

Digital Practice Platforms and Assessment Adjustment in English-Medium Data Science Education: A Four-Cohort Study

Tianzhen Chen^{1,*}

¹EIT Data Science and Communication College, Zhejiang Yuexiu University, Shaoxing, Zhejiang, China

*Corresponding author: terrychen17@zyufl.edu.cn

Abstract: Digital practice platforms can support repeated coding practice, but their value in English-medium undergraduate courses depends on how feedback, language demands, and assessment are aligned. This study analysed 293 valid assessment records from four successive offerings of an introductory data science course in a Sino-foreign cooperative programme. The cohorts represented Chinese-medium traditional teaching (2021), bilingual platform-supported teaching (2022), English-medium instruction with increased textbook and assessment difficulty (2023), and English-medium instruction with data-informed assessment adjustment (2024). Welch's one-way analysis of variance showed substantial cohort differences in overall grades, $F(3, 160.14) = 48.03$, $p < .001$, $\eta^2 = .331$. The 2022 cohort exceeded the 2021 cohort by 5.47 points ($p = .011$; Hedges' $g = 0.51$). The overall mean then fell by 18.18 points in 2023 before recovering by 15.38 points in 2024 ($p < .001$; $g = 1.68$). The overall pass rate recovered from 69.44% to 97.22%. Correlations between formative and final-examination grades remained positive but were numerically lower in the later cohorts. The results support a coordinated model in which digital practice, language scaffolding, formative tasks, and summative assessment difficulty are treated as one instructional system rather than independent reforms.

Keywords: Data Science Education, Digital Practice Platform, English-Medium Instruction, Formative Assessment, Assessment Adjustment

1. Introduction

Undergraduate data science is not a single-skill subject. Students are expected to move between statistical reasoning, programming, data management, visualisation, and interpretation in context. Evidence from recent data science teaching suggests that regular work with digital exercises, particularly when students use the feedback rather than merely receive it, is associated with stronger course performance ^[1]. Online project-based courses offer a related lesson: authentic tasks are more useful when they are arranged in a sequence that helps students connect technical work with later academic and professional choices ^[2]. Automated grading has a clear place in programming education because it provides rapid correctness checks and allows repeated attempts. Its limits are equally clear, however; it is generally better at judging whether code runs than at supporting explanation, interpretation, or code quality ^[3]. A systematic review of technology-supported formative assessment found that the tool itself is only part of the picture. Clear purposes, a close fit between assessment and pedagogy, and teachers who can interpret process data are also needed ^[4]. The form of feedback matters too. More complex error messages may increase cognitive processing demands without helping students understand what to do next ^[5]. Under English-medium instruction (EMI), these demands are layered onto differences in language preparation. Studies in Chinese higher education show that access to language and academic support is uneven across students ^[6]. In a first-year EMI chemistry course, English proficiency alone was not enough to ensure a smooth transition; some difficulties remained and required sustained support ^[7].

However, few empirical studies have examined how digital practice, an EMI transition, and subsequent assessment adjustment interact across successive cohorts within the same practice-intensive undergraduate course.

The course examined here was offered under four successive instructional conditions. The annual

offerings are treated as cohort-level packages, since technology use, teaching language, and assessment could not be separated experimentally. The analysis asks three questions: (1) how did grade levels and score distributions differ across the four cohorts; (2) how were the 2023 EMI transition and the 2024 assessment adjustment related to performance; and (3) what association was observed between formative grades and final-examination grades?

2. Course Context and Reform Design

2.1. Course and Learners

Introduction to Data Science is a core subject in the Big Data Management and Application major of a Sino-foreign cooperative undergraduate programme. Its content ranges from data acquisition and cleaning to visualisation, basic statistics, Python programming, and introductory modelling. The students were not a uniform group. Some arrived with prior coding experience, while others had only limited preparation in programming or mathematics. During the same period, the programme was also introducing more English-language materials into the curriculum.

2.2. Four Teaching Conditions

The four offerings differed in several practical respects, as summarised in Table 1. In 2021, teaching was mainly face to face and in Chinese, with little structured online practice. EduCoder was introduced in 2022 to organise coding tasks in stages, and UMOOC carried course resources and learning records. The 2023 offering changed more sharply: instruction moved fully into English, an English textbook was adopted, and the assessment became more demanding. Full EMI was retained in 2024, but the final assessment and the process-evaluation scheme were revised after the 2023 results had been considered. These were overlapping changes, so the dataset cannot show how much of the observed difference came from language, platform use, textbook choice, or assessment adjustment on its own.

Table 1: Teaching conditions across the four cohorts.

Cohort	Language	Digital tools	Materials	Assessment condition
2021	Chinese	Limited online practice	Chinese textbook	Theory-oriented
2022	Bilingual	EduCoder + UMOOC	Chinese textbook; bilingual slides	Practice-driven; stronger formative assessment
2023	English-medium	UMOOC	English textbook	Higher linguistic and assessment difficulty
2024	English-medium	UMOOC	English textbook	Data-informed assessment adjustment

Note. EduCoder is an online programming-practice platform; UMOOC is the institutional learning platform used for resources and learning records.

3. Data and Methods

3.1. Data and Measures

The analysis excluded records associated with deferred examinations, suspension of study, or an invalid academic status. This left 293 students: 75 in 2021, 74 in 2022, and 72 in each of 2023 and 2024. For each student, the dataset contained a formative grade, a final-examination grade, and the recorded overall grade, each expressed on a 0-100 scale. In the notation used below, F_{ic} , E_{ic} , and G_{ic} refer respectively to the formative, final-examination, and overall grades of student i in cohort c . The university calculated the overall grade using the following rule:

$$G_{ic} = \text{round}(0.50 F_{ic} + 0.50 E_{ic}) \quad (1)$$

where $\text{round}(\cdot)$ denotes conventional half-up rounding to the nearest integer. For each score component S_{ic} (formative, final-examination, or overall grade), the cohort pass and excellence rates were calculated as:

$$P_c(S) = \frac{1}{n_c} \sum_{i=1}^{n_c} I(S_{ic} \geq 60) \quad (2)$$

$$X_c(S) = \frac{1}{n_c} \sum_{i=1}^{n_c} I(S_{ic} \geq 85) \quad (3)$$

where $I(\cdot)$ is an indicator function. A score of at least 60 was defined as passing, and a score of at least 85 as excellent, for each component. The analyses used recorded raw scores; no cohort-wise z-score transformation was applied.

3.2. Analysis

Descriptive statistics were first calculated separately for each cohort, including the mean, median, sample standard deviation, pass rate, and excellence rate. The shape of the overall-grade distributions was examined with boxplots and Gaussian kernel density curves. Welch's one-way analysis of variance was chosen because the cohort sizes and variances were not identical. Games-Howell comparisons were then used for pairwise contrasts, with *Hedges' g* reported to indicate effect size. Pearson chi-square tests were applied to overall pass and excellence rates, and *Cramér's V* was used to describe the strength of those differences. Finally, Pearson's correlation coefficient was calculated within each cohort to estimate the linear association between formative and final-examination grades:

$$r_c = \frac{\sum_{i=1}^{n_c} (F_{ic} - \bar{F}_c)(E_{ic} - \bar{E}_c)}{\sqrt{\sum_{i=1}^{n_c} (F_{ic} - \bar{F}_c)^2 \sum_{i=1}^{n_c} (E_{ic} - \bar{E}_c)^2}} \quad (4)$$

95% confidence intervals for these correlations were derived using Fisher's z transformation and back-transformed to the correlation scale, and Spearman rank correlations were computed as a sensitivity analysis. For visualisation, an ordinary least-squares regression line was fitted to each cohort's formative-final scatter. Correlation was interpreted as association rather than agreement or causation, and all tests were two-sided with $\alpha=.05$.

4. Results and Discussion

4.1. Cohort Outcome Patterns

The cohort profiles in Table 2 are quite distinct. The highest overall mean was observed in 2022, when the mean reached 81.16 and 41.89% of students met the excellence threshold. Performance declined sharply in 2023: the mean fell to 62.99, only 69.44% passed, and no student obtained an excellent overall grade. In 2024, the mean rose to 78.36 and the pass rate returned to 97.22%. Scores were also less dispersed that year, with an overall standard deviation of 8.45, although the formative grades were concentrated within a narrow range.

Table 2: Main assessment outcomes by cohort.

Indicator	2021 (n = 75)	2022 (n = 74)	2023 (n = 72)	2024 (n = 72)
Formative mean (SD)	85.16 (9.46)	87.81 (9.33)	69.08 (11.14)	85.94 (4.54)
Formative pass rate	97.33%	100.00%	80.56%	100.00%
Final-exam mean (SD)	66.23 (12.84)	73.99 (14.27)	56.90 (11.39)	70.78 (14.72)
Final-exam pass rate	76.00%	86.49%	43.06%	81.94%
Overall mean (SD)	75.69 (10.44)	81.16 (10.79)	62.99 (9.72)	78.36 (8.45)
Overall pass rate	97.33%	97.30%	69.44%	97.22%
Overall excellence rate	9.33%	41.89%	0.00%	20.83%

Note. SD = sample standard deviation. Pass = score ≥ 60 ; excellence = score ≥ 85 .

Table 3 shows that the differences were not only visible in the raw summaries. The omnibus effect was largest for formative grades, and the effect for overall grades was also large. The pairwise results

tell a straightforward story: a moderate gain from 2021 to 2022, a pronounced fall in 2023, and a sizeable recovery in 2024.

Table 3: Inferential comparisons across cohorts.

Analysis	Outcome / contrast	Estimate	<i>p</i>	Effect size
Welch ANOVA	Formative grade	$F(3, 149.73) = 51.19$	$< .001$	$\eta^2 = .413$
Welch ANOVA	Final examination	$F(3, 159.80) = 25.33$	$< .001$	$\eta^2 = .189$
Welch ANOVA	Overall grade	$F(3, 160.14) = 48.03$	$< .001$	$\eta^2 = .331$
Games-Howell	2022 - 2021	+5.47 [0.95, 9.99]	.011	$g = 0.51$
Games-Howell	2023 - 2022	-18.18 [-22.59, -13.76]	$< .001$	$g = -1.76$
Games-Howell	2024 - 2023	+15.38 [11.43, 19.32]	$< .001$	$g = 1.68$

Note. η^2 is reported descriptively; Hedges' g is reported for pairwise effects.

Overall pass rates differed significantly across cohorts, $\chi^2(3) = 48.70$, $p < .001$, *Cramér's V* = .408. Overall excellence rates also differed significantly, $\chi^2(3) = 48.44$, $p < .001$, *Cramér's V* = .407.

4.2. Platform-Supported Improvement in 2022

The 2022 results were better than those of 2021, both in the overall mean and in the number of students reaching the upper score range. The boxplot in Figure 1 does not suggest that scores simply became more uniform; rather, the centre of the distribution moved upward. One practical difference in that year was the opportunity to work through coding tasks repeatedly. Students could run a solution, see where it failed, make a correction, and submit again. UMOOC, meanwhile, provided a stable place for materials and learning records.

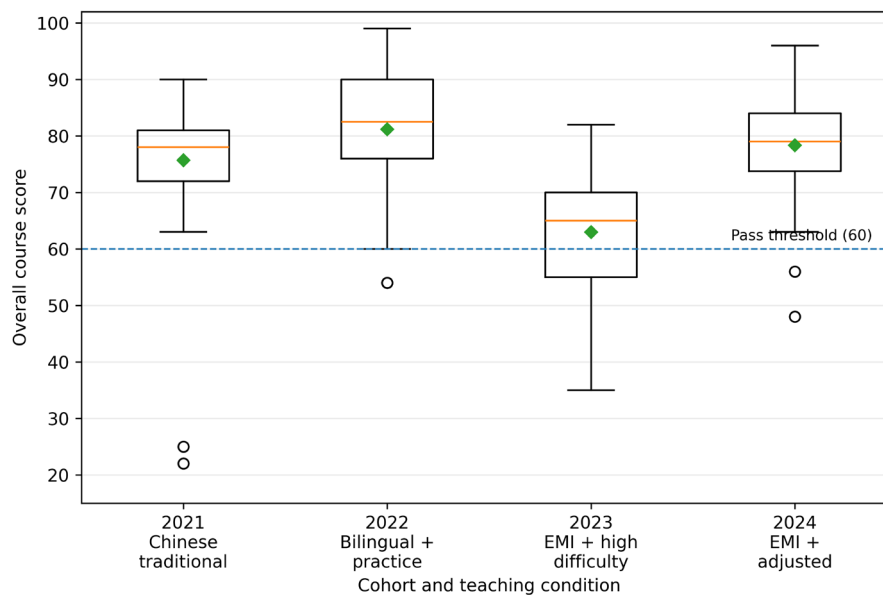


Figure 1: Boxplot of overall course grades across the four teaching conditions.

Recent data science research reports that students who complete more feedback-supported exercises tend to achieve stronger outcomes, although feedback must remain focused enough to avoid unnecessary cognitive effort ^[1].

The present result is also consistent with the programming-education literature, which identifies immediacy and repeated submission as important affordances of automated systems ^[3].

However, the 2022 condition also included bilingual instruction and stronger formative assessment. The observed improvement therefore reflects the combined instructional design and cannot be attributed to the platform alone.

4.3. EMI Transition and the 2023 Decline

The 2023 decline appeared at the same time as three major changes: full EMI, an English textbook, and a harder assessment. Students therefore had to read disciplinary English while simultaneously debugging code and interpreting output. It is reasonable to view this combined demand as part of the explanation for the lower scores. Still, the design cannot tell us whether language, examination difficulty, or differences in the cohort contributed most.

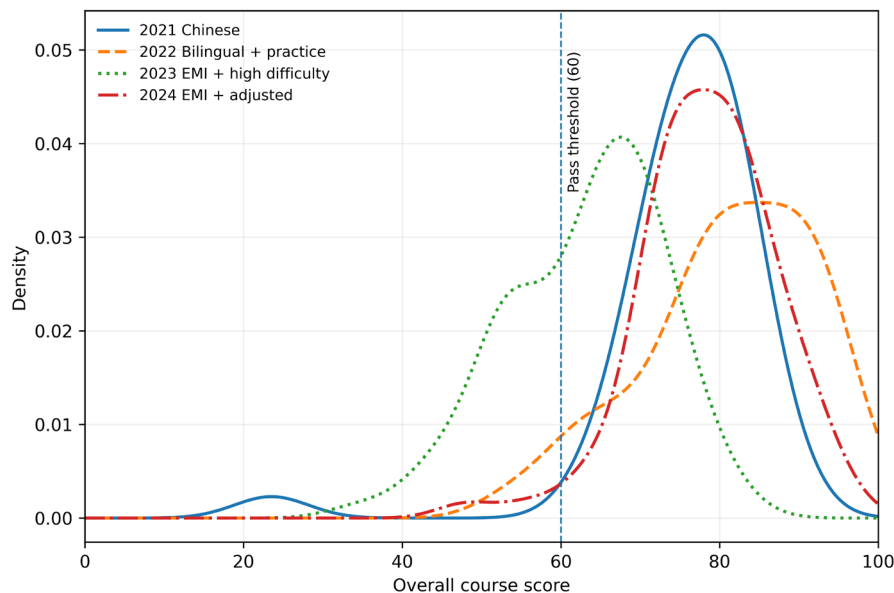


Figure 2: Kernel density distributions of overall course grades across teaching conditions.

The density curves in Figure 2 add some detail to the mean comparisons. The 2023 score distribution shifted toward the lower end of the scale, while the 2024 curve moved back toward the distributions seen before the full-EMI transition and was relatively compact. This pattern should be read alongside earlier EMI research. In Chinese higher education, students do not enter English-medium programmes with equal academic and linguistic support [6]. Difficulties can also persist after the transition even when students have adequate general English proficiency [7]. In EMI lectures, deeper understanding has been linked to motivation and learning strategy as well as to language proficiency [8]. For this course, a gradual transition would therefore be more defensible than an abrupt one. Bilingual terminology, annotated code, short diagnostic tasks, and low-stakes debugging can prepare students before they face open-ended assessment entirely in English.

4.4. Assessment Adjustment and the 2024 Recovery

Although EMI continued in 2024, the cohort's overall performance returned to a range comparable with 2021 and 2022. The Games-Howell results support this interpretation: the 2024 mean did not differ significantly from the 2021 mean ($p = .324$) or the 2022 mean ($p = .303$). A plausible interpretation is that language demands, formative support, and the final assessment were better matched than in 2023. However, the evidence does not support a more specific causal attribution. Since several elements were adjusted together and item-level results were unavailable, the recovery cannot be attributed to a single change in the examination.

Recent higher-education assessment research emphasises that authentic assessment should preserve meaningful performance demands while making criteria and progression visible to students [9].

A review of emerging assessment technologies similarly concludes that technology is most useful when assessment, feedback, and evaluation design are studied together rather than as separate technical functions [10].

For future course offerings, assessment adjustment should be documented through common anchor items, item difficulty, question type, English reading load, coding-task proportion, and formative-task completion. These records would allow stronger conclusions about whether later cohorts improved because of learning support, assessment design, or both.

4.5. Formative-Summative Association

The Pearson correlation coefficients were $r = .747$ for 2021 (95% $CI [.627, .833]$), $r = .658$ for 2022 (95% $CI [.505, .771]$), $r = .488$ for 2023 (95% $CI [.289, .646]$), and $r = .359$ for 2024 (95% $CI [.139, .545]$). Spearman's ρ values were .490, .677, .499, and .443, respectively. Thus, the linear coefficients were lower in the later cohorts, but they should not be treated as evidence that the two assessment components agreed less well. Correlation describes association, not interchangeability. The cohort-specific scatterplots and fitted regression lines are shown in Figure 3.

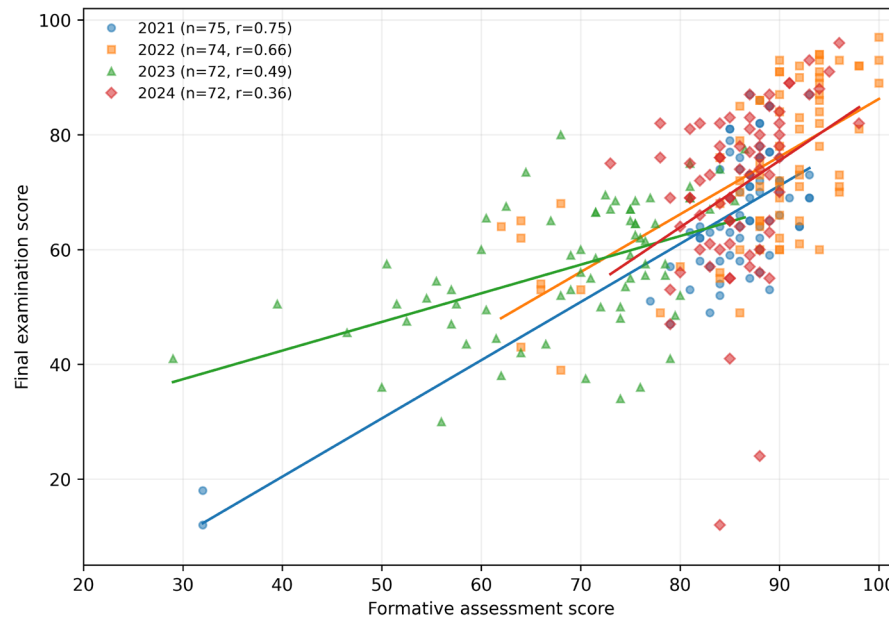


Figure 3: Formative grades and final-examination grades by cohort.

The 2024 formative scores deserve separate attention. Their standard deviation was only 4.54, and many students were graded near the top of the scale. When scores occupy such a narrow range, a Pearson correlation coefficient can be attenuated even when a genuine association remains. The concentration may also indicate that the process assessment had limited discriminative capacity. A stronger design would give more weight to tasks that combine problem interpretation, coding, debugging, evaluation of output, and a brief explanation of the reasoning used.

The combined evidence suggests four practical design priorities for future offerings (Table 4).

Table 4: Data-informed implications for course redesign.

Observed pattern	Instructional interpretation	Recommended response
2022 upward shift	Repeated practice within a bilingual design	Retain staged coding and require explanation
2023 broad decline	Language and assessment demands rose together	Stage EMI with bilingual terminology and diagnostics
2024 recovery	Support and assessment were better aligned	Use anchor items; document difficulty and reading load
Compressed 2024 formative scores	Possible ceiling effect in process assessment	Add integrated debugging and justification tasks

Note. Recommendations are derived from the cohort patterns and should be evaluated prospectively.

4.6. Robustness and Limitations

Mean differences, pass-rate changes, effect sizes, boxplots, and density curves showed the same broad sequence: improvement in 2022, decline in 2023, and recovery in 2024. Separately, the Spearman sensitivity analysis confirmed that formative and final-examination grades remained positively associated in every cohort. These converging indicators reduce the likelihood that the main

findings were produced only by a small number of extreme scores.

The annual grades were not based on a common equated examination, which limits any direct comparison of achievement across years. The study also concerns one course at one institution. Over the four offerings, the teaching language, textbook, platform configuration, formative work, teacher experience, cohort composition, and examination difficulty did not remain fixed. Information on prior English proficiency, programming experience, student demographics, individual test items, and platform activity was also unavailable. For these reasons, the results are best understood as differences in recorded cohort outcomes, not as causal estimates on a stable achievement scale. A future study would be stronger if it retained a common assessment blueprint, included a small set of anchor items, recorded English reading load and item difficulty, and checked formative tasks for both alignment and ceiling effects.

5. Conclusions

The four cohorts show a clear, though not causal, sequence. Performance was higher in the bilingual, platform-supported 2022 offering than in 2021. It then dropped substantially when full EMI, an English textbook, and a more demanding assessment were introduced together in 2023. Following the 2024 adjustment, scores recovered to a level that was not statistically different from the earlier cohorts.

The practical message is that these course elements cannot be planned independently. Digital practice may give students useful opportunities to act on feedback, but it needs to prepare them for the reasoning required in the final assessment. The language transition also requires dedicated scaffolding rather than being treated as a simple change of textbook or classroom language. Future research should record instructional and assessment changes in detail, retain common anchor tasks, and combine grades with learning logs and measures of students' language background.

Acknowledgements

This work was supported by the Fujian Provincial Educational Research Project for Young and Middle-aged Teachers (Science and Technology) (JAT220739) and the University-level Higher Education Teaching Reform Research Project of Zhejiang Yuexiu University (2405140006).

Data Availability

The anonymised student-level records are not publicly available because they contain academic records. Aggregate results and analysis code may be made available by the corresponding author on reasonable request, subject to institutional and privacy requirements.

Ethics Approval

The study used retrospective de-identified course assessment records, and no personally identifiable student information was reported.

References

- [1] Letteri I, Vittorini P. Enhancing student feedback in data science education: Harnessing the power of AI-generated approaches[J]. *International Journal of Artificial Intelligence in Education*, 2025, 35: 3071-3094. DOI: 10.1007/s40593-025-00492-8.
- [2] Castellanos-Reyes D, Leppard T. Online project-based learning to foster students' course choices in data science: A longitudinal case study using Sankey visualization[J]. *International Journal of Educational Technology in Higher Education*, 2025, 22: 63. DOI: 10.1186/s41239-025-00560-y.
- [3] Messer M, Brown N C C, Kölling M, Shi M. Automated grading and feedback tools for programming education: A systematic review[J]. *ACM Transactions on Computing Education*, 2024, 24(1): 1-43. DOI: 10.1145/3636515.
- [4] Børte K, Lillejord S, Chan J, Wasson B, Greiff S. Prerequisites for teachers' technology use in formative assessment practices: A systematic review[J]. *Educational Research Review*, 2023, 41: 100568. DOI: 10.1016/j.edurev.2023.100568.

- [5] Kuklick L, Greiff S, Lindner M A. *Computer-based performance feedback: Effects of error message complexity on cognitive, metacognitive, and motivational outcomes*[J]. *Computers & Education*, 2023, 200: 104785. DOI: 10.1016/j.compedu.2023.104785.
- [6] Yuan L, Fang F, Hu G. *Are students prepared and supported for English medium instruction in Chinese higher education to promote educational equality?*[J]. *System*, 2024, 125: 103423. DOI: 10.1016/j.system.2024.103423.
- [7] Aizawa I. *Tracking the first-year experience in English medium instruction: A pre-post study of transitional challenges*[J]. *English for Specific Purposes*, 2024, 73: 20-32. DOI: 10.1016/j.esp.2023.09.002.
- [8] Zhou S, Fung D, Thomas N. *Towards deeper learning in EMI lectures: The role of English proficiency and motivation in students' deep processing of content knowledge*[J]. *Journal of Multilingual and Multicultural Development*, 2025, 46(7): 1749-1764. DOI: 10.1080/01434632.2023.2248078.
- [9] Vlachopoulos D, Makri A. *A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations*[J]. *Studies in Educational Evaluation*, 2024, 83: 101425. DOI: 10.1016/j.stueduc.2024.101425.
- [10] Sembey R, Hoda R, Grundy J. *Emerging technologies in higher education assessment and feedback practices: A systematic literature review*[J]. *Journal of Systems and Software*, 2024, 211: 111988. DOI: 10.1016/j.jss.2024.111988.