AI-Assisted Grading and Feedback in International Baccalaureate Middle Years Programme Physics: A Mixed-Methods Investigation of Learning Outcomes, Feedback Quality, and the Moderating Role of English Language Proficiency

Aminurrashid Bin Abu Bakar

University College Fairview

Abstract

This sequential explanatory mixed-methods study examined the impact of two artificial intelligence (AI) assessment modules on student learning outcomes in International Baccalaureate (IB) Middle Years Programme (MYP) Physics at an IB World School in Shenzhen, China. Participants comprised 101 Grade 9 and Grade 10 students, of whom 91% were non-native English speakers, alongside three science teachers, over a 24-week implementation period. Revision Village was deployed for Criterion A assessments, while ChatGPT supported Criteria B, C, and D. Quantitative data from pre- and post-intervention MYP criterion-referenced assessments, two Feedback Quality Assessment Rubrics, and Technology Acceptance Surveys were analysed using paired-samples *t*-tests, one-way ANOVA, and Hayes' PROCESS macro (Models 1 and 4). Qualitative data from semi-structured teacher interviews, student focus groups, and classroom observations were analysed using Braun and Clarke's six-phase thematic analysis. Results revealed statistically significant improvements across all four MYP criteria, with effect sizes ranging from Cohen's $d = 0.43$ (Criterion B) to $d = 0.74$ (Criterion D), and overall MYP grade improvement from 4.58 to 5.27 ($d = 0.64$). Feedback quality characteristics significantly mediated 48.3% of the relationship between AI implementation and learning outcomes (bootstrap 95% CI [0.802, 2.047]). English language proficiency significantly moderated implementation effectiveness ($B = -0.782$, $p = .010$), with Low- and Mid-proficiency students demonstrating significantly greater gains than High-proficiency students. The study contributes evidence-based guidance for integrating AI assessment modules in linguistically diverse IB Physics classrooms and proposes a revised AI–Human Feedback Partnership Model in which teacher mediation functions as the central bridging mechanism between AI-generated feedback and learner outcomes.

Keywords: AI-assisted grading, feedback quality, International Baccalaureate physics, English language proficiency, mixed-methods research, educational technology

AI-Assisted Grading and Feedback in IB MYP Physics: A Mixed-Methods Investigation

The integration of artificial intelligence (AI) into educational assessment represents one of the most consequential developments in contemporary classroom practice (Crompton & Burke, 2023; Wongvorachan et al., 2022). Systematic analyses indicate that publications on AI in education in 2021–2022 nearly tripled the volume of preceding years (Wang et al., 2024), and AI-assisted grading platforms have demonstrated capabilities extending well beyond simple answer verification—including automated analysis of reasoning, real-time identification of misconceptions, and generation of personalised feedback (Luckin et al., 2016). Within this expanding landscape, however, the application of AI assessment to inquiry-based science education, and to the International Baccalaureate (IB) Physics curriculum in particular, remains critically underexamined.

The IB Middle Years Programme (MYP) Sciences curriculum emphasises conceptual understanding, inquiry-based methodology, and criterion-referenced assessment across four dimensions: Knowing and Understanding (Criterion A), Inquiring and Designing (Criterion B), Processing and Evaluating (Criterion C), and Reflecting on the Impacts of Science (Criterion D) (International Baccalaureate Organization, 2023). Effective formative assessment in this context requires feedback that is timely, specific, and aligned with the cognitive demands of each criterion (Brookhart, 2013; Hattie & Timperley, 2007). Yet traditional teacher-led assessment imposes substantial cognitive and time burdens on educators, with feedback turnaround in pre-implementation contexts often approaching ten days (Carless, 2006; Molin et al., 2021). This temporal gap undermines the formative purpose of assessment because students lose the cognitive context required to act productively on feedback (Wisniewski et al., 2020). Emerging research suggests that AI tools can deliver scalable, criterion-referenced feedback in hours rather

than days (Kortemeyer, 2023; Latif et al., 2024), but the conditions under which such tools translate into measurable learning gains remain poorly understood.

A critical research gap exists at the intersection of three under-examined dimensions: AI-assisted grading in inquiry-based science assessment, the IB MYP Physics curriculum, and the linguistic-cognitive realities of students learning in English-medium international school contexts where most learners are non-native English speakers. The present study addresses this gap through a sequential explanatory mixed-methods investigation conducted at an IB World School in Shenzhen, China, where approximately 91% of students are non-native English speakers. The study examines two AI assessment modules—Revision Village, an IB-calibrated platform, and ChatGPT, a generative AI tool—and evaluates their impact on student learning outcomes, the feedback quality mechanisms through which any effects operate, and the contextual factors that condition their effectiveness.

Four research questions guided the investigation:

RQ1. To what extent does AI-assisted grading implementation improve student learning outcomes in MYP Physics education?

RQ2. How do feedback quality characteristics mediate the relationship between AI-assisted grading implementation and student learning outcomes in MYP Physics?

RQ3. What challenges and opportunities do physics teachers and students encounter when implementing AI-assisted grading tools?

RQ4. How do contextual factors, including English language proficiency, shape the effectiveness of AI assessment modules on student learning outcomes?

Literature Review

Theoretical Framework

This study integrates four theoretical perspectives. Hattie and Timperley's (2007) feedback model provides the foundation, identifying three critical feedback functions—feed-up (clarification of learning objectives), feedback (progress monitoring), and feed-forward (guidance for subsequent learning)—and four feedback levels (task, process, self-regulation, and self), with self-regulation feedback yielding the strongest effects on durable learning (Mandouit & Hattie, 2023). Self-Determination Theory (Deci & Ryan, 2017) explains how AI feedback systems may support or undermine the basic psychological needs for autonomy, competence, and relatedness; immediate AI feedback can support competence by enabling observable mastery progression, while the absence of relational warmth may undermine relatedness for some learners (Fathali & Okada, 2018). The Technology Acceptance Model (Davis, 1989) and its extensions (Venkatesh et al., 2012) provide a framework for understanding teacher and student adoption of AI grading systems through perceived usefulness, perceived ease of use, and facilitating conditions (Dwivedi et al., 2019). Finally, Task–Technology Fit theory (Goodhue & Thompson, 1995) frames the alignment between AI-tool affordances and the cognitive demands of MYP criterion-based assessment.

AI Grading Effectiveness in STEM Contexts

Empirical research on AI grading in physics has produced promising but context-dependent results. Kortemeyer (2023) demonstrated that AI systems could grade introductory physics problem solutions with correlation coefficients of $R^2 = 0.84$ using a MathPix and GPT-4 workflow compared to human graders. A subsequent psychometric validation by Kortemeyer and Nöhl (2024) developed a threshold-based human–AI collaboration

model achieving $R^2 \approx 0.91$ when AI handled half the grading load. At the secondary level, Altal and Abo Ehsaiyan (2025) reported a large effect (Cohen's $d = 1.21$) for AI-assisted instruction incorporating GPT-4 in Grade 11 Advanced Physics in the United Arab Emirates. A recent meta-analysis by Alqahtani et al. (2023) reported a significant positive overall effect for AI integration (Hedges' $g = 0.86$), with chatbots and generative AI showing the largest impact on student learning outcomes (effect size = 1.02).

However, studies also reveal important limitations. Kortemeyer (2023) found that AI systems demonstrated high accuracy for structured physics problems but struggled with open-ended investigations and creative problem-solving approaches valued in inquiry-based physics education. Chen and Wan (2024) showed that GPT-3.5 could achieve human-level accuracy for conceptual physics questions through careful prompt engineering, but struggled with multi-step problems requiring complex mathematical reasoning. Crucially, these studies have been conducted predominantly in contexts where students are native English speakers; the performance of AI grading tools in linguistically diverse classrooms remains largely uncharted.

English Language Proficiency and Science Learning Outcomes

Students learning physics through English-medium instruction face dual cognitive–linguistic demands. Cummins' (2000) framework distinguishes Basic Interpersonal Communicative Skills (BICS), which typically develop within two years, from Cognitive Academic Language Proficiency (CALP), which requires five to seven years of sustained exposure. This timeline has direct implications for AI-mediated assessment, since AI-generated feedback is invariably produced in academic English calibrated against English-language criteria. Research on technology-mediated feedback for non-native English speakers reveals complex relationships between English proficiency, digital literacy, and engagement with

AI-mediated assessment. Yaseen et al. (2025) found that the impact of adaptive learning technologies and personalised feedback on student engagement is shaped by digital literacy, suggesting that English proficiency and digital literacy interact in shaping students' access to AI-generated feedback. Warschauer and Matuchniak (2010) further cautioned that educational technologies may exacerbate inequalities by favouring already-advantaged learners, a concern that motivates the present study's explicit attention to language as a moderating variable.

The Research Gap

Despite the growing volume of AI-in-education research, three critical gaps remain. First, physics-specific AI grading research is sparse, particularly studies evaluating scientific reasoning in inquiry-based contexts (Zawacki-Richter et al., 2019). Second, considerations of non-native English speakers are notably absent from AI grading research, despite the rapid global growth of English-medium instruction. Third, implementation research examining authentic classroom contexts, rather than controlled experimental tasks, remains limited (Cavalcanti et al., 2021). The present study addresses these converging gaps within the underexamined context of IB MYP Physics.

Methodology

Research Design

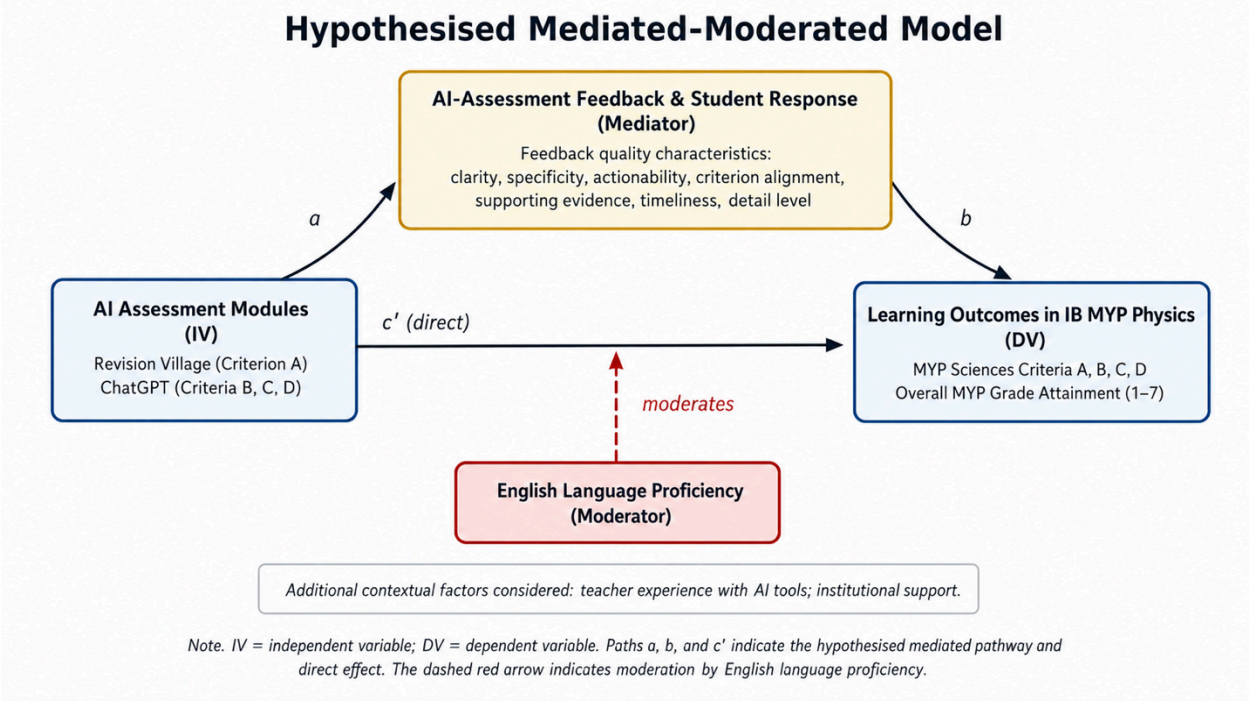
The study employed a sequential explanatory mixed-methods design embedded within a case study framework (Creswell & Plano Clark, 2018). Quantitative data were collected and analysed first, followed by qualitative data collection to explain and elaborate quantitative findings. Integration occurred through joint displays at the level of each research question. This

design was appropriate for examining both the extent and the mechanisms of AI implementation effects in an authentic classroom context.

Figure 1 presents the hypothesised conceptual model that guided the inferential analyses, depicting AI-assisted grading implementation as the independent variable, feedback quality characteristics as a hypothesised mediator (RQ2), student learning outcomes as the dependent variable (RQ1), and English language proficiency as a hypothesised moderator (RQ4).

Figure 1

Hypothesised Mediated-Moderated Model



Research Setting and Participants

The study was conducted at an IB World School in Shenzhen, China, serving a diverse international community. The MYP Sciences and Physics pathway encompassed Grade 9 MYP Sciences (52 students across three classes) and Grade 10 MYP Physics (49 students across three classes). Of 111 enrolled students, 101 (91%) were non-native English speakers and constituted

the study sample. Participants' first languages included Mandarin (*n* = 81, 80.2%), Cantonese (*n* = 11, 10.9%), Korean (*n* = 5, 5.0%), and other languages (*n* = 4, 4.0%). English language proficiency was classified as Low (*n* = 32, 31.7%), Mid (*n* = 43, 42.6%), or High (*n* = 26, 25.7%) based on course placement records and triangulated teacher evaluation. Three MYP Science teachers participated: Teacher A (male, 10 years IB experience, extensive prior AI experience), Teacher B (male, 4 years experience, moderate AI experience), and Teacher C (female, 6 years experience, minimal prior AI exposure). Table 1 summarises participant characteristics.

Table 1

Student Participant Characteristics by English Language Proficiency Group (N = 101)

Characteristic	Low (n = 32)	Mid (n = 43)	High (n = 26)	Full Sample (N = 101)
Grade 9 (MYP Sciences)	19	22	11	52 (51.5%)
Grade 10 (MYP Physics)	13	21	15	49 (48.5%)
First language: Mandarin	25	35	21	81 (80.2%)
First language: Cantonese	4	5	2	11 (10.9%)
First language: Korean	2	2	1	5 (5.0%)
First language: Other	1	1	2	4 (4.0%)

Note. English language proficiency was classified using course placement records and triangulated teacher evaluation. Percentages are computed against the full sample (*N* = 101).

AI Tools Under Investigation

Revision Village was deployed for Criterion A (Knowing and Understanding) assessments. This IB-calibrated platform provides structured, algorithm-driven feedback aligned with MYP Physics criterion descriptors, including adaptive questioning, progress tracking, and detailed solution explanations.

ChatGPT (OpenAI GPT-4) was deployed for Criteria B, C, and D, which require open-ended, conversational feedback suitable for evaluative and reflective tasks. Standardised prompt templates and grading instructions were developed to ensure consistent application across teachers and assessments.

Data Collection Instruments

MYP Sciences Criterion-Based Assessments. Pre- and post-intervention assessments measured student learning outcomes across the four MYP criteria (A–D, each scored 0–8), producing a total score out of 32 and an MYP grade of 1–7.

Feedback Quality Assessment Rubric. Two separate rubrics—one for Revision Village (Criterion A) and one for ChatGPT (Criteria B, C, and D)—assessed AI-generated and matched human teacher feedback across four dimensions: clarity (linguistic accessibility), specificity, actionability, and alignment with learning objectives. Each dimension was scored on a 4-point scale. Inter-rater reliability for rubric scoring was established at Cohen's $\kappa = 0.84$ across 25% of randomly selected feedback pairs.

Technology Acceptance Surveys. Separate surveys for teachers and students, grounded in the Technology Acceptance Model (Davis, 1989), measured perceived usefulness, perceived ease of use, learning effectiveness, and behavioural intention on 5-point Likert scales. Cronbach's alpha values ranged from .82 (Perceived Usefulness) to .88 (Behavioural Intention).

Qualitative Instruments. Semi-structured teacher interviews (45–60 minutes), student focus groups (60–75 minutes; 6–8 students per group), and 90-minute classroom observations were conducted using standardised protocols.

Procedure

The 24-week implementation was structured into three phases: quantitative data collection (Weeks 1–12), comprising pre-intervention assessments, surveys, the active intervention introducing AI tools, and post-intervention assessments; qualitative data collection (Weeks 13–20), comprising interviews, focus groups, and observations; and integration (Weeks 21–24). Feedback turnaround was tracked across 808 assignments (eight per student).

Data Analysis

Quantitative analysis used IBM SPSS Statistics (Version 28) with paired-samples *t*-tests, one-way ANOVA with Tukey HSD post hoc comparisons, Pearson correlations, and Hayes' PROCESS macro (Model 4 for mediation; Model 1 for moderation) with 5,000 bootstrap resamples. Effect sizes (Cohen's *d*) and partial eta-squared (η^2) were reported. Qualitative data were analysed using Braun and Clarke's (2021) six-phase thematic analysis, with inter-rater reliability for qualitative coding established at $\kappa = 0.81$.

Ethical Considerations and Trustworthiness

The study received ethics approval from the host institution, and parental consent and student assent were obtained for all participants. Pseudonymisation was applied throughout data handling and reporting. Methodological rigour was addressed through the criteria appropriate to each strand. For the quantitative strand, internal validity was supported through standardised assessment protocols and pre–post within-subjects design, reliability through internal-consistency checks (Cronbach's α = .82–.88), and inferential robustness through bootstrap-based confidence intervals. For the qualitative strand, trustworthiness was addressed using Lincoln and Guba's (1985) criteria of credibility (member checking with two teacher participants), transferability (thick description of context), dependability (audit trail of coding

decisions), and confirmability (independent inter-rater coding, $\kappa = 0.81$). Inferential quality at the mixed-methods level (Creswell & Plano Clark, 2018) was supported through joint-display integration of quantitative and qualitative strands for each research question.

Results

RQ1: AI Implementation and Student Learning Outcomes

Paired-samples *t*-tests revealed statistically significant improvements across all four MYP criteria following AI-assisted grading implementation (Table 2). The largest effect was observed in Criterion D, Reflecting on the Impacts of Science ($d = 0.74$), approaching a large effect size and suggesting that AI-assisted feedback particularly supported students in developing evaluative and reflective thinking. Criterion C, Processing and Evaluating ($d = 0.58$), and Criterion A, Knowing and Understanding ($d = 0.54$), demonstrated medium effects. Criterion B, Inquiring and Designing, showed the most modest but still statistically significant improvement ($d = 0.43$). The pre-intervention mean MYP grade of 4.58 improved to 5.27 post-intervention, with a total effect size of $d = 0.64$—a medium effect both statistically and practically meaningful within a single academic term.

Table 2

Pre- and Post-Intervention MYP Criterion Scores for Full Sample (N = 101)

Measure	Pre *M*	Pre *SD*	Post *M*	Post *SD*	Gain	*t*(100)	Cohen's *d*
Criterion A: Knowing & Understanding	4.62	1.58	5.35	1.45	0.73	−5.11***	0.54
Criterion B: Inquiring & Designing	4.45	1.55	5.00	1.48	0.55	−4.32***	0.43
Criterion C: Processing & Evaluating	4.28	1.52	5.04	1.44	0.76	−5.83***	0.58

Criterion D: Reflecting on Impacts	4.49	1.48	5.27	1.41	0.78	−7.12***	0.74
Total Score (0–32)	17.84	4.48	20.66	4.12	2.82	−6.43***	0.64
MYP Grade (1–7)	4.58	0.84	5.27	0.79	0.69	—	—

Note. Criteria were scored on a 0–8 scale. Total score range = 0–32; MYP grade range = 1–7. All *t* values reflect paired-samples *t*-tests (*df* = 100). Cohen's *d* was computed as the mean difference divided by the pooled *SD* of the difference scores. All tests two-tailed. *** $p < .001$.

Differential analysis by English language proficiency group revealed that Mid-proficiency students recorded the largest absolute gain (3.35 total score points), followed closely by Low-proficiency students (3.16 points). High-proficiency students recorded the smallest gain (1.50 points), consistent with a ceiling effect given their high pre-intervention mean (24.12 out of 32). These descriptive patterns are formally tested in the moderation analysis presented for RQ4 below.

RQ2: Feedback Quality as Mediator

Feedback Quality Comparison

Paired-samples *t*-tests compared AI-generated and human teacher feedback across the four rubric dimensions for Revision Village (40 matched pairs, Criterion A) and ChatGPT (120 matched pairs, Criteria B, C, and D). Table 3 presents the results.

Table 3

Feedback Quality Rubric Scores: AI Tools vs. Human Teacher Feedback

Rubric Dimension	AI *M*	Human *M*	Mean Diff.	*t*	*p*	Cohen's *d*
Revision Village (n = 40 pairs, Criterion A)						
Specificity	3.42	2.61	+0.81	7.77	< .001	1.23
Alignment with Objectives	3.56	2.47	+1.09	9.38	< .001	1.48

Clarity (Linguistic Accessibility)	2.65	3.38	−0.73	−6.89	< .001	−1.09
Actionability	3.14	3.06	+0.08	1.04	.305	0.16
ChatGPT (n = 120 pairs, Criteria B, C, D)						
Specificity	3.18	2.76	+0.42	7.43	< .001	0.68
Alignment with Objectives	3.08	2.72	+0.36	5.71	< .001	0.52
Clarity (Linguistic Accessibility)	2.71	3.24	−0.53	−10.07	< .001	−0.92
Actionability	3.09	3.01	+0.08	1.57	.119	0.14

Note. Scores on a 1–4 scale: 1 = Poor, 2 = Developing, 3 = Proficient, 4 = Exemplary. Positive *d* indicates AI advantage; negative *d* indicates human teacher advantage. Cohen's *d* reflects the mean difference in matched pairs divided by the *SD* of the difference scores. The Clarity dimension specifically measures linguistic accessibility for non-native English speakers.

Both AI tools scored significantly higher than human teacher feedback on specificity and alignment with learning objectives, and significantly lower on clarity. Effect sizes for Revision Village were substantially larger than for ChatGPT (specificity d = 1.23 vs. 0.68; alignment d = 1.48 vs. 0.52), consistent with Revision Village's explicit calibration to MYP Physics curriculum standards. No statistically significant differences were found for actionability for either tool. It is worth noting that the Clarity dimension specifically measures linguistic accessibility for the study's predominantly non-native English speaker sample; AI-generated feedback produced grammatically well-formed English but was less linguistically accessible to the target student population than teacher-mediated alternatives.

Feedback Timeliness

Pre-implementation teacher grading across 808 assignments averaged 9.9 days turnaround, with zero assignments returned within 48 hours. Post-implementation, Revision Village reduced mean turnaround to 0.3 days (100% within 48 hours), while ChatGPT reduced turnaround to 2.4 days (16.2% within 48 hours; 94.7% within one week). This dramatic compression of the feedback cycle has direct implications for the mediation analysis presented

next, as feedback timeliness contributes to the overall feedback quality composite operationalised in the rubric.

Mediation Analysis

Hayes' PROCESS Model 4 was used to test whether feedback quality characteristics mediated the relationship between AI implementation and learning outcomes. AI implementation was coded as a dichotomous variable (0 = pre-intervention, 1 = post-intervention), feedback quality characteristics were operationalised as a composite of the four rubric dimensions, and student learning outcomes were measured as total MYP score change. Bootstrap confidence intervals were computed using 5,000 resamples. Table 4 presents the results.

Table 4

Mediation Analysis: Feedback Quality Characteristics as Mediator (Hayes PROCESS Model 4, N = 101)

Path	*B*	*SE*	*t*	95% CI (Bootstrap)
Path *a*: AI → Feedback Characteristics	0.847	0.142	5.96***	[0.565, 1.129]
Path *b*: Feedback Characteristics → Outcomes	1.624	0.287	5.66***	[1.055, 2.193]
Direct Effect (*c′*): AI → Outcomes	1.474	0.458	3.22**	[0.564, 2.384]
Indirect Effect (*ab*): via Feedback Characteristics	1.376	—	—	[0.802, 2.047]*
Total Effect (*c*): AI → Outcomes	2.850	0.441	6.46***	[1.975, 3.725]
Proportion Mediated (*Pm*)	48.3%	—	—	—

Note. Bootstrap 95% CIs based on 5,000 resamples. R^2 for mediator model = .286; R^2 for outcome model = .412. *Indirect effect CI excludes zero, confirming significant mediation. ** $p < .01$. *** $p < .001$.

The mediation analysis confirmed statistically significant partial mediation. Path *a* was significant ($B = 0.847$, $p < .001$), confirming that AI implementation significantly improved

feedback quality characteristics. Path *b* was also significant ($B = 1.624$, $p < .001$), confirming that improved feedback characteristics were significantly associated with improved learning outcomes, controlling for AI implementation. The bootstrap 95% confidence interval for the indirect effect ($ab = 1.376$) excluded zero [0.802, 2.047], and the direct effect ($c' = 1.474$) remained significant ($p = .002$), indicating partial rather than full mediation. Feedback quality characteristics accounted for 48.3% of the total effect of AI implementation on learning outcomes.

RQ3: Implementation Challenges and Opportunities

Thematic analysis of teacher interviews, student focus groups, and classroom observations revealed three superordinate themes.

Opportunities. Teachers reported substantial reductions in grading time, with Teacher A estimating a 60% reduction. This time gain enabled redirection of professional energy towards one-on-one student support and instructional design. AI tools also provided more consistent application of MYP rubric standards across large numbers of students, reducing fatigue-driven criterion drift. Students described self-directed revision cycles enabled by immediate feedback: *"I can try again the same night. Before, I had to wait, and by then we were already on a new topic"* (Student G9-08, Mid proficiency).

Challenges. All three teachers described a verification burden, particularly during early implementation. Teacher C noted: *"At first, I checked every single comment line by line. It took almost as long as writing the feedback myself."* Linguistic and cultural blind spots in AI feedback required teacher mediation; Teacher B reported instances where ChatGPT used culturally Western examples or idiomatic expressions that confused Mandarin-speaking students.

Low-proficiency students consistently reported difficulty understanding the academic English in AI-generated feedback, requiring peer translation or dictionary use.

Institutional gaps. All three teachers identified the absence of formal professional development specifically focused on AI-assisted assessment as the most significant institutional constraint. Teacher C's self-reported trajectory from cautious verification in early weeks to efficient implementation by week six illustrated that the initial productivity deficit was not inevitable but reflected the absence of pre-implementation training—a modifiable institutional condition.

RQ4: English Language Proficiency as a Moderator

One-way ANOVA revealed a statistically significant effect of English proficiency group on learning gains, $F(2, 98) = 5.67$, $p = .005$, $\eta^2 = .104$, representing a medium effect. Tukey HSD post hoc comparisons showed that both Low- and Mid-proficiency groups demonstrated significantly greater gains than the High-proficiency group (Low vs. High, $p = .012$; Mid vs. High, $p = .003$); the difference between Low and Mid groups was not significant ($p = .891$). Hayes' PROCESS Model 1 confirmed a statistically significant interaction ($B = -0.782$, $SE = 0.298$, $t = -2.62$, $p = .010$, 95% CI [−1.373, −0.191]), accounting for an additional 4.8% of variance beyond main effects ($\Delta R^2 = .048$). The conditional effect of AI implementation was statistically significant at all three proficiency levels but was largest for Low-proficiency students ($B = 3.16$, $p < .001$), followed by Mid- ($B = 2.38$, $p < .001$), and smallest for High-proficiency students ($B = 1.59$, $p < .001$).

Technology Acceptance Survey results showed statistically significant differences across proficiency groups for all four constructs, with High-proficiency students consistently reporting

the highest acceptance and Low-proficiency students the lowest. Table 5 presents gain scores and TAM results by proficiency group.

Table 5

Learning Gains and Technology Acceptance by English Language Proficiency Group (N = 101)

Measure	Low (*n* = 32)	Mid (*n* = 43)	High (*n* = 26)	ANOVA *p*
Learning Gains				
Gain Total Score, *M* (*SD*)	3.16 (2.84)	3.35 (2.61)	1.50 (2.12)	.005**
Technology Acceptance (1–5 scale)				
Perceived Usefulness, *M* (*SD*)	3.52 (0.78)	3.81 (0.69)	4.08 (0.61)	.025*
Perceived Ease of Use, *M* (*SD*)	3.61 (0.82)	3.88 (0.71)	4.09 (0.58)	.007**
Learning Effectiveness, *M* (*SD*)	3.41 (0.84)	3.72 (0.74)	3.94 (0.62)	.022*
Behavioural Intention, *M* (*SD*)	3.68 (0.76)	3.95 (0.66)	4.17 (0.54)	.034*

Note. Gain Total Score = post-intervention minus pre-intervention MYP total score (range 0–32). Technology Acceptance scores on 1–5 Likert scale (1 = *Strongly Disagree*; 5 = *Strongly Agree*). Cronbach's α: Perceived Usefulness = .82; Perceived Ease of Use = .85; Learning Effectiveness = .83; Behavioural Intention = .88. * $p < .05$. ** $p < .01$.

Qualitative data illuminated the mechanism underlying this moderation. Teacher A captured the central asymmetry: *"My Low proficiency students benefited the most from AI feedback in absolute terms, because they had the most room to grow. But they also needed the most support from me to actually use the feedback. The Mid-level students were the sweet spot: they could read the AI feedback mostly on their own, and they had enough room to improve that the gains were visible."* This statement encapsulates the qualitatively grounded interpretation of the negative interaction term: a ceiling effect limited measurable growth for High-proficiency

students, while scaffolded teacher mediation enabled Low-proficiency students to access AI feedback benefits despite linguistic barriers.

Discussion

This study examined the implementation and impact of AI-assisted grading in IB MYP Physics, yielding four principal findings that extend theoretical understanding and inform educational practice. First, AI implementation produced statistically significant improvements across all four MYP criteria. Second, approximately half of this effect was attributable to enhanced feedback quality characteristics, with the remainder operating through additional pathways. Third, AI tools and human teachers demonstrated complementary rather than substitutable strengths. Fourth, contrary to one plausible prediction, lower-proficiency students gained more than higher-proficiency students.

AI-Assisted Grading Improves Learning Outcomes Through Feedback Quality

The finding that AI implementation produced statistically significant improvements across all four MYP criteria, with effect sizes from $d = 0.43$ to $d = 0.74$, aligns with the broader literature on automated feedback in STEM education (Alqahtani et al., 2023; Cavalcanti et al., 2021; Zawacki-Richter et al., 2019), although the effect sizes observed here are more modest than those reported in some quasi-experimental studies (e.g., Altal & Abo Ehsaiyan, 2025). The largest effect for Criterion D, Reflecting on the Impacts of Science, is theoretically coherent: this criterion requires higher-order evaluative thinking, and AI feedback operating at Hattie and Timperley's (2007) self-regulation level may be particularly effective for developing metacognitive awareness of assessment criteria.

The partial mediation finding (48.3%) confirms that feedback quality characteristics constitute a primary mechanism of AI implementation effects, consistent with Hattie and

Timperley's feedback model. The remaining direct effect, however, suggests additional pathways including the motivational benefits of iterative revision cycles described by students. This is consistent with Self-Determination Theory's (Deci & Ryan, 2017) proposition that competence experiences—such as observing mastery progression through multiple submission attempts—generate intrinsic motivation that extends beyond feedback content per se.

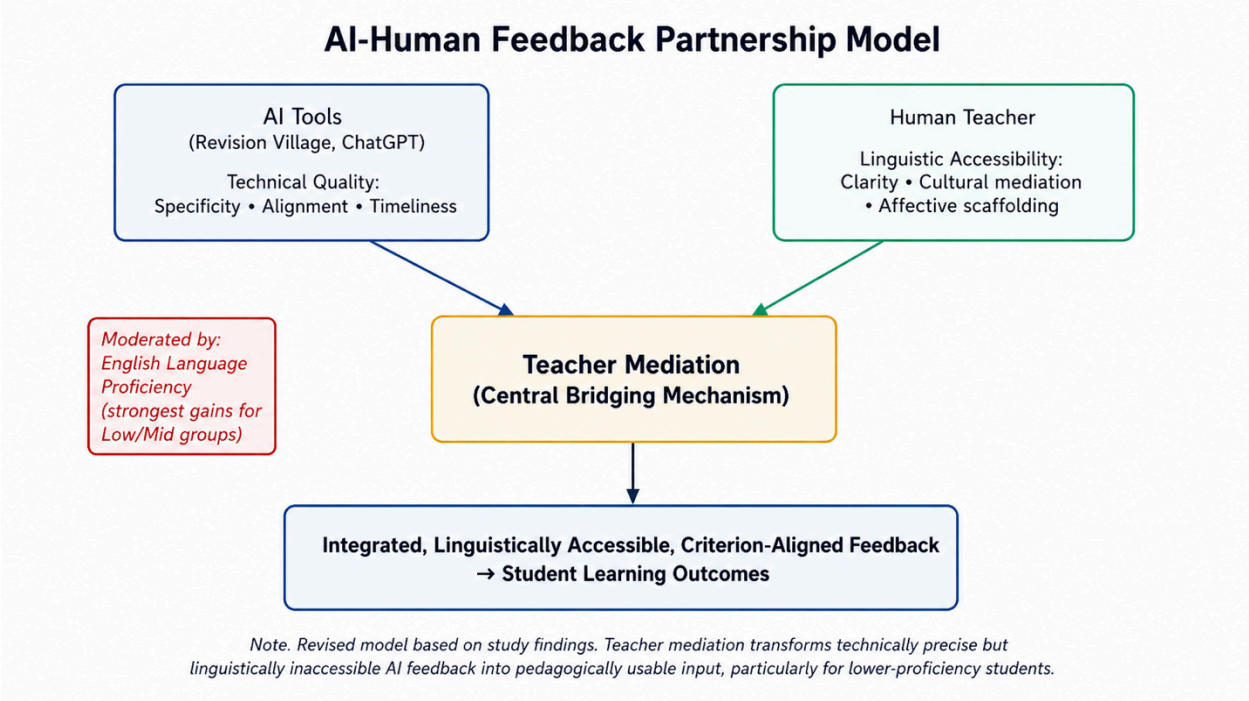
The AI–Human Feedback Partnership

The rubric comparison findings reveal complementary strengths rather than AI superiority or deficiency. AI tools excelled at specificity and criterion alignment—dimensions where consistency and scale matter most—while human teachers excelled at linguistic clarity, the dimension most consequential for non-native English speakers. Importantly, the Clarity dimension in this study specifically measured linguistic accessibility for students whose first language was not English. AI-generated feedback was syntactically well-formed in English and would likely be perceived as clear by fluent English speakers; the clarity gap therefore reflects a context-specific limitation rather than a general deficit. This complementary pattern suggests that the effective unit of AI-assisted grading is not the AI tool alone but the AI–human partnership, in which teacher mediation transforms technically precise but linguistically inaccessible feedback into pedagogically usable input.

This partnership extends Task–Technology Fit theory (Goodhue & Thompson, 1995) by specifying that fit is not solely a property of tool–task alignment but is co-constructed through teacher mediation. Figure 2 presents the revised AI–Human Feedback Partnership Model derived from this study's findings, in which teacher mediation functions as the central bridging mechanism connecting AI-generated feedback (strong in technical quality) and the linguistic accessibility required for non-native English speakers to translate feedback into learning gains.

Figure 2

Revised AI–Human Feedback Partnership Model



Note. The revised model decomposes feedback quality into technical quality (specificity, alignment, timeliness)—where AI tools consistently outperform human teachers—and linguistic accessibility (clarity, cultural mediation, affective scaffolding), where human teacher mediation provides indispensable complementary value, particularly for students with lower English language proficiency. Teacher mediation is positioned as the central bridging mechanism.

English Language Proficiency as a Contextual Moderator

The significant moderation effect of English language proficiency (p = .010) addresses a critical gap in the AI assessment literature. Contrary to concerns that AI tools may exacerbate educational inequality by favouring already-advantaged learners (Warschauer & Matuchniak, 2010), this study found that Low- and Mid-proficiency students demonstrated significantly larger gains than High-proficiency students. This finding is theoretically generative because it

implicates two simultaneous mechanisms: a ceiling effect that constrained measurable growth for High-proficiency students whose pre-intervention scores were already high, and a scaffolded-mediation mechanism through which lower-proficiency students gained substantial benefit when teacher mediation made AI feedback linguistically accessible. Crucially, these gains for lower-proficiency students were not frictionless; they required substantially more teacher scaffolding than gains for higher-proficiency students. The equity implication is that proportional investment in teacher mediation for lower-proficiency students is necessary for equitable outcomes, rather than a uniform tool deployment.

This finding also extends the Technology Acceptance Model (Davis, 1989) by demonstrating that language proficiency functions as a boundary condition for perceived usefulness and ease of use in English-medium instructional contexts. Technology acceptance research has rarely examined contexts where the technology delivers content in a language other than users' first language; the present study suggests that language proficiency moderates the core TAM relationships in such contexts.

Limitations

Several limitations warrant consideration. First, the single-site case study design limits the generalisability of findings to other IB World Schools with different demographic, linguistic, or infrastructural characteristics. Second, the absence of a control group receiving only human teacher feedback constrains causal inference; although the within-subjects pre–post design provides reasonable internal validity, maturation, curriculum-alignment effects, and Hawthorne effects cannot be entirely ruled out. Third, the teacher sample of three participants precluded formal quantitative testing of teacher-level variables, although triangulated qualitative evidence provides convergent support for the role of teacher experience and institutional factors. Fourth,

the 24-week implementation period does not address the durability of gains over multiple academic terms, leaving open important questions about long-term outcomes. Finally, the proficiency classification used in this study, while triangulated, did not employ a standardised English language proficiency instrument; future replications would benefit from formal CALP measurement to support stronger inferences regarding linguistic mechanisms.

Implications for Practice and Policy

For physics teachers, these findings support the adoption of differentiated feedback protocols calibrated to student proficiency levels, the integration of AI feedback into lesson cycles as structured revision windows, and the explicit cultivation of student AI literacy as part of assessment practice. For school administrators, the findings support investment in pre-implementation professional development for AI-assisted assessment—identified as the most significant institutional gap by all three teacher participants—and the establishment of living AI assessment policies subject to annual review. For IB curriculum developers and AI tool providers, IB-specific calibration of feedback tools, alongside the development of linguistically accessible feedback generation suitable for English-medium contexts with substantial multilingual populations, represents a high-leverage priority.

Conclusion

This mixed-methods study provides empirical evidence that AI-assisted grading, implemented as an AI–human partnership, can improve student learning outcomes in IB MYP Physics while reducing feedback turnaround time from days to hours. Feedback quality characteristics partially mediated this effect, with AI tools providing superior specificity and criterion alignment, and human teachers providing superior linguistic accessibility. English language proficiency significantly moderated implementation effectiveness, with

lower-proficiency students showing larger gains but requiring greater teacher mediation to access them. The proposed AI–Human Feedback Partnership Model positions teacher mediation as the central bridging mechanism between AI-generated feedback and learner outcomes, particularly in linguistically diverse classrooms.

Future research should pursue multi-site replication across diverse IB World Schools, longitudinal designs examining the durability of gains, and experimental comparisons of AI-only, human-only, and integrated feedback conditions. The partnership model proposed here generates testable hypotheses for future investigation and offers a principled framework for responsible AI integration in linguistically diverse STEM classrooms. As AI assessment tools continue to evolve and their use in international school contexts continues to grow, the findings of this study suggest that the most consequential design choice is not which AI tool to deploy but how to embed teacher mediation within its deployment.

References

Alqahtani, T., Badreldin, H. A., Alrashed, M., Alshaya, A. I., Alghamdi, S. S., Bin Saleh, K., & Albekairy, A. M. (2023). The emergent role of artificial intelligence, natural learning processing, and large language models in higher education and research. *Research in Social and Administrative Pharmacy, 19*(8), 1236–1242. https://doi.org/10.1016/j.sapharm.2023.05.016

Altal, A., & Abo Ehsaiyan, H. (2025). Evaluating AI-assisted instructional systems in secondary physics: A quasi-experimental study from the United Arab Emirates. *Inquisiva Open*.

Braun, V., & Clarke, V. (2021). *Thematic analysis: A practical guide*. SAGE.

Brookhart, S. M. (2013). *How to create and use rubrics for formative assessment and grading*. ASCD.

Carless, D. (2006). Differing perceptions in the feedback process. *Studies in Higher Education, 31*(2), 219–233. https://doi.org/10.1080/03075070600572132

Cavalcanti, A. P., Barbosa, A., Costa, F. A., Tucci, C., & Gaona, M. (2021). Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence, 2*, 100038. https://doi.org/10.1016/j.caeai.2021.100038

Chen, Z., & Wan, T. (2024). Achieving human-level partial credit grading of written responses to physics conceptual questions using GPT-3.5 with only prompt engineering. *arXiv*. https://arxiv.org/abs/2407.15251

Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education, 20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Cummins, J. (2000). *Language, power, and pedagogy: Bilingual children in the crossfire*. Multilingual Matters.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319–340. https://doi.org/10.2307/249008

Deci, E. L., & Ryan, R. M. (2017). *Self-determination theory: Basic psychological needs in motivation, development, and wellness*. Guilford.

Dwivedi, Y. K., Rana, N. P., Jeyaraj, A., Clement, M., & Williams, M. D. (2019). Re-examining the unified theory of acceptance and use of technology (UTAUT): Towards a revised theoretical model. *Information Systems Frontiers, 21*, 719–734. https://doi.org/10.1007/s10796-017-9774-y

Fathali, S., & Okada, T. (2018). Technology acceptance model in technology-enhanced OCLL contexts: A self-determination theory approach. *Australasian Journal of Educational Technology, 34*(4). https://doi.org/10.14742/ajet.3629

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213–236. https://doi.org/10.2307/249689

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

International Baccalaureate Organization. (2023). *MYP sciences guide*. International Baccalaureate Organization.

Kortemeyer, G. (2023). Toward AI grading of student problem solutions in introductory physics: A feasibility study. *Physical Review Physics Education Research, 19*(2), 020163. https://doi.org/10.1103/PhysRevPhysEducRes.19.020163

Kortemeyer, G., & Nöhl, J. (2024). Assessing confidence in AI-assisted grading of physics exams through psychometrics: An exploratory study. *Physical Review Physics Education Research, 21*(1), 010136.

Latif, E., Lee, G. G., Neumann, K., Kastorff, T., & Zhai, X. (2024). G-SciEdBERT: A contextualized LLM for science assessment tasks in German. *arXiv*. https://arxiv.org/abs/2402.06584

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. SAGE.

Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Mandouit, L., & Hattie, J. (2023). Revisiting "The power of feedback" from the perspective of the learner. *Learning and Instruction, 84*, 101718. https://doi.org/10.1016/j.learninstruc.2022.101718

Molin, F., Haelermans, C., Cabus, S., & Groot, W. (2021). Do feedback strategies improve students' learning gain? Results of a randomized experiment using polling technology in physics classrooms. *Computers & Education, 175*, 104339. https://doi.org/10.1016/j.compedu.2021.104339

Venkatesh, V., Thong, J. Y. L., & Xu, X. (2012). Consumer acceptance and use of information technology: Extending the unified theory of acceptance and use of technology. *MIS Quarterly, 36*(1), 157–178. https://doi.org/10.2307/41410412

Wang, S., Wang, F., Zhu, Z., Wang, J., Tran, T., & Du, Z. (2024). Artificial intelligence in education: A systematic literature review. *Expert Systems with Applications, 252*, 124167. https://doi.org/10.1016/j.eswa.2024.124167

Warschauer, M., & Matuchniak, T. (2010). New technology and digital worlds: Analyzing evidence of equity in access, use, and outcomes. *Review of Research in Education, 34*, 179–225. https://doi.org/10.3102/0091732X09349791

Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology, 10*, 487662. https://doi.org/10.3389/fpsyg.2019.03087

Wongvorachan, T., Lai, K. W., Bulut, O., Tsai, Y. S., & Chen, G. (2022). Artificial intelligence: Transforming the future of feedback in education. *Journal of Applied Testing Technology, 23*, 95–116.

Yaseen, H., Mohammad, A. S., Ashal, N., Abusaimeh, H., Ali, A., & Sharabati, A. A. (2025). The impact of adaptive learning technologies, personalized feedback, and interactive AI tools on student engagement: The moderating role of digital literacy. *Sustainability, 17*(3), 1133. https://doi.org/10.3390/su17031133

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education—Where are the educators? *International Journal of Educational Technology in Higher Education, 16*, 39. https://doi.org/10.1186/s41239-019-0171-0