OpenMind Journal of Humanities Arts & Creative Studies
Vol. 2, Issue – 6, (Nov-Dec) 2026
ISSN: 3139-2466 (Online)

Research Article

Developing and Validating a Rubric for Assessing 21st-Century Skills Creativity, Collaboration and Critical Thinking) in Project-Based Learning: A Many-Facet Rasch Measurement Approach in Imo State, South-East Nigeria

Nathaniel TimAhanna Ukwuoma, PhD[*1], Joy Onyeka Nkeiruka[2] & Celestine Kelvins Ihekwoaba, PhD[3]

[1, 2 & 3] *Department of Educational Psychology/Guidance and Counselling, Faculty of Education, Alvan Ikoku Federal University of Education, Owerri, Imo State, Nigeria*

***Related declarations are provided in the final section of this article*

ARTICLE INFORMATION

Received: 15 Oct 2026

Accepted: 10 Nov 2026

Published: 19 Nov 2026

Correspondence Author:

Nathaniel TimAhanna Ukwuoma, PhD *

Copyright © 2026 The Author(s). The Authors retain copyright. Licensed under the **Creative Commons Attribution 4.0 International License** (CC BY 4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Project-based learning (PBL) is promoted as a way of developing 21st-century skills, yet teachers need valid and fair tools to assess skills such as creativity, collaboration and critical thinking. This study developed a nine-criterion, four-level analytic rubric and validated it with a many-facet Rasch measurement (MFRM) approach in Imo State, South-East Nigeria. Content validity was established through review by eight experts, and the rubric was piloted before the main study. In the main study, six trained teachers rated 120 Senior Secondary 2 students' PBL work in a linked design, producing 3,240 ratings. The rating-scale many-facet model showed that the rubric defined a coherent variable: all criteria and raters fell within acceptable fit limits (infit and outfit mean squares 0.88 to 1.29), and the first residual contrast (eigenvalue = 1.58) gave no sign of a second dimension. The rubric separated students reliably (separation = 3.71; reliability = .93; five strata). Raters differed significantly in severity, spanning about one logit (−0.49 to 0.52), and raw scores shifted the rankings of 22 students (18.3%) by ten or more places relative to rater-adjusted measures. Only three of 54 rater-by-criterion interactions were flagged, close to chance. Rasch measures correlated with teachers' project grades (r = .78). The findings support the rubric as a valid and reliable school-based tool and show why rater effects should be modelled rather than ignored.

Keywords: Rubric Validation; Many-Facet Rasch Measurement; 21st-Century Skills; Project-Based Learning; Rater Effects; Nigeria

INTRODUCTION

Education systems worldwide have sought to develop 21st-century skills, such as creativity, collaboration and critical thinking, alongside subject knowledge (van Laar et al., 2020). Project-based learning (PBL) is a widely used approach for this purpose, because students work on authentic problems over time and produce public products (Guo et al., 2020; Krajcik & Shin, 2022). Meta-analytic evidence suggests PBL can improve learning outcomes, although effects depend on design and implementation (Zhang & Ma, 2023).

A major difficulty is assessment. Skills such as creativity or collaboration cannot be captured by multiple-choice tests, so teachers rely on rubrics and judgement. Rubrics can improve performance, self-regulation and transparency when well designed (Panadero & Jonsson, 2020; Panadero et al., 2023), but ratings are mediated by raters, whose severity and consistency vary. If this variation is ignored, scores may reflect who rated the work as much as what students can do (Eckes, 2023; Wind, 2023). Validity, rather than mere convenience, should drive assessment design (Dawson et al., 2024).

Many-facet Rasch measurement (MFRM) extends the Rasch model to rater-mediated assessments by estimating student proficiency, rater severity and criterion difficulty on one logit scale, and by testing whether the rating scale functions as intended (Aryadoust et al., 2021; Eckes, 2023). It is well suited to developing and validating rubrics, yet it has seldom been applied to 21st-century skills in African secondary schools.

In Nigeria, assessment in secondary schools is heavily shaped by external examinations such as those of the West African Examinations Council (WAEC) and the National Examinations Council (NECO), and tools for assessing skills through projects are scarce. In Imo State, in the South-East geopolitical zone, a validated rubric could support teachers who wish to broaden assessment.

This study developed and validated a rubric for creativity, collaboration and critical thinking in PBL, and addressed five questions:

i. **RQ1.** Does the rubric define a coherent measurement variable, with acceptable fit and one-dimensionality?
ii. **RQ2.** Do the rating scale categories function as intended?
iii. **RQ3.** How much do raters differ in severity and consistency, and do they interact with criteria?
iv. **RQ4.** How reliably does the rubric separate students?
v. **RQ5.** What validity evidence supports the interpretation of rubric measures?

LITERATURE REVIEW

21st-Century Skills in Project-Based Learning

Creativity, collaboration and critical thinking are consistently identified as core 21st-century skills (van Laar et al., 2020). In PBL, students generate ideas for solving real problems, work in

groups and evaluate options, which makes the three skills observables in the same task (Guo et al., 2020; Krajcik & Shin, 2022). Research on assessing them is developing. Collaborative problem solving has been assessed through social and cognitive dimensions (Andrews-Todd & Forsyth, 2020; Herborn et al., 2020), computational approaches are emerging for creativity (Beaty & Johnson, 2021), and instructors differ in how they define and prioritise critical thinking (Bellaera et al., 2021). These findings indicate a need for clear, shared criteria.

Rubrics and Rater-Mediated Assessment

Analytic rubrics specify criteria and performance levels, which supports fairness and feedback (Panadero & Jonsson, 2020; Panadero et al., 2023). However, rubrics do not remove rater effects such as severity, central tendency and inconsistency (Eckes, 2023). Classroom assessment guidance also emphasises that the quality of teacher judgement depends on training and moderation (Brookhart & McMillan, 2020; Schildkamp et al., 2020). Teachers' assessment literacy is therefore central to using rubrics well (Ayanwale et al., 2022; Pastore, 2023).

Many-Facet Rasch Measurement

MFRM models the probability of a rating as a function of student proficiency, rater severity, criterion difficulty and rating scale thresholds, placing all facets on a common logit scale (Eckes, 2023; Khine, 2020). Evidence from MFRM includes separation and reliability indices, fit statistics, rating scale diagnostics, bias (interaction) analysis and variable maps (Aryadoust et al., 2021; Boone & Staver, 2020; Wind, 2023). Validation frameworks require evidence about content, response processes, internal structure and relations to other variables (AERA et al., 2014; Dawson et al., 2024). Authentic, inclusive and valid assessment design remains a priority (Ajjawi et al., 2024; Brown, 2022; Sokhanvar et al., 2021; Tai et al., 2023), and the arrival of generative AI strengthens the case for assessing process and collaboration (Lodge et al., 2023; Swiecki et al., 2022).

METHODOLOGY

Research Design

The study used an instrument-development and validation design in three phases: rubric development and expert review, a pilot, and a main validation study analysed with MFRM (Khine, 2020).

Rubric Development

A draft rubric was built from the literature on creativity, collaboration and critical thinking in PBL (Andrews-Todd & Forsyth, 2020; Bellaera et al., 2021; van Laar et al., 2020). Each construct had three criteria, giving nine criteria rated on four levels (1 = beginning, 2 = developing, 3 = proficient, 4 = exemplary) with descriptors for each level. Eight experts in assessment, teacher education and secondary teaching judged the relevance of each criterion; item-level content validity indices ranged from .88 to 1.00 and the scale-level average was .96. A pilot with 24 students and three raters led to wording changes for two criteria. Table 1 shows the final rubric structure.

Table 1

Structure of the Rubric

Construct	Code	Criterion
Creativity	CR1	Originality of ideas
	CR2	Elaboration of ideas
	CR3	Fit of solution to the problem
Collaboration	CO1	Sharing responsibilities
	CO2	Contributing and listening
	CO3	Managing disagreement
Critical thinking	CT1	Analysing the problem and evidence
	CT2	Evaluating options and reasoning
	CT3	Reflecting on process and outcome

Participants, Task and Raters

The main study involved 120 Senior Secondary 2 students (66 female, 55.0%; 54 male, 45.0%) from 12 public and private schools across the three senatorial zones of Imo State (Owerri, Okigwe and Orlu), selected by stratified random sampling. Over six weeks, students worked in groups of four to five on a local problem project (for example, school waste management or water supply) and produced a report, a product or model and a presentation. Six experienced teachers served as raters after a one-day training using anchor samples. Each student was rated on all nine criteria by three raters in a linked, rotating design, producing 3,240 ratings (Table 2).

Table 2

Design of the Rating Study

Facet	Elements	Detail
Students	120	66 female, 54 male; 12 schools
Raters	6	Each student rated by 3 raters (linked design)
Criteria	9	3 creativity, 3 collaboration, 3 critical thinking
Rating scale	4 levels	1 = beginning to 4 = exemplary
Total ratings	3,240	120 students × 3 raters × 9 criteria

Measurement Model and Analysis

A rating-scale many-facet Rasch model was specified:

$$\ln\left(P_{nijk} / P_{nijk-1}\right) = \theta_n - \rho_j - \delta_i - \tau_k$$

where θ_n is the proficiency of student n, ρ_j is the severity of rater j, δ_i is the difficulty of criterion i, and τ_k is the threshold between categories $k - 1$ and k. Rater and criterion facets were centred at zero logits. Analyses followed recommended practice (Aryadoust et al., 2021; Eckes, 2023): separation, strata and reliability for each facet; fixed chi-square tests of homogeneity; infit and outfit mean squares judged against the 0.5 to 1.5 guideline; category and threshold diagnostics (Wind, 2023); principal component analysis of standardised residuals for unidimensionality; rater-by-criterion bias analysis; and correlation of measures with an external criterion, the teacher's end-of-term project grade (0–100). School authorities approved the study, guardians consented, students assented, and all data were anonymised.

RESULTS

Overall Model Quality and Reliability

Table 3 summarises the three facets. Students spread over a wide range (*SD* = 1.05 logits) and the rubric separated them reliably (separation = 3.71, strata = 5.28, reliability = .93), meaning it distinguished about five statistically distinct levels of proficiency (RQ4). Raters and criteria also differed reliably, as the significant chi-square tests show, which is expected and desirable for criteria but signals real severity differences among raters.

Table 3

Summary Statistics for the Three Facets

Facet	*n*	Mean (logits)	*SD*	Mean *SE*	Separation	Strata	Reliability	Fixed χ^2 (*df*)
Students	120	0.30	1.05	0.27	3.71	5.28	.93	—
Raters	6	0.00	0.38	0.06	6.51	9.02	.98	215.9 (5)*
Criteria	9	0.00	0.40	0.07	5.50	7.67	.97	246.8 (8)*

Note. * $p < .001$. Student infit mean square averaged 1.00 (*SD* = 0.24); 5.0% of students had infit above 1.5.

Criterion Functioning and Dimensionality (RQ1)

All nine criteria fitted the model (Table 4): infit ranged from 0.88 to 1.25 and outfit from 0.91 to 1.29. CO3 (managing disagreement) had the highest mean squares (infit 1.25, outfit 1.29), indicating slightly more unpredictable responses but still within limits. Difficulty ranged from −0.52 (CO1) to 0.53 logits (CT1). The hardest criteria were analysing the problem and evidence (CT1) and originality of ideas (CR1), whereas sharing responsibilities (CO1) and reflecting on the process (CT3) were easiest (Figure 1). The first contrast in the residuals had an eigenvalue of 1.58, below the usual 2.0 benchmark, and the second and third were 1.38 and 1.31, so there was

no evidence of a second dimension: the nine criteria worked together to define a single skills variable.

Table 4

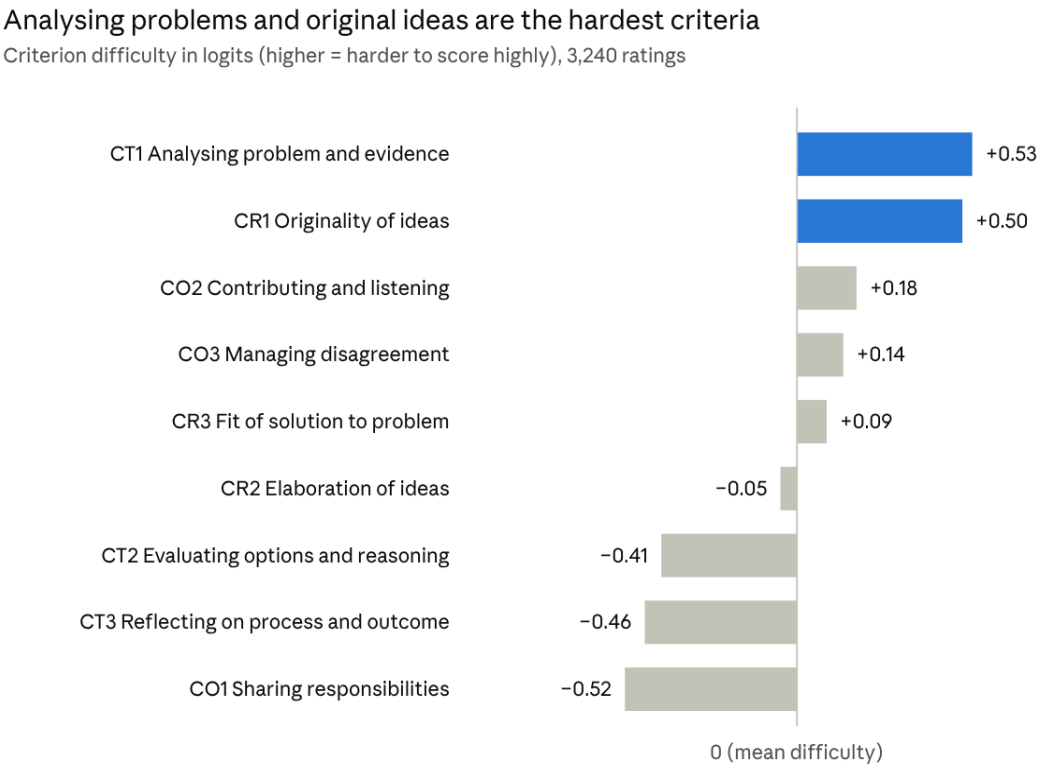
Criterion Measurement Report

Criterion	Difficulty (logits)	*SE*	Infit MNSQ	Outfit MNSQ
CT1 Analysing problem and evidence	0.53	0.07	0.94	0.95
CR1 Originality of ideas	0.50	0.07	0.94	0.92
CO2 Contributing and listening	0.18	0.07	1.03	1.04
CO3 Managing disagreement	0.14	0.07	1.25	1.29
CR3 Fit of solution to problem	0.09	0.07	0.92	0.92
CR2 Elaboration of ideas	−0.05	0.07	0.88	0.91
CT2 Evaluating options and reasoning	−0.41	0.07	1.02	0.99
CT3 Reflecting on process and outcome	−0.46	0.07	0.95	0.94
CO1 Sharing responsibilities	−0.52	0.07	1.04	1.03

Note. MNSQ = mean square. Higher difficulty means students were less likely to score highly.

Figure 1

Criterion Difficulty Measures (Logits)



N3,240 ratings

Rating Scale Functioning (RQ2)

The four categories functioned in order (Table 5). Average measures increased steadily from −0.94 to 1.38 logits, thresholds were ordered (−1.26, −0.04 and 1.30), and each category was used by at least 14% of ratings. Thresholds advanced by 1.22 and 1.34 logits, close to the 1.4 logit guideline. Outfit was elevated for category 1 (1.82) and category 4 (1.41), indicating some unexpected extreme ratings, which suggests that the descriptors for the lowest level could be sharpened.

Table 5

Rating Scale Category Statistics

Category	Count	%	Average measure	Outfit MNSQ	Threshold	Step distance
1 Beginning	470	14.5	−0.94	1.82	—	—
2 Developing	920	28.4	−0.25	0.68	−1.26	—
3 Proficient	1,089	33.6	0.55	0.63	−0.04	1.22
4 Exemplary	761	23.5	1.38	1.41	1.30	1.34

Rater Severity and Consistency (RQ3)

Raters differed significantly in severity (Table 6; Figure 2), spanning about one logit from −0.49 (Rater 1, most lenient) to 0.52 (Rater 6, most severe). In raw terms, Rater 1 averaged 3.01 and Rater 5 averaged 2.32 on the four-point scale. Fit was acceptable for all raters (infit 0.92 to 1.22; outfit 0.93 to 1.24), with Rater 5 the least consistent but within limits. Observed exact agreement between raters on the same student and criterion was 37.0%, almost identical to the 37.7% expected by the model, so raters used the scale as the model predicts, though they did not agree identically. Bias analysis (Table 7) flagged three of 54 rater-by-criterion cells (5.6%), approximately the chance rate, with one strong interaction: Rater 1 scored CO3 (managing disagreement) more severely than expected (−0.62 logits, $z = -3.50$).

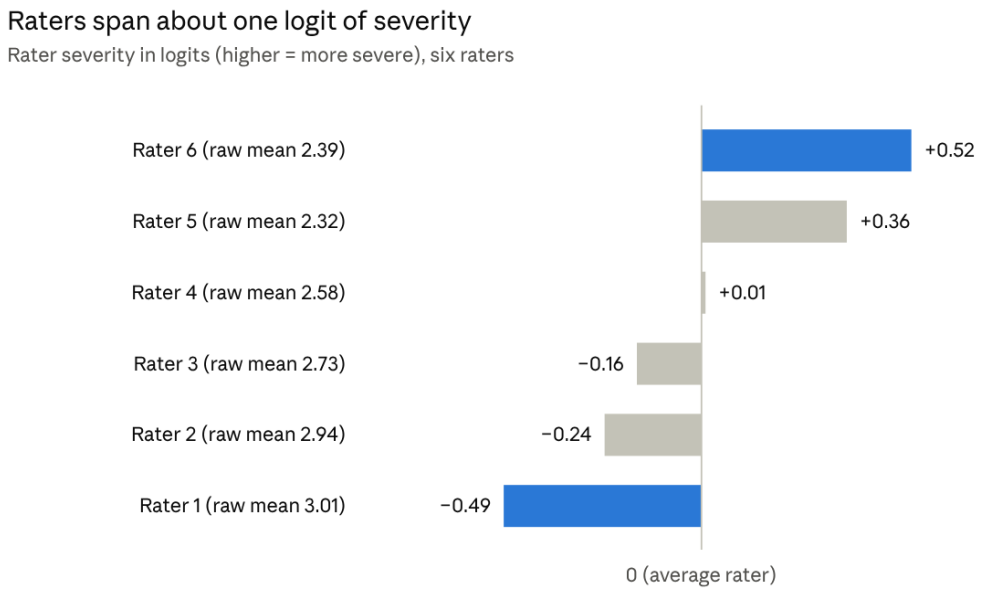
Table 6

Rater Measurement Report

Rater	Severity (logits)	*SE*	Infit MNSQ	Outfit MNSQ	Raw mean (1–4)
Rater 6	0.52	0.06	0.92	0.93	2.39
Rater 5	0.36	0.06	1.22	1.24	2.32
Rater 4	0.01	0.06	0.94	0.93	2.58
Rater 3	−0.16	0.06	0.93	0.94	2.73
Rater 2	−0.24	0.06	0.98	1.00	2.94
Rater 1	−0.49	0.06	0.99	0.96	3.01

Figure 2

Rater Severity Measures (Logits)



N3,240 ratings

Table 7

Flagged Rater-by-Criterion Interactions (|z| > 2)

Rater	Criterion	Ratings	Bias (logits)	*z*
Rater 1	CR1 Originality of ideas	60	0.47	2.75
Rater 1	CO3 Managing disagreement	60	−0.62	−3.50
Rater 5	CR1 Originality of ideas	60	−0.45	−2.60

Note. Positive bias means ratings were higher than the model expected. 54 rater-by-criterion cells were tested; 3 were flagged and 1 had |*z*| > 3.

Why Modelling Raters Matters, and Validity Evidence (RQ5)

Raw average scores and Rasch measures were strongly correlated (*r* = .97; Spearman = .98), but differences were consequential for individuals: the mean rank shift was 5.2 places, and 22 students (18.3%) moved ten or more places once rater severity was adjusted, with a maximum shift of 18.5 places (Table 8). Rasch measures correlated .78 with teachers' project grades, slightly higher than raw scores (.76), providing relations-to-other-variables evidence (RQ5).

Table 8

Comparison of Raw Scores and Rasch Measures, and Criterion Validity

Indicator	Value
Correlation, raw mean and Rasch measure (Pearson / Spearman)	.97 / .98

Indicator	Value
Mean absolute rank shift (places)	5.2
Students shifting 10 or more places, *n* (%)	22 (18.3)
Maximum rank shift (places)	18.5
Correlation with project grade: Rasch measure	.78
Correlation with project grade: raw mean	.76

Summary of Validity Evidence

Table 9

Summary of Validity Evidence for the Rubric

Source of evidence	Finding	Judgement
Content	I-CVI .88–1.00; S-CVI/Ave = .96; pilot refinement	Supported
Internal structure: fit	Criteria infit 0.88–1.25; raters 0.92–1.22	Supported
Internal structure: dimensionality	First contrast eigenvalue 1.58	Supported
Rating scale	Ordered thresholds; monotonic average measures	Supported, with minor category 1 misfit
Reliability	Student separation 3.71, reliability .93	Supported
Rater effects	Severity range ~1 logit; 3 of 54 interactions flagged	Rater adjustment needed
Relations to other variables	*r* = .78 with project grade	Supported

DISCUSSION

The MFRM analysis provided broad support for the rubric. The nine criteria worked together as a single variable, fitted the model, and separated students into about five reliable levels, which is sufficient for formative decisions and, with moderation, for reporting (Aryadoust et al., 2021; Eckes, 2023). The ordered category structure and the correlation with project grades add further

evidence in line with a validity argument that is built from several sources (AERA et al., 2014; Dawson et al., 2024).

The rater findings carry the main practical message. Severity differed by about one logit, equivalent to roughly 0.7 raw points between the most lenient and most severe raters, and 18.3% of students changed rank by ten or more places after adjustment. This confirms that raw rubric scores can be unfair when raters differ, even after training (Eckes, 2023; Wind, 2023). Raters were internally consistent, so the problem is not random error but systematic leniency and severity, which can be corrected statistically and reduced through moderation (Brookhart & McMillan, 2020; Schildkamp et al., 2020). The strongest interaction involved managing disagreement (CO3), the criterion with the highest mean squares, suggesting that collaboration is harder to observe and that descriptors or evidence sources for this criterion could be improved (Andrews-Todd & Forsyth, 2020; Herborn et al., 2020).

Criterion difficulty also has pedagogical meaning. Analysing problems and generating original ideas were hardest, in line with concerns that critical thinking and creativity are demanding and unevenly taught (Bellaera et al., 2021; van Laar et al., 2020), which can guide teachers on where to target instruction in PBL (Guo et al., 2020; Zhang & Ma, 2023). Finally, the elevated misfit at the extreme categories suggests sharpening the lowest-level descriptors.

CONCLUSION

The findings of the study indicate that the rubric is a valid and reliable instrument for assessing creativity, collaboration, and critical thinking in Project-Based Learning (PBL) within the study setting. However, its effectiveness in ensuring fair and accurate assessment depends on appropriate monitoring of rater severity and consistent application of the scoring criteria. The rubric therefore provides a useful framework for supporting evidence-based assessment of students' learning outcomes in PBL, particularly in Imo State and similar educational contexts in South-East Nigeria.

RECOMMENDATIONS

Based on the findings of the study, the following recommendations are made:

i. **Adopt the Rubric for School-Based Assessment of PBL:** Schools should adopt the rubric for assessing creativity, collaboration, and critical thinking in PBL activities. It should be used for formative feedback and, with appropriate moderation, for reporting students' performance.

ii. **Train and Moderate Raters:** Teachers and assessors should receive adequate training in rubric application. Anchor samples, regular moderation meetings, and monitoring of rater severity should be employed to promote scoring consistency and fairness, as raw scores alone may disadvantage some students (Eckes, 2023).

iii. **Use Multiple Raters for Important Decisions:** Two or more raters should be involved in important assessment decisions where feasible. Linked rating designs should be employed to adjust for differences in rater severity, particularly where resources permit.

iv. **Refine Collaboration Descriptors:** The collaboration dimension of the rubric should be further refined by providing clearer descriptors of observable behaviours, particularly managing disagreements, resolving conflicts, and defining the lowest performance level.

v. **Build Teacher Capacity:** Professional development should be provided to strengthen teachers' competence in rubric use, scoring, moderation, and Rasch-informed interpretation of assessment results (Ayanwale et al., 2022; Pastore, 2023).

LIMITATIONS AND FUTURE RESEARCH

The study was restricted to one project type, grade level, state, and six raters limiting the generalizability of the findings. Additionally, the analysis treated the nine criteria as a single composite variable; future work should test a multidimensional model, examine differential item functioning by gender and school type, evaluate response processes and stability over time, and compare findings using larger rater samples alongside AI-assisted scoring. (Lodge et al., 2023; Swiecki et al., 2022).

Article Publication Details
This article is published in the **OpenMind Journal of Humanities Arts & Creative Studies**, ISSN: 3139-2466 (Online). In Volume 2 (2026), Issue 6 (Nov - Dec) - 2026
The journal is published and managed by **OMR PUBLICATION** .

Acknowledgements

We sincerely thank the editors and the reviewers for their valuable suggestions on this paper.

Authors' contributions

All author(s) read and approved the final manuscript.

Funding

The author(s) declare that no funding was received for this work.

Data availability

No datasets were generated or analyzed during the current study.

Declarations

Ethics approval and consent to participate

The author(s) declare that not applicable.

Consent for publication

The author(s) declare that not applicable.

Competing interests

The author(s) declare no competing interests.

REFERENCES

1. Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education, 49*(4), 499–510.
2. American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.
3. Andrews-Todd, J., & Forsyth, C. M. (2020). Exploring social and cognitive dimensions of collaborative problem solving in an open online simulation-based task. *Computers in Human Behavior, 104*, Article 105759.
4. Aryadoust, V., Ng, L. Y., & Sayama, H. (2021). A comprehensive review of Rasch measurement in language assessment: Recommendations and guidelines for research. *Language Testing, 38*(1), 6–40.
5. Ayanwale, M. A., Sanusi, I. T., Adelana, O. P., Aruleba, K. D., & Oyelere, S. S. (2022). Teachers' readiness and intention to teach artificial intelligence in schools. *Computers and Education: Artificial Intelligence, 3*, Article 100099.

6. Beaty, R. E., & Johnson, D. R. (2021). Automating creativity assessment with SemDis: An open platform for computing semantic distance. *Behavior Research Methods, 53*(2), 757–780.
7. Bellaera, L., Weinstein-Jones, Y., Ilie, S., & Baker, S. T. (2021). Critical thinking in practice: The priorities and practices of instructors teaching in higher education. *Thinking Skills and Creativity, 41*, Article 100856.
8. Boone, W. J., & Staver, J. R. (2020). *Advances in Rasch analyses in the human sciences*. Springer.
9. Brookhart, S. M., & McMillan, J. H. (Eds.). (2020). *Classroom assessment and educational measurement*. Routledge.
10. Brown, G. T. L. (2022). The past, present and future of educational assessment: A transdisciplinary perspective. *Frontiers in Education, 7*, Article 1060633.
11. Dawson, P., Bearman, M., Dollinger, M., & Boud, D. (2024). Validity matters more than cheating. *Assessment & Evaluation in Higher Education, 49*(7), 1005–1016.
12. Double, K. S., McGrane, J. A., & Hopfenbeck, T. N. (2020). The impact of peer assessment on academic performance: A meta-analysis of control group studies. *Educational Psychology Review, 32*(2), 481–509.
13. Eckes, T. (2023). *Introduction to many-facet Rasch measurement: Analyzing and evaluating rater-mediated assessments* (3rd ed.). Peter Lang.
14. Guo, P., Saab, N., Post, L. S., & Admiraal, W. (2020). A review of project-based learning in higher education: Student outcomes and measures. *International Journal of Educational Research, 102*, Article 101586.
15. Herborn, K., Stadler, M., Mustafić, M., & Greiff, S. (2020). The assessment of collaborative problem solving in PISA 2015: Can computer agents replace humans? *Computers in Human Behavior, 104*, Article 105624.

16. Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education, 8*, Article 1270700.
17. Khine, M. S. (Ed.). (2020). *Rasch measurement: Applications in quantitative educational research*. Springer.
18. Krajcik, J. S., & Shin, N. (2022). Project-based learning. In R. K. Sawyer (Ed.), *The Cambridge handbook of the learning sciences* (3rd ed., pp. 72–92). Cambridge University Press.
19. Lodge, J. M., Howard, S., Bearman, M., Dawson, P., & Associates. (2023). *Assessment reform for the age of artificial intelligence*. Tertiary Education Quality and Standards Agency.
20. Panadero, E., & Jonsson, A. (2020). A critical review of the arguments against the use of rubrics. *Educational Research Review, 30*, Article 100329.
21. Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic review. *Educational Psychology Review, 35*, Article 113.
22. Pastore, S. (2023). Teacher assessment literacy: A systematic review. *Frontiers in Education, 8*, Article 1217167.
23. Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, Article 101602.
24. Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation, 70*, Article 101030.

25. Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, Article 100075.
26. Tai, J., Ajjawi, R., Bearman, M., Boud, D., Dawson, P., & Jorre de St Jorre, T. (2023). Assessment for inclusion: Rethinking contemporary strategies in assessment design. *Higher Education Research & Development, 42*(2), 483–497.
27. Van Laar, E., van Deursen, A. J. A. M., van Dijk, J. A. G. M., & de Haan, J. (2020). Determinants of 21st-century skills and 21st-century digital skills for workers: A systematic literature review. *SAGE Open, 10*(1).
28. Wind, S. A. (2023). Detecting rating scale malfunctioning with the partial credit model and generalized partial credit model. *Educational and Psychological Measurement, 83*(5), 953–983.
29. Zhang, L., & Ma, Y. (2023). A study of the impact of project-based learning on student learning effects: A meta-analysis study. *Frontiers in Psychology, 14*, Article 1202728.

PUBLISHER'S NOTE

OMR PUBLICATION remains neutral with regard to jurisdictional claims in published maps and institutional affiliations. The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of OMR PUBLICATION and/or the editor(s). OMR PUBLICATION disclaims responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.