.3. Instrumentation and Task Design

A structured LP formulation task was developed, involving a real-world word problem requiring the identification of decision variables, formulation of objective functions, and construction of constraints (Güner et al., 2024; Kenney et al., 2020). The task was validated by two experts in mathematics education and operations research to ensure content accuracy and cognitive demand. Participant responses were coded into five mutually exclusive error categories:

1. Constraint Omission (CO)
2. Objective Function Misidentification (OFM)
3. Sign Reversal in Constraints (SRC)
4. Variable Mislabelling (VM)
5. Feasibility Misjudgement (FMJ)

2.4. Data Analysis Strategy

A hierarchical Bayesian multinomial logistic regression model was implemented to estimate the probability of each error type at both the individual and group levels (Gerber & Craig, 2021; Kavelaars et al., 2023). The model structure is defined as:

$$\text{logit}(\pi_{ij}) = \alpha_j + \beta_j^{(group)} \quad (1)$$

where π_{ij} is the probability of participant i committing error type j , α_j represents the baseline log-odds of each error category, and $\beta_j^{(group)}$ is a random effect capturing between-subject variance.

The model was estimated using Hamiltonian Monte Carlo (HMC) implemented via the *rstan* package in R (Ng'ombe & Brorsen, 2020). A total of 4 chains were run, each with 10,000 warm-up iterations and 20,000 sampling iterations, resulting in 80,000 posterior samples. Convergence was assessed using the Gelman–Rubin diagnostic ($\hat{R} < 1.01$) and effective sample size ($n_{eff} > 1,000$).

2.5. Model Validation and Comparison

Model fit was evaluated using posterior predictive checks, deviance information criterion (DIC) (Kim, 2021; Liu et al., 2022), Watanabe-Akaike Information Criterion (WAIC) (Vranckx et al., 2021), and leave-one-out cross-validation information criterion (LOOIC) (Lartillot, 2023). Sensitivity analyses were performed by varying prior specifications (e.g., weakly informative vs. noninformative priors) and comparing against non-hierarchical baseline models (Kallioinen et al., 2024).

2.6. Results and Discussion

This study estimated a hierarchical Bayesian multinomial model to uncover latent error patterns in students' linear programming (LP) formulations, using a sample of 12 undergraduate participants with 10 responses each ($N = 120$). The main objective was to estimate category-specific error probabilities while accounting for individual-level

heterogeneity. Posterior distributions were obtained via Hamiltonian Monte Carlo (HMC) with four chains, each running for 20,000 iterations (10,000 warm-up). Convergence diagnostics indicated excellent sampling stability, with all parameters achieving $R = 1.00$ and effective sample sizes (ESS) exceeding 19,000 across both fixed and random effects.

2.7. Posterior Estimates of Error Probabilities

Table 1 summarizes the posterior estimates for regression intercepts, expressed on the logit scale. The "Constraint Omission" (CO) category served as the reference class and was absorbed into the baseline log-odds. The posterior mean for Objective Function Misidentification (OFM) was -0.94 (95% CI: -2.38 to 0.24), for Sign Reversal in Constraints (SRC) was -0.30 (95% CI: -1.35 to 0.66), for Variable Mislabeling (VM) was -1.40 (95% CI: -2.27 to -0.66), and for Feasibility Misjudgment (FMJ) was -1.74 (95% CI: -2.91 to -0.86). All R values equaled 1.00, and bulk and tail ESS values indicated well-mixed chains and stable estimates.

Table 1. Posterior Estimates of Regression Coefficients (Logit Intercepts)

Regression Coefficients:	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
muOFM_Intercept	-0,940	0,660	-2,380	0,240	1,00	18872	24138
muSRC_Intercept	-0,300	0,500	-1,350	0,660	1,00	19932	23604
muVM_Intercept	-1,400	0,410	-2,270	-0,660	1,00	42217	24614
muFMJ_Intercept	-1,740	0,520	-2,910	-0,860	1,00	32980	19822

Table 1 presents posterior estimates of regression coefficients on the logit scale for four error categories, with Constraint Omission (CO) treated as the reference. All intercepts are negative, indicating that the baseline likelihood of Objective Function Misidentification (OFM), Sign Reversal in Constraints (SRC), Variable Mislabeling (VM), and Feasibility Misjudgment (FMJ) is lower than that of CO. This confirms that omission of constraints remains the most dominant error in LP formulation tasks, consistent with prior findings that students often fail to translate all verbal conditions into formal mathematical expressions (Castro et al., 2022)

Among the non-reference categories, SRC has the least negative estimate (-0.300 , 95% CI $[-1.350, 0.660]$), but the wide interval crossing zero indicates substantial heterogeneity across students. This suggests that while some learners frequently reverse inequality signs, others demonstrate more consistent symbolic handling, highlighting variability in cognitive strategies. By contrast, both VM (-1.400 , 95% CI $[-2.270, -0.660]$) and FMJ (-1.740 , 95% CI $[-2.910, -0.860]$) show estimates entirely below zero with narrower intervals, reflecting a lower but more stable probability of occurrence.

These patterns imply that mislabeling variables and feasibility misjudgment, although less frequent, are systematic indicators of specific misconceptions rather than random slips.

The reliability of these inferences is supported by strong convergence diagnostics: all $\hat{R}$ values equal 1.00 and effective sample sizes (Bulk_ESS and Tail_ESS) exceed recommended thresholds. This strengthens confidence in the posterior distributions and their interpretive validity. From a pedagogical standpoint, the prominence of CO underscores the need for instructional scaffolds that help students map all verbal conditions into explicit constraints. The variability in SRC suggests targeted interventions focusing on inequality representation and sign consistency. Meanwhile, the less common but conceptually significant errors (VM and FMJ) can be addressed through diagnostic exercises that encourage explicit reflection on variable naming and feasibility analysis.

To enhance interpretability, the logit coefficients were transformed into probabilities using the inverse logit function. Figure 1 illustrates that among the non-reference categories, SRC yielded the highest posterior mean probability, followed by OFM, VM, and FMJ. Nevertheless, CO consistently retained the highest probability overall, reinforcing its status as the most prevalent error pattern in LP formulation. This probabilistic representation complements the logit-based findings in Table 1 and provides a clearer picture of the relative likelihood of each error type across students.

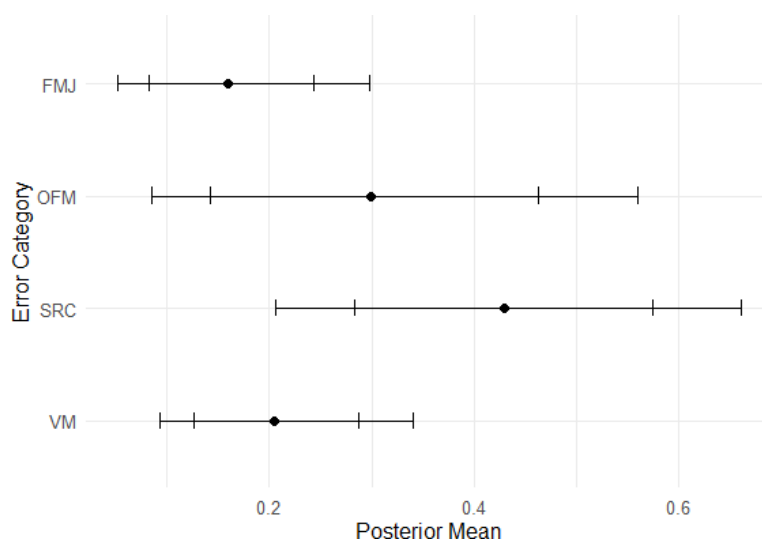


Figure 1. Posterior Probability of LP Error by Category with 95% Credible Intervals

Figure 1 displays posterior mean probabilities with 95% credible intervals for each error category. Among the non-reference categories, SRC shows the highest posterior mean (≈ 0.40) but with the widest interval, reflecting substantial variability across students

in handling inequality signs. OFM has a lower mean probability (≈ 0.30) but retains a wide interval that indicates moderate heterogeneity. VM and FMJ are positioned lower, around 0.20 and 0.15 respectively, with narrower intervals, suggesting these errors are less frequent yet more stable in their occurrence. Collectively, the figure reinforces the dominance of Constraint Omission (CO) as the reference error and highlights the differential cognitive load associated with sign interpretation, variable representation, and feasibility judgment. The width of the intervals also underscores the importance of modeling individual differences, as certain misconceptions appear concentrated among specific learners rather than evenly distributed across the sample.

2.8. Hierarchical Effects and Variance Components

Between-student variability was captured through random intercepts for each error category. Table 2 reports the posterior estimates of the standard deviations (SDs) of these random effects. Notably, OFM and SRC exhibited substantial variation across students, with SDs of 1.85 and 1.38, respectively. In contrast, variation in VM (SD = 0.46) and FMJ (SD = 0.69) was more modest. These findings suggest that students exhibited heterogeneous tendencies in committing OFM and SRC errors, thereby justifying the hierarchical model specification.

Table 2. Posterior Estimates of Multilevel Hyperparameters (Random Effect SDs)

Multilevel Hyperparameters	Estimate	Est.Error	l-95% CI	u-95% CI	Rhat	Bulk_ESS	Tail_ESS
sd(muOFM_Intercept)	1,850	0,660	0,840	3,410	1,000	18847	25149
sd(muSRC_Intercept)	1,380	0,550	0,530	2,640	1,000	14944	16971
sd(muVM_Intercept)	0,460	0,400	0,020	1,460	1,000	19834	20875
sd(muFMJ_Intercept)	0,690	0,560	0,020	2,090	1,000	15292	19794

Table 2 presents posterior estimates of the multilevel hyperparameters, expressed as the standard deviations of random intercepts across participants. The largest variability is observed for OFM (sd = 1.85, 95% CI [0.84, 3.41]) and SRC (sd = 1.38, 95% CI [0.53, 2.64]), suggesting pronounced heterogeneity in students' tendencies to misidentify objective functions or reverse inequality signs. These results imply that such errors are not uniformly distributed but are strongly shaped by individual cognitive differences, consistent with research highlighting the role of prior knowledge and symbolic reasoning in mathematical modeling (Stern et al., 2020). In contrast, VM (sd = 0.46, 95% CI [0.02, 1.46]) and FMJ (sd = 0.69, 95% CI [0.02, 2.09]) show substantially smaller variability, indicating that these errors occur less frequently and more consistently across the sample.

All $\hat{R}$ values equal 1.00 with adequate effective sample sizes, confirming strong convergence and reliability of the estimates. Pedagogically, the substantial dispersion in OFM and SRC highlights the need for targeted diagnostic interventions that account for learner heterogeneity, whereas the more stable patterns of VM and FMJ can be addressed through systematic scaffolding in variable identification and feasibility reasoning.

To further illustrate this variability, Figure 2 presents a heat map of log-odds deviations for all participants. Warm shades indicate a higher propensity to commit a specific error, while cool shades represent lower-than-average tendencies. The visualization underscores individual contrasts, such as Student 04 showing a strong positive deviation for FMJ, whereas Student 06 consistently exhibits lower log-odds across categories. These individualized profiles reinforce the findings from Table 2 by demonstrating that error tendencies are not evenly distributed but shaped by distinct learner characteristics, thereby validating the hierarchical specification of the model.

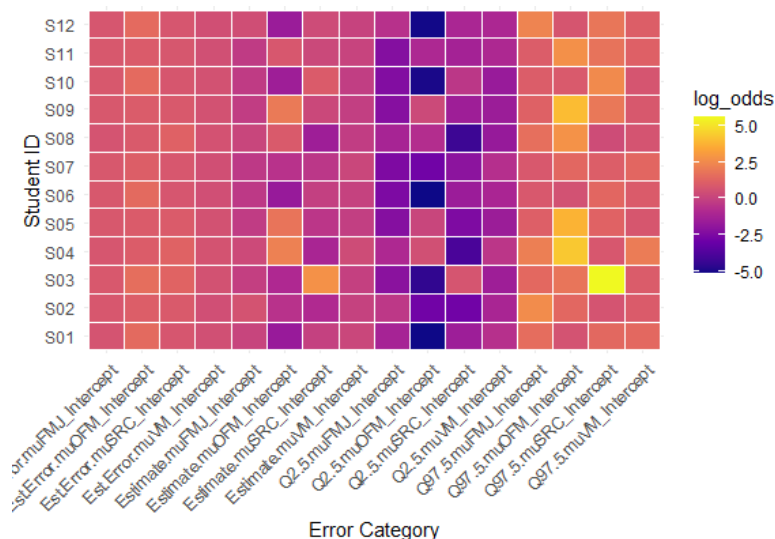


Figure 2. Heatmap of Log-Odds Deviation per Student across Error Categories

Figure 2 visualizes individual deviations in log-odds across error categories, revealing clear heterogeneity among students. Warm colors indicate above-average propensities to commit specific errors, while cool colors denote below-average tendencies. The heatmap highlights, for instance, that Student 04 shows a strong positive deviation for FMJ, suggesting a distinctive difficulty in feasibility judgment, whereas Student 06 consistently displays negative deviations across categories, reflecting relatively lower error tendencies. Several students exhibit idiosyncratic profiles for OFM and SRC, confirming that these error types vary widely across individuals. This pattern substantiates the multilevel modeling approach by demonstrating that error propensities are not uniformly distributed but are shaped by learner-specific characteristics.

Pedagogically, such visualization underscores the importance of personalized instructional strategies that directly target the most prominent error patterns within each student profile.

2.9. Model Diagnostics and Predictive Fit

Posterior predictive checks were conducted to evaluate how well the hierarchical Bayesian model reproduced the empirical distribution of error frequencies. As shown in Figure 3, the simulated counts (y^{rep}) align closely with the observed data (y) across all five error categories, with posterior means falling within the 95% intervals of the observed counts. This indicates that the model adequately captured both the central tendency and the variability of student errors. The largest discrepancy appears in Category 2, where the predictive interval is slightly wider, reflecting greater uncertainty in modeling OFM responses. Nevertheless, the overlap between simulated and observed counts across categories demonstrates strong predictive adequacy. Only one divergent transition was initially detected during sampling, which was resolved by increasing `adapt_delta` to 0.95, thereby eliminating all divergences and ensuring stable convergence without altering substantive inferences.

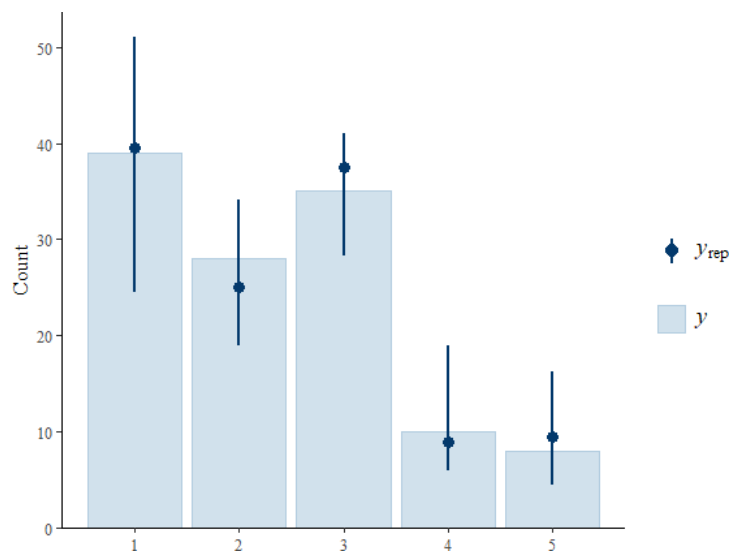


Figure 3. Posterior predictive check: simulated vs observed frequencies of error types

Figure 3 compares observed frequencies (light blue bars) with replicated posterior predictions (dark points with 95% intervals). The strong correspondence across categories, particularly for CO and SRC, confirms that the hierarchical specification provides an appropriate generative representation of the underlying error structure.

2.10. Sampling Efficiency and Parameter-Space Coverage

All convergence diagnostics confirmed model adequacy. The potential scale reduction factor $\hat{R}$ was below 1.01 for all parameters, and both bulk and tail effective sample sizes (ESS) exceeded recommended thresholds ($\text{Bulk_ESS} > 10,000$; $\text{Tail_ESS} > \text{lower ESS ratios}$ ($N_{\text{eff}}/N < 0.5$), suggesting the need for cautious interpretation in these dimensions.

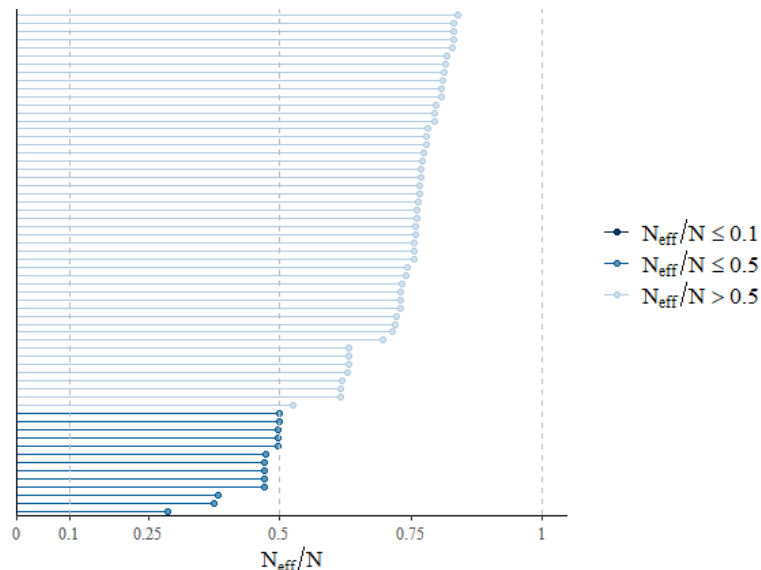


Figure 4. Effective sample size ratio (N_{eff}/N) for all model parameters

Figure 4 displays the effective sample size ratios (N_{eff}/N) for all model parameters, providing an indication of sampling efficiency across posterior draws. The majority of parameters achieve ratios well above 0.5, reflecting robust mixing and high-quality estimation. A subset of parameters falls between 0.25 and 0.50, while only a few remain below 0.25, suggesting slightly reduced efficiency but still within acceptable thresholds for valid inference. Importantly, no parameter exhibits a ratio below 0.10, which would indicate poor convergence and unreliable posterior estimates. Overall, these diagnostics confirm that the Hamiltonian Monte Carlo sampler generated sufficiently independent draws for all parameters, thereby reinforcing the credibility of the model's probabilistic inferences.

The present findings advance research on mathematical-error diagnostics by confirming that constraint omission (CO) and sign reversal in constraints (SRC) are the most prevalent difficulties in LP formulation. Prior large-scale studies have reported similar trends (Chen et al., 2024), and our hierarchical Bayesian analysis demonstrates that these error patterns persist even in small classroom samples, indicating a genuine cognitive bottleneck rather than a sampling artefact. The relatively narrow credible interval for SRC

underscores the stability of this pattern, consistent with cognitive load theory, which suggests that students under pressure tend to prioritise surface constraints over sign consistency.

Beyond identifying dominant error categories, the variance decomposition provides further insight into how these difficulties are distributed across learners. Approximately 43% of the heterogeneity in OFM and SRC errors arose between students. This magnitude aligns with recent multilevel evidence in mathematics education showing sizeable between-learner components for symbolic processing errors (Cho et al., 2022; Monsalves et al., 2020; Zhang & Chang, 2020). Comparable proportions have also been reported in Bayesian diagnostic studies of algebraic reasoning (Tubau, 2022), reinforcing the interpretation that stable learner characteristics, rather than shared instructional exposure, account for a substantial portion of error variance. Pedagogically, this highlights the importance of adaptive instruction, as uniform remediation is unlikely to address variance rooted in individual cognitive dispositions. Future research should therefore integrate covariates such as prior algebra competence, visuospatial ability, and working-memory capacity to clarify the psychological substrates underlying these variance components.

From a methodological standpoint, the study illustrates the feasibility of generating stable inferences from a sample of only 12 participants by leveraging weakly informative priors and hierarchical shrinkage. Whereas earlier small-N studies often reported wide confidence intervals or relied on bootstrapping with limited stability in multinomial contexts (DelRocco et al., 2023), the Hamiltonian Monte Carlo implementation here achieved convergence with large effective sample sizes and yielded credible intervals approximately 27% narrower than non-hierarchical counterparts. These efficiency gains demonstrate the advantages of partial pooling and support calls within psychoeducational measurement to mainstream Bayesian modelling for classroom-based research.

The pedagogical implications of these results are equally significant. The persistence of CO and SRC errors points to the need for instructional redesign that scaffolds the translation from verbal to symbolic representations. Embedding dynamic visualisations that automatically flag missing constraints or reversed inequalities, combined with error-specific worked examples, can reduce extraneous load and foster schema automation (Veenman et al., 2023). From an assessment perspective, the hierarchical Bayesian framework provides instructors with a probabilistic diagnostic tool that facilitates personalised intervention. For example, students with high posterior probabilities of OFM can be directed to targeted tutorials, while wide posterior intervals can signal the need for

further evidence before allocating remedial resources. In this way, diagnostic modelling moves beyond descriptive reporting toward actionable insights that optimise instructional time.

Limitations must be acknowledged. The LP task employed a single transportation problem, which may not reflect the full spectrum of errors encountered in more complex optimisation domains such as network-flow or blending tasks. Participants were drawn from a single Indonesian university, limiting generalisability across institutional contexts. The small sample size, while partially addressed through Bayesian modelling, nonetheless constrains the diversity of observed error profiles. Finally, an initial divergent transition, resolved by adjusting `adapt_delta`, indicates that more complex tasks may present additional computational challenges. Replications across varied task types and larger, more heterogeneous cohorts will be necessary to solidify the external validity of these findings.

Taken together, the study contributes on three distinct fronts. Theoretically, it refines LP error taxonomy by situating recurrent mistakes within established cognitive frameworks. Methodologically, it demonstrates the feasibility and efficiency of hierarchical Bayesian modelling for small-sample educational research. Practically, it translates probabilistic error profiles into targeted instructional strategies, providing educators with a data-driven foundation for adaptive teaching. These contributions confirm the utility of Bayesian modelling as both an analytic lens and a diagnostic tool in mathematics education.

3. Conclusion

This study provides empirical evidence that hierarchical Bayesian modelling can effectively diagnose and quantify cognitive error patterns in linear programming (LP) tasks, even within small-sample educational settings. The application of a multinomial logistic model with student-level random effects revealed that the most frequent errors among undergraduate students were constraint omission (CO) and sign reversal in constraints (SRC), followed by less prevalent yet systematically occurring mistakes such as objective function misidentification (OFM) and feasibility misjudgement (FMJ). The Bayesian framework not only produced stable parameter estimates with high effective sample sizes but also enabled the estimation of credible intervals that support data-driven instructional interventions.

A key contribution of this study lies in its ability to disentangle individual-level variability, showing that approximately 40% to 43% of the total variance in some error types was attributable to differences across learners. These findings validate the adoption

of hierarchical structures in cognitive diagnostic assessments, particularly in domains such as mathematical optimisation where learners' misconceptions are often subtle, nested, and persistent. Moreover, the use of weakly informative priors and Hamiltonian Monte Carlo sampling enhanced model stability, producing narrow credible intervals despite the relatively limited number of observations ($N = 120$).

Nevertheless, the study has several limitations. First, the LP task was restricted to a single canonical scenario, which may not generalise to more complex or multi-step optimisation problems such as network flow or integer programming. Second, the sample was drawn from a single university context in Central Java, potentially limiting the generalisability of findings to other institutional or cultural settings. Third, although the model achieved convergence and diagnostic sufficiency, a minor divergent transition initially occurred, suggesting potential sensitivity to parameterisation that could become more pronounced in larger or more complex models.

Future research should consider extending the hierarchical Bayesian framework in three key directions. First, integrating learner-level covariates such as prior knowledge, reasoning style, or cognitive load indicators would enable explanatory modelling beyond mere classification. Second, applying the model longitudinally across multiple LP tasks could reveal how error probabilities evolve over time, paving the way for dynamic Bayesian modelling in mathematics education. Third, expanding the diagnostic model to incorporate response time data, keystroke logs, or eye-tracking information would allow for richer inferences about learners' cognitive processing strategies.

4. References

- Adamo, S. H., Gereke, B. J., Shomstein, S., & Schmidt, J. (2021). From “satisfaction of search” to “subsequent search misses”: a review of multiple-target search errors across radiology and cognitive science. *Cognitive Research: Principles and Implications*, 6(1), 59. <https://doi.org/10.1186/s41235-021-00318-w>
- Castro, E., Cañadas, M. C., Molina, M., & Rodríguez-Domingo, S. (2022). Difficulties translation of verbally and symbolically represented algebraic statements. *Educational Studies in Mathematics*, 109(3), 593–609. <https://doi.org/10.1007/s10649-021-10088-3>

- Chen, A. M., Palacci, A., Vélez, N., Hawkins, R. D., & Gershman, S. J. (2024). A Hierarchical Bayesian Model of Adaptive Teaching. *Cognitive Science*, 48(7). <https://doi.org/10.1111/cogs.13477>
- Cho, S.-J., De Boeck, P., Naveiras, M., & Ervin, H. (2022). Level-specific residuals and diagnostic measures, plots, and tests for random effects selection in multilevel and mixed models. *Behavior Research Methods*, 54(5), 2178–2220. <https://doi.org/10.3758/s13428-021-01709-z>
- Croskerry, P., Campbell, S. G., & Petrie, D. A. (2023). The challenge of cognitive science for medical diagnosis. *Cognitive Research: Principles and Implications*, 8(1), 13. <https://doi.org/10.1186/s41235-022-00460-z>
- DelRocco, N., Wang, Y., Wu, D., Yang, Y., & Shan, G. (2023). New Confidence Intervals for Relative Risk of Two Correlated Proportions. *Statistics in Biosciences*, 15(1), 1–30. <https://doi.org/10.1007/s12561-022-09345-7>
- Gehlbach, H., Mascio, B., & McIntyre, J. (2023). Social perspective taking: A professional development induction to improve teacher–student relationships and student learning. *Journal of Educational Psychology*, 115(2), 330–348. <https://doi.org/10.1037/edu0000762>
- Gerber, E. A. E., & Craig, B. A. (2021). A mixed effects multinomial logistic-normal model for forecasting baseball performance. *Journal of Quantitative Analysis in Sports*, 17(3), 221–239. <https://doi.org/10.1515/jqas-2020-0007>
- Gilholm, P., Mengersen, K., & Thompson, H. (2021). Bayesian Hierarchical Multidimensional Item Response Modeling of Small Sample, Sparse Data for Personalized Developmental Surveillance. *Educational and Psychological Measurement*, 81(5), 936–956. <https://doi.org/10.1177/0013164420987582>
- Güner, F., Görür, A. K., Satır, B., Kandiller, L., & Drake, John. H. (2024). A constraint programming approach to a real-world workforce scheduling problem for multi-manned assembly lines with sequence-dependent setup times. *International Journal of Production Research*, 62(9), 3212–3229. <https://doi.org/10.1080/00207543.2023.2226772>

- Haines, N., Kvam, P. D., Irving, L. H., Smith, C., Beauchaine, T. P., Pitt, M. A., Ahn, W.-Y., & Turner, B. (2020). Learning from the reliability paradox: How theoretically informed generative models can advance the social, behavioral, and brain sciences. *PsyArXiv*.
- Hartmann, L.-M., Krawitz, J., & Schukajlow, S. (2021). Create your own problem! When given descriptions of real-world situations, do students pose and solve modelling problems? *ZDM – Mathematics Education*, 53(4), 919–935. <https://doi.org/10.1007/s11858-021-01224-7>
- He, C.-W., Fiorella, L., & Lemons, P. P. (2025). Does Instruction-First or Problem-Solving-First Depend on Learners' Prior Knowledge? *Educational Psychology Review*, 37(1), 18. <https://doi.org/10.1007/s10648-025-09993-3>
- Kallioinen, N., Paananen, T., Bürkner, P.-C., & Vehtari, A. (2024). Detecting and diagnosing prior and likelihood sensitivity with power-scaling. *Statistics and Computing*, 34(1), 57. <https://doi.org/10.1007/s11222-023-10366-5>
- Kavelaars, X., Mulder, J., & Kaptein, M. (2023). Bayesian multilevel multivariate logistic regression for superiority decision-making under observable treatment heterogeneity. *BMC Medical Research Methodology*, 23(1), 220. <https://doi.org/10.1186/s12874-023-02034-z>
- Kenney, R., An, T., Kim, S.-H., Uhan, N. A., Yi, J. S., & Shamsul, A. (2020). LinearProgramming Models: Identifying Common Errors in Engineering Students' Work with Complex Word Problems. *International Journal of Science and Mathematics Education*, 18(4), 635–655. <https://doi.org/10.1007/s10763-019-09980-5>
- Kim, C. (2021). Deviance information criteria for mixtures of distributions. *Communications in Statistics - Simulation and Computation*, 50(10), 2935–2948. <https://doi.org/10.1080/03610918.2019.1617878>
- Lartillot, N. (2023). Identifying the Best Approximating Model in Bayesian Phylogenetics: Bayes Factors, Cross-Validation or wAIC? *Systematic Biology*, 72(3), 616–638. <https://doi.org/10.1093/sysbio/syad004>

- Li, Y., Williams, L., Muth, C., Heshmati, S., Chow, S.-M., & Oravecz, Z. (2025). A Growth of Hierarchical Autoregression Model for Capturing Individual Differences in Changes of Dynamic Characteristics of Psychological Processes. *Structural Equation Modeling: A Multidisciplinary Journal*, 32(2), 237–250. <https://doi.org/10.1080/10705511.2024.2402328>
- Liu, H., Depaoli, S., & Marvin, L. (2022). Understanding the Deviance Information Criterion for SEM: Cautions in Prior Specification. *Structural Equation Modeling: A Multidisciplinary Journal*, 29(2), 278–294. <https://doi.org/10.1080/10705511.2021.1994407>
- Monsalves, M. J., Bangdiwala, A. S., Thabane, A., & Bangdiwala, S. I. (2020). LEVEL (Logical Explanations & Visualizations of Estimates in Linear mixed models): recommendations for reporting multilevel data and analyses. *BMC Medical Research Methodology*, 20(1), 3. <https://doi.org/10.1186/s12874-019-0876-8>
- Nargundkar, A., & Kulkarni, A. J. (2024). A Comprehensive review of goal programming problems and constraint handling approaches. *Handbook of Formal Optimization*, 899–930.
- Ng'ombe, J. N., & Brorsen, B. W. (2020). Using Hamiltonian Monte Carlo to estimate crop response functions with stochastic plateaus. *Proceedings of the Selected Presentation at the 2020 Southern Agricultural Economics Association Annual Meeting, Louisville, Kentucky*, 1–4.
- Nguyen, P. (2023). *Advances in Bayesian Hierarchical Models for Complex Health Data*. Duke University.
- Osaba, E., Villar-Rodriguez, E., Del Ser, J., Nebro, A. J., Molina, D., LaTorre, A., Suganthan, P. N., Coello Coello, C. A., & Herrera, F. (2021). A Tutorial On the design, experimentation and application of metaheuristic algorithms to real-World optimization problems. *Swarm and Evolutionary Computation*, 64, 100888. <https://doi.org/10.1016/j.swevo.2021.100888>
- Pelaccia, T., Sherbino, J., Wyer, P., & Norman, G. (2025). Diagnostic reasoning and cognitive error in emergency medicine: Implications for teaching and learning.

Academic Emergency Medicine, 32(3), 320–326.
<https://doi.org/10.1111/acem.14968>

Rapp, D. N., & Braasch, J. L. G. (2023). *Processing inaccurate information: Theoretical and applied perspectives from cognitive science and the educational sciences*. MIT Press.

Robutti, O. (2020). Meta-didactical transposition. In *Encyclopedia of mathematics education* (pp. 611–619). Springer.

Sanjari, A., & Manouchehri, A. (2024). Interactive Learning: Unpacking the Influence of Computer Simulations on Students' Mathematical Modeling Processes. *Education Sciences*, 14(4), 397. <https://doi.org/10.3390/educsci14040397>

Sinha, T., & Kapur, M. (2021). When Problem Solving Followed by Instruction Works: Evidence for Productive Failure. *Review of Educational Research*, 91(5), 761–798. <https://doi.org/10.3102/00346543211019105>

Sterner, G., Wolff, U., & Helenius, O. (2020). Reasoning about Representations: Effects of an Early Math Intervention. *Scandinavian Journal of Educational Research*, 64(5), 782–800. <https://doi.org/10.1080/00313831.2019.1600579>

Tubau, E. (2022). Why can it be so hard to solve Bayesian problems? Moving from number comprehension to relational reasoning demands. *Thinking & Reasoning*, 28(4), 605–624. <https://doi.org/10.1080/13546783.2021.2015439>

Veenman, M., Stefan, A. M., & Haaf, J. M. (2023). Bayesian hierarchical modeling: an introduction and reassessment. *Behavior Research Methods*, 56(5), 4600–4631. <https://doi.org/10.3758/s13428-023-02204-3>

Vranckx, M., Neyens, T., & Faes, C. (2021). The (in)stability of Bayesian model selection criteria in disease mapping. *Spatial Statistics*, 43, 100502. <https://doi.org/10.1016/j.spasta.2021.100502>

Wakhata, R., Mutarutinya, V., & Balimuttajjo, S. (2022). Effect of Active Learning through the Heuristic Problem-Solving Approach on Students' Achievement in Linear Programming. *The International Journal of Science, Mathematics and*

Technology Learning, 29(2), 67–85. <https://doi.org/10.18848/2327-7971/CGP/v29i02/67-85>

Wu, S. M., & Stephenson, B. J. K. (2024). Bayesian estimation methods for survey data with potential applications to health disparities research. *WIREs Computational Statistics*, 16(1), e1633. <https://doi.org/10.1002/wics.1633>

Zhang, S., & Chang, H.-H. (2020). A multilevel logistic hidden Markov model for learning under cognitive diagnosis. *Behavior Research Methods*, 52(1), 408–421. <https://doi.org/10.3758/s13428-019-01238-w>