Assessment Contexts: Oral vs Written Performance in Mathematics Education

Sima Hajilarova
Federation University
s.hajilarova@federation.edu.au

Daya Weerasinghe
Federation University
d.weerasinghe@federation.edu.au

Generative AI has intensified concerns about assessment integrity in both face-to-face and online education, prompting interest in oral examinations as an alternative to written assessments. While oral assessment research is established in universities, little is known about enabling programs. This study compares written and oral mathematics assessments in an enabling cohort. Although group-level performance was similar across delivery modes, individual results differed markedly. Weak, non-significant correlations suggest the two formats capture complementary aspects of student understanding, supporting the value of multimodal assessment for diverse learners.

Introduction

Generative Artificial Intelligence (Gen AI) has significantly disrupted higher education, prompting academics to reassess assessment design due to growing concerns about authenticity, integrity, and graduates' preparedness for the workplace (Dawson, 2024; Nallaya et al., 2024). As online education expands, these problems intensify. These challenges are especially salient in enabling programs, pre-university pathway programs created to help students who do not meet standard university admission criteria to transition into higher education. Oral assessments are gaining renewed attention as they can support academic integrity, identity verification, communication skills, and relational engagement, particularly for diverse and EAL (English as an additional language) learners (Fenton, 2025; Iannone & Vondrová, 2024). However, they also raise concerns around anxiety, fairness, and reliability. The present study offers a timely and novel contribution by directly comparing written and oral mathematics assessments in an enabling cohort, providing new empirical insight into assessment design, student experience, and the diverse ways in which enabling students demonstrate mathematical understanding.

Oral and written assessments

The concept of assessment has always been central to education, but contemporary conditions have introduced new and complex challenges. Generative AI is not the only force driving the need for new assessment approaches. The speedy expansion of online education has also created challenges for assessment modes that should adhere to the principles of validity, reliability and fairness. Akimov et al. (2020) argue that assessment should also have multiple parameters, and its components maintaining academic integrity should be both formative and summative. The authors broaden the requirements for online assessments' design to include considerations about issues of accessibility and legality, identity security and, inevitably, academic integrity. Assessing mathematical understanding is challenging because students' reasoning is not always evident in written work. Since mathematical proficiency involves both procedural skill and conceptual understanding (Videnovic, 2017), oral assessment can offer valuable insight by allowing students to explain their thinking, address misconceptions, and demonstrate understanding in real time.

Online format of Higher Education obliged respective Institutions to find a way to prevent academic fraud and to stop or identify possible violations of the exam rules. Many universities worldwide started to use the technology of online proctoring (the Latin word "proctor" – follow, observe) which includes supervising an exam or its preparation remotely, using a webcam and/or specialised proctoring software. These systems typically collect data such as webcam

(2026). In N. Berger, M. Thompson, J. Ley & C. Wilsmore (Eds.), *Beyond boundaries: Strengthening futures in mathematics education. Proceedings of the 48th annual conference of the Mathematics Education Research Group of Australasia, Sydney* (pp. 170–177). MERGA.

images, microphone audio, the computer's screen display, keystrokes entered during the session and conduct some sort of identity verification, such as checking that the student looks like the person on their identity (ID) card (Dawson, 2024).

According to Puchkov et al. (2020), modern proctoring systems can automatically detect various violations, including opening unauthorized applications (some systems block these entirely), running programs that allow remote access, using external devices or books, looking away from the screen, absence or change of the examinee's face, presence of other individuals, or unauthorized sounds and non-verbal communication with people outside the camera's view. In this context, as oral assessments involve co-temporal interaction, they naturally verify the student's identity and comprehension without employing intrusive proctoring solutions. Oral assessments allow lecturers to observe students' real-time reasoning, error correction, and decision-making strategies, elements that are often hidden in written work. Such interactions reveal the student's conceptual frameworks, misconceptions, and strengths, providing rich diagnostic information. The oral assessment format offers a notable time-saving advantage for examiners as well. Although that type of examination demands considerable time and effort from instructors for preparation, scheduling, and administration, it offers a distinct practical advantage: grading occurs immediately and transparently during the session (Theobold, 2021).

As students learn in diverse ways, a uniform approach to mathematics assessment is insufficient. While educators widely acknowledge the importance of differentiated instruction to address student individuality and variability, they must likewise recognise the need for differentiated assessment practices to accurately represent the learning of a heterogenous student population (Videnovic, 2017). Some researchers argue that written assessment formats are preferable in situations when it is necessary to assess higher-order cognitive skills such as problem solving and critical thinking. The linguistic complexity of mathematical test items unrelated to the content being assessed may contribute to the performance disparities between native English speakers and EAL learners, potentially compromising the validity of achievement tests. In fact, oral assessments can be particularly advantageous for students who speak English as a second language or experience challenges with speech processing/comprehension. In particular, oral examinations provide an opportunity for examiner involvement, facilitating students' ability to interpret concepts, reach shared understanding, and demonstrate conceptual mastery, even when their written English proficiency is limited.

Alongside its benefits, oral assessment entails potential drawbacks. Videnovic (2020) suggests that there is a common perception that oral assessments may provoke greater anxiety in students compared to other assessment formats. In addition, the author attributes this anxiety to two main factors: the unfamiliar nature of oral examinations and the belief that such tasks demand a deeper level of understanding, along with the ability to clearly explain concepts to others. Nevertheless, Iannone & Vondrova (2024) suggest that the novelty of the assessment method can encourage deeper learning. Because students perceive the unfamiliar assessment as more demanding, they may engage in study practices that promote stronger conceptual understanding. However, Ferdoush et al. (2025) express their disappointment, noting that oral assessment can be susceptible to errors such as halo effects, central tendency errors, a general leniency bias, and contrast errors. These biases can impact the accuracy and fairness of the assessment process. The authors also emphasise that studies have shown that scores are directly proportional to the number of words spoken by the examiner and the time taken. Overall, oral assessments are time-consuming and can become a resource-intensive exercise.

Nallaya et al. (2024) note that another major concern for the academic community regarding oral assessment is fairness. The authors refer to the opinion that without interrater reliability, oral assessments risk being perceived as subjective and potentially biased, which undermines their credibility. Moreover, even involving multiple assessors does not necessarily resolve the

challenges associated with achieving consistent scoring across evaluators. That said, Moothadeth (2024) underscores that the use of a rubric system can significantly enhance both the effectiveness and fairness of oral assessments, providing a structured mechanism to ensure objectivity and consistency in evaluation. Rubrics serve as a structured instrument that ensures objectivity and consistency in scoring, reducing the risk of subjectivity and bias often associated with oral evaluations.

Crawford et al. (2022) stress that the enabling environment can be seen as a collaborative community of teachers and learners. Learners arrive at this setting with aspirations for a transformed and enriched life, while teachers share a commitment to supporting these changes. However, working within enabling education also implies significant emotional labour. Teaching diverse and often vulnerable students imposes considerable emotional demands on educators, particularly in assessment contexts where students' anxieties are more visible. The oral assessment format is one such context: students' vulnerability frequently becomes evident, and teachers' capacity to provide emotional reassurance is brought to the forefront. The novelty of the oral assessment format can also amplify feelings of anxiety and discomfort, which in turn may influence student performance outcomes.

Although oral examinations and viva-voice assessments have been explored in undergraduate mathematics contexts, particularly in relation to academic integrity, reliability, and student anxiety (Fenton, 2025), there is currently no published research examining oral mathematics assessment within Australian enabling programs. Existing studies of enabling mathematics focus on curriculum, learner characteristics, critical pedagogy impacts, preparedness, and numeracy development rather than on assessment modality (Baker et al., 2022). Likewise, the enabling education literature acknowledges that oral presentations and performance-based tasks can heighten student vulnerability and anxiety (Crawford et al., 2022). Yet, these discussions are not discipline-specific and do not compare oral and written assessment outcomes. In other words, research offers little insight into enabling students' performance across different mathematical assessment types or the extent to which these modalities draw on separate competencies. This study explores the benefits and limitations of oral assessment as a context-sensitive response to contemporary assessment challenges.

Methodology

Oral dialogic assessments can be organised in several ways. In some cases, specific questions are not negotiated with students; instead, the examiner selects questions based on material previously covered in class. Alternatively, students may not know the exact questions in advance but could be asked to choose one from a list or draw a question from a set, with varying amounts of preparation time allowed (Iannone, 2023). Another approach involves providing students with a bank of theoretical questions, such as theorems and proofs, beforehand. The examiner then arranges these into "tickets," each containing two theoretical questions and one practical question unknown to the student. The tickets are numbered and placed face down on a table, with the student selecting one and being given time to prepare before responding. This paper discusses assessment based on interrogations, in which the student is quizzed by one or more examiners, and the short interview of students to confirm their authorship of a written paper (Joughin, 2010).

This is a cross-method study that compares the performance of 64 students on written tests and quizzes with their performance in oral interviews conducted as part of the assessment tasks in an Introductory Mathematics and Statistics unit offered through an enabling program. The program provides an alternative pathway into undergraduate degrees by equipping prospective students with the academic skills necessary for successful Higher Education in one of the regional Universities in Australia. The authors, both lecturers, are actively involved in teaching the unit. These raw scores were subsequently normalised to proportional values ranging from 0

to 1. Although results were expressed as percentages when drawing parallel boxplots to compare written and oral scores, other statistical analyses were performed using the normalised decimal scores.

Rahman (2011) emphasises that any well-structured examination demands considerable time and effort from examiners. The key challenge is to identify assessment instruments that justify this investment by delivering meaningful educational value. So, before implementing the oral interview, researchers carefully considered several key factors:

- Which skills and proficiencies should be assessed (e.g., conceptual knowledge, practical mastery, analytical reasoning)?
- What should be the size of the question bank?
- Should students receive the exact list of questions in advance?
- How many questions should each student answer, and how many attempts should be permitted?
- Should the assessment include theoretical questions (definitions, key concepts) or focus solely on practical ones?
- Should students be allowed to use calculators or consult notes?
- Should preparation time be provided, and should the interview be timed?

A 15-minute timed interview was ultimately selected, consisting of three questions (both theoretical and practical) drawn from previous assignments and coursework. Students were given one minute of preparation time, permitted to use notes but not calculators. The question bank comprised 30 questions, and students selected their own by choosing three numbers between 1 and 30 "on the spot". The answers were assessed against an elementary rubric shown in Table 1. Each question was scored out of four points. Zero marks were given to students who were unable to answer any questions. The interviews were conducted in two formats: face-to-face and online via Teams. All the interviews were recorded for reference.

Table 1

Rubric used for oral assessment

Mark	Descriptor
4	Answered the question straight away, clearly and accurately. No prompt required.
3	Answered correctly after one prompt or hint. Response is mostly accurate with adequate reasoning.
2	Answered only after several prompts or provided a partially correct/vague answer that shows incomplete understanding.
1	Required extensive prompting to get an answer.
0	Unable to give any answer.

The current study was primarily concerned with the following research questions:

- Do oral and written assessments' results correlate? If not, what does that discrepancy indicate?
- To what extent do the quantitative results differ between face-to-face and online students?

Data Analysis

To analyse the dataset, a combination of statistical methods was utilised. Spearman rank correlation coefficient (r_s) was used to examine the strength and direction of monotonic relationships between the written assignment and oral assignment scores, given its suitability for the data. This coefficient is a nonparametric, distribution-free statistic that measures both the strength and direction of a monotonic relationship between two ranked variables, or between one ranked variable and one continuous measurement variable (Xiao et al, 2015). Researchers frequently examine this association, as its value ranges from −1 (strong negative monotonic

relationship) to +1 (strong positive monotonic relationship). Values near zero suggest a weaker association.

Additionally, independent-samples *t*-tests were carried out to compare group means, with degrees of freedom calculated to determine the appropriate distribution for hypothesis testing. Statistical significance was assessed using *p*-values, which provided evidence about the null hypothesis. According to Jankowski et al. (2017), the modern *t*-test consists of two components: mathematical calculations of the *t*-statistic and a reference table that provides the distribution of possible *t*-values. The calculations quantify the difference between the sample means while accounting for variability within the data. The reference table represents all potential *t*-statistics and their associated probabilities (critical values or *p*-values) based on sample size.

The *p*-value represents the probability of rejecting or not rejecting the null hypothesis (H_0). The null hypothesis assumes that there is no effect, no relationship, or no difference between the two groups for a given variable. The "*p*" stands for probability, and the *p*-value is calculated as the likelihood of observing an effect as extreme as, or more extreme than, the one found, assuming H_0 is true. In fact, the *p*-value indicates the evidence weight against the Null Hypothesis, the smaller the *p*-value, the stronger the evidence suggesting H_0 is incorrect (Thiese et al., 2016). In this case, the null hypothesis (H_0) is the assumption that there is no relationship between oral and written assessment scores. The null hypothesis is the one that is test against.

No proctoring software was used to invigilate the written assignments submitted by online students; only time restrictions were imposed. Whereas all face-to-face students completed their written work in class under supervision. For this reason, the written assessment results of online and face-to-face students were analysed separately to avoid distortion related to supervision conditions.

The above multi-method approach ensured a rigorous examination of both associations and differences within the data. With all this in mind, the researchers have tried to determine the degree of convergence or divergence between the two assessment formats and to identify and qualitatively analyse cases where students' performances diverge, potentially revealing differences in theoretical insight, operational proficiency, test anxiety, verbal processing, or issues related to academic integrity.

Results and Discussion

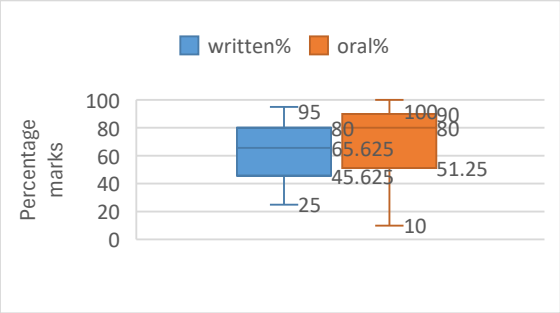
Written versus oral (face-to-face students)

The Spearman correlation (r_s) examination of the relationship between written and oral assessment scores has shown that the correlation is negligible and not statistically significant, with r_s = 0.05788 (essentially zero) and the *t*-statistic, $T(38) = 0.35741$. Because the *p*-value is 0.72276 (a very large value, indicating that the observed correlation does not differ statistically from zero), the null hypothesis of no association between the two assessment formats cannot be rejected. Statistically, this means that written and oral tasks behave as separate measures: students who perform well in the written task are not necessarily the ones who perform well in the oral task, and vice versa. There is no predictable pattern between the two.

To investigate the measures of spread, the researchers have also utilised parallel box plot analyses. For the face-to-face students, the overall oral scores are slightly higher than the written scores while written and oral median scores are 65.625% and 80% respectively (see Figure 1). Oral scores show a greater spread, especially toward the lower end. Written scores are a bit more tightly clustered, especially in the lower quartile. The middle 50% of oral scores sit slightly higher than the middle 50% of written scores. Oral scores also have a higher upper quartile, median and lower quartile, suggesting that more students scored relatively better on oral. Written scores are more compact, suggesting more uniform performance. This means that the oral task elicited a slightly higher but more variable level of achievement.

Figure 1

Written versus oral (face-to-face students) comparison



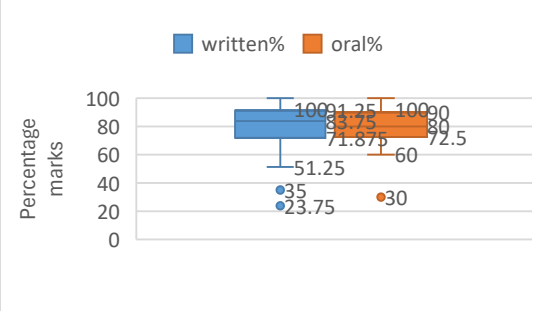
Written versus oral (online students)

A similar statistical analysis of the online students' outcomes has shown a stronger correlation than that observed in the face-to-face cohort, suggesting a somewhat more coherent performance pattern across modalities for online learners. The Spearman correlation for online students is positive, $r_s = 0.350677$, indicating a weak-to-moderate association in which stronger written performance tends to align with stronger oral performance. However, the evidence remains insufficient to reject the null hypothesis. The *t*-statistic, $T(19) = 1.632219$, indicates that the observed correlation, although positive, might not be large enough relative to the sample size to reach statistical significance. This is confirmed by the *p*-value of 0.119097, which again has shown that the null hypothesis of no association cannot be rejected.

For the online cohort, the parallel boxplot demonstrates that students performed comparably on written and oral tasks, with nearly identical medians and distributions as shown in Figure 2. The results do not contradict the outcomes of the Spearman correlation scrutiny. Although written and oral scores have similar medians, spread (IQR), and ranges and distributions of written and oral scores look alike overall, the boxplots do not tell the readers whether individual students who score high on one task score high on the other.

Figure 2

Written versus oral (online students) comparison



The comparison between oral and written performance reveals a clear pattern when the distributional characteristics and the interrelationship of paired scores are considered together. The oral scores display a slightly lower central tendency (median = 80%) than the written scores (median = 83.75%), suggesting that students tend to perform marginally better in the written task. In addition, the written assessment exhibits greater variability, with a wider range extending from 23.75% to 100%, compared to 30% to 100% in the oral task. The interquartile ranges of the two assessments overlap substantially, indicating broadly similar central distributions. Both formats reach the same maximum score of 100%, while the written assessment displays greater dispersion at the lower end of the distribution due to the presence of multiple low outliers.

Despite these written and oral scores differences of face-to-face cohort and similarities in central tendency and spread of online cohort, the degree to which written and oral performance align for the same student must be evaluated through correlation. In this dataset, there was no correlation and a weak correlation between written and oral scores of face-to-face students and online students respectively. In other words, the students who performed strongly in the written assessment were not consistently the same students who performed strongly in the oral assessment. Thus, while the oral task yielded slightly higher overall performance, the pattern of who succeeds and who struggles differs across modalities. Collectively, these findings point to the fact that the oral and written assessments are not interchangeable measures: although group-level outcomes show broadly similar performance, individual-level outcomes diverge. Such discrepancies may stem from the distinct demands of each medium. Written mathematics assessment emphasizes organized, formal solutions, whereas the oral format allows for interactive mathematical reasoning. Consequently, oral formats can reveal aspects of students' thinking that written solutions may not show. This implies that the two modalities activate skill repertoires and competencies that are partially distinct. The marginally higher performance scores observed in the oral assessment can be attributed to certain structural features of the task that differentiate it from the written format. First, the oral assessment generally evaluates a different subset of skills and competencies, such as swift information recall, oral clarification, and superficial analytical reasoning, rather than the extended problem-solving, structured argumentation, and deeper reasoning strategies required in written tasks. Second, unlike the written assessment, the oral task was not designed to be as score-heavy or cognitively demanding, leading examiners to select relatively straightforward questions that were intended to elicit quick responses rather than sustained analytical work. As a result, students could achieve higher performance without demonstrating the same level of reasoning required in the written assessment. Third, it is possible that examiners unknowingly provided guiding cues through facial expressions, tone, or body language, which may have supported or reassured students during their responses. This phenomenon is well-documented in the psychological literature, most famously in the case of Clever Hans, the "counting horse," whose apparent mathematical ability was later understood to arise from subtle, unconscious signals from human observers. As previously noted, such interpersonal unintentional hints can influence oral assessments despite examiners' best efforts to remain neutral, leading to inflated performance compared to written tasks.

Conclusion

From an assessment-design perspective, these results underscore the pedagogical importance of employing multimodal assessment strategies. Because the two formats cover complementary aspects of students' mathematical proficiency, relying exclusively on a written or oral assessment mode may misrepresent the competencies of particular learners. Using both written and oral components allows educators to obtain a more comprehensive view while ensuring that students with differing strengths, including neurodivergent learners who may prefer systematic written processing or dynamic oral reasoning, have equitable opportunities to demonstrate understanding. The results suggest that written assessments by themselves do not adequately capture diverse students' conceptual knowledge and relational understanding; consequently, the inclusion of oral assessments is an alternative assessment modality in mathematics. Further investigation is required into the use of oral assessments in contemporary educational contexts, particularly with respect to the perceptions of both students and educators. Additionally, the capacity of oral assessments to support academic integrity in the context of artificial intelligence may be regarded as a further advantage.

Acknowledgment

Ethics approval (No. 2026/008) was granted by the Federation University Human Research Ethics Committee and the research utilised de-identified assessment data.

References

Akimov, A., & Malin, M. (2020). When old becomes new: A case study of oral examination as an online assessment tool. *Assessment & Evaluation in Higher Education, 45*(8), 1205–1221.

Baker, S., Irwin, E., Hamilton, E., & Birman, H. (2022). What do we know about enabling education as an alternative pathway into Australian higher education, and what more do we need to know? A meta-scoping study. *Research Papers in Education, 37*(3), 321–343.

Crawford, N., Olds, A., Lisciandro, J. (2022). *Emotional labour demands in enabling education: A qualitative exploration of the unique challenges and protective factors*. University of the Sunshine Coast. Retrieved from https://research.usc.edu.au/esploro/outputs/journalArticle/Emotional-labour-demands-in-enabling-education/991122151902621

Dawson, P. (2024). Remote proctoring: Understanding the debate. In S. Eaton (Ed.), *Second handbook of academic integrity* (pp. 1511–1526). Springer. https://doi.org/10.1007/978-981-99-9374-8_77

Fenton, A. (2025). Reconsidering the use of oral exams and assessments: An old way to move into a new future. *Educational Researcher, 54*(5). https://doi.org/10.3102/0013189X251333638

Ferdoush, F., Alam, S. S., Jannifa, F. T., Haque, M. J., Johora, F., Parveen, M. F., & Rahman, M. S. (2025). Written versus oral assessment: Are both equally effective in pharmacology? *Bangladesh Journal of Medical Education, 16*(1), 1–8. https://doi.org/10.3329/bjme.v16i1.79095

Iannone, P. (2023). Collaborative evaluation of teaching and assessment interventions: Ideas from realistic evaluation. In *Mathematicians' reflections on teaching* (pp. 139–151). Springer. https://doi.org/10.1007/978-3-031-34250-9_10

Iannone, P., & Vondrová, N. (2024). The novelty effect on assessment interventions: A qualitative replication study of oral performance assessment in undergraduate mathematics. *International Journal of Science and Mathematics Education, 22*(2), 375–397. https://doi.org/10.1007/s10763-023-10399-4

Jankowski, K. R. B., Flannelly, K. J., & Flannelly, L. T. (2017). The *t*-test: An influential inferential tool in chaplaincy and other healthcare research. *Journal of Health Care Chaplaincy*, *24*(1), 30–39. https://doi.org/10.1080/08854726.2017.1335050

Joughin, G. (2010). *A short guide to oral assessment*. Leeds Met Press in association with University of Wollongong.

Moothadeth, M. (2024). Reliability of rubrics in the assessment of clinical oral presentation: A prospective controlled study. *Journal of Education and Health Promotion, 13*, 182.

Nallaya, S., Gentili, S., Weeks, S., & Baldock, K. (2024). The validity, reliability, academic integrity and integration of oral assessments in higher education: A systematic review. *Issues in Educational Research, 34*(2). Institutes for Educational Research in NSW, SA and WA.

Puchkov, E., Knyaginina, N., Novoselov, I., & Jankiewicz, S. (2020). COVID-19 and final (state) examination: Strategies of international and Russian universities. *Białostockie Studia Prawnicze, 25*(4), 257–274. https://doi.org/10.15290/bsp.2020.25.04.17

Rahman, G. (2011). Appropriateness of using oral examination as an assessment method in medical or dental education. *Journal of Education and Ethics in Dentistry*, *1*(2), 46–51. https://doi.org/10.4103/0974-7761.103674

Theobold, A. S. (2021). Oral exams: A more meaningful assessment of students' understanding. *Journal of Statistics and Data Science Education, 29*(2), 156–159. https://doi.org/10.1080/26939169.2021.1914527

Thiese, M. S., Ronna, B., & Ott, U. (2016). P value interpretations and considerations. *Journal of Thoracic Disease, 8*(9), E928–E931. https://doi.org/10.21037/jtd.2016.08.16

Videnovic, M. (2017). Oral vs. written exams: What are we assessing in mathematics? *IMVI Open Mathematical Education Notes, 7*(2017). https:// 10.7251/OMEN1701001V

Videnovic, M. (2020). *Mathematics professors' views on written and oral assessment in mathematics* (Doctoral dissertation, Simon Fraser University). Summit Research Repository. https://summit.sfu.ca/item/20438

Xiao, C., Ye, J., Esteves, R. M., & Rong, C. (2015). Using Spearman's correlation coefficients for exploratory data analysis on big dataset. *Concurrency and Computation: Practice and Experience*, *28*(14), 3866–3878.