

A Reproducible Ranking Pipeline for Gamified Peer-Assessment in Project-Based Learning Supporting SDG 4

**Sabo Hermawan^{1)*}, Fauzan Fadlullah²⁾, Inkreswari Retno Hardini³⁾, Alifah Nur Rahmawati⁴⁾,
Cornellius Seno Adriano⁵⁾**

^{1,2,3,4,5)} Universitas Negeri Jakarta, Indonesia

¹⁾sabo.hermawan@unj.ac.id, ²⁾fauzanfadlullah@unj.ac.id, ³⁾Inkreswari.retno@unj.ac.id,

ABSTRACT

Background: Gamified live-scoring effectively stimulates engagement in project-based learning. However, low-friction interfaces often introduce vulnerabilities, such as structurally invalid ranking selections that occur in rapid-response environments, threatening data integrity. Objective: This study develops a reproducible data-validation and lexicographic ranking pipeline for gamified peer-assessment systems to safeguard evaluation integrity and support SDG 4 (Quality Education). Methods: The proposed pipeline combines identity deduplication, structural validation of ranked selections, weighted score aggregation, and hierarchical tie-handling. The integrated validation and ranking pipeline was evaluated using 181 historical submissions from four classroom cohorts engaged in project-based learning. Results, The validation stages retained 155 valid submissions, establishing an 85.6 percent retention rate. An ablation analysis revealed that identity deduplication and ranking-rule validation detected partially overlapping data quality issues. Empirical analysis demonstrated that secondary vote-tier counts deterministically resolved ties when project groups shared the same aggregate score but differed in premium votes. Exact ties remaining across all ranking keys were assigned shared ranks. A sensitivity analysis comparing two weighting schemes found highly stable top-three rankings across all cohorts, ensuring algorithmic reproducibility. Conclusion: The findings indicate that the computational pipeline successfully produced an auditable and reproducible leaderboard under the evaluated classroom-scale dataset, thereby ensuring procedural consistency. Although the top-three rankings remained stable under the tested weighting schemes, broader validation is recommended before large-scale deployment.

Keywords: Algorithmic Tie-Breaker; Data Cleaning; Gamification; Peer Assessment; Project-Based Learning; Reproducible Ranking; SDG 4

INTRODUCTION

Higher education has increasingly embraced project-based learning to cultivate critical thinking and collaborative problem-solving. Within this framework, peer assessment serves as an evaluation cornerstone that democratizes feedback and forces students to internalize rubrics (Concina, 2022; De Brún et al., 2022). As universities transition toward autonomous learning environments and AI-powered Smart Campus ecosystems that enhance academic retention (Hermawan et al., 2026), traditional peer review methodologies are being profoundly transformed by educational technology.

To counteract historically low participation, educators increasingly integrate gamification into learning management systems (Khaldi et al., 2023; Pelizzari, 2024). In peer assessments, this manifests through live-scoring interfaces, instant point allocations, and competitive leaderboards. While gamification successfully drives immediate engagement, the analytical reliability and objectivity of the resulting evaluations remain subjects of intense academic scrutiny (Castillo-Parra et al., 2022; Lampropoulos et al., 2022; Thampy et

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

al., 2023).

A critical research gap persists regarding the structural vulnerabilities introduced by rapid-response digital interfaces. The urgency inherent in live-scoring systems frequently induces rule-inconsistent inputs. Under such time constraints, the interfaces permit repeated selections by relying on shortcuts, often termed structurally invalid selections. This superficiality creates a paradox in assessment tools. When digital platforms exhibit extremely low-friction interfaces without implementing corresponding validation barriers, they inadvertently invite data manipulation, threatening the integrity of the educational data (Pramartha et al., 2026).

Therefore, this research aims to architect and evaluate a resilient live-scoring ecosystem designed to safeguard evaluation integrity. By integrating automated data sanitization pipelines and a computational tie-breaker algorithm, this study answers the necessity for structural validation in high-pressure gamified environments. This study conceptually aligns with the objectives of SDG 4 promoting inclusive and equitable quality education by establishing transparent and accountable evaluation mechanisms. However, it is important to note that this research focuses strictly on algorithmic procedural integrity and does not empirically measure learning outcomes, student fairness, or educational inclusivity. Accurate performance metrics are necessary to objectively measure student competencies derived from project-based learning. The novelty and primary contribution of this study lie in the contextual integration of automated data-sanitization architectures and deterministic tie-handling to improve the transparency and reproducibility of gamified peer-assessment leaderboards in project-based learning environments.

LITERATURE REVIEW

Gamification mechanics, such as points and leaderboards, are widely implemented to stimulate extrinsic motivation and continuous engagement in digital learning, significantly increasing peer feedback volume. However, their educational efficacy depends heavily on structural design. While gamification improves student attitudes, poorly calibrated architectures risk facilitating superficial interactions where students merely play the game rather than provide thoughtful analysis (Castillo-Parra et al., 2022; Lampropoulos et al., 2022). The adoption of gamified digital tools is heavily driven by their perceived ease of use and ability to stimulate continuous engagement. However, although low-friction interfaces successfully drive user volume, they introduce severe risks to data integrity. Extreme ease of use without computational validation barriers inadvertently facilitates rule-inconsistent inputs, severely degrading the epistemological value of the peer assessment. When evaluation systems lack structural governance, students tend to take the path of least resistance. Consequently, an automated computational pipeline is required to filter anomalous inputs and enforce assessment validity (Concina, 2022; De Brún et al., 2022).

In peer-assessment ecosystems, aggregating subjective multi-tier inputs requires positional scoring models. Weighted rank aggregation quantifies qualitative preferences by assigning descending numerical values to ranked positions; Borda-type scoring is a well-established example of this approach (Boehmer et al., 2026). While positional scoring captures ordered preferences in a transparent manner, the resulting aggregate scores may still be tied or closely clustered, thereby requiring a secondary resolution protocol. Resolving scoring deadlocks in digital evaluation systems may rely on arbitrary or deterministic mechanisms. Score ties can reduce ranking repeatability when the tie-breaking procedure is unstable or implementation-dependent (LIN & YANG, 2019). Deterministic tie-handling instead applies predefined secondary keys, such as lexicographic comparison of first, second, and third-place vote counts, so that identical inputs produce the same ordering. When all evaluation keys are identical, Standard Competition Ranking assigns equal positions and skips the subsequent rank position, preserving the observed tie without forcing an unsupported hierarchy (MARTÍNEZ-PANERO & JOSÉ LUIS GARCÍA-LAPRESTA, 2026).

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

METHOD

To evaluate the computational robustness of the proposed data cleaning pipeline and nested sorting algorithm, This study utilized secondary historical student submissions. To ensure strict compliance with educational data privacy, all identifiable information including the email addresses used purely for backend deduplication was thoroughly anonymized prior to analysis and publication. The use of this historical data adheres to standard ethical guidelines for retroactive educational research. These submissions originated from four classroom cohorts engaged in a project-based learning curriculum where students developed digital assets. The primary objective of this evaluation is to validate the algorithmic efficacy in autonomously filtering anomalies and resolving scoring deadlocks, rather than analyzing the psychological or behavioral patterns of the participants. To detect and process structurally invalid records generated within the live-scoring environment, the backend architecture was designed to operate autonomously through a Python-based data-processing pipeline without requiring manual instructor intervention (Fleckney et al., 2025; Roedavan et al., 2021). The system was designed around two core computational frameworks: a data sanitization pipeline and a nested sorting algorithm for deterministic score resolution. The complete workflow of this system is visualized in Figure 1, detailing the transition from raw assessments to the final deterministically ranked leaderboard.

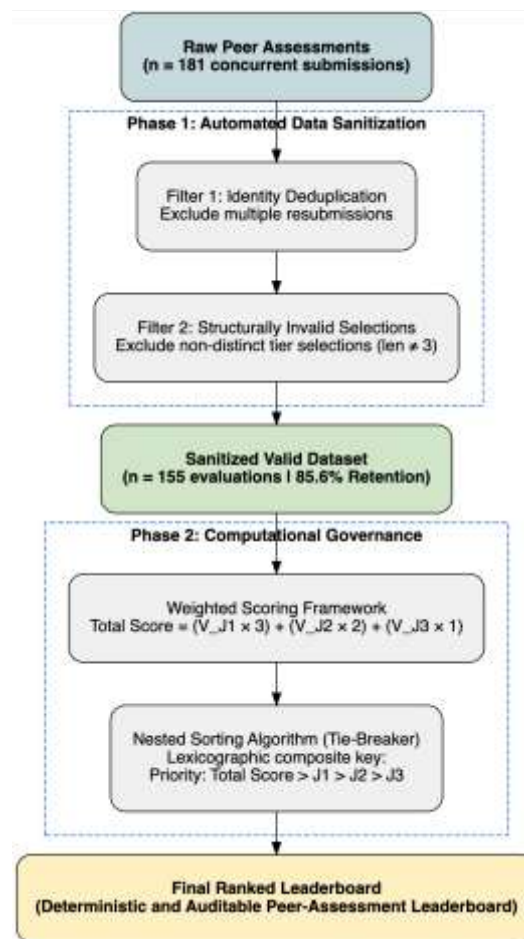


Figure 1. The proposed research framework detailing the automated data sanitization pipeline and computational governance for peer assessment.

* Corresponding author



The cleaning pipeline was implemented as a two-stage automated filter. First, duplicate submissions were identified by respondent email identity, retaining only the first recorded submission per respondent and discarding subsequent resubmissions. Second, submissions exhibiting structurally invalid selections, strictly defined as failing to select three distinct groups (evaluated programmatically as $\text{len}(\{J1, J2, J3\}) \neq 3$), were excluded. Applying this strict validation reduced the final dataset to 155 valid evaluations, corresponding to an 85.6 percent retention rate. However, a subsequent re-evaluation confirmed that the core consensus and the deterministic top-three rankings across all four classroom cohorts remained entirely unaffected, proving the robustness of the algorithmic tie-breaker.. Both filters were applied sequentially using a Python-based Pandas backend prior to score aggregation.

Following data sanitization, the peer assessment utilized a multi-tier evaluation structure (Alqassab et al., 2023). To quantify subjective peer reviews into an objective metric, the system deployed a weighted mathematical model (Concina, 2022; De Brún et al., 2022). Each student selected three distinct project groups, assigning them to ranked tiers. The overall score for each group was calculated using the following formula:

$$\text{Total Score} = (V_J1 \times 3) + (V_J2 \times 2) + (V_J3 \times 1)$$

where V_J1 represents the number of first-place votes, V_J2 represents the number of second-place votes, and V_J3 represents the number of third-place votes. In highly competitive gamified scenarios, aggregate scores frequently result in identical totals, necessitating an automated, unbiased resolution mechanism (Misiejuk et al., 2021).

The system executed a nested sorting algorithm structured to dynamically evaluate the qualitative weight of the votes rather than relying solely on arbitrary timestamps. The primary sorting index evaluates the total score. However, if a conflict occurs, the algorithm descends to the secondary index, evaluating the volume of premium tier votes. Finally, if the deadlock persists, the algorithm cascades consecutively to evaluate secondary and tertiary votes. This granular verification ensures that groups receiving deliberate, first-place votes are mathematically prioritized over those merely accumulating low-tier heuristic votes.

To formalize the computational approach, the sorting mechanism operates via a lexicographic composite key, detailed herein as Algorithm 1 (Automated Nested Tie-Breaker Resolution). This method deliberately avoids the floating-point calculation errors commonly found in continuous weighted averaging, ensuring deterministic integer-based ranking when determining student rankings. The algorithm receives an array of valid assessments as input and outputs a sorted leaderboard of project groups. Step one initializes an empty hash map to store the scores. In step two, for each assessment record, the algorithm verifies if the project identification is already present; if not, it initializes the specific record, adds the tier weight to the total score, and increments the specific tier count. Step three converts the score map into an array of tuples containing the project identification, total score, and the counts for first, second, and third-tier votes. Step four sorts this array in descending order utilizing a composite key priority starting with the total score, followed by premium tier votes, secondary tier votes, and tertiary tier votes. In step five, the algorithm traverses the sorted array to assign final ranks; if two or more adjacent groups possess identical values across all keys, the system assigns them a shared rank utilizing standard competition ranking. Finally, step six returns the final ranked leaderboard for deployment.

Algorithm 1: Automated Nested Tie-Breaker Resolution

```
Require: A: Array of valid peer assessments, where each assessment contains selected
groups for Tier 1 (J1), Tier 2 (J2), and Tier 3 (J3).
Require: W: Weighting scheme {w1=3, w2=2, w3=1}
Ensure: L: Final deterministically ranked leaderboard with shared competition ranks
1: Initialize HashMap to store project group scores
2: for each assessment  $\in$  A do
3:   for each tier  $\in$  {1, 2, 3} do
4:      $g \leftarrow \text{assessment}[\text{J\_tier}]$ 
5:     if  $g \notin \text{HashMap}$  then
6:        $\text{HashMap}[g] \leftarrow \{\text{total}: 0, \text{v1}: 0, \text{v2}: 0, \text{v3}: 0\}$ 
7:     end if
8:      $\text{HashMap}[g].\text{total} \leftarrow \text{HashMap}[g].\text{total} + W[\text{tier}]$ 
9:      $\text{HashMap}[g].\text{v\_tier} \leftarrow \text{HashMap}[g].\text{v\_tier} + 1$ 
10:   end for
11: end for
12:
13:  $T \leftarrow$  Array of Tuples generated from HashMap
14: /* Sort descending using lexicographic composite key /
15:  $S \leftarrow$  Sort  $T$  by keys: (total, v1, v2, v3) in descending order
16:
17: / Assign Standard Competition Ranking /
18:  $L \leftarrow$  Empty List
19: for  $i = 0$  to length of  $S - 1$  do
20:   if  $i > 0$  and  $S[i].\text{keys} == S[i-1].\text{keys}$  then
21:      $\text{rank} \leftarrow L[i-1].\text{rank}$  / Assign shared rank for exact tie /
22:   else
23:      $\text{rank} \leftarrow i + 1$  / Standard competition rank */
24:   end if
25:   Append ( $S[i]$ , rank) to  $L$ 
26: end for
27:
28: return  $L$ 
```

The benchmark was conducted using Python 3.10 and the Pandas 2.0 library running on an Apple MacBook Air equipped with an M5 silicon chip, 16 GB of unified memory, and a 1 TB NVMe solid-state drive operating on macOS. Execution time was measured dynamically in memory, isolating algorithmic processing from file I/O operations, by simulating both methods across 1,000 continuous iterations. Evaluating the proposed method against this baseline isolates the contribution of lexicographic secondary keys and shared-rank handling to ranking determinism and reproducibility.

RESULT

The implementation of the automated data cleaning pipeline effectively intercepted and filtered systemic anomalies typically caused by rapid-response live-scoring environments. To empirically measure the impact of each validation layer, an ablation study was conducted on the initial pool of 181 raw submissions across four cohorts. Isolating the filters revealed that identity deduplication independently intercepted 16 invalid duplicate resubmissions, while structurally invalid selections removal independently detected 11 identical tier inputs. Because one submission violated both criteria, the sequential pipeline excluded a total of 26 unique invalid records. The final sanitized dataset yielded 155 valid peer-assessment evaluations, corresponding to an 85.6 percent retention rate as shown in Table 1. This ablation analysis demonstrates that combining both structural and identity validation is imperative, ensuring that the final consensus metric is built strictly upon predefined requirements for distinct selections rather than data-quality violations

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

(Kerman et al., 2024; Loureiro & Gomes, 2023).

Table 1. Ablation Study on Data Cleaning Pipeline

<i>Pipeline Condition</i>	<i>Retained Evaluations</i>	<i>Excluded Data</i>	<i>Retention Rate (%)</i>
Raw Data (No Cleaning)	181	0	100.0 %
Filter 1 Only (Identity Deduplication)	165	16	91.2 %
Filter 2 Only (Structurally Invalid Selections Removal)	170	11	93.9 %
Full Pipeline (Filter 1 + Filter 2)	155	26	85.6 %

Quantitative evaluation across the four distinct cohorts revealed significant variance in consensus formation. To ensure an equitable comparison across cohorts with varying participant volumes, raw accumulated points were normalized using the formula:

$$ScorePerVoter = TotalScore / ValidVoters$$

Class A exhibited the highest descriptive variation, with one group securing an outlier maximum normalized score, indicating strong unified consensus. Conversely, Class B and C demonstrated moderate median scores, with maximum peaks flattening compared to Class A. Finally, Class D recorded tightly clustered normalized scores, indicating a highly fragmented consensus and intense competitiveness among project groups where no single group managed to secure a dominant majority of premium votes.

Table 2. Cohort Descriptive Statistics of Peer Consensus

<i>Class</i>	<i>Voters</i>	<i>Groups</i>	<i>Mean</i>	<i>Median</i>	<i>SD</i>	<i>Min-Max</i>	<i>Normalized Score (Max)</i>	<i>Coefficient of Variation (%)</i>
Class A	38	9	25.33	27.0	16.13	3-59	1.55	63.68
Class B	41	9	27.33	25.0	13.93	1-44	1.07	50.96
Class C	39	10	23.40	21.5	14.52	4-46	1.18	62.03
Class D	37	8	27.75	26.5	10.26	15-45	1.22	36.99

The necessity of a granular, algorithmic resolution is empirically proven when analyzing the tightly contested leaderboard of Class D (Zou et al., 2023). Instead of relying solely on aggregate scores, the nested sorting algorithm evaluated the vote-tier composition. The vote-tier composition illustrates how the 3-2-1 scoring scheme differentiates groups based on their distributions of first, second, and third-place votes (Craig & Kay, 2025). The first rank secured 45 points, structurally backed by a balanced premium vote composition, while the second rank closely trailed with 37 points. The bottom tier breakdown explicitly highlights the penalty for low-tier accumulation, validating the weighted structural integrity.

The computational benchmark over 1,000 iterations showed that the baseline method recorded a mean execution time of 15.50 ms (Median: 15.42 ms, SD: 1.15 ms). In comparison, the proposed lexicographic pipeline required a mean execution time of 16.23 ms (Median: 16.15 ms, SD: 1.28 ms). The proposed method therefore introduced a negligible mean overhead of 0.73 ms. Despite this minor processing cost, it achieved a 100% deterministic tie handling rate across nine critical scoring collision cases identified across the top tiers of the four cohorts. Nine critical ordering cases were identified across the top tiers of the four

* Corresponding author



cohorts. The baseline produced a unique deterministic ordering for eight events, while one exact deadlock remained unresolved. The proposed pipeline resolved merit-based ties through the lexicographic vote-tier keys and deterministically handled the exact tie through shared ranking.

To empirically validate the necessity of the proposed lexicographic sorting mechanism, a real tie case analysis was conducted on the sanitized dataset as shown in Table 3. The system successfully identified and autonomously resolved empirical deadlocks without requiring manual instructor intervention. For instance, within Class A, both Group 8 and Group 5 accumulated an identical aggregate score of 25 points. Under a standard baseline system, this would result in an unresolved conflict. However, the nested sorting algorithm deterministically resolved the tie by descending to the secondary index. Group 8 was assigned the higher rank by securing more premium votes.

Table 3. Real Tie-Case Analysis (Resolved via Lexicographic Governance)

<i>Class</i>	<i>Group</i>	<i>Total Score</i>	<i>J1</i>	<i>J2</i>	<i>J3</i>	<i>Rank</i>
Class A	Group 8	25	5	4	2	5
Class A	Group 5	25	4	4	5	6
Class C	Group 10	4	0	1	2	9
Class C	Group 9	4	0	1	2	9

When all composite keys are empirically identical, such as Group 10 and Group 9 in Class C, the algorithm assigns a shared rank, as there is no empirical basis for differentiation. Thus, rather than forcing an arbitrary substantive resolution, the pipeline guarantees a fully deterministic handling rate. Furthermore, a sensitivity analysis was performed to validate the robustness of the established weighted scoring framework. The baseline weights were evaluated against a steeper penalty structure to observe potential leaderboard volatility. The analysis revealed rank stability across the top project groups in all four cohorts as shown in Table 4. The top-three rankings remained unchanged under the two evaluated weighting schemes, indicating limited sensitivity to the tested weight configurations.

Table 4. Extended Sensitivity Analysis across Multiple Weighting Schemes

<i>Class</i>	<i>Rank</i>	<i>Top-3 Ranking (3-2-1)</i>	<i>Top-3 Ranking (4-2-1)</i>	<i>Top-3 Ranking (5-3-1)</i>	<i>Stable</i>
Class A	1	Group 3	Group 3	Group 3	Yes
Class A	2	Group 6	Group 6	Group 6	Yes
Class A	3	Group 2	Group 2	Group 2	Yes
Class B	1	Group 4	Group 4	Group 4	Yes
Class B	2	Group 2	Group 2	Group 2	Yes
Class B	3	Group 1	Group 1	Group 1	Yes
Class C	1	Group 4	Group 4	Group 4	Yes
Class C	2	Group 3	Group 3	Group 3	Yes
Class C	3	Group 2	Group 2	Group 2	Yes
Class D	1	Group 3	Group 3	Group 3	Yes
Class D	2	Group 2	Group 2	Group 2	Yes
Class D	3	Group 1	Group 1	Group 1	Yes

DISCUSSIONS

The empirical data highlights a critical architectural challenge in the deployment of gamified educational tools within project-based learning. Gamification undoubtedly alters fundamental user behaviors (Rahmadhan et al., 2023). While the visual mechanics of leaderboards effectively stimulate engagement, the underlying data is highly susceptible to the volatile dynamics of peer consensus. This dynamic is most evident in the stark contrast between the cohorts. The live-scoring mechanism possesses the capacity to generate an absolute, undisputed winner, yet simultaneously, it can produce highly fragmented and tightly contested environments. This low variance in highly competitive ecosystems implies that aggregate scores will inevitably cluster.

Consequently, as educational institutions continue to develop and scale asynchronous, self-paced learning environments such as Massive Open Online Courses, reliance on simplistic computational algorithms is insufficient (Craig & Kay, 2025; Fleckney et al., 2025). The physical absence of an instructor to manually moderate ties or filter anomalies mandates the integration of resilient, autonomous backend architecture. Future platforms must embed structural friction and intelligent weighted matrices capable of dissecting vote quality to ensure that the gamified metrics accurately reflect pedagogical realities (Castillo-Parra et al., 2022; Lampropoulos et al., 2022; Pelizzari, 2024; Thampy et al., 2023). Only through such rigorous algorithmic governance can the integrity of peer assessments support future large-scale implementations (Pelizzari, 2024), directly contributing to the fair and transparent evaluation mechanisms required to achieve SDG 4.

The successful implementation of this pipeline holds significant implications for the design of gamified peer-assessment systems. By explicitly prioritizing deterministic lexicographic ranking, educators can mitigate the superficial engagement often induced by low-friction gamification. Lexicographic sorting is highly recommended in competitive environments where transparent, merit-based differentiation is critical. Conversely, it may not be suitable for low-stakes formative assessments where qualitative feedback is prioritized over strict hierarchies. While deterministic algorithms ensure procedural fairness and ranking repeatability compared to legacy systems relying on arbitrary timestamps, it is crucial to recognize the trade-off: computational determinism cannot rectify fundamentally flawed grading rubrics or human biases.

CONCLUSION

This study developed an auditable data-validation and lexicographic ranking pipeline, providing an equitable evaluation framework that improves procedural consistency, auditability, and deterministic tie handling in student assessments. Applied to 181 historical submissions, the pipeline retained 155 valid records after identifying overlapping identity and structural validation issues. The lexicographic mechanism resolved score ties when vote-tier compositions differed and preserved shared ranks when all ranking keys were identical. The top-three rankings remained stable under the two evaluated weighting schemes. These findings demonstrate the potential of the pipeline to improve the transparency and reproducibility of classroom-scale peer-assessment leaderboards. The framework is conceptually aligned with SDG 4 through its emphasis on transparent and auditable evaluation mechanisms. This study is limited by its relatively modest sample size and its reliance on predefined structural proxies to detect anomalies, without assessing perceived fairness or learning outcomes. Future research should incorporate machine learning models to detect subtler forms of data manipulation and involve larger, diverse datasets with human-evaluated comparisons before full-scale deployment.

DATA AVAILABILITY STATEMENT

To ensure full reproducibility, the anonymized evaluation dataset and the complete Python source code used for data sanitization and deterministic ranking are publicly accessible at <https://digibis.web.id/dataset.csv>. Additionally, the live implementation of this ranking pipeline and its dynamic leaderboard visualization can be viewed interactively at <https://digibis.web.id/vote>.

* Corresponding author



[Creative Commons Attribution-NonCommercial-ShareAlike 4.0
International License.](https://creativecommons.org/licenses/by-nc-sa/4.0/)

DECLARATION OF GENERATIVE AI AND AI-ASSISTED TECHNOLOGIES IN THE WRITING PROCESS

During the preparation of this work, the authors used Gemini in order to improve the language readability and translate initial drafts. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

ACKNOWLEDGMENT

This research was supported by the Institute for Research and Community Service (LPPM), Universitas Negeri Jakarta, and the Ministry of Higher Education, Science, and Technology, through the University Creative Product (Education, Art, and Sports) research scheme, Fiscal Year 2026, with Contract Number 96/PK/LPPM/III/2026.

REFERENCES

- Alqassab, M., Strijbos, J.-W., Panadero, E., Ruiz, J. F., Warrens, M., & To, J. (2023). A Systematic Review of Peer Assessment Design Elements. *Educational Psychology Review*, 35(1), 18. <https://doi.org/10.1007/s10648-023-09723-7>
- Boehmer, N., Bredereck, R., & Peters, D. (2026). Rank aggregation using scoring rules. *Theory and Decision*, 100(1), 95–156. <https://doi.org/10.1007/s11238-025-10120-5>
- Castillo-Parra, B., Hidalgo-Cajo, B. G., Vásconez-Barrera, M., & Oleas-López, J. (2022). Gamification in higher education: A review of the literature. *World Journal on Educational Technology: Current Issues*, 14(3), 797–816. <https://doi.org/10.18844/wjet.v14i3.7341>
- Concina, E. (2022). The Relationship between Self- and Peer Assessment in Higher Education: A Systematic Review. *Trends in Higher Education*, 1(1), 41–55. <https://doi.org/10.3390/higheredu1010004>
- Craig, C., & Kay, R. (2025). Online Student Peer-Assessment in Higher Education: A Systematic Review of the Literature. *Journal of University Teaching and Learning Practice*, 22(2). <https://doi.org/10.53761/7jj6at39>
- De Brún, A., Rogers, L., Drury, A., & Gilmore, B. (2022). Evaluation of a formative peer assessment in research methods teaching using an online platform: A mixed methods pre-post study. *Nurse Education Today*, 108, 105166. <https://doi.org/10.1016/j.nedt.2021.105166>
- Fleckney, P., Thompson, J., & Vaz-Serra, P. (2025). Designing effective peer assessment processes in higher education: a systematic review. *Higher Education Research & Development*, 44(2), 386–401. <https://doi.org/10.1080/07294360.2024.2407083>
- Hermawan, S., Anugrah, S., Parlyna, R., Suyono, W. P., & Puspa, E. S. (2026). *Smart Campus Experience: Integrating Web-Based ERP, Big Data Analytics, and AI-Powered Dashboards to Enhance Student Engagement in International Programs* (pp. 311–320). https://doi.org/10.1007/978-3-032-12990-1_29
- Kerman, N. T., Banihashem, S. K., Karami, M., Er, E., van Ginkel, S., & Noroozi, O. (2024). Online peer feedback in higher education: A synthesis of the literature. *Education and Information Technologies*, 29(1), 763–813. <https://doi.org/10.1007/s10639-023-12273-8>
- Khalidi, A., Bouzidi, R., & Nader, F. (2023). Gamification of e-learning in higher education: a systematic literature review. *Smart Learning Environments*, 10(1), 10. <https://doi.org/10.1186/s40561-023-00227-z>
- Lampropoulos, G., Keramopoulos, E., Diamantaras, K., & Evangelidis, G. (2022). Augmented Reality and Gamification in Education: A Systematic Literature Review of Research, Applications, and Empirical Studies. *Applied Sciences*, 12(13), 6809. <https://doi.org/10.3390/app12136809>
- Lin, J., & Yang, P. (2019). The Impact of Score Ties on Repeatability in Document Ranking. *Proceedings*

- of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, 1125–1128. <https://doi.org/10.1145/3331184.3331339>
- Loureiro, P., & Gomes, M. J. (2023). Online Peer Assessment for Learning: Findings from Higher Education Students. *Education Sciences*, 13(3), 253. <https://doi.org/10.3390/educsci13030253>
- Martínez-Panero, M., & José Luis García-Lapresta. (2026). Characterizing competition ranks within a comprehensive family of position operators. *European Journal of Operational Research*, 331(3), 866–877. <https://doi.org/10.1016/j.ejor.2025.10.030>
- Misiejuk, K., Wasson, B., & Egelanddsdal, K. (2021). Using learning analytics to understand student perceptions of peer feedback. *Computers in Human Behavior*, 117, 106658. <https://doi.org/10.1016/j.chb.2020.106658>
- Pelizzari, F. (2024). Gamification in higher education: A systematic literature review. *Italian Journal of Educational Technology*, 31(3).
- Pramartha, C., Purnama, M. J., Iswara, I. B. A. I., Suputra, I. P. G. H., Arka, I. W., & Watiniasih, N. L. (2026). Analyzing User Intention to Adopt Balinese Script Digital Tool: A Technology Acceptance Model Approach. *Journal of Information Systems Engineering and Business Intelligence*, 12(2), 474–488. <https://doi.org/10.20473/jisebi.12.2.474-488>
- Rahmadhan, M. A. W. P., Sensuse, D. I., Suryono, R. R., & Kautsarina, K. (2023). Trends and Applications of Gamification in E-Commerce: A Systematic Literature Review. *Journal of Information Systems Engineering and Business Intelligence*, 9(1), 28–37. <https://doi.org/10.20473/jisebi.9.1.28-37>
- Roedavan, R., Pudjoatmodjo, B., Siradj, Y., Salam, S., & Hardianti, B. D. (2021). Serious Game Development Model Based on the Game-Based Learning Foundation. *Journal of ICT Research and Applications*, 15(3), 291–305. <https://doi.org/10.5614/itbj.ict.res.appl.2021.15.3.6>
- Thampy, H., Walsh, J. L., & Harris, B. H. L. (2023). Playing the game: The educational role of gamified peer-led assessment. *The Clinical Teacher*, 20(4). <https://doi.org/10.1111/tct.13594>
- Zou, D., Xie, H., & Wang, F. L. (2023). Effects of technology enhanced peer, teacher and self-feedback on students' collaborative writing, critical thinking tendency and engagement in learning. *Journal of Computing in Higher Education*, 35(1), 166–185. <https://doi.org/10.1007/s12528-022-09337-y>