ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Making IOAs work at scale: An evidence-based framework from six semesters of implementation

Hasti Samadi, Geela Venise Fabic Chee

University of Melbourne

The widespread availability of generative artificial intelligence (GenAI) has increased demand for assessment approaches that provide stronger assurance of student learning while maintaining authentic educational value. Interactive Oral Assessments (IOAs) have emerged as a promising solution, but practical guidance for implementing them sustainably in large undergraduate subjects remains limited. This paper proposes a framework grounded in six consecutive semesters of continuous IOA implementation in a large undergraduate computing subject enrolling between 300 and 560 students per offering. Evidence draws on cohort mark distributions, IOA–examination correlations, academic integrity outcomes, and examiners' consistency, as reflected in inter-rater reliability data. Repeated iterations of IOA design, delivery, evaluation, and refinement, informed by the literature and feedback from students and staff, produced four design principles critical to successful implementation of IOAs at scale. These principles address examiner calibration, rubric design, question depth and individual mark weighting. They are synthesised into an operational blueprint that covers assessment architecture, rubric design, question development, examiner preparation, and student preparation. Together, these provide transferable guidance for institutions seeking to implement IOAs sustainably at scale in the GenAI era.

Keywords: interactive oral assessment, academic integrity, generative AI, higher education, constructive alignment, longitudinal case study

Introduction

The increasing availability of generative artificial intelligence (GenAI) has intensified pressure on universities to design assessments that provide stronger assurance of individual understanding. Take-home assignments are more vulnerable than ever to authorship questions: students can feed course materials into GenAI tools to produce entire submissions, a risk that spans both individual and group work (Lodge et al., 2025; Tu, 2024). Interactive Oral Assessments (IOAs) have attracted growing interest as a viable response to this challenge. IOAs are live, structured question-and-answer sessions in which an examiner questions a student about work they have already submitted, requiring them to explain and justify their reasoning in real time and confirm that the understanding is genuinely their own (Sotiriadou & Logan-Fleming, 2025). Beyond assessment security, IOAs are grounded in constructive alignment and active learning: they require students to justify decisions, interpret outputs, and reflect on their reasoning rather than reproduce prepared content. Evidence supports their value for authentic learning, student engagement, and communication skill development, with students and staff alike reporting positive experiences (Krautloher, 2024; Sotiriadou & Logan-Fleming, 2025; Tan et al., 2022; Ward et al., 2024). Despite these advantages, translating IOA principles into a working model is not straightforward, particularly in large undergraduate subjects where multiple assessors must mark diverse cohorts within constrained timeframes while maintaining marking consistency and alignment with learning outcomes. The challenge is not simply to design an IOA assessment but to develop a sustainable process that supports students, prepares assessors, and ensures fairness at scale. Existing literature identifies useful supports (clear rubrics, practice activities, assessor training, calibration workshops) but stops short of explaining how these elements integrate into a coherent operational framework (Krautloher, 2024; Sotiriadou & Logan-Fleming, 2025).

This paper responds by proposing an evidence-based IOA framework developed over six consecutive semesters of implementation in a large undergraduate computing subject that enrols 300–560 students per offering. Rather than simply advocating for IOAs, the paper identifies four design principles that emerged through repeated cycles of implementation, failure and recovery, and synthesises them with an operational blueprint for institutions seeking to implement IOAs at scale. While a recent publication has proposed a related framework

(Kalman et al., 2026), its applicability to large-cohort undergraduate computing contexts remains unclear from publicly available information at the time of writing this paper.

Study context

The framework presented in this paper was developed over six consecutive semesters of IOA implementation in a large undergraduate data analytics (computing) subject at the University of Melbourne, which enrolled between 300 and 560 students per semester. Although the broader assessment structure evolved across offerings in response to student feedback and curriculum requirements, the IOA remained a component throughout and was always tied to a non-invigilated take-home assignment (a written report and/or a programming task). Students submitted their work first; the IOA followed shortly after, with questions drawn from each student's own submission. Framework development drew on multiple evidence sources across each iteration: cohort-level mark distributions, inter-rater reliability measures, comparisons between IOA and final examination results, student feedback, and teaching staff reflections. These informed successive cycles of design, delivery, and refinement rather than a fixed design from the outset. The IOA was first piloted with a smaller cohort of 30–35 students before full-scale implementation (Samadi & Chee, 2024); that pilot provided the starting point for the first full-cohort iteration and full mixed-methods evaluation (Samadi & Chee, 2025). This paper focuses on the resulting framework, operational blueprint, and design principles. Analyses for this study were conducted using de-identified data under the Teaching and Learning Laboratory (TLL) Teaching Evaluation Protocol, #29248.

Framework development

The IOA was piloted in a smaller data analytics subject with 30–35 students before full-scale rollout (Samadi & Chee, 2024); this six-semester iteration follows the pilot, which commenced in Semester 2 (S2) 2023. Every student was independently marked by two examiners, although reliable marker-level records survive only from S2 2024, so the marker-level analysis starts there. We quantified examiner agreement using the Intraclass Correlation Coefficient (ICC), specifically ICC(1) with one-way random effects, a statistical metric for assessing interrater reliability (Koo & Li, 2016). ICC(1) was used because examiners were not a fixed set across semesters. Accordingly, semester-level ICC estimates were treated as approximately independent. ICC(1) differences between semesters were evaluated using Fisher's z-transformation (McGraw & Wong, 1996) to enable statistical comparison between ICC values. We then applied the Holm–Bonferroni correction (Holm, 1979) across the three resulting comparisons ($\alpha = .05$), and only the S1 2026 improvement reached significance under this parametric test. Because this test treats each semester as fully independent, we subjected the same comparison to additional clustering-aware testing, reported in Phase 3, which revised this result.

Phase 1 (S2 2023–S2 2024): Building the foundations

Phase 1 was the foundational period in which the IOA was built and its core design problems diagnosed. Three iterations yielded two main lessons that shaped everything that followed. The first lesson was question design. To scale individual questioning across large cohorts, we trialled AI-generated questions using the university-provided AI tool (Samadi & Chee, 2024). Observations revealed that although the AI-generated questions were tailored to each submission using a well-developed prompt, they were not sophisticated enough to probe deep understanding; instead, they focused on recalling explanations already present in the submission. Our main finding was that an IOA's authenticity lies not in the "oral" element but in the "interaction," which requires questions that demand reasoning and are paired with well-curated follow-ups.

Table 1.
ICC(1)= inter-rater reliability between examiners' estimates across four reportable iterations. Band classifications follow Koo & Li (2016): poor <0.50, moderate 0.50–0.75, good 0.75–0.90, excellent >0.90. n = number of students assessed. 95% confidence intervals are reported in brackets

Semester (S) – Year	**ICC(1)**	**95% CI, n**	**Band**
S2 - 2024	0.701	[0.63, 0.76], 391	moderate
S1 - 2025	0.755	[0.71, 0.79], 517	(Low) Good
S2- 2025	0.754	[0.70, 0.80], 358	(Low) Good
S1 - 2026	0.845	[0.82, 0.87], 560	Good

From S2 2024, we abandoned AI-assisted generation, and all questions were examiner-generated and drawn directly from each student's work. Despite the time cost of manual question creation, using an AI tool proved slower overall: curating prompts, waiting for processing, and substantially revising generic output outweighed any efficiency gain and reinforced a core recommendation: IOA validity depends on question specificity to the individual student.

The second lesson was rubric design. The rubric oscillated across Phase 1, from an analytic question format to three explicitly detailed criteria per question, then to a simpler set of criteria. It was observed that the overly detailed rubric was practically hard to apply consistently when marking live IOAs. In Table 1, we show that simplifying the rubric in S1 2025 coincided with a nominal rise from ICC(1) = 0.701 to 0.755, crossing from the moderate to the good band on the Koo and Li (2016) scale (0.50–0.75 moderate, 0.75–0.90 good). However, this increase did not reach statistical significance (z = 1.44, p = .151), so it should be read as a trend rather than a confirmed shift in the reliability band.

Table 2.
IOA design and outcome data across six semesters (S2 2023–S1 2026). Itr = Iteration; Wt% = mark weight as percentage of total subject mark; M = mean; SD = standard deviation; NS% = non-attendance rate; CV% = within-group coefficient of variation; r = Pearson correlation with final exam; m/St = Minutes per Student; grp = group; indv. = individual; pres. = presentation; lect. = lecture

Itr	N	**Design**					**Outcome**		
		Wt%	Time	Rubric structure	Tutor Prep	Student Prep	M; SD; NS%	CV%	r (exam)
S2 2023	335	10% (grp)	10 min = 8 grp pres. + 1.5 grp Q&A	Analytic group questions	—	—	M=0.84, SD=0.17 NS=2.4%	0% of grps	r=0.275 p<.001
S1 2024	508	8% = 4% grp + 4% indv.	25 min = 9 min grp pres. + 3 m/St	Methods understanding	2 hr: rubric + logistics	1 hr pres. skills lect.; sample videos	M=0.83, SD=0.16 NS=1.2%	3.4% of grps	r=0.301 p<.001
S2 2024	391	13% = 6% grp + 7% indv.	30 min = 7 min grp pres. + 4 m/St	Explanation & interp.; Justification; Evaluation	2 hr: rubric + sample rec. practice	1 hr pres. skills lect.	M=0.76, SD=0.16 NS=1.0%	3.7% of grps	r=0.356 p<.001
S1 2025	517	17.5% = 3% grp + 14.5% indv.	6 m/St	A1: code justif.; grp res. analysis	2×2 hr: rubric + calibration samples	1 hr pres. skills lect.	M=0.86, SD=0.18 NS=2.3%	22% of grps	r=0.244 p<.001
			8 min grp pres. + 6 m/St	A2: Explanation; Justification			M=0.74, SD=0.19 NS=1.1%	26.8% of grps	r=0.387 p<.001

S2 2025	358	18% = 3% grp + 15% indv.	6 min/St	A1: code justif.; grp res. analysis	2×2 hr: rubric + calibration on samples	30 min pres. skills; 30 min interview skills lect.	M=0.78, SD=0.22 NS=1.7%	34.4% of grps	r=0.405 p<.001
			8-10 min grp pres. + 6 m/St	A2: Explanation; Justification			M=0.71, SD=0.20 NS=0%	18.6% of grps	r=0.591 p<.001
S1 2026	560	20% = 3% grp + 17% indv.	6 m/St	A1: code justif.; code modif.; result analysis	2×3 hr: rubric, sample scenarios, sample Q evaluation calibration	Q-bank; examiner intro videos; 1-hr AI practice workshop; 1-hr lect.	M=0.79, SD=0.26 NS=2.9%	36.8% of grps	— (exam pending
			8-10 min grp pres. + 6 m/St	A2: Justification; Analysis; Critical reflection			M=0.70, SD=0.20 NS=0.9%	23.1% of grps	— (exam pending

Phase 2 (S1 2025–S2 2025): Full scale implementation to both assignments (A1 and A2)

Phase 2 expanded the IOA to include both non-invigilated take-home assignments. In Phase 1, the IOA was attached only to Assignment 2 (A2), an open-ended analytics project that spanned most of the subject's learning outcomes. Extending it to Assignment 1 (A1), an early-semester programming task, enabled the use of GenAI in both assignments: students could use AI to develop the programming aspects of their assignments, with the IOA ensuring they could account for their own work. Since A1 IOA was worth only 5% of students' final marks and was deemed a minor assessment, it was conducted via Zoom. In contrast, A2 IOA was conducted in person (worth 15% of students' final marks). The two assignments correlated differently with the final examination. A2's exam correlation was numerically higher than A1's in both Phase 2 iterations (Table 2). Steiger's (1980) Z-test for dependent correlations found this difference non-significant in S1 2025 ($Z = 1.59$, $p = .112$) and, after Holm–Bonferroni correction (Holm, 1979), across the two semesters, found no significance in S2 2025 either ($Z = 2.12$, $p = .034$; corrected threshold 0.025). A bootstrap comparison of the two gap sizes (95% CI: [−0.13, 0.16]) likewise found no distinguishable widening over time (Good, 2005). The pattern is nonetheless directionally consistent with a coverage explanation: A2 mirrors the exam's syllabus-wide scope, while A1 addresses only a narrow slice; and A2's in-person, report-based format taps conceptual reasoning more directly than A1's programming code-focused, Zoom-based format. Both points offer a plausible, though unconfirmed, rationale for weighting report-anchored, in-person questioning more heavily where resources are constrained.

Marking reliability held steady through the expansion: ICC(1) = 0.755 and 0.754 in S1 and S2 2025 ($z = −0.03$, $p = .975$), indicating that adding a second assignment did not degrade agreement. This stability masked a criterion-level gradient: ICC(1) for Justification (0.693, 0.696) and Explanation (0.681, 0.640) stayed in the moderate-to-good range across both semesters, while its value for Delivery (0.584, 0.453) and Slides/Content (0.487, 0.439) were lower and declined further, with Slides/Content falling into the poor range (Koo & Li, 2016). It indicates that examiners agreed most on questions that demanded justification and explanation, and least on those where criteria rested on subjective judgement, foreshadowing Phase 3's shift toward higher-order reasoning. Within-group variability remained substantial (Table 2); since each group is assessed by one examiner pair per session, this spread is more likely to reflect genuine differences in individual performance than marking inconsistency, which an individual oral is designed to surface. Consistent with this, misconduct cases were identified through the IOA process across multiple semesters, illustrating its capacity to expose work that does not reflect individual understanding (see P4 for additional survey-based results).

Phase 3 (S1 2026): Refinement

Phase 3 directly addresses the rubric flaws of earlier phases. The rubric was explicitly grounded in Bloom's Taxonomy (Anderson & Krathwohl, 2001) and stabilised around three higher-order criteria – Justification, Analysis, and Critical Reflection – each mapped to a defined cognitive level, for a standardised vocabulary to provide more consistent performance descriptors across levels. Peer review was added to embed a metacognitive dimension: students evaluate peers' performances against the same criteria before their own assessment, on the premise that judging against the standard deepens understanding of what it requires. The reliability evidence is directionally consistent with the refinement having improved marking consistency, with

one caveat. After two statistically indistinguishable semesters (S1–S2 2025), agreement rose numerically to ICC(1) = 0.845 in S1 2026, the largest gain in the series. A cluster-level permutation test (10,000 permutations) (Good, 2005) and a linear mixed-effects model (Bates et al., 2015) with examiner pair as a random intercept, using Satterthwaite-approximated degrees of freedom (Satterthwaite, 1946), found no significant difference between the two semesters. We therefore treat the S1 2026 improvement as a substantial numerical trend coinciding with the Bloom-aligned rubric and calibration redesign, rather than a confirmed statistical effect. With the largest cohort to date (n = 560), examination results are pending; early IOA means are consistent with prior phases (Table 2), and whether the stronger reliability is matched by an improved IOA–examination correlation remains to be seen. The six-semester trajectory establishes the evidence base for what follows. The Operational Blueprint documents the most mature version of the IOA design reached through this process, a practical, tested model. The Design Principles distil the recurring findings that shaped it into transferable guidance for institutions.

Operational blueprint

The blueprint translates the design lessons from six semesters into a set of concrete, replicable components. The preceding section described our most mature IOA implementation, including the rationale for each decision, informed by both the literature and our practical experience. This blueprint covers IOAs at scale for group assessments (three to four members). The group format reflects the constraints of this context, not a universal recommendation.

Assessment architecture

The IOA described here in detail was applied to Assignment 2 (A2), a non-invigilated group assignment in which students submit a written analytic report based on a programming task. Each session consists of a group presentation (eight to 10 minutes) followed by individual Q&A, with the individual component contributing most of the assessment weighting. All questions are drawn exclusively from each group's own submission, requiring students to explain and defend their own work. The individual component mark grew from 0% to 85% of the IOA total across iterations (Table 2) to ensure every student demonstrates understanding of their entire project. Two examiners independently assess each student, and their scores are averaged to promote fairness and consistency at scale. The individual session requires a minimum of six minutes per student, with adjustments allowed under an approved special consideration arrangement. Our sessions run for two hours, with three groups assessed per session by a randomly allocated pair of examiners. All IOA sessions for A2 are conducted in person, in morning and afternoon slots throughout the week (on business days), to accommodate students' availability. Groups select their own session times, with priority given to those who select first.

Rubric design

In our experience, three criteria are the most likely ceiling for the individual Q&A rubric, each aligned with a distinct cognitive level; beyond this, complexity increases without improving marking consistency. For analytical assignments, our rubric draws from the upper levels of Bloom's revised taxonomy (Anderson & Krathwohl, 2001): Justification, Analysis, and Critical Reflection. Each criterion links to a defined cognitive level and is described through observable behaviours rather than subjective descriptors such as "clarity" or "depth," with performance bands distinguishing reasoning depth by whether the student demonstrates awareness of alternatives, trade-offs, or limitations. The group presentation was assessed using a separate, simplified rubric focused on communication and presentation skills.

Question design

Questions should be written directly from each group's submitted work, as they are intended to assess students' decision-making processes and understanding, rather than recall of general knowledge. Every question should align with a rubric criterion, with the complete set of questions covering the entire rubric. For example, questions may ask students to justify decisions made in their submission, interpret and analyse their own outputs, or critically evaluate the limitations of their approach; while specific criteria will vary with the assessment task, this principle holds across contexts. We recommend extracting eight to 10 questions per criterion, each coupled with one to two follow-up prompts. In our design, each student should receive at least

one question per criterion, spanning the full scope of the submission, including sections beyond those they personally completed in the group setting. As our current rubric comprises three criteria, this equates to a minimum of three substantive questions per student. Follow-up prompts are designed to probe incomplete responses, giving students the best opportunity to demonstrate their understanding. Although follow-ups can be premeditated, the interactive nature of the assessment demands some improvised questioning in the moment, underscoring the importance of examiners having a deep understanding of the rubric criteria and what each is actually intended to measure. For integrity purposes, before individual questioning begins, all phones, paper notes, and electronic devices must be set aside, and students must stand before the examiners without access to any materials.

Examiner preparation

Consistent marking at scale depends on structured examiner preparation. The aim is for all examiners to share a common understanding of the rubric criteria and their thresholds for a defensible outcome in any IOA practice. In our subject, before the IOA period, examiners must attend two mandatory training workshops, each lasting two to three hours, led by the lead academic and covering the rubric, cognitive expectations for each criterion, and marking calibration using exemplars. Scripted exemplars were developed deliberately to target cases where judgement is most likely to vary. The calibration aims to reach consensus among the examiners to ensure alignment with the rubric. Training should also include an explicit discussion of bias in marking oral assessments, reinforcing that marks are awarded only for the quality of students' responses, not for language fluency, accent, communication style, or presentation manner. This is both an equity principle and a validity requirement: if delivery influences marking, the assessment is no longer measuring what the rubric claims to assess. Exemplar responses at different proficiency levels include varied delivery styles to allow practice in separating content from manner. During the live IOA sessions, we conducted random quality-assurance observations to monitor consistency. In our latest design, student peer review also serves a quality-assurance function: in addition to evaluating peers' responses against the rubric, students rate the quality of the questions and students' perceptions of the examiner's fairness and performance, thereby providing a feedback loop across the cohort.

Student preparation

Students should receive structured preparation covering the assessment format, question types, marking criteria, and logistics. Our materials include sample questions drawn from an exemplar submission, sample answers at each band level to clarify what each band means in practice, and short introduction videos from all examiners to reduce uncertainty about who students will face. A one-hour lecture teaches presentation skills and strategies for managing the Q&A component, alongside a preparation workshop that teaches students how to simulate the IOA using GenAI tools against their own submitted work. However, it is emphasised that the assessment requires them to reason independently without external support, so devices are permitted only during practice and are prohibited during the actual IOA. We repeatedly emphasised that marking focuses on the quality of reasoning, not fluency or presentation style. Accessibility adjustments are always incorporated through standard institutional processes and may include live written questions, modified environments, or alternative settings, provided they preserve the assessment's intended learning outcomes.

Proposed design principles for IOA

The iterative development of our IOA implementation suggests that successful deployment at scale depends on a small set of interconnected design principles (Ps).

P1: Keep rubric criteria few and behaviourally anchored

IOA rubrics should minimise cognitive load by using a small number of student-level criteria rather than applying multiple criteria to every question. Our early implementation required independent scoring of several criteria for each question, resulting in a complex marking process with only moderate inter-rater agreement, as indicated by ICC(1). Simplifying the rubric coincided with a nominal rise in ICC (0.701 to 0.755, not independently significant; see Phase 1), but agreement remained limited because the criteria described subjective qualities, e.g., "rigorous rationale," leaving unclear standards. The largest improvement occurred when the rubric retained

a small number of criteria while redefining each using descriptors aligned with Bloom's revised taxonomy. Rather than judging abstract qualities, examiners assessed observable actions, such as "compares alternatives," "explains the limitation," and "qualifies a claim based on the data." This appears to have supported a more shared interpretation of performance standards and coincided with substantially higher inter-rater reliability. These findings suggest that reliable IOA rubrics depend on careful wording of criteria to clearly observable evidence of performance.

P2: Calibration is what makes marking consistent at scale

A well-designed rubric alone is insufficient to achieve consistent marking at scale. As the number of oral assessments scales, ensuring consistent application of performance standards becomes the primary challenge. This makes structured examiner calibration an essential component of IOA implementation, not an optional step. Calibration translates rubric criteria into consistent marking decisions by establishing common performance thresholds for student assessment. Our findings suggest that progressively structured examiner preparation strengthens marking consistency across successive iterations. The largest gain coincided with a simultaneous rubric redesign (P1), and the two cannot be statistically separated. We treat the ICC gains reported here as attributable to the combined calibration-and-rubric intervention rather than calibration alone. Inter-rater reliability rose from ICC(1) = 0.701 in S2 2024, through a statistically indistinguishable plateau of 0.755 (S1 2025) and 0.754 (S2 2025), to 0.845 in S1 2026, the largest numerical gain in the series, coinciding with structured calibration using exemplars aligned to Bloom's taxonomy. A parametric test flagged this gain as significant, but more conservative clustering-aware testing did not confirm it as statistically distinguishable from the preceding range (see Phase 3). We therefore treat this as an emerging trend consistent with improved marking consistency over time, not a demonstrated effect. We read it as a substantial trend consistent with improved marking consistency over time. The accompanying shift in the standard deviation of IOA marks (Table 2) suggests greater differentiation of student performance, with marks becoming less concentrated around the centre of the scale. The corresponding changes in the IOA average mark in our implementation iterations are also consistent with this interpretation (Table 2). Various strands of the literature have repeatedly found that less experienced or less confident examiners award higher marks, which is attributed to lower familiarity with marking thresholds and a weaker grasp of standards (Chong et al., 2017; McManus et al., 2006). For institutions deploying an IOA for the first time, organising examiner calibration sessions is the investment most likely to be underestimated. Structured calibration, practising on scripted exemplars, and reaching consensus on borderline cases before live marking begins turn a well-designed rubric into consistent marking decisions at scale.

P3: Question depth and correct follow-ups determine what the IOA measures in practice

An IOA question drawn from a student's own submission can still fail to assess individual understanding if it does not probe at sufficient depth. Questions must require students to explain, justify, interpret, or critically evaluate decisions in their own work; questions that merely ask students to describe their work simply assess recall rather than understanding. Our findings support this in two ways. First, AI-generated questions, though drawn from each submission, were too generic, mostly assessing memorisation rather than understanding, so we abandoned them in favour of examiner-written questions. This shift is evident in A2 gap rates (IOA marks more than 30 percentage points below the submission mark): low during the AI-generated semesters (0.61% in 2023 SM2; 1.39% in 2024 SM1), but consistently higher afterwards (3.65%–10.49%), with fluctuations likely reflecting evolving examiner training. This increase is unlikely to reflect declining student performance; rather, it is consistent with shallow, recall-based questions failing to probe genuine understanding and so being answerable regardless of whether it exists, whereas deeper questions are more likely to expose the gap. Second, this gap is not just an artifact of the AI-to-examiner transition: it reflects understanding invisible in the submission alone, exposed only when questions require students to reconstruct and defend their thinking, indicating that submitted work alone may overestimate knowledge, particularly given the non-invigilated, take-home format. This kind of increase in depth of questions is rarely achieved by a single well-designed set of questions; it typically emerges through follow-up, as examiners respond to what a student has just said. This is consistent with our earlier finding that IOA authenticity lies not in the "oral" element but in the "interaction," and it means follow-up questions matter as much as the main question itself. Survey data from A2 IOAs in S1 2026 suggest this adaptive questioning largely functioned as intended: when asked whether examiners properly adjusted questions based on students' answers, 82.4% agreed (41.65% "somewhat agree," 40.73% "strongly agree"), versus 11.0% neutral and 5.7% disagreeing (n = 540), though examiner variability in follow-up quality was not

directly measured. Students' own accounts point to why this interaction mattered: approximately 14% of substantive comments (*n* = 40/292) credited the exchange of questions, follow-ups, and clarifications, rather than the oral format itself, with deepening understanding, describing an interactive quality that revealed understanding a static report could not. This benefits both parties: students can clarify and expand their reasoning, while examiners can probe further to distinguish genuine understanding from recall. These findings suggest that questions, both initial and follow-up, should be designed to elicit reasoning rather than recall. Developing and calibrating high-quality questions aligned with the assessment criteria is a core component of IOA implementation.

P4: Weight the individual component enough to make it consequential

Group assessments promote collaboration skills but can obscure individual contributions when marks are shared, allowing some students to benefit disproportionately from others' work (Hall & Buzwell, 2013). An IOA addresses this limitation only if the individual component carries sufficient weight to make individual understanding consequential. Our findings are consistent with this principle. The individual weighting of the IOA rose from 0% (fully group-marked, S2 2023) to 85% of the total IOA mark by S1 2026 (Table 2). Early in this progression, reports of inactive or disengaged group members declined from about 10% under the group-only mark to 3.5% once individual weighting reached 50% (S1 2024) (Samadi & Chee, 2024). Within-group coefficient of variation (CV), chosen for its scale-independence and group-level interpretability, reinforces this from a different angle. This dispersion did not exist before the individual component was introduced: under the prior group-only mark, all members of a group received the same grade, so the within-group CV was zero. Across Phase 2, CV instead ranged from 18% to 34% depending on assignment and semester (Table 2). Increased variability in awarded marks is consistent with greater differentiation of student performance. However, iterative implementation means the effects of individual design changes cannot be isolated. The shift from zero to substantial dispersion after the introduction of individual weighting is nonetheless consistent with the accountability signal the IOA is intended to surface: a low individual weight is less likely to make substantial differences visible in the final mark.

IOA integrity outcomes

Our findings show that IOAs support academic integrity and are well-received by students. Results indicated that IOA performance diverged from the submitted work, leading to confirmed integrity cases through follow-up procedures. On students' opinions on IOAs, evidence comes from two sources. First, post-assignment surveys in 2025 S1 revealed that students self-reported having improved critical thinking skills at 54.0% (A1) and 61.3% (A2), and 52.8% (A1) and 62.5% (A2) felt the IOA encouraged reliance on their own understanding over GenAI. Although response rates were low, they are still notable (15.6%; *n* = 56 for A1 and 10.6%; *n* = 38 for A2). A more recent survey in S1 2026, with a 75.8% response rate (*n* = 413), asked whether students would recommend using oral assessment in other subjects to enhance learning. Results showed that agreement outweighed disagreement by roughly 2:1, with 48.9% agreeing or strongly agreeing vs. 25.4% disagreeing or strongly disagreeing (the remaining 25.7% were neutral). Furthermore, there was a decline in IOA non-attendance: in A2, non-attendance fell from 2.4% (S2 2023) to 0% (S2 2025), remaining low thereafter (range 0–1.1%; Table 2). This is an observed behavioural trend rather than a measured attitude; we interpret it as consistent with students increasingly treating the IOA as a fixed part of the assessment, though we did not survey students on this directly. This is consistent with the proposition that IOAs support the integrity of the assessment prior to the IOA session. Students were made aware that their understanding of their submitted work would later be examined through dialogue in the IOA. This possibly influences how students approach the assignment from the outset, encouraging greater reliance on their own understanding. Overall, these results suggest a broadly positive endorsement of the IOA as a learning experience.

Observations through several iterations of IOAs suggest that they should be designed primarily as learning-focused assessments that require students to explain, justify, analyse, and reflect on their own work. Because the questions target higher-order reasoning, the same assessment also serves a second function: to help identify submitted work that does not reflect the individual understanding exhibited during the live IOA. This could eliminate the need for a separate assessment focused primarily on upholding academic integrity. To run a successful IOA at scale, institutions should first focus on learning-centred IOAs, ensuring questions target deep

understanding and encourage critical thinking (P3). This must also be supported by marking calibration sessions (P2), anchoring rubric criteria to observable behaviours (P1), and carefully considering the weighting of individual marks (P4), with the latter applicable only to IOAs associated with group work.

Discussion

The primary contribution of this work is not another IOA implementation but an evidence-informed framework for designing and delivering IOAs at scale. Our findings suggest that the components of an IOA rubric design, question development, and examiner preparation sessions cannot be developed in isolation; each must be designed holistically as part of a single implementation. The resulting operational blueprint and design principles provide a transferable foundation for institutions seeking to implement IOAs in their own settings. This positions our contribution alongside Kalman et al. (2026), whose framework for institutional capacity-building around IOAs addresses a related but different problem; our six-semester iteration offers at-scale evidence for sustaining an IOA within a single high-enrolment subject, a context their framework does not directly address. A separate paper, currently in preparation, will report the full mixed-methods evaluation in greater depth; what we present here is the design framework, not the complete evaluative picture. These findings have implications beyond data analytics subjects and are likely to generalise to other large-cohort disciplines facing similar GenAI pressures. Our findings suggest that implementation quality depends not on the oral format itself but on the supporting system around it: examiner preparation sessions, rubric design, and question development function as core implementation components rather than peripheral logistics. Institutions adopting oral assessment in response to GenAI should plan for this system from the outset, rather than treating it as something to refine after the fact.

This work has several limitations. First, the framework was developed within a single-institution data analytics subject and may require adaptation in other contexts. Second, the iterative implementation introduced multiple concurrent changes, preventing causal attribution to any single modification; findings should be read as design evidence rather than controlled experimental evidence. Third, the IOA assesses understanding demonstrated at the time of the live assessment rather than direct contribution to the preceding group work: a student who contributed little could still have prepared extensively beforehand and performed well, so while this reflects individual learning, IOAs cannot conclusively determine who completed specific aspects of the submission. Fourth, reliable marker-level records exist only from S2 2024 onward, one semester after individual marking began. Fifth, the CV evidence for P4 cannot separate individual weighting from overall improved consistency in marking (P2) and cohort changes across semesters. Finally, differences in IOA–examination correlations were tested with Steiger's Z-test within each semester; after correcting for multiple comparisons across the two semesters, neither within-semester difference was statistically significant, and the apparent widening of the A2–A1 gap across semesters was not statistically distinguishable by bootstrap resampling (Good, 2005). Both the within-semester gap and its apparent widening over time should therefore be treated as descriptive rather than confirmed.

Conclusion

Scaling IOAs is not primarily a logistical challenge. It is a design challenge. The evidence across six semesters consistently shows that marking consistency, construct validity, and integrity assurance are not separate problems requiring separate solutions. They are products of an integrated set of design decisions, in which each component affects the others, and weakening any of them degrades the entire assessment. Institutions that approach IOA implementation by adding logistics around an existing task are likely to encounter the same failure modes that this study took several iterations to resolve. The broader implication is conceptual as much as practical. In the GenAI era, the question for assessment design is no longer whether a student submitted acceptable work, but whether the student exhibits genuine understanding of the work. IOAs make that question answerable. The shift is from ensuring authorship to ensuring learning, and the framework presented here supports it. This paper represents one output of ongoing research. A separate paper, currently in preparation, will report the full mixed-methods evaluation underlying these findings.

References

Anderson, L. W., & Krathwohl, D. R. (Eds.) (2001). *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. Addison Wesley Longman, Inc. https://lccn.loc.gov/00063423

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Chong, L., Taylor, S., Haywood, M., Adelstein, B. A., & Shulruf, B. (2017). The sights and insights of examiners in objective structured clinical examinations. *Journal of Educational Evaluation for Health Professions*, 14, 34. https://doi.org/10.3352/jeehp.2017.14.34

Good, P. (2005). *Permutation, parametric and bootstrap tests of hypotheses*. Springer New York.

Hall, D., & Buzwell, S. (2013). The problem of free-riding in group projects: Looking beyond social loafing as reason for non-contribution. *Active Learning in Higher Education*, 14(1), 37-49. https://doi.org/10.1177/1469787412467123

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, *6*(2), 65–70. https://www.jstor.org/stable/4615733

Kalman, E., Clarke, S. L., Miller, B., & Frawley, J. (2026). Developing academic capacity for interactive oral assessments. In K. Pechenkina (Ed.) *Design, delivery, and evaluation of professional learning for university staff* (pp. 329-364). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-3638-1.ch012

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, 15(2), 155-163. https://doi.org/10.1016/j.jcm.2016.02.012

Krautloher, A. (2024). Improving assessment equity using interactive oral assessments. *Journal of University Teaching and Learning Practice*, 21(4), 1-17. https://doi.org/10.53761/4hg1me11

McGraw, K. O., & Wong, S. P. (1996). Forming inferences about some intraclass correlation coefficients. *Psychological Methods*, *1*(1), 30–46. https://doi.org/10.1037/1082-989X.1.1.30

McManus, I. C., Thompson, M., & Mollon, J. (2006). Assessment of examiner leniency and stringency ('hawk-dove effect') in the MRCP (UK) clinical examination (PACES) using multi-facet Rasch modelling. *BMC Medical Education*, 6, 42. https://doi.org/10.1186/1472-6920-6-42

Samadi, H., & Chee, G. V. F. (2024). Leveraging oral assessments to enhance learning and integrity in teaching data analytics. *Proceedings ASCILITE 2024*. Melbourne (pp. 529–533). https://doi.org/10.14742/apubs.2024.1282

Samadi, H., & Chee, G. V. F. (2025). Interactive oral assessments: Combating ghostwriting in large-scale technical subjects. In *36th Australasian Association for Engineering Education Annual Conference (AAEE2025)* (pp. 929-937). Engineers Australia. https://aaee2025.org/microsite/pdf/full-paper_271.pdf

Sotiriadou, P., & Logan-Fleming, D. (2025). Harnessing the potential of interactive oral assessments in blended learning: Lessons from case studies. In P. K. Misra, S. Mishra & S. Panda (Eds.) *Case studies on blended learning in higher education: Design, development, and delivery* (pp. 193-210). Springer Nature Singapore. https://doi.org/10.1007/978-981-96-0722-8_11

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, *87*(2), 245–251. https://doi.org/10.1037/0033-2909.87.2.245

Tan, C. P., Howes, D., Tan, R. K., & Dancza, K. M. (2022). Developing interactive oral assessments to foster graduate attributes in higher education. *Assessment & Evaluation in Higher Education*, 47(8), 1183-1199. https://doi.org/10.1080/02602938.2021.2020722

Tu, X., Zou, J., Su, W., & Zhang, L. (2024). What should data science education do with large language models? *Harvard Data Science Review*, 6(1). https://doi.org/10.1162/99608f92.bff007ab

Ward, M., O'Riordan, F., Logan-Fleming, D., Cooke, D., Concannon-Gibney, T., Efthymiou, M., & Watkins, N. (2024). Interactive oral assessment case studies: An innovative, academically rigorous, authentic assessment approach. *Innovations in Education and Teaching International*, 61(5), 930-947. https://doi.org/10.1080/14703297.2023.2251967

ASCILITE 2026

The Future is Connection:

Collaboration, Creativity and Compassion in a Digital Society

Samadi, H. & Chee, G.V.F. (2026). Making IOAs work at scale: An evidence-based framework from six semesters of implementation. In Cochrane, T., Narayan, V., Poole, T., Lai, P., Lake, P., Moore, O., & Li, L. (Eds.), *The Future is Connection: Collaboration, Creativity and Compassion in a Digital Society*. Proceedings ASCILITE 2026. Brisbane (pp. 343–353). https://doi.org/10.65106/apubs.2026.3255

Note: All published papers are refereed, having undergone a double-blind peer-review process.
The author(s) assign a Creative Commons by attribution license enabling others to distribute, remix, tweak, and build upon their work, even commercially, as long as credit is given to the author(s) for the original creation.

© Samadi, H. & Chee, G.V.F. 2026