



From Flipped Classroom to AI-Enhanced Assessment: Sustained Pedagogical Innovations in Undergraduate Mathematics Education

Ka Ho Law

Department of Mathematics, University of Hong Kong, Hong Kong

Abstract

In response to the challenges posed by generative artificial intelligence and increasingly diverse student populations, this paper examines a decade-long programme of pedagogical innovation in undergraduate mathematics education. At a research-intensive university in Hong Kong, a coherent set of evidence-informed practices has been developed and scaled across courses of different sizes and natures. These include a fully flipped classroom model supported by animated lecture videos and an exercise generator, a deliberate shift towards open-book and oral assessments, and the adoption of an AI-assisted grading platform to provide timely and individualised feedback. Drawing on multi-year institutional student surveys, mid-term feedback and peer observations, the evaluation reveals consistently positive outcomes, as reflected in teacher and course effectiveness scores, while students also reported enhanced conceptual understanding, greater engagement, improved inclusivity and stronger communication skills. The integration of AI-assisted grading significantly improved the timeliness and quality of feedback, particularly in large-enrolment classes. Supported by competitive teaching development grants and recognised through institutional teaching awards, this sustained body of work contributes to the Scholarship of Teaching and Learning by demonstrating how practitioner-led innovation can be systematically developed, rigorously evaluated, and progressively institutionalised. Rather than offering a prescriptive model, the paper provides an adaptable example for STEM educators seeking to strengthen student learning and assessment practices in an era of rapid technological change.

Keywords: flipped classroom, AI-assisted assessment, mathematics education, oral assessment, pedagogical innovation

1. Introduction

This paper presents a longitudinal case study of sustained pedagogical innovation in undergraduate mathematics at a research-intensive university in Hong Kong. Over more than a decade, a coherent and evolving suite of evidence-informed practices has been developed,

implemented and scaled across a wide range of undergraduate courses. These practices include a fully flipped classroom model supported by animated lecture videos and an exercise generator, a deliberate redesign of assessment towards open-book examinations and oral tests, as well as the use of an AI-assisted grading platform. The innovations have been refined iteratively in response to changing student needs, institutional priorities and the rapid emergence of generative artificial intelligence.

The work is situated within the Scholarship of Teaching and Learning (SoTL) tradition, in which practitioners systematically investigate their own teaching, gather evidence of impact and share insights that can inform practice beyond their immediate context (Boyer, 1990; Hutchings & Shulman, 1999). Rather than presenting isolated classroom experiments, this paper documents a departmentally embedded programme of innovation that has reached thousands of students, influenced institutional infrastructure and contributed to broader conversations about teaching in the AI era.

The Department of Mathematics at the university concerned serves a highly diverse student population. Courses range from large service classes of 100–200 students (primarily from actuarial science, economics, engineering and business programmes) to smaller courses for mathematics majors and minors. Students enter with varying levels of mathematical preparation, and many face competing demands on their time. In this environment, traditional lecture-based approaches often fail to accommodate different learning paces or to develop the higher-order thinking and communication skills increasingly valued by employers and accreditation bodies.

The emergence of generative AI tools since 2022 has added urgency to these challenges. While such tools offer new possibilities for personalised support, they also threaten the validity of conventional assessments that rely heavily on standard procedural tasks (Bittle & El-Gayar, 2025; Francis et al., 2025). In response, the innovations described here have sought to shift emphasis from memorisation and routine calculation toward conceptual understanding, authentic problem-solving, and communication — capabilities that remain difficult for current AI systems to replicate convincingly.

This case study draws on multiple sources of evidence collected over more than ten years: official Student Feedback on Teaching and Learning (SFTL) surveys, mid-term course feedback, peer observations and reviews, grant outcomes, as well as documented patterns of resource adoption by colleagues. The paper examines the design and implementation of each major innovation, presents evidence of impact at course and institutional levels, discusses challenges encountered and lessons learned and considers the transferability of the approach to other STEM contexts. By framing the work as a longitudinal case of evidence-informed innovation, the paper aims to contribute to the growing body of practitioner scholarship on teaching and learning in mathematics in the age of artificial intelligence.

2. Methods and Sources of Evidence

This paper adopts a longitudinal practitioner case-study approach within the Scholarship of Teaching and Learning (SoTL) tradition. It documents the design, implementation, iterative refinement and institutional embedding of a suite of pedagogical innovations in undergraduate mathematics education over more than a decade at a research-intensive university in Hong Kong.

2.1. Context and participants

The innovations were implemented across a range of courses taught and coordinated by the author. These included large service courses with enrolment over 100 that target students from actuarial science, economics, engineering and business programmes, as well as smaller core and elective courses for mathematics majors and minors. Collectively, the work has reached several thousand undergraduate students.

2.2. Data sources

Evidence was drawn from multiple complementary sources. These comprised official Student Feedback on Teaching and Learning (SFTL) surveys administered at the end of each course offering (multi-year data with response rates typically over 50%), mid-term course feedback and comments arising from Staff-Student Consultative Committee meetings, peer observations and formal performance reviews, outcomes of competitive teaching development grants, and documented patterns of resource adoption by colleagues both within and beyond the Department of Mathematics.

2.3. Analysis

Quantitative SFTL data, specifically teacher and course effectiveness scores, were examined for consistency of trends across years and in relation to departmental averages. Qualitative data from open-ended survey comments and mid-term feedback were analysed thematically to identify recurring student perceptions. The overall analytical approach is primarily descriptive and interpretive. No experimental control groups or randomised designs were employed. Consequently, the emphasis of the analysis lies on the consistency of positive indicators over time and on evidence of institutional uptake, rather than on strong causal claims about learning gains.

2.4. Limitations

The absence of controlled pre- and post-intervention measures of conceptual understanding (such as concept inventories) and of detailed grade-distribution analyses limits the strength of claims that can be made about learning outcomes. These limitations are revisited in Section 7.3.

3. The Flipped Classroom Model

The flipped classroom model forms the foundational pedagogical strategy underpinning the innovations described in this paper. Rather than using limited face-to-face time primarily for content delivery, the approach shifts foundational learning to structured pre-class activities, thereby freeing class time for interactive activities with higher-order engagement. This model has been implemented consistently across all undergraduate courses taught by the author, encompassing both service courses for non-majors as well as core and elective courses for mathematics majors and minors, with class sizes ranging from a dozen to three-digit figures.

3.1. Pre-class Resources and Design

Effective implementation of the flipped classroom depends heavily on the quality and accessibility of pre-class materials. Two complementary resources have been developed and refined over several years to support student preparation: animated lecture videos and an automated parameterised exercise generator.

Lecture videos were produced for students to watch before class to replace traditional in-class content delivery. These videos were deliberately designed with a strong emphasis on visualisation. Animations and dynamic graphics were used extensively to illustrate abstract mathematical concepts that students often find difficult to grasp through static textbook explanations or whiteboard presentations. Examples include the geometric meaning of linear transformations in three-dimensional space, the behaviour of surfaces defined by functions of two variables, and the interpretation of double and triple integrals through successive cross-sections and volume accumulation. These videos provide a substantial library of high-quality resources that students can access asynchronously and revisit as needed.

Complementing the videos is an automated exercise generator, which creates fresh questions with varying numerical parameters or contextual details each time a student attempts an exercise. This design serves multiple pedagogical purposes. It reduces the incentive for students to copy solutions from peers or previous attempts, provides repeated practice with conceptual understanding rather than rote memorisation and generates a large pool of unique questions suitable for both formative practice and summative assessment. During the COVID-19 period, the generator proved particularly valuable in enabling the rapid creation of multiple versions of online quizzes and tests while maintaining academic integrity.

Together, these pre-class resources allow students to engage with foundational content at their own pace and according to their own schedules. This flexibility is especially beneficial for students with weaker mathematical background or those balancing heavy workloads across multiple disciplines.

3.2. In-class Active Learning

With foundational content addressed before class, face-to-face sessions are reorganised around active learning and supported practice. A substantial portion of the classes is typically devoted to problem-solving activities during which students work on tasks with immediate access to the instructor and their peers. This arrangement enables real-time clarification of misconceptions and reduces the likelihood that students will turn to generative AI tools for unsupervised homework.

The remaining class time is used for activities designed to promote deeper conceptual understanding, critical thinking and communication skills. The specific activities vary according to course nature and enrolment size. In service courses and/or classes with large enrolment, activities often include conceptual voting using classroom response systems, short group discussions of common mistakes identified from pre-class work, and guided problem-solving in pairs or small groups. In courses for mathematics majors and/or classes with small enrolment, more ambitious activities are possible. A notable example is a course on network theory, in which students (mostly pursuing a major or a minor in mathematics) are required to research and present real-world applications of network models in relation to the theory learned in the course. Since the author took up this course in 2021, course effectiveness scores on the official SFTL survey have consistently exceeded 90 out of 100, which are substantially above norms.

The active learning approach also serves an important role in addressing emerging challenges associated with generative artificial intelligence. By shifting a significant amount of problem-solving practice into the supervised classroom environment, the model reduces incentives for students to rely inappropriately on AI tools for routine homework. At the same time, it creates space for discussions about the appropriate and ethical use of AI in mathematical work.

3.3. Student Perceptions and Evidence

Student response to the flipped classroom model has been generally positive, although not without challenges. Quantitative data from multi-year SFTL surveys indicate that teacher effectiveness and course effectiveness scores in the courses concerned have frequently matched or exceeded departmental averages, with several courses achieving scores in the top decile of the department, such as the course on network theory previously mentioned.

Qualitative feedback from mid-term course surveys, Staff-Student Consultative Committee meetings, and open-ended SFTL comments reveals several recurring themes. Many students appreciate the flexibility of learning foundational material through videos at their own pace and the opportunity to review difficult materials multiple times. Students with weaker mathematical backgrounds or those returning from extended breaks particularly value this feature. In addition, students frequently comment positively on the interactive nature of class sessions and the opportunity to receive immediate help while working on problems.

However, some students initially express concern that the model increases their overall workload or requires greater self-discipline to complete pre-class activities consistently. These perceptions are well-documented in the literature on flipped learning (Lo, 2021). In the present context, such concerns are addressed through several strategies. At the beginning of each course, the pedagogical rationale is explained clearly, and students are shown how the model is designed to make more effective use of both their independent study time and class time. Clear weekly schedules are provided in advance, and low-stakes pre-class exercises are used to ensure that students get well-prepared before class while also informing the design of in-class activities. Over time, students gradually adapt to the expectations of the model.

Meta-analyses have consistently found that well-designed flipped classroom implementations produce moderate positive effects on student performance across disciplines (Strelan et al., 2020; Cheng et al., 2019). In mathematics education specifically, the approach has been associated with improvements in both learning outcomes and aspects of student engagement when appropriate support structures are in place (Lo, 2021). The experience reported here aligns with these broader findings. Furthermore, the resources developed have extended their impact beyond the author's own courses. The exercise generator and selected video segments have been adopted or adapted by other colleagues within the Department of Mathematics and, in some cases, in other faculties. This wider dissemination reflects a key aspiration of the Scholarship of Teaching and Learning: moving from private pedagogical practice towards publicly shared and improvable knowledge (Hutchings & Shulman, 1999).

4. Assessment Practices in the AI Era

Assessment design has been deliberately and systematically revised in parallel with the instructional changes. The primary goals have been to promote deeper conceptual understanding, develop communication skills, as well as maintain the validity and integrity of assessment in an environment where generative artificial intelligence tools are increasingly accessible to students. Two complementary approaches have been emphasised: open-book examinations and oral tests.

4.1. Open-book Tests and Examinations

Written assessments in courses taught by the author have all been implemented in an open-book manner as a response to both pedagogical and practical considerations. In these assessments, students are permitted to bring a limited set of notes or “cheat sheets” with them. This format fundamentally shifts the focus of assessment away from rote memorisation

of standard procedures and towards the understanding, application, interpretation and critical evaluation of mathematical ideas.

The design of examination questions has been adjusted accordingly. Rather than asking students to reproduce standard methods or solve routine problems that can be easily generated or solved by current AI tools, questions increasingly require students to explain why a particular approach is appropriate, to identify and correct errors in a given solution, to interpret mathematical results in context, or to compare different methods for solving the same problem. Examples of such items include proofreading a mathematical proof for logical gaps or errors, reading and interpreting a mathematical argument presented in textual form, and applying concepts to novel or unfamiliar situations. These question types are more resistant to simple generative AI solutions and better reflect the kinds of reasoning that mathematicians and professionals in quantitative fields actually employ in their work.

The shift to open-book examinations also supports inclusivity. Students with varying levels of prior preparation or those who experience examination anxiety often benefit from being able to refer to their own organised notes rather than relying solely on memory under timed conditions. At the same time, the format aligns well with the flipped classroom model. Because students are already accustomed to accessing resources independently during pre-class preparation, the open-book examination represents a natural extension of the skills they are developing throughout the course.

From an academic integrity perspective, open-book examinations reduce the incentive for students to engage in traditional forms of cheating while simultaneously making it more difficult for generative AI to produce complete and convincing answers without substantial human input and understanding with the careful design of questions. Research on assessment in the age of artificial intelligence has highlighted the growing vulnerability of traditional closed-book, procedural assessments and the need for assessment formats that emphasise higher-order thinking and authentic application (Bittle & El-Gayar, 2025). The open-book approach adopted here represents one practical response to this challenge.

4.2. Oral Tests

The author has also introduced oral tests in several of his undergraduate courses, especially in courses for mathematics majors and/or courses with small enrolment. The rationale for incorporating oral assessment is multifaceted. First, it provides direct evidence of students' conceptual understanding that is difficult to obtain through written work alone. Students are required to explain their reasoning, justify their choice of methods, and articulate their understanding of underlying concepts, often in response to follow-up questions from the examiner. This process reveals not only whether a student can arrive at a correct answer, but also the depth and coherence of their understanding.

Second, oral tests contribute directly to the development of communication skills. The university has placed increasing emphasis on communication-intensive courses (CICs), and oral assessment provides an authentic opportunity for students to practise explaining mathematical ideas clearly and precisely. This aligns with broader institutional goals of developing graduates who can communicate complex quantitative information effectively to both specialist and non-specialist audiences.

Third, oral assessment offers a degree of resilience against the challenges posed by generative artificial intelligence. While current AI tools can generate written solutions to many standard mathematical problems, they are less capable of engaging in real-time, adaptive dialogue that requires students to clarify, defend or extend their reasoning. Oral tests therefore serve both a pedagogical and an integrity function.

Each oral test comes with a prescribed coverage (usually a few chapters of the course) and a set of pre-assigned questions. Students are allowed to bring a piece of cheat sheet to the test as in written assessments. During the test, students will randomly draw one of the pre-assigned questions and start explaining it on the board. In the process they may be requested by the examiner to clarify or further explain some of the steps. After that, follow-up questions, which may or may not be related to the pre-assigned questions, will be asked. This allows the difficulty level to be adjusted according to the performance of the students, so that they can be tested at the right level.

Clear rubrics are provided to students in advance, outlining the criteria for assessment. These typically include the accuracy and clarity of explanations, the depth of conceptual understanding demonstrated, the ability to respond to follow-up questions, and the quality of mathematical communication and presentation.

The above procedural safeguards support the validity and reliability of the oral tests. Efforts to enhance inter-rater reliability include the use of shared, detailed rubrics and regular discussions among examiners involved in oral assessment. These calibration conversations help align expectations and improve consistency in the application of marking criteria. While formal inter-rater reliability statistics are still being developed as the use of oral assessment expands, the combination of structured rubrics, examiner dialogue and adaptive follow-up questioning contributes to a more transparent and consistent process.

Empirical support for the value of oral assessment in mathematics education can be found in recent studies. Research published in 2025 found that oral tests in undergraduate mathematics courses were associated with improved conceptual understanding and longer-term knowledge retention compared with written-only assessment approaches (Law & Chiang, 2025). Systematic reviews indicate that well-designed oral formats can encourage deeper engagement, although student anxiety remains a common challenge (Stephenson et al., 2025).

Scalability is a practical challenge for oral tests. Conducting individual oral assessments in a course with over 100 students, for instance, is time-intensive. For this reason, oral tests have been introduced selectively rather than universally, with greater emphasis placed on courses where communication skills are explicitly prioritised or where class sizes permit more feasible implementation. Future developments, including the use of interactive AI avatars for formative oral practice and to assist in the assessment process, may help address scalability concerns while preserving the pedagogical benefits of this assessment format.

5. AI-Assisted Grading for Scalable, High-Quality Feedback

A major practical challenge in scaling active learning and richer assessment is the provision of timely, individualised feedback to large numbers of students. This challenge was addressed by the introduction of an AI-assisted grading platform. The platform was first piloted in 2024 with the support of a Faculty Teaching Innovation Fund. Following a successful pilot, the author organised a series of workshops to introduce the tool to colleagues in the Faculties of Science and Engineering. The use of the platform was subsequently funded at higher levels for more widespread use.

The platform offers several features that significantly enhance both the accuracy and efficiency of assessment. It supports consistent application of marking rubrics across large numbers of scripts, reducing variability that can arise from manual marking. Automatic totalling of marks eliminates simple arithmetic errors and saves considerable time, particularly in examinations with multiple components. These features contribute to greater

accuracy and fairness in grading, especially important in large service courses with enrolments exceeding 100 students.

Beyond efficiency gains, the platform enables much richer and more individualised feedback than was previously feasible. Markers can provide detailed annotations, insert reusable comment banks, and highlight common mistakes across the cohort. Students receive timely, specific feedback that helps them understand their errors and improve their performance. In large classes, this capability has transformed what was often a time-consuming and relatively superficial feedback process into one that provides actionable guidance within a reasonable timeframe. Practitioner studies in STEM disciplines have similarly reported high levels of satisfaction with the platform's ability to improve both grading efficiency and the quality of feedback (Calamas, 2024; Abiad, 2025).

This AI-assisted grading initiative also illustrates how a course-level pedagogical innovation can evolve into a broader institutional development. What began as an attempt to address feedback challenges in the author's own large-enrolment courses in 2024 gradually gained traction through workshops and peer sharing. The demonstrated benefits in terms of grading consistency, efficiency and feedback quality led to growing interest from other teachers. This momentum resulted in the Faculties of Science and Engineering jointly supporting the licence in 2025–26, followed by the university's decision to provide central funding from 2026–27 onwards. This trajectory, from individual initiative to faculty-level collaboration and eventually university-wide infrastructure, demonstrates the potential for practitioner-led innovation to influence institutional practices when supported by evidence and active dissemination.

Furthermore, the adoption of the platform has contributed to a broader shift in teaching culture at the university. By making it easier to provide detailed, timely feedback at scale, the platform has encouraged more colleagues to reflect on their feedback practices and explore ways to enhance student learning. Workshops organised by the author have not only facilitated the technical adoption of the tool but have also sparked discussions about assessment design and the role of feedback in large classes. In this sense, the initiative has extended beyond a technical solution for grading efficiency to become part of ongoing efforts to improve teaching and learning across the institution.

Ethical considerations and potential sources of bias in AI-assisted grading have also been attended to. The platform is used primarily to support consistent application of human-designed rubrics, automatic totalling of marks, annotation and the identification of common errors. Final academic judgements remain the responsibility of human markers. Comment banks and feedback templates are regularly reviewed to reduce the risk of uneven or stereotypical comments. Students receive transparent, timely feedback intended to support learning rather than solely to rank performance. These safeguards help ensure that the efficiency gains of the platform do not come at the expense of fairness or pedagogical integrity.

These developments in feedback practices naturally raise questions about their overall effectiveness and institutional reach. The following section therefore examines the broader impact of the pedagogical innovations, drawing on student feedback, institutional recognition, grant success and leadership roles.

6. Evaluation, Impact and Institutional Embedding

The innovations described in this paper have been subject to ongoing evaluation over more than a decade. Evidence has been drawn from multiple sources, including official institutional

surveys, peer observations, formal performance reviews, grant outcomes and documented patterns of adoption by colleagues. This section examines the impact of the work at both course and institutional levels.

6.1. Student Feedback and Learning Outcomes

Student feedback has been collected systematically through official SFTL surveys, mid-term course evaluations and Staff-Student Consultative Committee meetings. Table 1 provides a concise summary of key quantitative indicators.

Table 1. Summary of selected SFTL and impact indicators

Indicator / Course context	Observation	Notes
Course on network theory (majors/minors)	Course effectiveness scores consistently >90/100 since 2021	Substantially above departmental averages
Large service courses	Teacher and course effectiveness scores frequently matched or exceeded departmental averages	Multi-year pattern
Qualitative themes	Students positively commented on the flexibility of video learning, interactive class sessions and timely feedback from AI-assisted grading	Recurring across mid-term and end-of-course comments

Qualitative feedback from open-ended SFTL comments and mid-term surveys reveals several recurring themes. Many students highlight the benefits of learning foundational material at their own pace through the lecture videos, noting that both stronger and weaker students could benefit. Students frequently comment on the value of interactive class sessions, stating that they feel more engaged when class time is used for problem-solving and discussion rather than passive listening. Feedback on the AI-assisted grading platform has been especially positive in large classes, with students appreciating the timeliness and detail of individualised feedback.

While overall responses have been favourable, some students initially express concern about workload or the self-discipline required for pre-class preparation. These perceptions tend to diminish as students become accustomed to the structure of the courses. Over time, the majority of feedback indicates that students recognise the pedagogical rationale and appreciate the opportunities for active engagement and supported practice.

Taken together, the quantitative and qualitative data suggest that the innovations have contributed to improved student engagement, greater inclusivity for students with diverse backgrounds, and enhanced opportunities for developing conceptual understanding and communication skills. These outcomes align with the broader educational aims of the university, which emphasise not only content mastery but also the development of transferable academic skills.

It should be noted that systematic pre- and post-intervention concept inventories or detailed analyses of grade distributions were not conducted as part of this practitioner study. Future work will seek to incorporate such measures to strengthen evidence of learning outcomes.

6.2. Institutional Recognition and Wider Impact

The work has received formal recognition at various levels. The author has received both faculty and university teaching awards for the innovation work mentioned in the present paper. Beyond formal recognition, the innovations have contributed to wider scholarly

exchange through activities such as presentations at international conferences, which have provided opportunities to share experiences and receive feedback from the broader academic community. These activities align with the emphasis of the Scholarship of Teaching and Learning on making pedagogical practice public and open to peer review.

6.3. Grant Success and Leadership Roles

The development and scaling of the innovations have been supported by success in securing competitive teaching development funding. As Principal Investigator, the author has obtained more than HK\$2.5 million in grants from internal and external sources. Notable awards include a major Teaching Development and Language Enhancement Grant (TDLEG) for an inter-institutional project focused on interactive learning videos and an AI-enhanced self-evaluative platform for flipped classrooms from the University Grants Council (UGC), as well as a Teaching Development Grant from the university to introduce interactive AI avatars for communication-intensive mathematics courses with oral assessment components. These grants have enabled the creation, evaluation and wider dissemination of teaching resources. The consistent success in securing funding reflects recognition of the quality and potential impact of the work.

The pedagogical innovations have also informed and been informed by the author's leadership roles within the university such as head of teaching and learning in the department as well as mentor and reviewer for the Advance Higher Education Fellowship Scheme in the university. These positions have provided opportunities to influence teaching and learning at a broader level while feeding insights back into ongoing pedagogical development.

7. Challenges, Limitations and Future Directions

While the pedagogical innovations described in this paper have produced positive outcomes, their development and implementation have not been without challenges. This section discusses key difficulties encountered, acknowledges limitations in the current evidence base, and outlines directions for future work.

7.1. Student Adaptation and Workload Perception

One recurring challenge has been students' initial adaptation to the flipped classroom model. Although many students eventually respond positively, some initially perceive the approach as increasing their overall workload. This perception often stems from the requirement to engage with pre-class videos and exercises, in contrast to more traditional models where students may attend lectures with minimal preparation. These concerns are addressed through clear communication of the pedagogical rationale, structured weekly schedules, and low-stakes pre-class activities that ensure students get well prepared. Over time, most students adapt to the expectations.

Oral tests have also presented challenges. Some students, particularly those more accustomed to written examinations, have expressed anxiety about this less familiar format. While efforts such as clear rubrics, low-stakes practice opportunities and supportive practices could have helped, it remains an area requiring ongoing attention, especially when scaling oral assessment to larger cohorts.

A further challenge lies in maintaining objective and consistent marking standards in oral assessments. Unlike written examinations, where scripts can be moderated more systematically, oral assessments involve real-time interaction, making standardisation more complex. Detailed rubrics and regular discussions among examiners help improve

consistency, although achieving high inter-examiner reliability remains an ongoing area of development.

7.2. Scalability and Resource Sustainability

A second set of challenges relates to scalability and the ongoing sustainability of the resources developed. Oral assessment, while pedagogically valuable for developing communication skills and verifying conceptual understanding, is inherently time-intensive. Conducting meaningful individual oral assessments becomes increasingly difficult as class sizes grow, say, beyond 50 students. For this reason, oral tests have been introduced selectively rather than across all courses, with greater emphasis placed on smaller classes and/or courses that emphasise more on conceptual understanding. The introduction of interactive AI avatars mentioned earlier represents one strategy for addressing scalability challenges. These avatars are intended to provide students with scalable, low-stakes opportunities to practise oral explanation and receive formative feedback, and possibly assist examiners. If successful, this development may help extend some of the benefits of oral practice without placing unsustainable demands on instructor time. However, the effectiveness of such tools will require careful evaluation, particularly regarding their ability to simulate meaningful dialogue and provide accurate feedback on mathematical reasoning.

The production and maintenance of high-quality lecture videos and the automated exercise generator also require long-term effort and resources. Creating effective animated videos demands both pedagogical expertise and technical skill, while keeping the exercise generator current and aligned with course content requires ongoing maintenance. Although internal and external grants have provided crucial support for initial development, continued funding is necessary to update materials, expand the resource library and incorporate new technological possibilities.

7.3. Methodological Limitations and Future Research

A further limitation concerns the nature of the evidence presented in this paper. The evaluation draws primarily on institutional student feedback data (SFTL surveys), mid-term evaluations, as well as peer observations and reviews. While these sources provide valuable indicators of impact and have been collected consistently over many years, they do not constitute controlled experimental evidence. It is therefore not possible to make strong causal claims about the effects of individual innovations on student learning outcomes. This is a common characteristic of practitioner scholarship conducted in authentic educational settings, where practical and ethical constraints often limit the feasibility of randomised controlled trials.

The strength of the present work lies in its longitudinal nature and the consistency of positive indicators across multiple data sources and course contexts. The ability to document the evolution of practice over more than a decade, alongside evidence of institutional adoption and external recognition provides a rich case study of pedagogical innovation. Nevertheless, future research could strengthen the evidence base through more systematic collection of learning outcome data, including pre- and post-intervention measures of conceptual understanding and longitudinal tracking of student performance in subsequent courses.

Several directions for future work are being pursued. The project on interactive AI avatars will include systematic evaluation of their effectiveness in supporting oral communication practice. The ongoing UGC TDLEG inter-institutional project offers opportunities for comparative study across institutions and for wider dissemination of resources and findings. In addition, further investigation into the long-term effects of oral assessment on conceptual

retention and communication skill development would contribute valuable insights to the field.

Finally, continued efforts to support colleagues in adopting and adapting elements of the model remain a priority. While resources such as the exercise generator and lecture videos have already been shared within the department, more structured support — through workshops, mentoring, and the development of implementation guides — could help extend the impact of this work more broadly.

8. Conclusion

This paper has documented a multi-year programme of pedagogical innovation in undergraduate mathematics education. A coherent set of practices — including a fully flipped classroom model, open-book and oral assessments, as well as the adoption of an AI-assisted grading platform — has been developed and scaled across large service courses and smaller major courses.

The work contributes to the Scholarship of Teaching and Learning by demonstrating how practitioner-led innovation can generate impact beyond individual classrooms. Evidence from multi-year institutional data, teaching awards and competitive grants shows that these practices have enhanced student engagement, conceptual understanding and communication skills, while also contributing to institutional developments such as the university-wide adoption of AI-assisted grading tools.

In the context of rapid advances in generative artificial intelligence, this case study illustrates how mathematics educators can redesign instruction and assessment to emphasise conceptual understanding, authentic reasoning and communication — capabilities that remain difficult for current AI systems to replicate. It offers a coherent and adaptable example for other practitioners in STEM disciplines, particularly those working in large-enrolment or resource-constrained contexts.

As higher education continues to navigate technological disruption, sustained engagement in the Scholarship of Teaching and Learning remains essential. This case study affirms the value of practitioner scholarship in improving student learning, building institutional capacity, and contributing to the wider educational community. Future work will continue to explore emerging technologies, such as interactive AI avatars, to further enhance scalable support for student learning in mathematics.

Acknowledgment

This work was supported by multiple sources of funding, including Teaching Development Grants from The University of Hong Kong as well as grants from the University Grants Committee of Hong Kong under the Teaching Development and Language Enhancement Grant scheme. The author thanks colleagues in the Department of Mathematics and students whose feedback has contributed to the development of the pedagogical innovations described in this paper.

References

- Abiad, A. (2025). AI-assisted grading increases quality feedback and grading efficiency in large undergraduate chemistry laboratory courses. *Journal of Chemical Education*. <https://doi.org/10.1021/acs.jchemed.5c00664>
- Bittle, K., & El-Gayar, O. (2025). Generative AI and academic integrity in higher education: A systematic review and research agenda. *Information*, 16(4), 296. <https://doi.org/10.3390/info16040296>
- Boyer, E. L. (1990). *Scholarship reconsidered: Priorities of the professoriate*. Carnegie Foundation for the Advancement of Teaching.
- Calamas, D. (2024). Student and instructor feedback on an AI-assisted grading tool. In *2024 ASEE Southeast Section Conference*. American Society for Engineering Education. <https://doi.org/10.18260/1-2--45560>
- Cheng, L., Ritzhaupt, A. D., & Antonenko, P. (2019). Effects of the flipped classroom instructional strategy on students' learning outcomes: A meta-analysis. *Educational Technology Research and Development*, 67, 793–824. <https://doi.org/10.1007/s11423-018-9633-7>
- Francis, N. J., Jones, S. & Smith, D. P. (2025). Generative AI in higher education: Balancing innovation and integrity. *British Journal of Biomedical Science*, 81, 14048. <https://doi.org/10.3389/bjbs.2024.14048>
- Hutchings, P., & Shulman, L. S. (1999). The scholarship of teaching: New elaborations, new developments. *Change: The Magazine of Higher Learning*, 31(5), 10–15. <https://doi.org/10.1080/00091389909604218>
- Law, K. H., & Chiang, S. L. (2025). Investigating the effectiveness of oral tests in mathematics in higher education. *PUPIL: International Journal of Teaching, Education and Learning*. <https://doi.org/10.20319/ictel.2025.8384>
- Lo, C. K. (2021). Student engagement in mathematics flipped classrooms: A systematic review. *Frontiers in Psychology*, 12, 672610. <https://doi.org/10.3389/fpsyg.2021.672610>
- Stephenson, Z., Jackson, A., & Wilkes, V. (2025). Interventions and facilitators of oral assessment performance in higher education: A systematic review. *Assessment & Evaluation in Higher Education*. <https://doi.org/10.1080/02602938.2025.2504621>
- Strelan, P., Osborn, A., & Palmer, E. (2020). The flipped classroom: A meta-analysis of effects on student performance across disciplines and education levels. *Educational Research Review*, 30, 100314. <https://doi.org/10.1016/j.edurev.2020.100314>