Reducing Subjectivity in Rubric-Based Assessment Through an Adaptive Weighted Penalty Model: A Computational Approach to Fair and Consistent Evaluation

Mohamad Irwan Pandapotan Harahap*

Fakulti Kejuruteraan Awam, UiTM Cawangan Pulau Pinang

***Corresponding Author**

DOI: https://doi.org/10.47772/IJRISS.2026.102100216

Received: 03 September 2026; Accepted: 08 September 2026; Published: 24 October 2026

ABSTRACT

Rubric-based assessment is widely used in higher education to evaluate student performance against predefined criteria. However, conventional rubric scoring remains susceptible to assessor subjectivity, inconsistent interpretation of performance levels, and variations in scoring decisions. This paper presents an Adaptive Weighted Penalty Model (AWPM) implemented through a computational assessment system to improve fairness, consistency, and transparency in rubric-based evaluation. The proposed approach converts qualitative assessor judgements into structured numerical scores by applying adaptive penalty values according to the degree of deviation from expected performance. The system integrates rubric configuration, assisted score entry, automated computation, and structured reporting within a unified assessment workflow. By maintaining academic judgement while standardising the computational treatment of assessment decisions, the approach has the potential to reduce scoring variability and improve the traceability of assessment outcomes. The proposed model is applicable to courses involving rubric-based evaluation, including projects, presentations, practical work, and other performance-based assessments. Its spreadsheet-based implementation also provides a practical and accessible basis for adoption across different educational settings without requiring specialised software infrastructure. The study contributes a structured computational approach to addressing persistent concerns regarding fairness and consistency in rubric-based assessment.

Keywords: Rubric-based Assessment, Adaptive Weighted Penalty, Assessment Fairness, Assessment Consistency, Computational Assessment

INTRODUCTION

Rubric-based assessment is an established approach for evaluating complex student work because it makes assessment criteria and performance expectations explicit. Analytic rubrics are particularly useful for projects, presentations, practical tasks and other performance-based assessments because they allow performance to be judged against several criteria rather than reduced to a single holistic impression. Recent research continues to associate well-designed rubrics with transparency, differentiated feedback and more defensible score interpretation (Großmann & Krüger, 2024; Xiong et al., 2023). However, the existence of a rubric does not by itself guarantee consistent scoring. Reliability can still vary when assessors interpret descriptors differently, apply different levels of strictness, or translate qualitative judgements into marks using different mental arithmetic and decision rules.

This problem becomes more consequential when several assessors evaluate the same type of work or when large cohorts require repeated application of the same rubric. Recent evidence confirms that rubric use does not automatically produce identical ratings. Xiong et al. (2023) found that reliability and validity in rubric-based peer assessment vary across contexts and are influenced by characteristics of the rubric, task and raters. Goertzen and Klaus (2023) likewise showed that inter-rater reliability in graduate rubric assessment

can vary substantially depending on the raters and the statistical method used to evaluate agreement. Performance-assessment research further demonstrates that rater effects remain an important source of score variation even when analytic rubrics are used (Arslan Mancar & Gülleroğlu, 2022). Thus, a central challenge is not simply how to write a rubric, but how to ensure that the scoring logic applied after an assessor makes a judgement remains stable, transparent and auditable.

Digital assessment technologies offer ways to standardise parts of this process. Automated and artificial-intelligence-based grading can reduce repetitive workload, yet recent studies also report variability, opacity and weaker performance for tasks requiring complex disciplinary judgement (Castelblanco et al., 2024; Flodén, 2025). These limitations suggest that fully replacing the assessor is neither necessary nor desirable for many rubric-based tasks. A more practical alternative is to retain the assessor as the source of academic judgement while computationally governing how that judgement is converted into a score. This paper therefore presents an Adaptive Weighted Penalty Model (AWPM), implemented in a spreadsheet-based assessment system, that separates human judgement from numerical computation. The model is intended to reduce avoidable scoring variation by applying the same configurable penalty logic to equivalent assessor inputs.

The study has three objectives: (1) to formulate a computational model that translates rubric-based judgements into criterion scores using adaptive weighted penalties; (2) to implement the model in an accessible assessment workflow comprising rubric configuration, assisted entry, automated computation and structured reporting; and (3) to examine, through functional evaluation, scenario-based comparison and parameter sensitivity analysis, how the model can support consistency, transparency and auditability without removing academic judgement from the assessment process.

LITERATURE REVIEW

Rubric-Based Assessment, Fairness and Consistency

Rubrics support assessment by defining criteria and describing levels of performance. Their value is strongest when descriptors are sufficiently specific to guide raters while remaining aligned with the construct being assessed. In a large cross-context study, Xiong et al. (2023) found that rubric-related variables, including description length and the presence of substantive anchor statements, were associated with the reliability and validity of online peer assessment. Großmann and Krüger (2024) similarly emphasised transparent criteria and evidence supporting score interpretation when developing an analytic rubric for science lesson plans. More recently, Xu and Liu (2026) reported acceptable inter-rater reliability and positive perceptions of fairness for a purpose-built analytic rubric, while also stressing the importance of iterative development and validation.

These findings show that fairness is not achieved merely by providing identical rubric sheets to assessors. A rubric defines the judgement space, but scoring still involves interpretation. Studies of performance-based and graduate assessment report that inter-rater agreement can vary even when raters use a common rubric, reinforcing the need for calibration, clear descriptors and explicit scoring procedures (Arslan Mancar & Gülleroğlu, 2022; Goertzen & Klaus, 2023). More recent evidence comparing analytic rubrics with rating scales also confirms that the scoring instrument and rater interaction can materially affect reliability (Nalbantoğlu Yılmaz, 2024). Consequently, the computational stage between judgement and final mark is itself an important design problem. Standardising this stage can improve procedural consistency even when professional judgement remains necessary. Educational measurement theory provides a useful basis for distinguishing scoring consistency from the broader reliability and validity of assessment outcomes. Generalizability theory (Nalbantoğlu Yılmaz, 2024) recognises that observed scores may be influenced by multiple sources of variation, including assessors, assessment tasks and their interactions. In this context, AWPM addresses the computational component of score determination by standardising the numerical consequences of equivalent assessor judgements. However, it does not independently control variation arising from the interpretation of performance criteria. Accordingly, the model should be regarded as a

procedural scoring mechanism whose contribution to inter-rater reliability requires subsequent empirical validation.

Digital and Automated Assessment

Recent digital assessment research has increasingly examined automated grading and artificial intelligence. Lin et al. (2024) demonstrated that automated grading with reliability detection can improve efficiency in large-scale peer assessment. Fernández-Sánchez et al. (2025) reported that AI-supported rubric development can improve precision and alignment while reducing the time required to construct assessment instruments. These developments demonstrate the growing role of computation in assessment design and processing.

At the same time, current evidence cautions against treating automation as a substitute for expert judgement. Flodén (2025) found that large-language-model grading could produce plausible university examination scores but also showed substantial differences from teacher grading and raised concerns about transparency and bias. In construction management capstone assessment, Castelblanco et al. (2024) found poor overall correspondence between GPT-4 and human evaluation for complex argumentative work. Such findings are especially relevant to performance-based assessment, where quality may depend on contextual, technical and creative considerations. The design position adopted in this study is therefore assessor-centred computation: the assessor determines the observed performance and error categories, while the computational model applies the scoring rules consistently.

Research Gap and Proposed Contribution

Recent literature provides strong support for clearer rubrics, rater calibration, digital assessment and automated scoring, but comparatively less attention has been given to transparent rule-based computation that operates between qualitative assessor judgement and the final numerical mark. Existing studies mainly examine rubric design and reliability, peer-assessment aggregation, or AI-based grading (Goertzen & Klaus, 2023; Lin et al., 2024; Flodén, 2025). The proposed AWPM addresses a narrower procedural gap by treating scoring as a governed transformation. It does not infer quality directly from student work. Instead, it accepts structured assessor inputs and applies a visible mathematical rule in which criterion weight, completion, error frequency, error severity and sensitivity parameters can be configured. This makes the computational treatment inspectable and repeatable while preserving the assessor's responsibility for judging performance.

METHODOLOGY

System architecture

The work followed a design-and-development approach in which an assessment problem was translated into a computational model and then implemented as a working spreadsheet-based prototype. The development process comprised four stages: requirements identification, mathematical model formulation, system implementation and functional/scenario-based evaluation. Requirements were derived from recurrent problems in manual rubric assessment identified in the project, namely inconsistent deductions, repetitive arithmetic, limited traceability and difficulty enforcing a common scoring logic across assessors. The prototype was implemented within a spreadsheet environment to maximise accessibility and minimise dependence on specialised software infrastructure. Figure 1 illustrates the four-component AWPM architecture. Rubric configuration defines criterion weights and penalty parameters, while assisted assessor entry records completion ratios and the frequencies of minor, intermediate and major errors. The AWPM scoring engine applies the configured rules to calculate criterion marks, which are subsequently consolidated into structured assessment records for reporting and audit. The sequence in Figure 1 also clarifies the separation between assessor judgement at the entry stage and deterministic score calculation at the scoring stage; the resulting records preserve the inputs and configured rules for subsequent moderation.

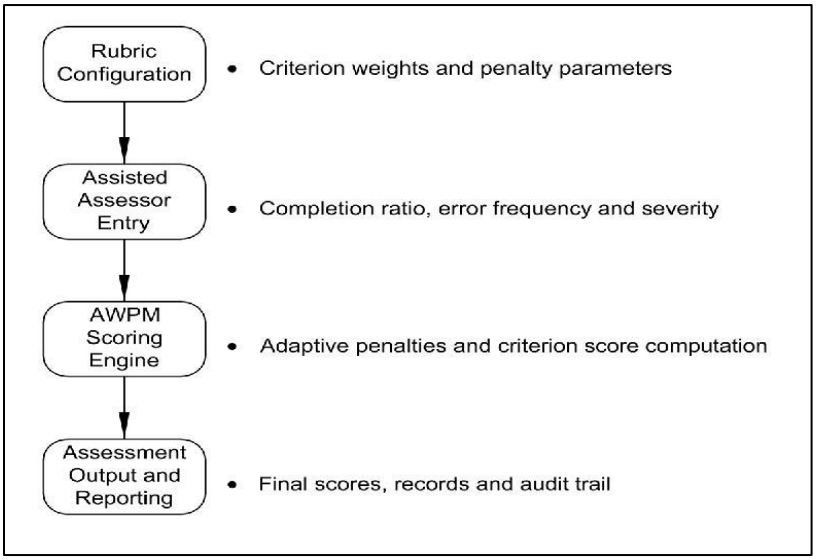


Figure 1 AWPM assessment system architecture

Adaptive Weighted Penalty Model

The Adaptive Weighted Penalty Model (AWPM) computes the final criterion score (S) by adjusting the maximum criterion marks for task completion and deducting severity-weighted error penalties according to configurable sensitivity parameters. The model is expressed as follows:

$$S = max\left\{0, C^{\beta}W - \lambda[1 + \gamma(1 - C)]\sum k_i E_i^{\alpha}\right\} + B$$

Where:	
$S =$	Final criterion score (marks)
$W =$	Maximum marks allocated to the assessment criterion
$k_1 =$	Minor error cost
$k_2 =$	Intermediate error cost
$k_3 =$	Major error cost
$\alpha =$	Error sensitivity exponent
$\beta =$	Completion sensitivity exponent
$\gamma =$	Error–completion interaction factor
$\lambda =$	Penalty scaling factor
$B =$	Bonus term
$C =$	Completion ratio
$E_1 =$	Minor error frequency
$E_2 =$	Intermediate error frequency
$E_3 =$	Major error frequency

The summation covers three error-severity categories, where $i = 1,2,3$ represents minor, intermediate and major errors, respectively. The zero lower bound prevents negative scores before the optional bonus is applied. Where bonus marks are permitted, their maximum value and treatment must be specified in the approved assessment rubric.

The AWPM parameters are configurable but should be systematically calibrated rather than assigned arbitrarily. The criterion weight, W is taken directly from the approved rubric, while the completion ratio, C ranges from 0 to 1, where 1 represents full completion. The exponent β controls sensitivity to incompleteness through C^{β}. A value of β =1 produces a linear reduction, while values above 1 impose progressively greater penalties for incomplete work. A practical initial range is 1.0–1.5.

Error severity is represented by k_1, k_2, and k_3 for minor, intermediate, and major errors, respectively, with $k_1 < k_2 < k_3$. An initial ratio of 1:2:4 provides a reasonable distinction between severity levels, although this should be calibrated according to the assessment context. The corresponding E_1, E_2, and E_3 represent the observed frequencies of these errors. The exponent α determines how repeated errors accumulate. A value of α=1 produces a linear penalty, while 1.0–1.5 allows repeated errors to have progressively greater consequences.

The parameter γ controls the interaction between incompleteness and error penalties through $\gamma(1 - C)$. A moderate initial range of 0.25–0.50 is appropriate to avoid excessive double penalisation, since incompleteness is already accounted for by C^{β}. The scaling factor λ, initially around 0.5–1.0, controls the overall magnitude of deductions and should be calibrated against the criterion weight and expected error patterns. The optional bonus term, B should normally be set to zero unless additional credit is explicitly permitted by the rubric.

These values should be regarded as initial calibration ranges rather than universal constants. Final parameter values should be established through expert moderation and sensitivity testing using representative assessment scenarios and then fixed for all students undertaking the same assessment. Importantly, AWPM does not determine whether an error is minor, intermediate, or major. This remains an academic judgement guided by the rubric. The model instead standardises the computational consequences of equivalent assessor judgements, thereby reducing avoidable scoring variability while preserving academic judgement.

To demonstrate the application of AWPM, a hypothetical project-assessment criterion worth $W = 20$ marks is considered. The assessed work has a completion ratio $C = 0.90$, with two minor errors, one intermediate error and no major errors. Illustrative severity costs of $k_1 = 1$, $k_2 = 2$ and $k_3 = 4$ are assigned to the respective error categories, while the parameters $\alpha = 1.0$, $\beta = 1.2$, $\gamma = 0.30$ and $\lambda = 0.75$, respectively. No bonus is awarded ($B = 0$). Substituting these values into the AWPM equation gives:

$$S = max\{0, 0.90^{1.2} \times 20 - 0.75[1 + 0.30(1 - 0.90)](1 \times 2^1 + 2 \times 1^1 + 4 \times 0^1)\} + 0 = 14.53$$

The resulting score of 14.53 out of 20 (72.67%) demonstrates how AWPM combines completion sensitivity with severity-weighted error deductions. The example also illustrates the interaction between incomplete work and error penalties, whereby the deduction increases as the completion ratio decreases. The selected parameter values are hypothetical and demonstrate the model's computational operation rather than empirically validated scoring standards.

A systematic calibration procedure is proposed to establish context-appropriate AWPM parameters before assessment implementation. The procedure comprises four stages: (1) defining criterion weights and performance expectations based on the approved rubric; (2) establishing error-severity weights through expert moderation; (3) selecting initial sensitivity parameters and examining their effects using representative assessment scenarios; and (4) refining and fixing the parameters following moderation and sensitivity testing. After checking whether the proposed deductions are proportionate and avoid unintended score reversals, the calibrated values should remain consistent across assessors and students undertaking the same assessment to ensure procedural comparability. This procedure provides a reproducible basis for parameter selection while allowing adaptation to different assessment contexts. Nevertheless, the proposed calibration process requires empirical verification before its effectiveness can be established.

Functional and Scenario-Based Evaluation

At the current stage, the prototype was evaluated functionally rather than through a controlled multi-rater experiment. Testing examined whether identical structured inputs generated identical outputs, whether increasing error severity or frequency produced non-increasing scores, whether reduced completion appropriately lowered the attainable score, whether negative marks were prevented, and whether outputs could be traced to the configured parameters and assessor inputs. Scenario testing also examined boundary conditions such as full completion with no errors, partial completion, repeated minor errors, and

intermediate or major errors. These tests establish computational correctness and behavioural plausibility, but they do not demonstrate improved inter-rater reliability. A subsequent validation study should compare conventional and AWPM-assisted scoring using common student artefacts and multiple independent assessors. Inter-rater reliability should be quantified using an appropriate agreement statistic, such as an intraclass correlation coefficient or generalizability analysis, rather than relying only on score correlations (Arslan Mancar & Gülleroğlu, 2022; Goertzen & Klaus, 2023). The scenario-based evaluation is further structured around two complementary analyses. Parameter sensitivity analysis examines changes in computed scores resulting from variations in model parameters, while comparative analysis evaluates AWPM against simple weighted scoring and fixed-deduction approaches under equivalent hypothetical assessment conditions. These analyses focus on computational behaviour and scoring transparency rather than empirical improvements in inter-rater reliability or marking efficiency. As detailed in Table 1, the corresponding controls address inconsistent arithmetic, severity weighting, completion, error–completion interaction, calculation errors and weak audit trails. The table distinguishes intended functional effects from outcomes that would require empirical verification.

Table 1 Functional requirements and corresponding AWPM controls

Assessment issue	AWPM/system control	Expected effect
Different arithmetic or deduction practices	Single configurable scoring equation	Same inputs receive the same numerical treatment
Unequal treatment of error severity	Severity weights (k_i)	Larger penalties for more consequential errors
Incomplete work treated inconsistently	Completion ratio (C) and sensitivity β	Explicit adjustment for degree of completion
Interaction between incompleteness and errors	Adaptive factor γ	Higher penalty when errors coexist with lower completion
Manual calculation mistakes	Automated scoring engine	Removal of repetitive assessor arithmetic
Weak audit trail	Structured input and output records	Traceable relationship between judgement, parameters and mark

FINDINGS AND DISCUSSION

Standardisation of the Computational Stage

The principal outcome of the prototype is the separation of academic judgement from numerical processing. In conventional rubric practice, two assessors may identify a comparable deficiency yet apply different numerical deductions because each uses an implicit scoring rule. AWPM makes that rule explicit. Once criterion weights and penalty parameters are approved and fixed, equivalent combinations of completion and error inputs are processed identically. The model therefore targets a specific source of avoidable variability: inconsistency in translating structured judgements into numerical marks. It does not eliminate differences arising from assessors' interpretation of student performance. This distinction is supported by previous studies showing that assessment reliability depends on rubric design, rater behaviour and assessment context (Xiong et al., 2023; Goertzen & Klaus, 2023). Furthermore, automated grading systems may also produce inconsistent evaluations, particularly when assessing complex performance tasks (Castelblanco et al., 2024; Flodén, 2025). AWPM therefore does not claim to eliminate assessment subjectivity but seeks to standardise the numerical consequences of assessor judgements. This approach preserves professional judgement while improving the reproducibility and transparency of the scoring procedure.

Adaptive Penalty Behaviour and Transparency

The weighted penalty structure offers greater discrimination than a fixed deduction approach. Minor, intermediate and major errors can carry different costs, while the frequency exponent α allows the designer to determine whether repeated errors should accumulate linearly or more strongly. Completion sensitivity β provides a separate mechanism for incomplete performance. The interaction term γ further recognises that the meaning of an error may differ when a task is substantially incomplete. Because each parameter has a

defined role, the scoring logic can be documented and discussed during moderation rather than remaining embedded in individual assessor habits. A one-at-a-time sensitivity analysis was conducted using the same hypothetical 20-mark assessment scenario, with baseline parameters of $\alpha = 1.0$, $\beta = 1.2$, $\gamma = 0.3$ and $\lambda = 0.75$. Each parameter was varied independently while the remaining parameters and assessment inputs were held constant. As presented in Table 2, increasing the sensitivity and scaling parameters generally reduced the criterion score, although the magnitude of change varied. The penalty scaling factor (λ) produced the largest variation, ranging from 15.56 to 13.50 marks, followed by completion sensitivity (β), ranging from 14.91 to 13.99 marks. Error-frequency sensitivity (α) produced scores between 14.53 and 13.89 marks, while the interaction factor (γ) had a comparatively smaller effect, with scores ranging from 14.55 to 14.47 marks. These findings indicate predictable score responses within the selected parameter ranges, although broader testing involving different error distributions and assessment conditions is required to establish model robustness.

Transparency is also strengthened by the spreadsheet implementation. The project prototype contains an input sheet, assisted-entry interface and output sheet, while the architecture connects rubric configuration to the scoring engine and reporting layer. The presentation identifies assisted entry, controlled workflow logic, automated computation and a transparent dashboard as core features. These functions are directly relevant to auditability because they make the path from assessment decision to final mark visible rather than relying on unrecorded mental calculations.

Table 2 One-at-a-time sensitivity of the illustrative AWPM criterion score

Parameter	Values tested	AWPM score (marks)
α	1.00 / 1.25 / 1.50	14.53 / 14.24 / 13.89
β	1.00 / 1.25 / 1.50	14.91 / 14.44 / 13.99
γ	0.25 / 0.375 / 0.50	14.55 / 14.51 / 14.47
λ	0.50 / 0.75 / 1.00	15.56 / 14.53 / 13.50

Comparative Assessment of Scoring Approaches

To examine differences between AWPM and conventional scoring approaches, three scoring rules were applied to the same hypothetical assessment scenario: simple completion-weighted scoring, fixed deductions and AWPM. The simple approach calculates the score directly from criterion weight and completion ratio, whereas fixed deductions subtract predefined error costs from the completion-adjusted score. AWPM additionally incorporates completion sensitivity, error-frequency sensitivity and an interaction-adjusted penalty. As presented in Table 3, the respective scores were 18.00, 14.00 and 14.53 marks. The differences demonstrate how the selected scoring rules respond to equivalent assessment inputs. Although all three approaches can produce reproducible scores when their rules are explicitly defined, AWPM provides additional flexibility in differentiating error frequency, severity and incomplete performance. Its explicit parameters also support the documentation and examination of scoring decisions. However, the comparison does not establish improvements in inter-rater reliability, assessor workload or assessment fairness, which require empirical evaluation.

Table 3 Illustrative comparison of alternative scoring rules for the same assessment scenario

Scoring rule	Specified computation	Score / 20
Simple completion-weighted	20×0.90	18.00
Fixed deduction	$20 \times 0.90 - (2 \times 1 + 1 \times 2)$	14.00
AWPM	$20(0.90)^{1.2} - 0.75(1.03)(4)$	14.53

Operational and Institutional Implications

The spreadsheet-based implementation is intentionally pragmatic. Universities commonly have access to spreadsheet software, and an institution can deploy a rule-based workbook without acquiring a dedicated assessment platform. The prototype can therefore be applied to coursework and laboratory assessment,

projects and presentations, studio or technical evaluations, and skills-based assessment. It may also support outcome-based education processes where common criteria must be applied across sections or assessors. The approach may reduce repetitive workload because assessors enter structured observations while the system performs the arithmetic and consolidates outputs. It can also support moderation by giving assessors a common computational reference. These benefits should nevertheless be treated as design propositions until measured empirically. Recent research on digital and AI-supported assessment shows that efficiency and consistency gains must be balanced against validity, transparency and context sensitivity (Banihashem et al., 2025; Flodén, 2025). AWPM responds to this concern by using an interpretable rule rather than a black-box prediction model. Implementation may also present challenges involving assessor familiarity, consistent classification of errors and acceptance of predefined penalty rules. Clear descriptors, initial moderation and periodic review can help address these issues; students might otherwise concentrate on avoiding specified penalties rather than demonstrating full competence. Configurations should therefore be adapted to learning outcomes and assessment contexts: technical and practical tasks may permit explicit error categories, whereas creative or interpretative tasks may require more flexible descriptors and greater assessor discretion. Assessor acceptance and student perceptions of fairness remain subjects for future empirical study. Adaptation may also be necessary across educational levels and institutional assessment practices. Introductory undergraduate assessments may employ simpler error classifications and parameter settings, whereas advanced or postgraduate assessments may require more differentiated criteria to reflect complex learning outcomes. Institutional adoption would consequently require alignment with approved assessment policies, assessor moderation procedures and established quality assurance practices.

Limitations and Validation Requirements

The current study is a model-development and prototype paper. It demonstrates a computational mechanism and a functioning assessment workflow, but it does not yet provide experimental evidence that inter-rater reliability, marking time or student perceptions of fairness improve in actual courses. Parameter selection also requires disciplinary calibration. Future validation should use authentic student work, multiple independent assessors and a comparison between conventional and AWPM-assisted scoring. Inter-rater agreement should be assessed with statistics appropriate to the scoring design, while score dispersion, marking time, usability and perceived transparency should also be measured (Arslan Mancar & Gülleroğlu, 2022; Goertzen & Klaus, 2023). Although the illustrative sensitivity analysis demonstrates predictable computational responses within the selected parameter ranges, it does not establish robustness across alternative error distributions, severity-cost configurations or authentic assessment conditions. Further empirical investigation is therefore required to determine the reliability, usability and educational defensibility of AWPM under actual assessment practices.

CONCLUSION

This paper presents the Adaptive Weighted Penalty Model as a computational approach to improving the consistency and transparency of rubric-based assessment. The model retains the assessor's academic judgement while standardising the numerical treatment of that judgement through configurable criterion weights, completion ratios, severity-weighted errors and adaptive sensitivity parameters. Its spreadsheet-based implementation integrates rubric configuration, assisted entry, automated computation and structured reporting in a single workflow.

The main contribution is therefore procedural rather than fully automated judgement. AWPM does not attempt to decide what constitutes good performance. It makes the consequences of assessor decisions explicit, repeatable and auditable. This positioning is important because recent research shows that both human and automated assessment can exhibit variability, particularly for complex performance tasks. A transparent rule-based layer can reduce avoidable computational inconsistency without transferring academic authority to an opaque algorithm.

The next research stage should validate the model empirically across assessors, courses and assessment types. Particular attention should be given to inter-rater reliability, parameter calibration across disciplines,

assessor workload and student perceptions of fairness. If these studies confirm the expected benefits, the model could provide a low-cost foundation for wider institutional use in project, practical, presentation, studio and competency-based assessment. The worked numerical example, parameter sensitivity analysis and comparative evaluation demonstrate the model's computational behaviour under defined hypothetical conditions, providing an initial basis for further investigation. However, empirical validation involving multiple assessors and authentic student assessments remains necessary to establish its effectiveness in improving assessment reliability, fairness and operational efficiency.

ACKNOWLEDGEMENTS

The authors acknowledge the support of Universiti Teknologi MARA Cawangan Pulau Pinang and colleagues at the Faculty of Civil Engineering who contributed to the development and refinement of the assessment workflow.

REFERENCES

1. Arslan Mancar, S., & Gülleroğlu, H. D. (2022). Comparison of inter-rater reliability techniques in performance-based assessment. International Journal of Assessment Tools in Education, 9(2), 515–533.
2. Banihashem, S. K., Gašević, D., & Noroozi, O. (2025). A critical review of using learning analytics for formative assessment: Progress, pitfalls and path forward. Journal of Computer Assisted Learning, 41(3), e70056. https://doi.org/10.1111/jcal.70056
3. Castelblanco, G., Cruz-Castro, L., & Yang, Z. (2024). Performance of a large-language model in scoring construction management capstone design projects. Computer Applications in Engineering Education, 32(6), e22796. https://doi.org/10.1002/cae.22796
4. Fernández-Sánchez, A., Lorenzo-Castiñeiras, J. J., & Sánchez-Bello, A. (2025). Navigating the future of pedagogy: The integration of AI tools in developing educational assessment rubrics. European Journal of Education, 60, e12826. https://doi.org/10.1111/ejed.12826
5. Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. British Educational Research Journal, 51(1), 201–224. https://doi.org/10.1002/berj.4069
6. Goertzen, B. J., & Klaus, K. (2023). Is it actually reliable? Examining statistical methods for inter-rater reliability of a rubric in graduate education. Research & Practice in Assessment, 18(2), 31–41.
7. Großmann, L., & Krüger, D. (2024). Assessing the quality of science teachers' lesson plans: Evaluation and application of a novel instrument. Science Education, 108(1), 153–189. https://doi.org/10.1002/sce.21832
8. Lin, Z., Yan, H., & Zhao, L. (2024). Exploring an effective automated grading model with reliability detection for large-scale online peer assessment. Journal of Computer Assisted Learning, 40(4), 1535–1551. https://doi.org/10.1111/jcal.12970
9. Nalbantoğlu Yılmaz, F. (2024). Comparing the reliability of performance task scores obtained from rating scale and analytic rubric using the generalizability theory. Studies in Educational Evaluation, 83, 101413. https://doi.org/10.1016/j.stueduc.2024.101413
10. Xiong, Y., Schunn, C. D., & Wu, Y. (2023). What predicts variation in reliability and validity of online peer assessment? A large-scale cross-context study. Journal of Computer Assisted Learning, 39(6), 2004–2024. https://doi.org/10.1111/jcal.12861
11. Xu, Z., & Liu, P. (2026). Assessing argument quality in L2 argumentative writing: Integrating argument elements and intellectual standards. International Journal of Applied Linguistics. Advance online publication. https://doi.org/10.1111/ijal.70219