Biometric Technology Today
ISSN (online): 1873-1880

AI-Supported Evaluation of Reflective Learning in Creative Practice Courses

Élodie Martin[1], Julien Moreau[1], Camille Laurent[1*]

[1] Faculty of Fine Arts and Music, The University of Melbourne, Southbank, Australia.

[1] Faculty of Fine Arts and Music, The University of Melbourne, Southbank, Australia.

[1] Faculty of Fine Arts and Music, The University of Melbourne, Southbank, Australia.

camille.laurent@ucl.ac.uk

Abstract: This study examines how AI-supported text analysis can be used to evaluate reflective learning in creative practice courses. The study is designed to collect approximately 2,400 reflective journals, 1,200 learning logs, 600 project statements, and 480 teacher feedback records from 400 students enrolled in creative practice courses over a 16-week semester. The research measures reflective depth, conceptual development, emotional expression, self-assessment quality, problem identification, revision awareness, and learning strategy use. Natural language processing methods, including topic modeling, sentiment analysis, semantic similarity analysis, text classification, and large language model-assisted rubric scoring, are used to analyze students' written reflections. The study further compares AI-generated scores with teacher-rated scores through correlation analysis, inter-rater reliability testing, mean absolute error, and intraclass correlation coefficient. Regression analysis is used to test whether reflective depth and revision awareness significantly predict project performance and learning improvement. The innovation of this study is that it does not treat reflective writing as purely qualitative evidence, but converts reflective learning into measurable textual indicators while preserving teacher interpretation as the final evaluative reference.

Keywords: AI-Supported Assessment, Reflective Learning, Creative Practice, Natural Language Processing, Text Analysis, Rubric Scoring, Formative Evaluation, Learning Process

Introduction

Reflective learning is an essential component of creative practice education because students are expected not only to produce artistic outcomes but also to understand and articulate how their ideas, methods, judgments, and decisions develop throughout the learning process. In studio-based, design, music, performance, and other practice-oriented courses, learning often occurs through repeated cycles of experimentation, evaluation, revision, and reinterpretation [1, 2]. Reflective journals, learning logs, process notes, and project statements provide important records of these cycles by documenting problem identification, decision making, emotional responses, technical difficulties, failed attempts, and changes in artistic judgment. Such records can reveal aspects of learning that are difficult to infer from final products alone. Previous research in arts and higher education has shown that structured reflection can support self-assessment, decision making, metacognitive awareness, and professional development [3, 4]. Recent research in public art education has also examined collaborative teaching models and the optimization of teaching-resource efficiency, indicating growing attention to structured and process-oriented approaches to supporting creative learning [5]. These developments suggest that the evaluation of creative learning should consider not only final performance but also the processes through which students interpret experience, revise decisions, and develop more informed creative strategies. Despite its educational value, reflective writing remains difficult to assess consistently. Students differ substantially in how they represent their learning processes. Some texts provide mainly descriptive accounts of completed activities, whereas others explain why a particular approach was adopted, how difficulties were interpreted, why revisions were made, and what was learned from unsuccessful attempts. These differences make reflective depth difficult to capture through conventional grading procedures. Rubrics can improve transparency and scoring consistency by defining criteria such as critical analysis, self-evaluation, evidence of revision, and connection between experience and future action [6, 7]. However, fixed criteria may also overlook individualized forms of expression, especially in creative disciplines where learning can be nonlinear and where artistic experimentation does not always follow standardized patterns. A reflective entry

that is linguistically simple may still demonstrate substantial conceptual change, while a fluent and well-structured text may remain largely descriptive. Assessment therefore needs to distinguish the quality of reflection from surface-level writing proficiency and to connect reflective evidence with the artistic decisions, revisions, and learning strategies that characterize creative practice.

Natural language processing provides new possibilities for examining reflective learning at a larger scale. Topic modeling, sentiment analysis, semantic representation, and text classification have been used to identify recurring themes, emotional patterns, conceptual development, and different levels of reflection in educational texts. Context-sensitive language models can capture relationships between words and sentences more effectively than simple keyword-based methods and can therefore improve the identification of reflective elements that depend on contextual meaning [8, 9]. This is particularly important because reflection is rarely expressed in isolated sentences. A student may describe a problem in one part of a journal, explain its cause several sentences later, and discuss a revision strategy in another section [10, 11]. Sentence-level analysis may separate these related components and weaken the representation of the complete reflective process [12]. Document-level and context-aware analysis may therefore be more suitable for creative learning records in which ideas develop across multiple entries and stages of a project. Automated assessment research has nevertheless concentrated primarily on essays, short-answer responses, and subject-specific academic tasks, while reflective writing in creative practice courses has received comparatively less attention [13, 14]. Large language models have recently expanded the possibilities for rubric-based educational assessment because they can interpret longer passages, follow detailed scoring instructions, and generate rapid evaluations across large datasets. Existing studies indicate that such models can support automated or semi-automated scoring, but their performance varies according to the model, prompt design, assessment criteria, and type of writing being evaluated [15,16]. Research using large writing datasets has also reported moderate to strong agreement between AI-generated scores and human ratings under some conditions, although reliability remains sensitive to text type and scoring context [17, 18]. These findings suggest that high agreement in one assessment setting cannot automatically be generalized to reflective writing in creative education. Reflection often includes ambiguity, personal interpretation, incomplete ideas, emotional responses, and unconventional forms of reasoning that may be meaningful within a particular course or project but difficult for an automated system to evaluate without contextual information.

The role of teachers therefore remains central in the assessment of reflective learning. Teachers understand the objectives of the course, the sequence of project activities, the student's previous work, and the significance of particular artistic choices. They can distinguish between an unusual but meaningful creative decision and a poorly justified response in ways that may be difficult for a general-purpose language model. At the same time, manual assessment of large quantities of journals and learning logs requires considerable time and may itself be affected by scorer fatigue and differences in interpretation. AI-assisted assessment may therefore be most useful when it supports, rather than replaces, teacher judgment. Such a model could help identify patterns across large collections of reflective texts, provide preliminary rubric scores, and highlight entries that require closer human review. However, concerns about algorithmic bias, privacy, transparency, unclear scoring criteria, and the possible reduction of students' individual voices remain important barriers to fully automated evaluation [19, 20]. These concerns are particularly relevant in creative education, where assessment should preserve diversity of expression rather than encourage students to reproduce standardized forms of reflective writing. Important methodological gaps also remain in the existing literature. Many studies rely on relatively small samples drawn from a single course, assignment, or type of written response, making it difficult to determine whether automated indicators remain reliable across different stages of creative work [21, 22]. Reflective journals, learning logs, and project statements serve related but different purposes: journals may capture immediate reactions and difficulties, learning logs may summarize progress and strategy use, and project statements may provide more integrated explanations of conceptual development and final artistic decisions [23,24]. Analyzing only one form of writing may therefore provide an incomplete representation of the student's reflective process. Teacher feedback is also rarely incorporated into the same analytical framework, even though it provides an important external reference for understanding how students respond to instruction and revise their work. In addition, much of the automated assessment literature focuses on agreement between machine-generated and human-generated scores. Less attention has been given to whether the reflective characteristics detected in student writing are educationally meaningful—that is, whether they are associated with subsequent project quality, improvement over time, or evidence of stronger learning development.

To address these limitations, this study develops a multi-source analytical framework for examining reflective learning in creative practice education. The dataset includes approximately 2,400 reflective journals, 1,200 learning logs, 600 project statements, and 480 teacher feedback records collected from 400 students over a 16-week semester.

Topic modeling, sentiment analysis, semantic similarity analysis, text classification, and LLM-assisted rubric scoring are integrated to examine reflective depth, conceptual development, emotional expression, self-assessment quality, problem identification, revision awareness, and learning strategy use. Rather than treating these indicators as isolated textual features, the study investigates how they collectively represent changes in students' creative thinking and learning processes across the semester. AI-generated scores are compared with teacher ratings using correlation analysis, inter-rater reliability, mean absolute error, and intraclass correlation coefficients to determine the extent to which automated assessment can reproduce teacher judgments across different forms of reflective writing. Regression analysis is further used to examine whether reflective depth, revision awareness, and related learning indicators are associated with project performance and learning growth. By linking automated text analysis with teacher evaluation and educational outcomes, the study moves beyond the simple question of whether AI can reproduce a score and examines whether AI-assisted measures capture forms of reflection that are meaningful for creative learning. The study therefore positions AI as an analytical and decision-support tool rather than an independent evaluator, with teacher ratings serving as the primary reference for interpretation. This framework is intended to provide a more scalable and context-sensitive approach to reflective assessment while preserving the central role of human judgment in evaluating individualized creative development.

Materials and Methods

Participants and Study Settings

The study included 400 students from 16 creative practice courses at four higher education institutions. The courses covered visual arts, design, digital media, creative writing, and performance. Stratified cluster sampling was used to include different course types, study levels, and levels of prior creative experience. Students were included when they completed at least 12 weeks of the 16-week semester and submitted at least four reflective texts. The sample included 168 males and 232 females aged 18–32 years, with a mean age of 22.4 years. During the semester, 2,400 reflective journals, 1,200 learning logs, 600 project statements, and 480 teacher feedback records were collected. All texts were written in English and linked to course projects through coded student numbers.

Study Design and Comparison Groups

A quasi-experimental design with matched courses was used. Eight courses formed the AI-supported group, and eight formed the comparison group. Courses were matched by subject area, class size, study level, and mean baseline project score. Students in the AI-supported group received weekly feedback on reflective depth, problem identification, revision awareness, and learning strategies. The text analysis system produced the first feedback draft, which was checked by the course teacher before release. Students in the comparison group received standard teacher feedback without AI-based text scores. Teachers used the same rubric to grade final projects in both groups. The AI system did not assign final grades. This design tested the effect of AI-supported feedback while keeping teachers responsible for assessment.

Measures and Quality Control

Reflective learning was measured in seven areas: reflective depth, concept development, emotional expression, self-assessment, problem identification, revision awareness, and learning strategy use. Each area was scored on a five-point rubric. Two trained teachers independently rated a stratified sample of 600 texts. These scores were used as the reference data for model training and testing. Twenty percent of the texts were rated by both teachers to measure agreement. Intraclass correlation values above 0.75 were accepted. Cohen’s kappa was used to test agreement for topic labels and reflection categories, with values above 0.80 considered reliable. Student names, course codes, and personal details were removed before analysis. Very short texts, copied content, duplicate files, and records with more than 20% missing text were excluded. Ten percent of the records were checked against the original files.

Text Processing and Statistical Models

File headers, repeated symbols, and unrelated platform content were removed before analysis. Sentence boundaries were marked, and all texts were converted into a common format. Topic modeling was used to find common themes. Sentiment analysis measured positive, negative, and mixed emotional language. Semantic similarity was used to compare journals with project statements and teacher feedback. A text classifier and a rubric-based language model produced scores for the seven reflection areas. Agreement between AI and teacher scores was measured using mean absolute error (1):

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|A_i-T_i| \quad (1)$$

where A_i is the AI score and T_i is the teacher score for text i. The link between reflection and project performance was tested using the following regression model (2):

$$P_i=\beta_0+\beta_1 RD_i+\beta_2 RA_i+\beta_3 CD_i+\beta_4 SA_i+\gamma X_i+u_c+\varepsilon_i \quad (2)$$

where P_i is project performance, RD_i is reflective depth, RA_i is revision awareness, CD_i is concept development, and SA_i is self-assessment quality. X_i includes age, study level, prior creative experience, and baseline performance. u_c represents differences among courses.

Model Testing and Research Ethics

The dataset was divided into training, validation, and test sets at a ratio of 70:15:15. All texts from the same student were kept in the same set to prevent data leakage. AI scores were compared with teacher scores using Pearson correlation, intraclass correlation, mean absolute error, and score difference tests. Model performance was also checked across course type, gender, study level, and text length. This step was used to identify possible scoring bias. Teachers reviewed cases in which the AI score differed from the teacher score by more than one rubric point. AI results were used only for research and formative feedback. They were not used for final grading. All participants provided written consent, and ethical approval was obtained before data collection.

Results and Discussion

Agreement between AI and Teacher Scores

After data cleaning, 2,352 reflective journals, 1,174 learning logs, 589 project statements, and 472 teacher feedback records were retained. The two teachers showed good agreement when scoring the 600 reference texts, with an overall intraclass correlation coefficient of 0.84. Agreement across the seven areas ranged from 0.78 to 0.89. AI scores were also close to teacher scores. The mean correlation was 0.81, the intraclass correlation coefficient was 0.79, and the mean absolute error was 0.42 on the five-point scale. Revision awareness and problem identification showed the best agreement, with mean absolute errors of 0.34 and 0.36. Emotional expression showed the lowest agreement, with an error of 0.51. The score distributions are shown in Figure 1. These results indicate that clear actions, such as identifying a problem or explaining a revision, are easier to detect than personal feelings [25]. Zhang et al. (2026) reported correlations of 0.545–0.722 between human raters across several writing areas. The higher agreement in this study may be related to the use of a clear rubric and course-based examples [26]. However, the lower result for emotional expression shows that teacher review is still needed when scores depend strongly on personal context.

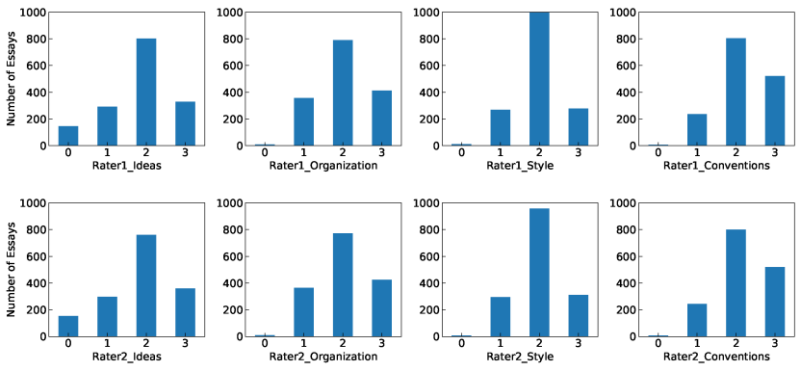

Figure 1. Comparison of Model and Teacher Scores across Seven Areas of Reflective Learning

Model Performance and Training Data Size

The text classifier reached a macro F1 score of 0.85 on the test set. Revision awareness had the highest F1 score at 0.89, followed by problem identification at 0.87 and reflective depth at 0.85. Emotional expression had the lowest score at 0.76. Model performance improved quickly when the training sample increased from 20% to 60% of the full dataset. The macro F1 score rose from 0.74 to 0.83. It increased by only 0.02 after the remaining data were added, as shown in Figure 2. This pattern suggests that a well-rated training set may be more useful than a larger set with unclear labels [27]. Francese et al. (2026) also found that BERT performed better than simpler models when more training data were added, with a final weighted F1 score of about 0.81. The present model reached a slightly higher score, although the two studies used different subjects, text units, and scoring rules. The weaker result for

emotional expression also shows that more data alone cannot solve all scoring problems [28]. Clearer labels and more course context are still needed.

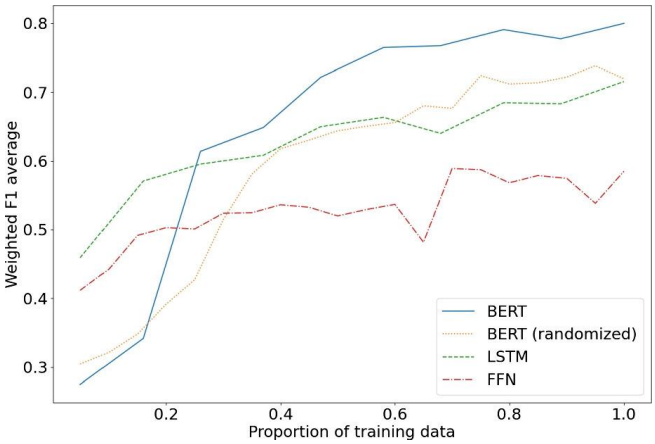

Figure 2. Classification Performance at Different Training Set Sizes

Changes in Reflective Learning and Project Performance

Students in the AI-supported group showed greater improvement than those in the comparison group. Their mean reflective depth score increased from 3.18 to 3.91, while the comparison group increased from 3.20 to 3.52. Revision awareness rose from 3.06 to 3.88 in the AI-supported group and from 3.08 to 3.43 in the comparison group. Similar gains were found for self-assessment and learning strategy use. Significant group-by-time effects were found for reflective depth, problem identification, revision awareness, and self-assessment quality ($p<0.001$). Effect sizes ranged from 0.56 to 0.74. The mean project score increased by 6.8 points in the AI-supported group and by 3.1 points in the comparison group. These results suggest that weekly feedback helped students explain their decisions more clearly and connect earlier problems with later changes [29]. Lin et al. (2026) also found that teacher and analytics feedback improved students' use of reflective writing. The present study extends this finding by linking feedback with both reflection quality and project performance. However, the group difference cannot be linked to AI alone because teachers checked the feedback before it was given to students.

Predictors of Learning and Limits of Automated Scoring

Regression analysis showed that reflective depth was the strongest predictor of final project performance, with a standardized coefficient of 0.31. Revision awareness followed at 0.27, while concept development and learning strategy use had coefficients of 0.19 and 0.13. Together, these variables explained 48% of the variance in project scores. Reflective depth and revision awareness also predicted learning improvement after baseline performance, course type, study level, and prior creative experience were controlled. Emotional expression was not a significant predictor. This does not mean that emotion is unimportant. Negative feelings may reflect confusion, failure, or useful struggle, and their meaning depends on the stage of the project. Teacher review found 86 texts in which AI and teacher scores differed by more than one point. AI often gave high scores to long texts that used reflective terms but offered little analysis. It sometimes gave low scores to short texts that clearly linked a problem with a design choice [30]. Wenger et al. (2025) also noted that automated systems may miss implied meaning that teachers can understand from context. The results therefore support the use of AI as a review tool rather than a final scorer. AI can help teachers screen large text sets, but teachers should review unclear cases and make final assessment decisions.

Conclusion

This study found that text analysis can help assess reflective learning in creative practice courses. Model scores were close to teacher scores for revision awareness and problem identification, but emotional expression was harder to rate. Students who received model-supported feedback showed greater gains in reflective depth, self-assessment, revision awareness, and project performance than those who received standard feedback. Reflective depth and revision awareness were also the strongest predictors of project outcomes. The study combines journals, learning logs, project statements, and teacher feedback to measure reflection across the learning process. This method can help teachers review large text sets, identify weak areas, and provide timely feedback. However, the model sometimes gave high scores to long texts with little analysis and missed clear meaning in short or indirect writing.

The study was limited to one semester and four institutions. Future research should test the method in more subjects, languages, and course settings, while keeping teachers responsible for final assessment

References

[1] H. Gui, Y. Fu, Z. Wang, and W. Zong, "Research on dynamic balance control of Ct gantry based on multi-body dynamics algorithm," 2025.

[2] K. G. D. Waduge, "A digital cognitive training for design education: Implications on spatial skills," Doctoral dissertation, Oklahoma State University, 2026.

[3] F. Chen, S. Li, H. Liang, P. Xu, and L. Yue, "Optimization study of thermal management of domestic SiC power semiconductor based on improved genetic algorithm," 2025.

[4] S. E. Banner, A. J. Rock, S. M. Cosh, N. Schutte, and K. Rice, "Self-reflection on competence: Metacognitive process and barriers of self-assessment in psychologists," *Advances in Health Sciences Education*, vol. 31, no. 1, pp. 103–123, 2026.

[5] T. Lin, A. Chen, and Z. Spivack, "Research on collaborative piano teaching models and teaching resource efficiency optimization in public art education systems," SSRN 6284818, 2026.

[6] B. Akkoyunlu, S. Bardakci, S. A. Yildirim, M. D. Avşaroğlu, G. Uludağ, A. Koçer, and M. Elmas, "Developing and validating a rubric-based approach to quality assurance in Turkish higher education," *Evaluation and Program Planning*, p. 102673, 2025.

[7] Y. Wang, J. Chen, R. Arias, Y. Wang, and X. Yin, "Development and validation of a patient-friendly digital assessment platform for precision screening of oral anti-obesity medications (AOMs)," 2026.

[8] W. Zhu, Y. Yao, and J. Yang, "Optimizing financial risk control for multinational projects: A joint framework based on CVaR-robust optimization and panel quantile regression," 2025.

[9] I. A. N. Arachchige, "Computational analysis of historical narratives through large language models," Doctoral dissertation, Lancaster University (United Kingdom), 2026.

[10] Y. Du, "Research on digital quality traceability system for temperature-controlled supply chain of foreign trade wine driven by blockchain and IoT," *Business and Social Sciences Proceedings*, vol. 4, pp. 57–65, 2025.

[11] Y. Wang, J. Chen, Y. Wang, and X. Yin, "Application of obtainable biological agent characteristics in efficacy stratification of oral anti-obesity drugs," 2026.

[12] S. R. A. Shah, T. Elyas, and C. Mitchell, "Sentence complexity in Arab EFL learners' academic writing: A qualitative case study," *Cogent Education*, vol. 13, no. 1, p. 2675017, 2026.

[13] Y. Jiao, B. Zhao, A. Wang, and T. Shi, "Construction and empirical study of a modularized teaching system for art courses based on a unified training pathway," 2026.

[14] J. Yang, "Stage-coupled computational framework for stratified accessibility and equity analysis in community-based elderly care services," 2026.

[15] M. Abd Jal, S. N. Kew, and G. Bunari, "ChatGPT as an automated essay scoring system in comparison with human raters in Malaysia," in *Rethinking the Pedagogy of Sustainable Development in the AI Era*, IGI Global Scientific Publishing, 2025, pp. 253–276.

[16] S. Hendawi, T. Kanan, M. Elbes, A. Mughaid, and S. Alzu'bi, "Automated prompt engineering pipelines: Fine-tuning LLMs for enhanced response accuracy," *Expert Systems with Applications*, p. 129689, 2025.

[17] Z. Zhang, "A study on return optimization in e-commerce for complex consumer goods driven by installation information quality," SSRN 6734720, 2026.

[18] C. Qi and X. Qiao, "Building and operating a large scale multi-agent system: A case study from industry," SSRN 6795198, 2026.

[19] D. Deepshikha, "A systematic review on the future of educational assessment: AI-driven grading and personalised feedback in higher education," in *Artificial Intelligence in Education*, Emerald Publishing Limited, Nov. 2025, pp. 1–41.

[20] D. Su and Y. Wu, "Multilevel growth and social network modeling for functional communication enhancement in students with intellectual disabilities," 2026.

[21] A. Mizumoto and M. F. Teng, "Large language models fall short in classifying learners' open-ended responses," *Research Methods in Applied Linguistics*, vol. 4, no. 2, p. 100210, 2025.

[22] G. Gao, R. Gao, C. Lu, R. Gao, and Y. Kuang, "Security governance methods and quantitative evaluation for enterprise SMS and verification code systems," in *2026 International Conference on Generative Artificial Intelligence and Information Security (GAIIS)*, Mar. 2026, pp. 455–458, IEEE.

[23] T. Xu, W. Zhu, and J. Zhang, "A study on the application of alternative data in credit assessment for the unbanked population," 2026.

[24] S. Chardonnens, "Adapting educational practices for Generation Z: Integrating metacognitive strategies and artificial intelligence," in *Frontiers in Education*, vol. 10, p. 1504726, Jan. 2025, Frontiers.

[25] Z. Zhang, Y. Tong, and Y. Gao, "Causal representation learning for explainable early warning in high-stakes decision systems," 2026.

[26] L. Yin, Y. Tong, Z. H. Islam, K. Zhang, R. Xie, J. Burger, *et al.*, "Targeted NAD+ delivery for intimal hyperplasia and re-endothelialization: A novel anti-restenotic therapy approach," *bioRxiv*, 2024-02, 2024.

[27] R. Francese, L. De Santis, F. Iannotta, and F. Iasevoli, "Prompt-based justification and scoring with large language models for thought and language disorder assessment," *Multimedia Tools and Applications*, vol. 85, no. 2, p. 170, 2026.

[28] W. Xiong, Y. Guo, and Y. Zeng, "A study on the energy efficiency of MEP systems and coordinated dispatch mechanisms for multi-energy systems in high-density urban buildings," SSRN 7120518, 2026.

[29] T. Lin, Z. Spivack, and A. Chen, "Monitoring teaching quality and validating educational equity in public art education," SSRN 6512778, 2026.

[30] E. Wenger and Y. Kenett, "We're different, we're the same: Creative homogeneity across LLMs," *arXiv preprint arXiv:2501.19361*, 2025.