



Monte Carlo Simulation of Critical-Thinking Assessment Categories from University Presentation Records

Diva Amanda Putri^{1*}, Firdaus Annas²

^{1,2}Universitas Islam Negeri Sjech M. Djamil Djambek Bukittinggi, Bukittinggi, Indonesia

Article Information

Article History:

Submitted: July 6, 2026

Revised: October 28, 2026

Accepted: October 28, 2026

Published: October 31, 2026

Keywords

Monte Carlo simulation, critical-thinking assessment, presentation records, learning analytics, historical distribution

Correspondence

E-mail: divaamandaputri11@gmail.com*

A B S T R A C T

Critical-thinking performance is difficult to infer directly from presentation grades, but historical grade categories can be modelled as an aggregate distribution. This study applied a quantitative descriptive-analytical Monte Carlo approach to 474 eligible presentation records from a population of 537 Informatics and Computer Engineering Education students in the 2023 and 2024 cohorts at UIN Sjech M. Djamil Djambek Bukittinggi. Observed counts were 241 high, 207 moderate, and 26 low. The reported 474-draw simulation produced 241, 213, and 20, respectively. Absolute count differences were 0, 6, and 6, giving MAE = 4. The aggregate agreement calculated from the manuscript's stated formula was 97.47%, while the previously reported 91.34% is the unweighted mean of category-specific agreements. Retrospective reconstruction of the sample pseudorandom sequence suggested a candidate linear congruential generator with a 127-value period; its actual parameters require confirmation. Because the same records informed both simulation probabilities and evaluation, these figures describe internal distributional agreement rather than accuracy on new cohorts. The study contributes a context-specific demonstration of category simulation and highlights the importance of instrument validity, generator transparency, and independent validation.

Abstrak

Kinerja berpikir kritis sulit disimpulkan secara langsung dari nilai presentasi, namun kategori nilai historis dapat dimodelkan sebagai distribusi agregat. Studi ini menerapkan pendekatan kuantitatif deskriptif-analitis berbasis metode Monte Carlo terhadap 474 catatan presentasi yang memenuhi syarat, yang diambil dari populasi 537 mahasiswa Pendidikan Teknik Informatika dan Komputer angkatan 2023 dan 2024 di UIN Sjech M. Djamil Djambek Bukittinggi. Hasil observasi menunjukkan 241 data berkategori tinggi, 207 sedang, dan 26 rendah. Simulasi dengan 474 kali pengambilan sampel menghasilkan masing-masing 241, 213, dan 20 data untuk kategori tersebut. Selisih jumlah absolut adalah 0, 6, dan 6, yang menghasilkan nilai MAE sebesar 4. Tingkat kesesuaian agregat yang dihitung berdasarkan rumus dalam naskah adalah 97,47%, sementara angka 91,34% yang dilaporkan sebelumnya merupakan rerata tidak berbobot dari tingkat kesesuaian untuk setiap kategori. Rekonstruksi retrospektif terhadap urutan semu acak (pseudorandom) dari sampel tersebut mengindikasikan penggunaan *linear congruential generator* dengan periode 127 nilai; parameter aktualnya masih memerlukan konfirmasi. Karena data yang sama digunakan baik untuk menentukan probabilitas simulasi maupun evaluasi, angka-angka ini lebih mencerminkan kesesuaian distribusi internal daripada akurasi untuk kelompok mahasiswa yang baru. Studi ini memberikan demonstrasi simulasi kategori dalam konteks tertentu serta menekankan pentingnya validitas instrumen, transparansi generator, dan validasi independen.

This is an open access article under the CC-BY-SA license



1. Introduction

Critical thinking is widely regarded as an important outcome of higher education, but it is not a single, directly observable attribute. Performance-based assessments distinguish interpretation, analysis, evaluation, and inference from students' dispositions and task-specific performance. Recent syntheses have identified substantial variation in the constructs, assessment formats, and theoretical

frameworks used to measure these abilities [1], [2]. Cross-national performance-based evidence demonstrates the feasibility of assessing such competencies while highlighting variation across contexts [3]. Studies of instrument structure and test-taking behaviour further show why scores must be interpreted in light of assessment design and measurement conditions [4], [5].

Presentations and classroom discussions may reveal how students explain evidence, defend arguments, respond to questions, and formulate conclusions. Nevertheless, a presentation grade is not automatically a validated measure of critical thinking; scoring may also reflect communication, preparation, and discipline-specific expectations. Recent studies of teacher ratings and learning discourse [6], [7] and reviews of authentic assessment [8] emphasise the need for clear indicators and scoring rules. Institutional and review-based accounts likewise caution against treating administrative grades as interchangeable with validated competence measures [9], [10]. The present investigation therefore regards the available presentation records as a proxy rather than as individual diagnoses of critical-thinking ability.

Learning analytics increasingly uses institutional records to describe learning patterns and support decisions. Reviews of higher education analytics document both potential uses and persistent problems of data quality, pedagogical alignment, and transfer into practice [11]–[15]. Research on engagement, distance learning, and self-regulated learning illustrates that indicators extracted from educational records require explicit operational definitions and careful interpretation [16]–[19]. These findings are relevant to presentation assessments because aggregated scores can be informative for monitoring a cohort while remaining insufficient to establish why individual students performed as they did.

Prediction studies go a step further by using historical observations to estimate outcomes for subsequently observed students or periods. Contemporary learning analytics frameworks integrate prediction with educational action [20], [21], while educational data-mining studies demonstrate the importance of input features, sample composition, and evaluation design [22]–[26]. Such studies usually distinguish modelling data from evaluation data. A simulation generated from a distribution and then compared with the very records used to define that distribution should instead be described as an internal distributional check, not as externally validated forecasting.

The methodological gap addressed in this paper is deliberately narrow. School- and university-level studies have modelled student performance through a range of data-driven approaches, yet published reviews emphasise that educational relevance, interpretability, and validation must accompany the numerical output [27]–[29]. This study demonstrates how historical presentation-score categories can be translated into empirical probabilities, cumulative intervals, and a Monte Carlo-generated category distribution for the Informatics and Computer Engineering Education programme (PTIK) at UIN Sjech M. Djamil Djambek Bukittinggi. Its contribution is a transparent, context-specific assessment of distributional agreement and methodological limitations, not a claim of superior forecasting or individual-level classification.

The guiding question is whether a single Monte Carlo draw, based on empirical category probabilities, reproduces the frequency distribution observed in the historical records and how that agreement should be quantified. The article separates the verified descriptive quantities from limitations that cannot be resolved without the underlying assessment rubric, generator configuration, and independent evaluation data. No additional data collection or external validation is represented as having occurred.

2. Method

2.1. Research Design

This study applied a quantitative descriptive-analytical design using a discrete Monte Carlo simulation. The unit of modelling was the frequency of three presentation-score categories among a

cohort of university students. The modelling workflow comprised empirical frequency estimation, cumulative probability intervals, pseudorandom number generation, category assignment, and comparison with the original frequency distribution. Stochastic modelling has precedent in research on university academic performance [30], and educational applications of Monte Carlo methods emphasise transparent assumptions, adequate numbers of draws, and sensitivity to simulation settings [31]. This design did not include a held-out cohort, a prospective outcome period, or an individual diagnostic test.

2.2. Research Setting and Data Source

The study drew on secondary records from presentation assessments in courses offered within the Informatics and Computer Engineering Education Study Programme (PTIK), Faculty of Tarbiyah and Teacher Training, UIN Sjech M. Djamil Djambek Bukittinggi. Research activities were reported for May–July 2026. Lecturers’ presentation-assessment records for students from the 2023 and 2024 cohorts were consolidated for category-level analysis. The exact number of records by course and cohort, the dates on which the presentations were assessed, and the procedure for resolving multiple presentation scores for one student were not specified in the available records; consequently, equivalence across courses cannot be assumed.

2.3. Population and Sample

The stated population comprised 537 PTIK students in the 2023 and 2024 cohorts. A purposively selected set of 474 student records met the study’s data-availability criteria and was included in the simulation. The unit of analysis was the consolidated student-level assessment record, as described by the authors; the method for consolidating multiple course scores, if present, requires confirmation.

Records were eligible when they concerned a member of either cohort, included a presentation assessment from a relevant course, and could be compiled sufficiently for category analysis. The remaining 63 students were not represented in the analytical dataset because their records did not meet the stated requirements. More specific reasons for exclusion and a comparison of included and excluded students were not documented, so the risk of selection bias and the representativeness of the 474 records cannot be established.

2.4. Critical-Thinking Assessment Indicators

The original presentation scoring was said to draw on four indicators associated with critical thinking: interpretation, analysis, evaluation, and inference. Presentations may provide opportunities to demonstrate these behaviours, but the available document does not reproduce the scoring rubric or describe how much each indicator contributed to the final grade. The term “critical-thinking assessment category” is therefore used cautiously for a category derived from presentation records, not as evidence of independently validated critical-thinking proficiency.

Interpretation concerns understanding and explaining information; analysis examines relationships among claims, reasons, and evidence; evaluation concerns assessing credibility and argument quality; and inference involves drawing supported conclusions. The study did not report item-level scores or inter-rater agreement for these dimensions. It is not possible to separate the contribution of the four indicators from presentation delivery or course-specific grading practices.

The secondary scores were processed only as aggregated category counts. The simulation did not classify unidentified students into new individual categories and did not establish an individual diagnostic instrument. This distinction is essential because a distributional reproduction and an assessment of individual critical-thinking competence are different methodological objectives.

Three predefined score intervals were used: high (80–100), moderate (60–79), and low (0–59). These cut-offs were applied before simulation. The original manuscript did not provide the official institutional rubric or a validation study supporting those boundaries; accordingly, their status is

treated as an operational categorisation of the recorded grades rather than as an externally validated standard of critical thinking.

2.5. Data Processing Procedure

The analytical sequence was: compiling historical presentation scores, assigning the predefined categories, calculating category frequencies and probabilities, accumulating probabilities, mapping pseudorandom numbers onto the resulting intervals, and comparing the simulated and observed frequency distributions. Figure 1 retains the original workflow diagram. The simulation describes potential category frequencies under the empirical distribution; it does not incorporate an estimated transition or causal mechanism for changes in student ability.

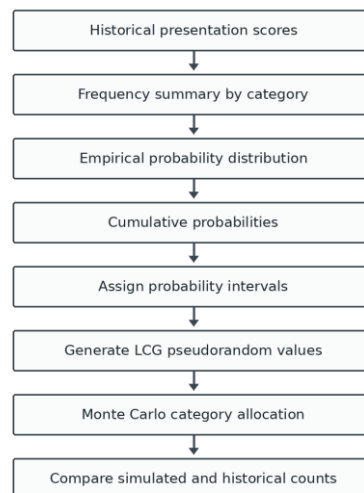


Figure 1. Workflow for the Monte Carlo distribution simulation

The empirical probability of category i was calculated from its observed frequency and the total number of included records (Equation 1).

$$p_i = \frac{f_i}{n} \quad (1)$$

In Equation 1, p_i denotes the probability of category i , f_i is its historical frequency, and $n = 474$ is the total number of eligible records. The same empirical probabilities were subsequently used to construct the cumulative intervals.

2.6. Monte Carlo Simulation Procedure

A linear congruential generator (LCG) was reported as the source of pseudorandom values. Its standard recurrence is shown in Equation 2; a generated integer z_i is normalised to $r_i = z_i/m$ in Equation 3. The original methods did not state the multiplier a , increment c , modulus m , or initial seed. Generator selection matters because pseudorandom sequences are deterministic and their period and statistical quality can influence simulation outputs [32], [33].

Pseudorandom integers were described as being generated using a linear congruential recurrence (Equation 2).

$$z_i = (az_{i-1} + c) \bmod m \quad (2)$$

In Equation 2, z_i is the generated integer at iteration i , z_{i-1} is the previous integer, a is the multiplier, c is the increment, and m is the modulus. The parameters actually selected by the authors were not documented. A retrospective reconstruction from the 20 displayed integers is consistent with

$a = 21$, $c = 13$, $m = 509$, and $z_0 = 125$; this is an audit inference rather than a confirmed author-reported configuration. ^{$Z_i Z_i$}

The generated integer was normalised to the interval $[0, 1)$ according to Equation 3.

$$R_i = \frac{Z_i}{m_{(3)}}$$

In Equation 3, r_i denotes the normalised value, z_i the generated integer, and m the modulus. Under the reconstructed candidate settings, the 20 normalised values shown in the original results correspond to $z_i/509$.

Each normalised value was assigned to one mutually exclusive cumulative-probability interval. This procedure was reported for 474 draws. The candidate LCG settings reconstructed from the example have a period of 127, substantially shorter than 474 draws, so the candidate sequence repeats within the reported simulation. The reconstruction reproduces the category frequencies in Table 4, but the actual implementation settings and whether other seeds or generators were examined must be confirmed by the authors. ^{R_i}

2.7. Model Validation

The authors compared simulated category frequencies with the same historical records used to estimate the model probabilities. This “historical data validation” measures in-sample distributional agreement only. It does not provide an out-of-sample test, a future-cohort forecast, or evidence that the simulation predicts individual critical-thinking outcomes.

For the three categories, the mean absolute error (MAE) was calculated as shown in Equation 4.

$$MAE = \frac{1}{K} \sum_{i=1}^k |H_i - S_i| \quad (4)$$

Here, H_i is the historical count in category i , S_i is the simulated count, and $K = 3$ is the number of categories. Absolute differences of 0, 6, and 6 produce $MAE = 12/3 = 4$ students per category.

The originally specified overall distributional-agreement formula is shown in Equation 5.

$$Akurasi = \left(1 - \frac{\sum_{i=1}^k |H_i - S_i|}{\sum_{i=1}^k H_i} \right) \times 100\% \quad (5)$$

With the observed counts, Equation 5 gives $100 \times (1 - 12/474) = 97.47\%$. The previously reported figure of 91.34% is obtained from a different, unweighted mean of category-specific agreements: $[100\% + (1 - 6/207) \times 100\% + (1 - 6/26) \times 100\%]/3 = 91.34\%$. These are different descriptive measures: 97.47% weights discrepancies by the total number of observations, whereas 91.34% weights the three categories equally.

Neither agreement index establishes prospective predictive accuracy because the historical distribution served both as model input and benchmark. A smaller MAE reflects closer simulated counts to the source counts for this particular run, but no between-run variability, confidence interval, or independent validation set was reported.

Expert judgment was additionally reported as 93.33% for the modelling approach. However, the number of experts, their evaluation criteria, individual ratings, and calculation were not available in the manuscript. The reported percentage is retained as an author-reported appraisal rather than independently verified evidence of computational validity.

2.8. Ethical Considerations

The analysed material consisted of pre-existing student assessment records obtained through coordination with the lecturers or institutional units responsible for them. The manuscript reports institutional permission and limits analysis to aggregate counts. The exact authorisation arrangements, data storage safeguards, access rules, and record-retention period were not specified.

Student names and identifying attributes were excluded from the displayed analysis. The simulation was not used to assign public labels or diagnoses to individual students. Access to any underlying data should remain subject to the institution's confidentiality requirements and appropriate anonymisation or de-identification procedures.

The use of secondary educational records requires clear accountability for permission, confidentiality, and permitted secondary uses. The available documentation supports the stated use of aggregated results but does not independently establish approval by a formal ethics committee; no such approval is claimed here.

3. Results and Discussion

3.1. Results

The results comprise the historical category distribution, its cumulative-probability representation, an example pseudorandom sequence, one simulated category distribution, and a comparison of observed and simulated frequencies. These findings describe a single reported run and should be distinguished from an externally evaluated forecasting model.

3.1.1. Historical Distribution of Presentation Assessment Scores

A total of 474 eligible student records were classified into high, moderate, and low presentation-score categories. No individual identifiers are reported. The historical distribution forms both the empirical input to the Monte Carlo model and the benchmark against which its output was compared.

The observed frequencies were 241 (high), 207 (moderate), and 26 (low). These correspond to 50.84%, 43.67%, and 5.49%, respectively. High-category scores accounted for the largest share of the available presentation records.

Table 1. Historical distribution of presentation-score categories

No.	Category	Frequency	Percentage
1	High	241	50.84%
2	Moderate	207	43.67%
3	Low	26	5.49%
	Total	474	100.00%

3.1.2. Empirical Probabilities and Cumulative Intervals

The probabilities estimated from the observed counts were approximately 0.5084 (high), 0.4367 (moderate), and 0.0549 (low). The cumulative boundaries were 0.5084 and 0.9451; values were allocated to the three categories using a non-overlapping interval convention.

High was assigned to $0 \leq r_i < 0.5084$, moderate to $0.5084 \leq r_i < 0.9451$, and low to $0.9451 \leq r_i < 1$. The rounded boundaries are display approximations; classification at boundary values must follow a consistent rule to prevent overlap.

Table 2. Empirical probabilities, cumulative probabilities, and intervals

No.	Category	Frequency	Probability	Cumulative probability	Interval
1	High	241	0.5084	0.5084	$0 \leq r_i < 0.5084$
2	Moderate	207	0.4367	0.9451	$0.5084 \leq r_i < 0.9451$
3	Low	26	0.0549	1.0000	$0.9451 \leq r_i < 1$
	Total	474	1		

3.1.3. Pseudorandom Sequence and Category Assignment

Table 3 retains the first 20 generated integers and normalised values reported in the source. Dividing these integers by 509 reproduces the six-decimal normalised values, and the resulting categories agree with the cumulative intervals.

The original manuscript described 474 simulated draws but did not provide all 474 generated values or the parameter set. Retrospective reconstruction with $a = 21$, $c = 13$, $m = 509$, and $z_0 = 125$ reproduces the 20 examples and the reported 474-draw category totals. Since this candidate generator has period 127, these findings also signal potential repetition rather than 474 distinct draws. The reconstruction is an audit of reported numbers, not evidence of an independent replication study conducted by the authors.

Table 3. Examples of LCG output and assigned categories

Draw	z_i	r_i	Probability interval	Category
z1	93	0.182711	$0 \leq r_i < 0.5084$	High
z2	439	0.862475	$0.5084 \leq r_i < 0.9451$	Moderate
z3	70	0.137525	$0 \leq r_i < 0.5084$	High
z4	465	0.913556	$0.5084 \leq r_i < 0.9451$	Moderate
z5	107	0.210216	$0 \leq r_i < 0.5084$	High
z6	224	0.440079	$0 \leq r_i < 0.5084$	High
z7	136	0.267191	$0 \leq r_i < 0.5084$	High
z8	324	0.636542	$0.5084 \leq r_i < 0.9451$	Moderate
z9	200	0.392927	$0 \leq r_i < 0.5084$	High
z10	141	0.277014	$0 \leq r_i < 0.5084$	High
z11	429	0.842829	$0.5084 \leq r_i < 0.9451$	Moderate
z12	369	0.724951	$0.5084 \leq r_i < 0.9451$	Moderate
z13	127	0.249509	$0 \leq r_i < 0.5084$	High
z14	135	0.265226	$0 \leq r_i < 0.5084$	High
z15	303	0.595285	$0.5084 \leq r_i < 0.9451$	Moderate
z16	268	0.526523	$0.5084 \leq r_i < 0.9451$	Moderate
z17	42	0.082515	$0 \leq r_i < 0.5084$	High
z18	386	0.75835	$0.5084 \leq r_i < 0.9451$	Moderate
z19	484	0.950884	$0.9451 \leq r_i < 1$	Low
z20	506	0.994106	$0.9451 \leq r_i < 1$	Low

For example, $z_1 = 93$ gives $r_1 = 93/509 = 0.182711$ and maps to the high category, while $z_2 = 439$ gives $r_2 = 0.862475$ and maps to moderate. This mapping illustrates the algorithmic procedure without establishing the randomness quality of the complete simulated sequence.

The single reconstructed series extends to 474 iterations and matches the displayed aggregate output. No uncertainty interval, independently seeded replication, or comparison with a higher-period generator was reported.

3.1.4. Distribution Produced by the Monte Carlo Run

The reported simulation allocated 241 values to the high category, 213 to the moderate category, and 20 to the low category. These counts describe one realisation generated from probabilities fitted to the 474 historical records.

Table 4. Category frequencies in the reported Monte Carlo run

No.	Category	Simulated frequency	Percentage
1	High	241	50.84%
2	Moderate	213	44.94%
3	Low	20	4.22%
	Total	474	100.00%

The corresponding simulated proportions were 50.84% (high), 44.94% (moderate), and 4.22% (low). The totals sum to 474 and 100%, respectively. They should not be interpreted as observed outcomes of a new cohort or as classifications assigned to specific students.

3.1.5. Comparison of Historical and Simulated Frequencies

Table 5 compares the empirical counts with the reported simulation. The high category coincided exactly, while the simulated moderate category exceeded the observed count by six and the low category fell short by six.

Table 5. Historical and simulated category frequencies

No.	Category	Historical count	Simulated count
1	High	241	241
2	Moderate	207	213
3	Low	26	20
	Total	474	474

Figure 2 illustrates the same three-category comparison using a grouped bar chart. The chart is descriptive rather than an independent confirmation of forecasting accuracy.

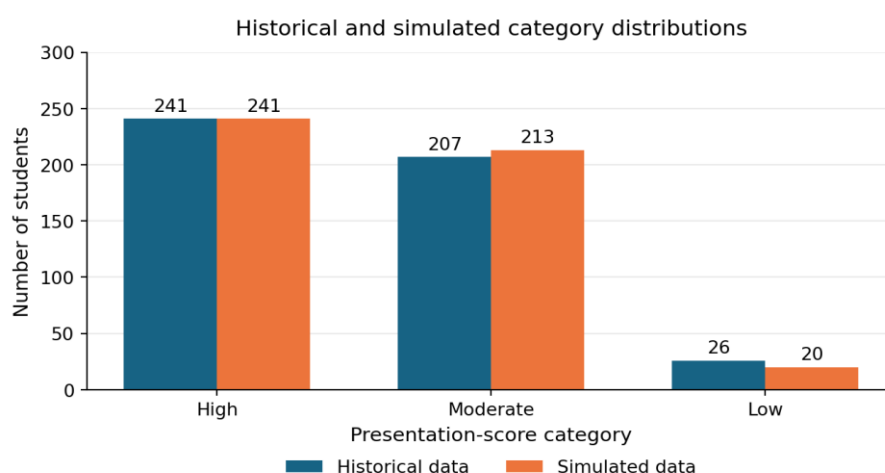


Figure 2. Comparison of historical and simulated presentation-score categories

The absolute differences were 0, 6, and 6, yielding a total absolute difference of 12. Because the simulation was conditioned on these same historical frequencies, close agreement is expected in principle; whether the degree of agreement is typical for repeated simulations was not tested.

3.1.6. Historical Distributional Agreement

Historical data validation was operationalised as a comparison of category frequencies in the simulated series with those observed in the original dataset. The calculation was an internal descriptive consistency check and not a performance evaluation against previously unseen records.

Table 6. Historical distributional agreement and error summary

No.	Category	Historical count	Simulated count	Absolute difference
1	High	241	241	0
2	Moderate	207	213	6
3	Low	26	20	6
	Total	474	474	12
	MAE			4
	Aggregate agreement (Eq. 5)			97.47%

As shown in Table 6, the differences were 0 for high, 6 for moderate, and 6 for low, and their mean absolute value was 4. The aggregate-count agreement defined by Equation 5 is 97.47%; the unweighted mean of category-specific agreements is 91.34%. Reporting both values with their distinct denominators resolves the numerical inconsistency of the original manuscript. Neither percentage is an estimate of future-cohort predictive accuracy.

3.2. Discussion

The reproduced distribution largely reflects the historical probabilities used to configure the model. The reported 241/213/20 allocation is close to the historical 241/207/26 allocation, but a close fit is not by itself a new substantive finding about students' cognitive development. Reviews of learning analytics caution that useful educational interpretations require an explicit connection between the underlying construct, the available data, and the decisions supported [11], [14]. Here, the safest interpretation is the behaviour of a category-frequency simulation based on an institution-specific grading record.

The difference between the two agreement indices is important. MAE = 4 is a mean absolute count difference across three categories, 97.47% is an aggregate-count agreement measure, and 91.34% gives equal weight to each category's relative discrepancy. In particular, the difference of six records in the small low category has a much larger proportional impact than the same difference in the moderate category. Reporting only the unweighted mean as "prediction accuracy" would conceal both its definition and the fact that the same source data were used twice.

An earlier Indonesian application of Monte Carlo simulation to student-understanding categories also relied on historical educational records [34], illustrating the relevance of stochastic category modelling beyond business applications. By contrast, work on predictive analytics in educational performance has used multiple analytical approaches and explicit evaluation strategies [35], [36]. Accordingly, the present work should be positioned as a transparent distribution-modelling demonstration rather than as evidence of superior predictive performance.

From a learning-analytics perspective, aggregated presentation records may help instructors inspect the balance of operational score categories. They cannot, without a validated rubric, be equated with a complete profile of critical-thinking skills. Investigations of students' and teachers' conceptions of critical thinking highlight how task context shapes such judgments [37]. Research on simulation-based collaborative learning similarly distinguishes traceable indicators from claims about underlying learning processes [38]. A cohort-level count may therefore support exploratory monitoring, but it is not a substitute for detailed cognitive assessment.

The study's contribution is a reproducible account of how a categorical empirical distribution can be used for a Monte Carlo run and how differing agreement metrics change interpretation. It does not establish that Monte Carlo is more accurate than a direct display of empirical proportions. Recent investigations of critical-thinking assessment and predictive monitoring show that changes in scores, learning outcomes, and intervention effects need their own evidence [39], [40]. More defensible future evaluations would use held-out periods, a baseline such as empirical proportions, sufficiently long-period random number generation, and multiple independent runs.

Limitations include records from one programme, unspecified cross-course grading comparability, 63 non-included students, uncertain rubric validity, missing author-reported LCG settings, a candidate generator period shorter than the draw count, and the absence of prospective validation and replication intervals. The expert appraisal percentage of 93.33% is not independently auditable from the submitted documentation. These constraints preclude individual diagnosis, causal educational inference, and claims of generalisable predictive accuracy.

4. Conclusion

This study organised 474 historical university presentation-assessment records into three predefined score categories and applied a Monte Carlo distribution simulation. The observed counts were 241 high, 207 moderate, and 26 low; the reported simulated counts were 241, 213, and 20. The resulting MAE was 4 students per category. The aggregate distributional agreement under the displayed formula was 97.47%, whereas an unweighted average of category-specific agreements was 91.34%. These describe two distinct measures of within-dataset agreement, not independently validated fore-

casting accuracy. Retrospective reconstruction of the reported pseudorandom sequence also indicated a potentially short generator period that requires author confirmation. The practical contribution is a transparent example of aggregate category modelling and of the reporting checks it requires. Because the scores are only a proxy for critical thinking and the model was not tested on new cohorts, future work should validate the assessment rubric, document algorithmic settings, and test generalisation using independent data.

Declarations

Author Contributions

Diva Amanda Putri contributed to conceptualisation, methodology, data acquisition, analysis, simulation, data visualisation, and the original draft. Firdaus Annas contributed to supervision, methodological review, critical revision, and final approval of the manuscript. The authors remain responsible for verifying these contribution statements.

Funding

The authors report that the research received no specific external funding and that its costs were covered independently.

Data Availability

The manuscript reports only aggregated category counts and selected pseudorandom-number outputs. Availability of anonymised individual assessment records, complete simulation sequences, and the implementation code is subject to author confirmation and institutional data-protection requirements. These materials should not be described as publicly available without a verified access route.

Conflicts of Interest

The authors report no financial or nonfinancial conflicts of interest affecting the study or its interpretation.

Declaration of Generative AI Use

The authors must confirm the tools and purposes, if any, of generative AI use in preparing this manuscript before submission. No single disclosure could be established from the mutually inconsistent alternatives in the submitted document.

Additional Information

Requests concerning the study or supporting materials may be addressed to Diva Amanda Putri and Firdaus Annas through the correspondence details provided above.

ORCID

[Diva Amanda Putri], <https://orcid.org/0009-0006-3054-8490>

[Firdaus Annas], <https://orcid.org/0009-0007-5585-1100>

References

- [1] H. Hyytinen, M. Jämsä, T. Tuononen, and K. Kleemola, "A systematic-narrative review of performance-based assessments of critical thinking in higher education," *Assessment & Evaluation in Higher Education*, vol. 50, no. 8, pp. 1293–1310, 2025, doi: [10.1080/02602938.2025.2553341](https://doi.org/10.1080/02602938.2025.2553341).
- [2] P. Andreucci-Annunziata et al., "Conceptualizations and instructional strategies on critical thinking in higher education: A systematic review of systematic reviews," *Frontiers in Education*, vol. 8, art. no. 1141686, 2023, doi: [10.3389/feduc.2023.1141686](https://doi.org/10.3389/feduc.2023.1141686).
- [3] D. Van Damme, D. Zahner, O. Cortellini, T. Dawber, and K. Rotholz, "Assessing and developing critical-thinking skills in higher education," *European Journal of Education*, vol. 58, no. 3, pp. 369–386, 2023, doi: [10.1111/ejed.12563](https://doi.org/10.1111/ejed.12563).
- [4] K. Kleemola, H. Hyytinen, and A. Toom, "Exploring internal structure of a performance-based critical thinking assessment for new students in higher education," *Assessment & Evaluation in Higher Education*, vol. 47, no. 4, pp. 556–569, 2022, doi: [10.1080/02602938.2021.1946482](https://doi.org/10.1080/02602938.2021.1946482).

- [5] H. Hyytinen, K. Nissinen, K. Kleemola, J. Ursin, and A. Toom, "How do self-regulation and effort in test-taking contribute to undergraduate students' critical thinking performance?," *Studies in Higher Education*, vol. 49, no. 1, pp. 192–205, 2024, doi: [10.1080/03075079.2023.2227207](https://doi.org/10.1080/03075079.2023.2227207).
- [6] H. Galindo-Domínguez, M.-J. Bezanilla, L. Campo, D. Fernández-Nogueira, and M. Poblete, "A teachers' based approach to assessing the perception of critical thinking in Education university students based on their age and gender," *Frontiers in Education*, vol. 8, art. no. 1127705, 2023, doi: [10.3389/educ.2023.1127705](https://doi.org/10.3389/educ.2023.1127705).
- [7] A. Ostendorf and E. Kaczak, "Critical thinking in the community of inquiry framework: An analysis of the theoretical model and cognitive presence coding schemes," *Computers & Education*, vol. 193, art. no. 104662, 2023, doi: [10.1016/j.compedu.2022.104662](https://doi.org/10.1016/j.compedu.2022.104662).
- [8] D. Vlachopoulos and A. Makri, "A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations," *Studies in Educational Evaluation*, vol. 83, art. no. 101425, 2024, doi: [10.1016/j.stueduc.2024.101425](https://doi.org/10.1016/j.stueduc.2024.101425).
- [9] O. Jabali, B. Hamamra, and A. Ayyoub, "Critical thinking, assessment, and educational policy in Palestinian universities," *International Journal for Educational Integrity*, vol. 20, art. no. 12, 2024, doi: [10.1007/s40979-024-00160-9](https://doi.org/10.1007/s40979-024-00160-9).
- [10] T. Li, Z. Zhan, Z. Lin, and X. Zhang, "Assessment of higher-order thinking in education – A systematic review of empirical studies," *Studies in Educational Evaluation*, vol. 90, art. no. 101655, 2026, doi: [10.1016/j.stueduc.2026.101655](https://doi.org/10.1016/j.stueduc.2026.101655).
- [11] A. Stojanov and B. K. Daniel, "A decade of research into the application of big data and analytics in higher education: A systematic review of the literature," *Education and Information Technologies*, vol. 29, pp. 5807–5831, 2024, doi: [10.1007/s10639-023-12033-8](https://doi.org/10.1007/s10639-023-12033-8).
- [12] L. Paulsen and E. Lindsay, "Learning analytics dashboards are increasingly becoming about learning and not just analytics – A systematic review," *Education and Information Technologies*, vol. 29, pp. 14279–14308, 2024, doi: [10.1007/s10639-023-12401-4](https://doi.org/10.1007/s10639-023-12401-4).
- [13] L. Márquez, V. Henríquez, H. Chevreux, E. Scheihing, and J. Guerra, "Adoption of learning analytics in higher education institutions: A systematic literature review," *British Journal of Educational Technology*, vol. 55, no. 2, pp. 439–459, 2024, doi: [10.1111/bjet.13385](https://doi.org/10.1111/bjet.13385).
- [14] E. Drugova, I. Zhuravleva, U. Zakharova, and A. Latipov, "Learning analytics driven improvements in learning design in higher education: A systematic literature review," *Journal of Computer Assisted Learning*, vol. 40, no. 2, pp. 510–524, 2024, doi: [10.1111/jcal.12894](https://doi.org/10.1111/jcal.12894).
- [15] M. Afzaal and J. Nouri, "A systematic review of software for learning analytics in higher education," *International Journal of Emerging Technologies in Learning*, vol. 19, no. 7, pp. 17–43, 2024, doi: [10.3991/ijet.v19i07.50313](https://doi.org/10.3991/ijet.v19i07.50313).
- [16] N. Bergdahl, M. Bond, J. Sjöberg, M. Dougherty, and E. Oxley, "Unpacking student engagement in higher education learning analytics: A systematic review," *International Journal of Educational Technology in Higher Education*, vol. 21, art. no. 63, 2024, doi: [10.1186/s41239-024-00493-y](https://doi.org/10.1186/s41239-024-00493-y).
- [17] A. Palanci, R. M. Yilmaz, and Z. Turan, "Learning analytics in distance education: A systematic review study," *Education and Information Technologies*, vol. 29, pp. 22629–22650, 2024, doi: [10.1007/s10639-024-12737-5](https://doi.org/10.1007/s10639-024-12737-5).
- [18] R. Kleimola, L. Hirsto, and H. Ruokamo, "Promoting higher education students' self-regulated learning through learning analytics: A qualitative study," *Education and Information Technologies*, vol. 30, pp. 4959–4986, 2025, doi: [10.1007/s10639-024-12978-4](https://doi.org/10.1007/s10639-024-12978-4).
- [19] P. G. de Barba, E. A. Oliveira, and N. English, "Development and validation of a learning analytics rubric for self-regulated learning," *Educational Technology Research and Development*, vol. 73, pp. 3223–3245, 2025, doi: [10.1007/s11423-025-10521-x](https://doi.org/10.1007/s11423-025-10521-x).

- [20] K. Alalawi, R. Athauda, R. Chiong, and I. Renner, "Evaluating the student performance prediction and action framework through a learning analytics intervention study," *Education and Information Technologies*, vol. 30, pp. 2887–2916, 2025, doi: [10.1007/s10639-024-12923-5](https://doi.org/10.1007/s10639-024-12923-5).
- [21] K. Alalawi, R. Athauda, and R. Chiong, "An extended learning analytics framework integrating machine learning and pedagogical approaches for student performance prediction and intervention," *International Journal of Artificial Intelligence in Education*, vol. 35, pp. 1239–1287, 2025, doi: [10.1007/s40593-024-00429-7](https://doi.org/10.1007/s40593-024-00429-7).
- [22] A. Angeioplastis, J. Aliprantis, M. Konstantakis, and A. Tsimpiris, "Predicting student performance and enhancing learning outcomes: A data-driven approach using educational data mining techniques," *Computers*, vol. 14, no. 3, art. no. 83, 2025, doi: [10.3390/computers14030083](https://doi.org/10.3390/computers14030083).
- [23] Z. Alamgir, H. Akram, S. Karim, and A. Wali, "Enhancing student performance prediction via educational data mining on academic data," *Informatics in Education*, vol. 23, no. 1, pp. 1–24, 2024, doi: [10.15388/infedu.2024.04](https://doi.org/10.15388/infedu.2024.04).
- [24] D. Khairy et al., "Prediction of student exam performance using data mining classification algorithms," *Education and Information Technologies*, vol. 29, pp. 21621–21645, 2024, doi: [10.1007/s10639-024-12619-w](https://doi.org/10.1007/s10639-024-12619-w).
- [25] S. Malik et al., "Advancing educational data mining for enhanced student performance prediction: A fusion of feature selection algorithms and classification techniques with dynamic feature ensemble evolution," *Scientific Reports*, vol. 15, art. no. 8738, 2025, doi: [10.1038/s41598-025-92324-x](https://doi.org/10.1038/s41598-025-92324-x).
- [26] V. Balachandar and K. Venkatesh, "A multi-dimensional student performance prediction model (MSPP): An advanced framework for accurate academic classification and analysis," *MethodsX*, vol. 14, art. no. 103148, 2025, doi: [10.1016/j.mex.2024.103148](https://doi.org/10.1016/j.mex.2024.103148).
- [27] F. J. García-García, M. I. Gómez-Núñez, and C. Molla-Esparza, "Applications of learning analytics in the study of academic performance in higher education: A meta-review with an educational perspective," *Higher Education*, published online June 2026, doi: [10.1007/s10734-026-01709-y](https://doi.org/10.1007/s10734-026-01709-y).
- [28] Q. Li, Y. Jung, and A. F. Wise, "How instructors use learning analytics: The pivotal role of pedagogy," *Journal of Computing in Higher Education*, vol. 38, pp. 227–255, 2026, doi: [10.1007/s12528-025-09432-w](https://doi.org/10.1007/s12528-025-09432-w).
- [29] D. Šimić et al., "HELA-CMM: Capability maturity model for adoption of learning analytics in higher education," *International Journal of Educational Technology in Higher Education*, vol. 22, art. no. 25, 2025, doi: [10.1186/s41239-025-00509-1](https://doi.org/10.1186/s41239-025-00509-1).
- [30] E. Caro, C. González, and J. M. Mira, "Student academic performance stochastic simulator based on the Monte Carlo method," *Computers & Education*, vol. 76, pp. 42–54, 2014, doi: [10.1016/j.compedu.2014.03.007](https://doi.org/10.1016/j.compedu.2014.03.007).
- [31] A. Pritchard, D. Taylor, and M. Belford, "Teaching data-driven decision making for inventory analysis with Monte Carlo simulation," *Decision Sciences Journal of Innovative Education*, vol. 23, art. no. e12328, 2025, doi: [10.1111/dsji.12328](https://doi.org/10.1111/dsji.12328).
- [32] M. Biezuński, M. Dulek, M. Jessa, and Ł. Matuszewski, "Statistical quality of number sequences produced by pseudorandom number generators built into contemporary programming languages," *Applied Sciences*, vol. 16, no. 20, art. no. 9979, 2026, doi: [10.3390/app16209979](https://doi.org/10.3390/app16209979).
- [33] X. Wu et al., "Pseudorandom number generators based on neural networks: A review," *Journal of King Saud University – Computer and Information Sciences*, vol. 37, art. no. 18, 2025, doi: [10.1007/s44443-025-00007-4](https://doi.org/10.1007/s44443-025-00007-4).
- [34] L. G. Suhaidir, Sumijan, and Y. Yunus, "Prediksi tingkat pemahaman siswa terhadap data nominatif," *Jurnal Sistim Informasi dan Teknologi*, vol. 2, 2020, doi: [10.37034/jsisfotek.v2i3.28](https://doi.org/10.37034/jsisfotek.v2i3.28).
- [35] A. Y. Vladova and K. M. Borchyk, "Predictive analytics of student performance: Multi-method and code," *JRAMathEdu*, vol. 9, no. 4, 2024, doi: [10.23917/jramathedu.v9i4.4643](https://doi.org/10.23917/jramathedu.v9i4.4643).
- [36] Y. Qi, "A multi-dimensional prediction system for students' academic performance driven by deep learning," *Discover Artificial Intelligence*, vol. 6, art. no. 69, 2026, doi: [10.1007/s44163-025-00744-5](https://doi.org/10.1007/s44163-025-00744-5).

- [37] J. Gimenez et al., "Critical thinking on a 10-week EAP course: Teachers' and international pre-university students' views and experiences," *Journal of English for Academic Purposes*, vol. 76, art. no. 101534, 2025, doi: [10.1016/j.jeap.2025.101534](https://doi.org/10.1016/j.jeap.2025.101534).
- [38] C. Sellberg and A. Sharma, "Toward multimodal learning analytics in simulation-based collaborative learning: A design ethnography of maritime training," *International Journal of Computer-Supported Collaborative Learning*, vol. 20, pp. 201–221, 2025, doi: [10.1007/s11412-024-09435-2](https://doi.org/10.1007/s11412-024-09435-2).
- [39] X. Yin and K. Dou, "An AI-assisted critical thinking intervention to enhance undergraduate EFL learners' writing proficiency," *Studies in Educational Evaluation*, vol. 86, art. no. 101480, 2025, doi: [10.1016/j.stueduc.2025.101480](https://doi.org/10.1016/j.stueduc.2025.101480).
- [40] B. Flanagan, R. Majumdar, and H. Ogata, "Early-warning prediction of student performance and engagement in open book assessment by reading behavior analysis," *International Journal of Educational Technology in Higher Education*, vol. 19, 2022, doi: [10.1186/s41239-022-00348-4](https://doi.org/10.1186/s41239-022-00348-4).