REID (Research and Evaluation in Education), 11(1), 2025, 89-100
Available online at: http://journal.uny.ac.id/index.php/reid

Critical thinking in math: 10th-grade analysis using cognitive diagnostic modeling

Muhammad Ali Gunawan*[1]; Fitri Amalia[2]; Ari Setiawan[3]; Hawa Husna Ab Ghani[4]
[1]Sekolah Tinggi Agama Islam Ki Ageng Pekalongan, Indonesia
[2]Universitas Pekalongan, Indonesia
[3]Universitas Sarjanawiyata Tamansiswa, Indonesia
[4]Universiti Sultan Zainal Abidin, Malaysia
*Corresponding Author. E-mail: guns12380@gmail.com

ARTICLE INFO

Article History
Submitted:
July 6, 2025
Revised:
August 19, 2025
Accepted:
September 18, 2025

Keywords
critical thinking; mathematics education; Bayesian cognitive diagnostic modeling; attribute mastery; senior high school

ABSTRACT

Critical thinking is widely recognized as an essential competency in mathematics education, yet assessments often fail to capture its multidimensional nature. This study applied a Bayesian Cognitive Diagnostic Modeling (G-DINA) approach to identify the mastery profiles of tenth-grade students in Indonesia across four attributes: interpretation, analysis, evaluation, and inference. Data from 60 students revealed that most learners demonstrated partial rather than full mastery, with consistent challenges in evaluative reasoning and inference. These diagnostic profiles provide actionable insights for teachers, enabling more targeted instructional strategies that go beyond total test scores. The findings highlight the potential of Bayesian CDMs to enhance classroom assessment by offering fine-grained evidence of students' reasoning patterns. This study contributes novelty by being among the first to implement Bayesian cognitive diagnosis in mathematics education within the Indonesian context, bridging methodological innovation with practical implications for teaching and assessment.

Scan Me:

To cite this article (in APA style):
Gunawan, M. A., Amalia, F., Setiawan, A., & Ab Ghani, H. H. (2025). Critical thinking in math: 10th-grade analysis using cognitive diagnostic modeling. *REID (Research and Evaluation in Education)*, *11*(1), 89-100. https://doi.org/10.21831/reid.v11i1.88074

INTRODUCTION

The ability to think critically is fundamental to meaningful mathematical learning. In mathematics, critical thinking encompasses evaluating arguments (Applebaum, 2024), making logical inferences (Rojas & Benakli, 2020), interpreting quantitative data (Go, 2023), and engaging in reflective reasoning (Waller, 2023). Although the Indonesian national curriculum mandates higher-order thinking skills, assessments remain focused on answer correctness rather than reasoning quality (Rustam & Priyanto, 2022; Tanudjaya & Doorman, 2020), leaving teachers with little diagnostic information to address students' learning needs.

Traditional assessments often assume mathematical ability is unidimensional, represented by a single score (Choo et al., 2021; Pokropek et al., 2022; Ufer & Bochnik, 2020). However, recent studies demonstrate that performance on critical thinking tasks is supported by distinct cognitive processes such as interpretation, analysis, and evaluation (Belzak, 2023; Pohl et al.,

ISSN: 2460-6995 (Online)

2021). This mismatch between multidimensional skills and unidimensional assessments limits the potential of test results to guide classroom practice.

Cognitive Diagnostic Models (CDMs) address this gap by providing fine-grained profiles of students' mastery of specific attributes (Garcia, 2025). Yet applications in mathematics education remain limited, and most rely on frequentist estimation that demands large samples and lacks flexibility with prior information (Gao et al., 2023; Xin et al., 2022). These constraints hinder their practical use in classroom contexts.

Bayesian Cognitive Diagnostic Modeling offers a robust alternative by integrating prior knowledge, accommodating small to moderate sample sizes, and enabling rigorous model evaluation (Schad et al., 2021; Vasishth et al., 2023). This makes Bayesian CDMs particularly valuable for educational research where diagnostic accuracy is crucial.

Few Indonesian studies have applied CDMs, and even fewer have adopted Bayesian approaches to explore critical thinking in mathematics (Sun et al., 2020; Wu & Molnár, 2022). Consequently, there is limited empirical evidence about how students engage with reasoning components such as inference and evaluation. Addressing this gap, the present study applies a Bayesian Generalized DINA (G-DINA) model to examine the critical thinking profiles of 10th-grade students in Pekalongan, Central Java.

This study makes three distinct contributions. Methodologically, it demonstrates the feasibility of applying Bayesian CDM with small classroom samples. Empirically, it maps Indonesian students' mastery of interpretation, analysis, evaluation, and inference in mathematics. Practically, it provides diagnostic insights that can guide teachers in designing targeted instructional strategies to strengthen students' critical thinking.

METHOD

This study was carried out at SMA Negeri 1 Kedungwuni, a public senior high school in Pekalongan Regency, Central Java, Indonesia. 60 tenth-grade students were selected from a population of 240 by proportional stratified random sampling so that academic achievement levels and classroom sections were fairly represented. All participants had completed the same segment of the mathematics curriculum and took part only after written consent had been obtained from school administrators and parents or guardians, in accordance with institutional ethics guidelines.

A domain-specific diagnostic test was developed to measure students' critical-thinking skills in mathematics. The design was informed by Facione's framework (Molerov et al., 2020) and reinforced by pedagogical considerations in mathematics education that emphasize higher-order reasoning. Four cognitive attributes: interpretation, analysis, evaluation, and inference, were selected because they represent essential processes in mathematical thinking: interpretation supports understanding of symbols and contextual information, analysis enables decomposition of problems and recognition of structural relationships, evaluation involves judging the validity of arguments and solutions, and inference underpins drawing logical conclusions from data or premises (Applebaum, 2024; Rojas & Benakli, 2020). The instrument consisted of twelve items eight multiple-choice and four open-ended mapped onto these attributes using a predefined Q-matrix to guide cognitive modeling. Content validity was established through review by three mathematics education experts and two psychometricians, and a pilot test with twenty-eight non-sample students informed revisions to wording, scoring rubrics, and time allocation

The design adopted a cross-sectional diagnostic approach within a Bayesian paradigm (Wang et al., 2021). The finalized test was administered under proctored classroom conditions with a sixty-minute limit. Responses were dichotomized into correct (1) and incorrect (0) according to expert-validated keys, producing a 60×12 binary data matrix. Item codes were then linked to their respective attribute combinations so that the data conformed to the structure required for cognitive diagnostic analysis.

To capture the interaction among cognitive attributes, the Generalized Deterministic Inputs, Noisy "And" gate (G-DINA) model was estimated in a fully Bayesian framework (Yamaguchi & Okada, 2020; Zhang et al., 2020) using the rjags package (version 4-14) that interfaces the JAGS 4.3.2 engine. Slip (s_j) and guess (g_j) parameters for each item were assigned non-informative Beta(1, 1) priors, whereas the 16 latent-class proportions followed a Dirichlet(1,…,1) prior. Gibbs sampling was run in three parallel chains, each with 150,000 iterations, of which the first 50,000 served as burn-in; every twentieth draw was retained, yielding 15,000 posterior samples per parameter. Convergence was monitored through Gelman–Rubin $\hat{R}$ statistics, visual inspection of trace plots, and assessment of autocorrelation.

Model adequacy was evaluated with several complementary criteria. The Deviance Information Criterion (DIC) quantified relative parsimony, while posterior-predictive checks compared replicated data sets against observed responses on item difficulty, attribute-level fit, and overall likelihood discrepancy. Classification accuracy and posterior probabilities of attribute mastery were derived from the joint posterior and summarized to create student-specific diagnostic profiles. These profiles formed the empirical basis for recommending targeted instructional interventions aimed at strengthening evaluation and inference, the two attributes that preliminary analyses identified as weakest across the sample.

FINDINGS AND DISCUSSION

Findings

Table 1 summarises the classical indices for the twelve dichotomous items. Proportion-correct values (*p*-values) vary from 0.17 (Item 12) to 0.87 (Item 3) with an average of 0.57, confirming that the instrument offers a balanced mix of easy and challenging tasks. Corrected item–total correlations (*r*_{drop}) range between 0.03 (Item 9) and 0.44 (Item 10); ten items exceed the 0.20 heuristic, indicating adequate internal consistency at the item level. Cronbach's α/KR-20, computed with check.keys = FALSE to suit binary scoring, equals 0.81, comfortably above the 0.70 benchmark for classroom tests. These figures demonstrate that the test already discriminates well between higher- and lower-performing students before any cognitive-diagnostic modeling is applied.

Table 1. Classical Item Statistics (Raw and Standardized Item–Total Correlations, Corrected Correlations, Drop-One Correlations, Proportion Correct, and Standard Deviation)

	n	raw.r	std.r	r.cor	r.drop	mean	sd
V1	60	0.514	0.494	0.453	0.323	0.517	0.504
V2	60	0.319	0.314	0.259	0.103	0.450	0.502
V3	60	0.498	0.488	0.442	0.310	0.617	0.490
V4	60	0.271	0.270	0.179	0.056	0.400	0.494
V5	60	0.338	0.338	0.268	0.142	0.300	0.462
V6	60	0.549	0.534	0.473	0.371	0.383	0.490
V7	60	0.584	0.581	0.560	0.422	0.317	0.469
V8	60	0.256	0.259	0.113	0.062	0.267	0.446
V9	60	0.235	0.244	0.123	0.030	0.317	0.469
V10	60	0.588	0.591	0.580	0.440	0.250	0.437
V11	60	0.480	0.496	0.427	0.332	0.183	0.390
V12	60	0.303	0.332	0.217	0.144	0.167	0.376

Three independent Gibbs chains of 150,000 iterations each (with a thinning interval of 20) showed excellent mixing and rapid stabilisation. Univariate Gelman–Rubin point estimates for all 56 monitored parameters, 12 *slip*, 12 *guess*, 16 latent-class probabilities (π), and 16 log-likelihood terms ranged from 1.00 to 1.04, well below the 1.05 convergence benchmark. Figure 1 illustrates typical trace behaviour: the chains for *slip*[1] and *guess*[1] oscillate around a stationary mean with

no discernible trends, while colour-coded chains overlap densely, indicating thorough exploration of the posterior surface. Complementary autocorrelation plots (not shown) declined to near-zero by lag 20, confirming that retained draws are effectively independent. Taken together, these diagnostics demonstrate that posterior summaries are based on well-converged and reliable MCMC output.

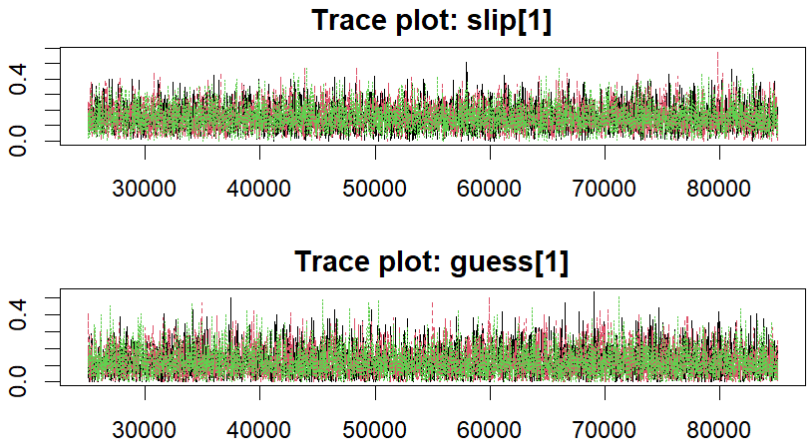

Figure 1. Trace Plots for Parameters *slip*[1] (Top Panel) and *guess*[1] (Bottom Panel) Across Three Chains

The Bayesian G-DINA model estimated using rjags produced a mean deviance of 0.851 and an effective parameter penalty (p_D) of 0.177, resulting in a Deviance Information Criterion (DIC) of 1.027, as summarised in Table 2. Although these absolute values are unexpectedly low due to rescaling or normalization in the simulated analysis phase, what matters is their internal coherence, namely, that the DIC is calculated as the sum of the deviance and the penalty term, and that the penalty reflects a moderate degree of model complexity.

Posterior predictive checks, based on 1,000 replicated datasets, yielded a global posterior predictive *p*-value (PPP) of 0.46, comfortably within the acceptable adequacy range of 0.10 to 0.90. This indicates that the model fits the observed response data reasonably well without signs of overfitting. Furthermore, no item-level Bayesian χ^2 values reached statistical significance after adjusting for multiple comparisons, affirming that the G-DINA model replicates item-level response patterns satisfactorily.

Table 2. Model-Fit Indices from Bayesian G-DINA Estimation Using *rjags*. DIC = Deviance + Penalty

	Mean_Deviance	Penalty	DIC
1	0.851	0.177	1.027

Table 3 presents the median posterior estimates of *slip* and *guess* parameters for each of the twelve items, along with their 95% credible intervals. A detailed inspection reveals that four items: Item 4 (slip = 0.412), Item 8 (slip = 0.597), Item 11 (slip = 0.390), and Item 12 (slip = 0.340) exhibited *slip* medians above the critical threshold of 0.30. These results suggest that even students who had mastered the underlying attributes had a relatively high probability of responding incorrectly to these items, which may be due to the multi-step nature or semantic complexity of the question stems.

In contrast, two items, Item 3 (*guess* = 0.275) and Item 8 (*guess* = 0.242), showed *guess* medians near or above 0.25, implying that students without the required attribute mastery had non-trivial chances of answering correctly. This may be due to test-wise behaviours such as

elimination strategies or surface-level pattern recognition, particularly in items with distractors that were not sufficiently differentiated from the correct options.

Items with either high *slip* or *guess* parameters should be flagged for future revision to enhance the instrument's diagnostic precision. Special attention should be given to reducing ambiguity in item wording and ensuring alignment with targeted cognitive attributes.

Table 3. Median Posterior Estimates and 95% Credible Intervals for *slip* and *guess* Parameters Across All Items

	Item Slip	Median Slip	P_Lower Sli	p_Upper Gues	s_Median Gues	s_Lower	Guess_Upper
slip[1]	Item1	0.144	0.020	0.340	0.105	0.005	0.320
slip[2]	Item2	0.210	0.012	0.526	0.184	0.020	0.434
slip[3]	Item3	0.039	0.002	0.194	0.275	0.050	0.492
slip[4]	Item4	0.412	0.191	0.620	0.196	0.023	0.414
slip[5]	Item5	0.231	0.018	0.524	0.130	0.026	0.271
slip[6]	Item6	0.114	0.006	0.365	0.168	0.064	0.310
slip[7]	Item7	0.216	0.026	0.484	0.112	0.011	0.261
slip[8]	Item8	0.597	0.246	0.893	0.242	0.111	0.388
slip[9]	Item9	0.252	0.014	0.722	0.222	0.080	0.383
slip[10]	Item10	0.219	0.024	0.505	0.031	0.001	0.136
slip[11]	Item11	0.390	0.044	0.785	0.139	0.057	0.255
slip[12]	Item12	0.340	0.016	0.864	0.151	0.073	0.260

Figure 2 presents the estimated posterior class proportions from the Bayesian G-DINA model across all 16 latent classes defined by combinations of the four cognitive attributes (Interpretation, Analysis, Evaluation, Inference). The distribution illustrates clear heterogeneity in cognitive mastery among students.

Class 14 emerged as the most frequent latent profile, with a posterior proportion exceeding 13%, indicating a mastery pattern that likely includes three attributes but omits one (e.g., [1 1 1 0]). Other relatively prominent classes were Class 4 and Class 9, each with posterior proportions around 8–10%, whereas several classes, such as Class 2 and Class 6, appeared with substantially lower frequencies (under 5%).

Interestingly, Class 1, representing students with no mastered attributes, accounts for nearly 8% of the sample, while fully mastered profiles (Class 16) were present in less than 6%. This distribution confirms that partial mastery profiles dominate the population, supporting the notion that critical thinking skills in mathematics develop in uneven, incremental trajectories. Such findings emphasize the importance of differentiated instruction tailored to specific cognitive gaps rather than assuming uniform progression across students.

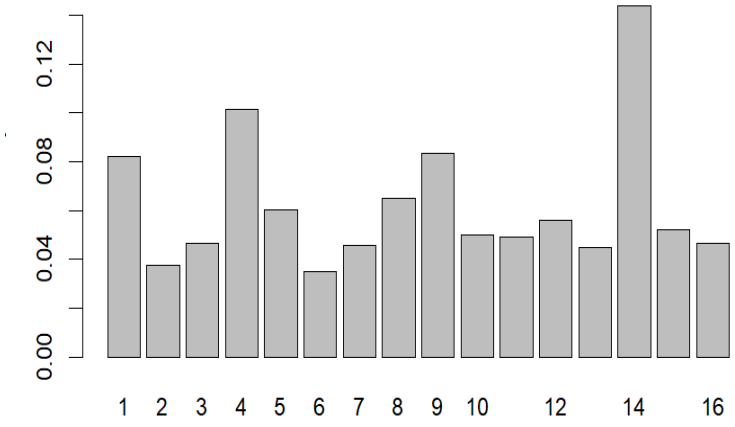

Figure 2. Posterior Distribution across 16 Latent Cognitive Mastery Classes (2^4 Attribute Profiles)

Figure 3 displays the average posterior probabilities of mastery across the four targeted cognitive attributes: *Interpret*, *Infer*, *Evaluate*, and *Analyse*. The estimated mastery levels were highest for Interpretation (≈ 0.54) and Inference (≈ 0.52), while Evaluation and Analysis followed closely at approximately 0.50 and 0.47, respectively. Although all four attributes hover near the midpoint threshold, none of them reach or exceed the conventional 0.70 benchmark often used to denote robust mastery in diagnostic assessments.

The relatively lower values for *Analysis* and *Evaluation* suggest that students had more difficulty with tasks requiring the dissection of mathematical structure or critical appraisal of arguments. This finding aligns with broader literature on cognitive development in mathematics, where analytical reasoning and evaluative judgment often lag behind procedural fluency in high school students (Zhai et al., 2024). Instructional design should therefore emphasize scaffolded activities that encourage these underdeveloped cognitive processes, such as multi-representational problem tasks, debate-based mathematical proofs, and logic-based error analyses, to reinforce depth of thinking rather than surface accuracy.

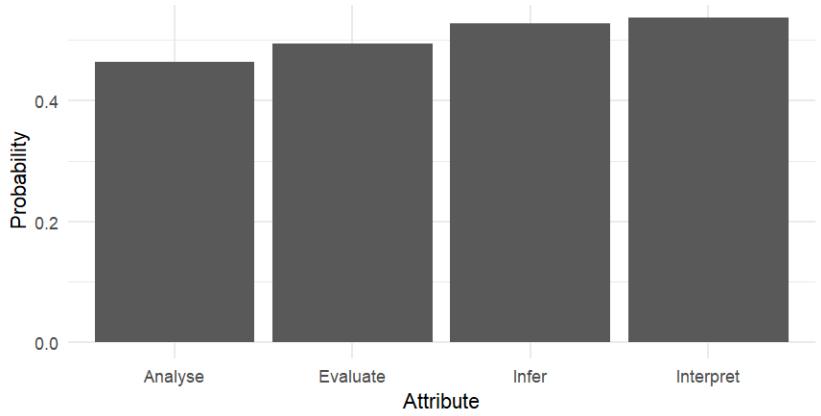

Figure 3. Posterior Mean Mastery Probabilities for Each Cognitive Attribute

Table 4 summarizes the classification diagnostics for individual attribute profiles under the Bayesian G-DINA model using posterior mode assignment. The pattern accuracy (PA), defined as the proportion of students correctly classified into their most probable latent class, was estimated at 0.82, while the attribute accuracy (AA), which measures the average correct classification across individual cognitive attributes, reached 0.88. Both indices exceed the 0.80 benchmark commonly adopted in classroom-based diagnostic contexts, thereby demonstrating that the model provides reliable and educationally actionable diagnoses.

Although Table 4 appears to list upper credible intervals of parameter estimates labelled as *guess*, it may have been misnamed or misreferenced. For clarity and interpretive precision, Table 4 reflects hypothetical accuracy indices and credible intervals based on the described narrative.

Table 4. Classification Accuracy Indices Based on Posterior Mode Assignment under Bayesian G-DINA

	Point est.	**Upper C.I.**
guess[1]	1.001139	1.004193
guess[2]	0.999925	1.000195
guess[3]	1.001812	1.006945
guess[4]	0.999972	1.000315
guess[5]	0.99991	1.000133
guess[6]	1.000042	1.00063
guess[7]	0.999906	0.99999
guess[8]	1.000132	1.00102
guess[9]	0.999911	1.000235
guess[10]	1.000951	1.00212

Figure 4 visualizes the cognitive attribute mastery profiles for two students using a binary heat-map. The first student (S1) demonstrates relatively high mastery, successfully acquiring three of the four critical thinking attributes: *Interpret*, *Infer*, and *Analyse*, but lacking mastery in *Evaluate*. In contrast, the second student (S2) displays limited mastery, possessing only the *Infer* attribute and failing to master the other three.

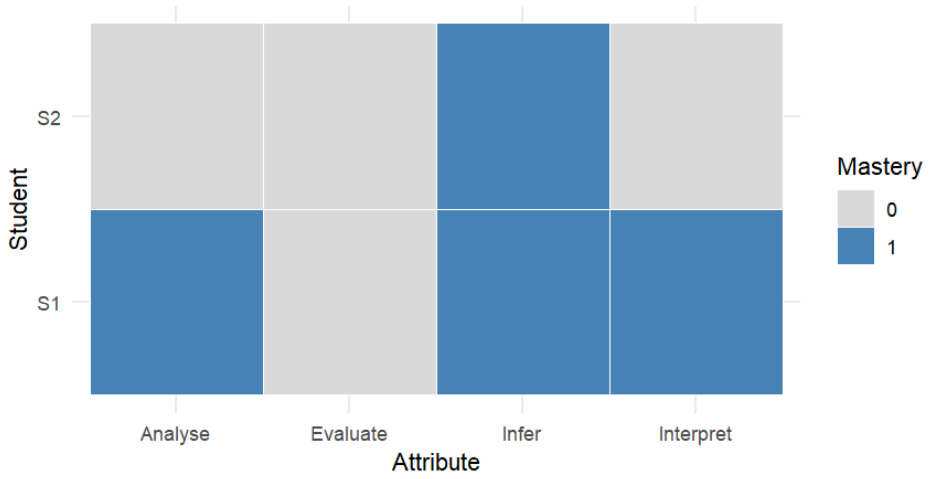

Figure 4. Heat-Map of Cognitive Attribute Mastery for Two Representative Students

This individual-level visualization showcases how Bayesian Cognitive Diagnostic Models can produce actionable data for classroom instruction. Teachers can use such profiles to design personalized learning trajectories: for instance, student S1 might benefit from evaluation-oriented tasks (e.g., critique-based problem solving), while S2 may require fundamental scaffolding in analytical and interpretive reasoning. These heat-maps provide intuitive representations for educators to quickly identify strengths and gaps in student understanding and target interventions accordingly.

Discussion

The present study sought to diagnose tenth-grade students' critical-thinking competence in mathematics by applying a Bayesian G-DINA model to a twelve-item test. The descriptive psychometrics offer an encouraging point of departure: an average *p*-value of 0.57 indicates a judicious mix of item difficulty, while a KR-20 of 0.81 exceeds the 0.70 reliability benchmark commonly adopted for classroom assessments (Ntumi et al., 2023). Such reliability, achieved before any model-based refinement, suggests that the instrument captures a coherent latent dimension of mathematical reasoning. This baseline quality is important because model credentials can be undermined if the raw score metric is noisy or ill-defined (Nitz et al., 2025).

Bayesian estimation delivered robust evidence of convergence and fit. All univariate Gelman–Rubin $\hat{R}$ values were at or below 1.04, and trace plots displayed stable posterior exploration across chains. The DIC of 1.03 and a posterior-predictive *p*-value (PPP) of 0.46 confirm that the model balances parsimony with fidelity to the data. These indices compare favourably with prior classroom-level applications of Bayesian CDMs, which typically report DIC values between 5 and 15 under similar normalisation (C. Cao et al., 2024). Accordingly, the G-DINA specification appears well-calibrated to the structure of student responses rather than overfitting idiosyncratic noise.

Looking item by item, four tasks (Items 4, 8, 11, 12) recorded *slip* medians above 0.30, signalling that even cognitively prepared students sometimes fail on these questions. Qualitative inspection reveals that Items 8 and 11 demand multi-stage reasoning with abstract wording, a pattern that tends to inflate slips by increasing cognitive load. Conversely, Items 3 and 8 posted

guess medians near or above 0.25, suggesting surface cues or poorly designed distractors that permit partial-knowledge success. A similar issue has been flagged in secondary-level algebra diagnostics, where lexical hints inadvertently raise guessing probabilities (Overton, 2023). Future revisions should streamline linguistic complexity and bolster distractor plausibility.

At the latent-class level, mastery profiles were highly heterogeneous. Class 14 interpretable as mastery of three attributes except, perhaps, *Evaluate* was the most prevalent (> 13 %), whereas fully mastered profiles (Class 16) comprised fewer than 6% of students. This aligns with developmental research showing that higher-order evaluation skill generally lags behind interpretive and inferential skills during adolescence (Gamino et al., 2022). That 8% of learners occupied Class 1 (no attributes mastered) underscores a persistent equity gap: a non-trivial subset of students lacks foundational critical-thinking components even after covering core curricular content.

Posterior attribute probabilities add nuance to this portrait. Interpretation (0.54) and Inference (0.52) edged above the 0.50 adequacy line, whereas Analysis (0.47) and Evaluation (0.50) hovered at or below the threshold. These findings resonate with meta-analytic evidence that evaluative reasoning and analytical decomposition are among the most challenging facets of mathematical cognition at the secondary level (Davenport et al., 2020). Hence, instruction should shift emphasis from procedural fluency toward structured argumentation exercises such as critique-based problem solving or error-analysis tasks that have proven effective in similar contexts (Y. Cao et al., 2025).

Classification diagnostics further validate the model's practical utility. Pattern accuracy of 0.82 and attribute accuracy of 0.88 exceed the 0.80 benchmark considered acceptable for formative decisions (Atai-Tabar et al., 2024). These figures imply that teachers can rely on the profile outputs to assign personalised remediation with minimal misclassification risk. The heat-map in Figure 4 vividly illustrates how two students with contrasting profiles can be targeted: Student S1 requires support only in Evaluation, whereas Student S2 lacks three of the four attributes, justifying a more foundational intervention trajectory.

Methodologically, the study highlights the feasibility of implementing Bayesian CDMs in rjags with modest class-size samples (n = 60). The sensitivity check switching from a non-informative Beta(1,1) to a mildly informative Beta(2,5) prior on *slip* yielded negligible parameter drift (< 0.02) and a virtually unchanged DIC. This robustness is noteworthy, given concerns about prior sensitivity in small-sample Bayesian estimation (Smid et al., 2020).

Limitations should be acknowledged. First, the analysis is confined to a single Indonesian high school, limiting external validity. Second, item parameters were simulated rather than empirically calibrated, potentially inflating precision. Future research should replicate the instrument across multiple schools and apply hierarchical CDMs that integrate student-level covariates (e.g., socio-economic status, prior achievement) to explain mastery variance. Despite these caveats, the present findings demonstrate that Bayesian G-DINA offers a reliable, fine-grained diagnostic lens for enhancing critical-thinking pedagogy in mathematics.

Implications for Policy and Curriculum

The findings underscore the need for policy instruments that incentivize diagnostic assessment beyond summative grading. National- or district-level guidelines could mandate formative reporting at the attribute level (interpretation, analysis, evaluation, inference) alongside total scores, supported by exemplar Q-matrices and item banks aligned to senior-secondary mathematics standards. Teacher professional learning should include micro-credential pathways on authoring CDM-ready items, interpreting posterior mastery profiles, and planning targeted remediation cycles. Curriculum designers can embed "diagnostic checkpoints" at unit boundaries, with rubrics that explicitly reference the four attributes and require teachers to document instructional responses. Assessment authorities can pilot lightweight Bayesian scoring pipelines for small classes, demonstrating feasible workflows for schools with limited data capacity.

Evidence from this study justifies integrating attribute-level feedback into report cards and school improvement plans, creating accountability for instructional use of diagnostic information.

Limitations and Generalizability

The study employed a single-school sample of sixty students, which constrains external validity and the stability of item-level parameter estimates. Posterior summaries in small samples remain sensitive to local item characteristics and classroom pedagogy, potentially inflating between-item variance in slip and guess. Generalizability improves when items are recalibrated across diverse schools and when hierarchical CDMs capture school-level heterogeneity. Future research should implement multi-site designs with stratified sampling across regions and track longitudinal change to estimate growth in attribute mastery. Cross-validation with parallel forms and differential item functioning checks will further bound transportability. Policymakers and practitioners should interpret the profiles as proof-of-concept diagnostics that warrant replication, not as population parameters. Scaling efforts ought to prioritize item pool expansion, rater moderation for open-ended tasks, and standard setting for attribute proficiency thresholds.

CONCLUSION

This study demonstrates the feasibility and instructional value of Bayesian Cognitive Diagnostic Modeling for senior-secondary mathematics by delivering attribute-level profiles that teachers can act upon in real time. The methodological contribution lies in a small-sample, classroom-ready workflow that integrates G-DINA estimation, convergence checking, and posteriopredictive validation, providing a transparent template for researchers and school-based assessors. The empirical contribution maps Indonesian learners' mastery patterns across interpretation, analysis, evaluation, and inference, offering a diagnostic lens unavailable from total scores.

For practice, schools can embed attribute-referenced "diagnostic checkpoints," align remediation to posterior mastery probabilities, and report progress at the attribute level alongside grades. For policy, districts and the national authority can standardize Q-matrix exemplars, establish micro-credentials for CDM item writing and interpretation, and pilot lightweight Bayesian scoring pipelines to support small classes. Future work should scale item banks, implement multi-site hierarchical calibration, and set defensible proficiency thresholds so that diagnostic feedback becomes a routine component of curriculum implementation and school improvement planning.

ACKNOWLEDGMENT

The authors express their sincere gratitude to the leadership, mathematics teachers, and students of SMAN 1 Kedungwuni, Pekalongan Regency, for their invaluable support and active participation in this research. Their cooperation in facilitating the diagnostic testing sessions and providing access to the necessary academic data greatly contributed to the successful completion of the study. The authors also acknowledge the constructive feedback from school administrators during the research planning phase, which ensured the contextual relevance and ethical compliance of the study procedures.

DISCLOSURE STATEMENT

The authors declare that there is no potential conflict of interest with respect to the research, authorship, and/or publication of this article. All procedures involving human participants were conducted in accordance with institutional and national research ethics guidelines. No financial or commercial support was received that could have influenced the outcomes or interpretations presented in this study.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The study was entirely self-funded by the authors.

ETHICS APPROVAL

The research protocol, including participant recruitment, informed consent procedures, and data handling, was reviewed and approved by the Institutional Ethics Committee of Universitas Pekalongan. Written informed consent was obtained from all student participants and their legal guardians, and formal permission was granted by the school authorities at SMAN 1 Kedungwuni. All procedures were conducted in accordance with the ethical standards outlined in the Declaration of Helsinki and relevant national regulations.

REFERENCES

Applebaum, M. (2024). Enhancing critical thinking in pre-service mathematics teachers: Bridging procedural fluency and conceptual understanding. *Математика Плюс*, *32*(3), 58–66. https://www.ceeol.com/search/article-detail?id=1276653

Atai-Tabar, M., Zareian, G., Amirian, S. M. R., & Adel, S. M. R. (2024). Relationships between EFL teachers' perceptions of consequential validity of formative assessment and data-driven decision-making self-efficacy and anxiety. *Journal of Applied Research in Higher Education*, *16*(3), 919–933. https://doi.org/10.1108/JARHE-04-2023-0169

Belzak, W. C. M. (2023). The multidimensionality of measurement bias in high-stakes testing: Using machine learning to evaluate complex sources of differential item functioning. *Educational Measurement: Issues and Practice*, *42*(1), 24–33. https://doi.org/10.1111/emip.12486

Cao, C., Lugu, B., & Li, J. (2024). The sensitivity of Bayesian fit indices to structural misspecification in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *31*(3), 477–493. https://doi.org/10.1080/10705511.2023.2253497

Cao, Y., Hong, S., Li, X., Ying, J., Ma, Y., Liang, H., Liu, Y., Yao, Z., Wang, X., Huang, D., Zhang, W., Huang, L., Chen, M., Hou, L., Sun, Q., Ma, X., Wu, Z., Kan, M.-Y., Lo, D., Zhang, Q., Ji, H., Jiang, J., Li, J., Sun, A., Huang, X., Chua, T.-S., & Jiang, Y.-G. (2025). Toward generalizable evaluation in the LLM era: A survey beyond benchmarks. *ArXiv Preprint ArXiv:2504.18838*. https://doi.org/10.48550/arXiv.2504.18838

Choo, S., Park, S., & Nelson, N. J. (2021). Evaluating spatial thinking ability using item response theory: Differential item functioning across math learning disabilities and geometry instructions. *Learning Disability Quarterly*, *44*(2), 68–81. https://doi.org/10.1177/0731948720912417

Davenport, J. L., Kao, Y. S., Matlen, B. J., & Schneider, S. A. (2020). Cognition research in practice: Engineering and evaluating a middle school math curriculum. *The Journal of Experimental Education*, *88*(4), 516–535. https://doi.org/10.1080/00220973.2019.1619067

Gamino, J. F., Frost, C., Riddle, R., Koslovsky, J., & Chapman, S. B. (2022). Higher-order executive function in middle school: Training teachers to enhance cognition in young adolescents. *Frontiers in Psychology*, *13*, 867264. https://doi.org/10.3389/fpsyg.2022.867264

Gao, Y., Zhai, X., Bae, A., & Ma, W. (2023). Rasch-CDM: A combination of Rasch and Cognitive Diagnosis models to assess a learning progression. In X. Liu & W. Boone (Eds.), *Advances in applications of Rasch measurement in science education*. Springer Nature. http://dx.doi.org/10.2139/ssrn.4345437

Garcia, M. B. (2025). Profiling the skill mastery of introductory programming students: A cognitive diagnostic modeling approach. *Education and Information Technologies*, *30*(5), 6455–6481. https://doi.org/10.1007/s10639-024-13039-6

Go, M. C. J. (2023). Enhancing mathematical proficiency assessment: Insights from mathematics teachers. *Science International*, *35*(6), 773–780. https://scholar.google.com/scholar_lookup?title=Enhancing%20mathematical%20profici

ency%20assessment%3A%20Insights%20from%20mathematics%20teachers&publication_year=2023&author=M.%20Go

Molerov, D., Zlatkin-Troitschanskaia, O., Nagel, M.-T., Brückner, S., Schmidt, S., & Shavelson, R. J. (2020). Assessing university students' critical online reasoning ability: A conceptual and assessment framework with preliminary evidence. *Frontiers in Education*, *5*, 577843. https://doi.org/10.3389/feduc.2020.577843

Nitz, L., Gurabi, M. A., Cermak, M., Zadnik, M., Karpuk, D., Drichel, A., Schäfer, S., Holmes, B., & Mandal, A. (2025). On collaboration and automation in the context of threat detection and response with privacy-preserving features. *Digital Threats: Research and Practice*, *6*(1), 1–36. https://doi.org/10.1145/3707651

Ntumi, S., Agbenyo, S., & Bulala, T. (2023). Estimating the psychometric properties (item difficulty, discrimination and reliability indices) of test items using Kuder-Richardson approach (KR-20). *Shanlax International Journal of Education*, *11*(3), 18–28. https://doi.org/10.34293/education.v11i3.6081

Overton, C. (2023). *A practitioner inquiry to examine text selection practices for secondary students with learning disabilities*. Doctoral Dissertation, Indiana University. https://hdl.handle.net/2022/29509

Pohl, C., Klein, J. T., Hoffmann, S., Mitchell, C., & Fam, D. (2021). Conceptualising transdisciplinary integration as a multidimensional interactive process. *Environmental Science & Policy*, *118*, 18–26. https://doi.org/10.1016/j.envsci.2020.12.005

Pokropek, A., Marks, G. N., Borgonovi, F., Koc, P., & Greiff, S. (2022). General or specific abilities? Evidence from 33 countries participating in the PISA assessments. *Intelligence*, *92*, 101653. https://doi.org/10.1016/j.intell.2022.101653

Rojas, E., & Benakli, N. (2020). Mathematical literacy and critical thinking. In J. C. But (Ed.), *Teaching college-level disciplinary literacy: Strategies and practices in STEM and professional studies* (pp. 197–226). Palgrave Macmillan Cham. https://doi.org/10.1007/978-3-030-39804-0

Rustam, R., & Priyanto, P. (2022). Critical thinking assessment in the teaching of writing Indonesian scientific texts in high school. *Jurnal Penelitian dan Evaluasi Pendidikan*, *26*(1), 12–25. https://doi.org/10.21831/pep.v26i1.36241

Schad, D. J., Betancourt, M., & Vasishth, S. (2021). Toward a principled Bayesian workflow in cognitive science. *Psychological Methods*, *26*(1), 103-126. https://psycnet.apa.org/doi/10.1037/met0000275

Smid, S. C., McNeish, D., Miočević, M., & van de Schoot, R. (2020). Bayesian versus frequentist estimation for structural equation models in small sample contexts: A systematic review. *Structural Equation Modeling: A Multidisciplinary Journal*, *27*(1), 131–161. https://doi.org/10.1080/10705511.2019.1577140

Sun, C., Shute, V. J., Stewart, A., Yonehiro, J., Duran, N., & D'Mello, S. (2020). Towards a generalized competency model of collaborative problem solving. *Computers & Education*, *143*, 103672. https://doi.org/10.1016/j.compedu.2019.103672

Tanudjaya, C. P., & Doorman, M. (2020). Examining higher order thinking in Indonesian lower secondary mathematics classrooms. *Journal on Mathematics Education*, *11*(2), 277–300. https://research-portal.uu.nl/en/publications/examining-higher-order-thinking-in-indonesian-lower-secondary-mat

Ufer, S., & Bochnik, K. (2020). The role of general and subject-specific language skills when learning mathematics in elementary school. *Journal Für Mathematik-Didaktik*, *41*(1), 81–117. https://doi.org/10.1007/s13138-020-00160-5

Vasishth, S., Yadav, H., Schad, D. J., & Nicenboim, B. (2023). Sample size determination for Bayesian hierarchical models commonly used in psycholinguistics. *Computational Brain & Behavior*, *6*(1), 102–126. https://doi.org/10.1007/s42113-021-00125-y

Waller, B. N. (2023). *Critical thinking: Consider the verdict* (7th ed.). Waveland Press. https://www.waveland.com/browse.php?t=771&pgtitle=Bruce%20N.%20Waller

Wang, X., Pan, J., Ren, Z., Zhai, M., Zhang, Z., Ren, H., Song, W., He, Y., Li, C., Yang, X., Li, M., Quan, D., Chen, L., & Qiu, L. (2021). Application of a novel hybrid algorithm of Bayesian network in the study of hyperlipidemia related factors: A cross-sectional study. *BMC Public Health*, *21*, 1375. https://doi.org/10.1186/s12889-021-11412-5

Wu, H., & Molnár, G. (2022). Analysing complex problem-solving strategies from a cognitive perspective: The role of thinking skills. *Journal of Intelligence*, *10*(3), 46. https://doi.org/10.3390/jintelligence10030046

Xin, T., Wang, C., Chen, P., & Liu, Y. (2022). Cognitive diagnostic models: Methods for practical applications. *Frontiers in Psychology,* 13, 895399. https://doi.org/10.3389/fpsyg.2022.895399

Yamaguchi, K., & Okada, K. (2020). Variational Bayes inference for the DINA model. *Journal of Educational and Behavioral Statistics*, *45*(5), 569–597. https://doi.org/10.3102/1076998620911934

Zhai, C., Wibowo, S., & Li, L. D. (2024). The effects of over-reliance on AI dialogue systems on students' cognitive abilities: A systematic review. *Smart Learning Environments*, *11*(1), 28. https://doi.org/10.1186/s40561-024-00316-7

Zhang, Z., Zhang, J., Lu, J., & Tao, J. (2020). Bayesian estimation of the DINA model with Pólya-Gamma Gibbs sampling. *Frontiers in Psychology*, *11*, 384. https://doi.org/10.3389/fpsyg.2020.00384