

From Question Design to Descriptive Feedback: Operationalising the IAT Model for AI-Assisted Assessment Automation in Teacher Practice

Hossein talebzadeh

talebzadehdr@gmail.com

Farhangian University <https://orcid.org/0009-0009-4312-1244>

Research Article

Keywords: assessment automation, generative AI, formative feedback, teacher agency, pedagogical stimuli, IAT model

Posted Date: August 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10567053/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

The integration of generative artificial intelligence into educational assessment presents both opportunities and challenges for teacher practice. While AI offers potential for automating time-consuming evaluation tasks, there remains a notable gap in operational frameworks that guide teachers through the complete assessment cycle—from question design to personalised feedback—while maintaining pedagogical integrity and human agency. This study addresses this gap by examining a practical instantiation of the Integrated AI Triad (IAT) model within a professional development programme involving 138 teachers across four cohorts. Analysing a dataset of 70 constructed-response questions, over 330 AI-simulated student responses, AI-driven error diagnoses, and AI-generated feedback messages, the study employed qualitative content analysis guided by formative assessment theory, feedback models, cognitive load theory, and self-regulated learning frameworks. Findings reveal three key patterns: (a) teachers predominantly designed scenario-based questions targeting higher-order thinking (analysis, application, evaluation = 80%), with science as the most frequent subject (45%); (b) six recurrent conceptual error types were identified—causal confusion, oversimplification, foundational misconceptions, failure to recognise causal relationships, narrowing of abstract concepts, and confusion of physical and chemical changes; and (c) AI-generated feedback consistently followed a three-part structure—positive opening, guiding question, encouraging closure—aligning with theoretical principles of effective formative feedback. Synthesising these findings, the study proposes a five-step operational framework for assessment automation and introduces the concept of pedagogical stimuli—AI-simulated responses that serve as catalysts for teacher reflection and error diagnosis. This framework preserves teacher agency while leveraging AI for efficiency, contributing both theoretical insights and practical guidance for AI-assisted assessment in K-12 education.

1. Introduction

The transformation of educational assessment stands as one of the most pressing challenges confronting contemporary education systems. Assessment is not merely a technical procedure for assigning grades; it is a fundamental process of reasoning from evidence about student learning (Pellegrino, 2018). When designed effectively, assessment serves both summative and formative purposes, providing information that can guide instructional decisions and support student growth (Black & Wiliam, 1998). However, the practical realities of classroom assessment—particularly the generation of meaningful feedback for each student—remain persistently challenging for educators worldwide.

Central to effective assessment is the quality of feedback provided to learners. Hattie and Timperley (2007) demonstrated that feedback is among the most powerful influences on student achievement, yet its impact depends critically on its nature and timing. Shute (2008) further emphasised that formative feedback, when delivered in a timely and specific manner, can significantly enhance learning outcomes by helping students identify gaps in their understanding and take corrective action. Despite this evidence, teachers face considerable constraints in providing personalised, descriptive feedback to every student. The cognitive and temporal demands of designing quality assessment tasks, evaluating student responses, diagnosing conceptual errors, and crafting individualised feedback often overwhelm teachers' capacity, particularly in large classes or when managing multiple subjects and grade levels.

The emergence of generative artificial intelligence (AI) offers a potentially transformative solution to these persistent challenges. Luckin and Holmes (2016) argued that AI in education should not be viewed as a replacement for teachers but as a tool to augment human expertise, enabling more personalised, flexible, and inclusive learning experiences. Similarly, Cukurova (2024) proposed that the future of AI in education lies in hybrid human-AI intelligence systems, where AI extends human cognition rather than replacing it. However, Selwyn (2019) cautioned that the integration of any technology into education must be presented as a choice, requiring careful consideration of values, judgments, and the social and emotional dimensions of teaching that AI cannot replicate. This tension between technological potential and human agency lies at the heart of contemporary efforts to operationalise AI in educational settings.

The application of AI to assessment, in particular, has gained increasing attention. Scholars have begun to explore how generative AI can support the design of reflective assignments (Samuel, Soh, & Jung, 2025), the development of assessment rubrics (Fernández-Sánchez, Lorenzo-Castiñeiras, & Sánchez-Bello, 2025), and the automation of certain evaluation processes (Topping et al., 2025). Yet, as Corbin et al. (2025) recently argued, the research landscape on assessment after AI remains fragmented. There is little shared sense of what kinds of studies are most needed or how their findings might connect. Specifically, there is a notable gap in operational frameworks that guide teachers through the complete assessment cycle—from question design to the generation of personalised feedback—using AI as a collaborative partner while maintaining human agency and pedagogical integrity (Corbin et al., 2025; Molenaar, 2022). The challenge is not merely technical but pedagogical: how can AI be operationalised to support teachers in their existing workflows without diminishing their professional judgment or distancing them from their students?

This study addresses this gap by examining a practical instantiation of such an operational framework. The Integrated AI Triade (IAT) model (Talebzadeh, 2025a) provides a systemic framework for redefining the educational ecosystem by positioning AI as a mediator between three core pillars: teachers, managers, and families. Within this model, the teacher transitions from a content deliverer to a "Bilingual Learning Designer" (AI-LDP)—one fluent in both human pedagogy and AI technology—and the school manager becomes a "Data-Driven Strategic Leader." The IAT framework is grounded in a series of implemented training modules for both teachers and school leaders (Talebzadeh, 2025a), collectively demonstrating how the concept of the "bilingual professional" can ensure human agency remains central in an increasingly automated world. The present study operationalises the IAT model specifically within the domain of assessment, demonstrating how its principles can be translated into a coherent, replicable workflow for teachers.

The dataset analysed in this study originates from the third session of the bilingual teacher training programme, titled "Assessment Automation" (Talebzadeh, 2025b). This session introduced a five-step workflow for AI-assisted assessment: (1) content generation and (2) objective/qualitative test design, followed by (3) generating hypothetical student responses, (4) diagnosing conceptual errors, and (5) producing constructive feedback. The resulting dataset is organised into four interconnected sections that directly mirror this workflow: Section 1: Smart Question Design (encompassing the creation of base content and the design of constructed-response questions); Section 2: AI-Generated Hypothetical Student Responses (comprising over 330 responses at three performance levels—excellent, average, and weak, with the latter containing conceptual errors); Section 3: AI-Driven Diagnosis of Common Conceptual Errors (identifying patterns in student misconceptions); and Section 4: Feedback and Error Analysis (generating constructive feedback messages, each structured with a positive opening, a guiding question, and an encouraging closure). These AI-simulated responses function as pedagogical stimuli—a methodological innovation that provides teachers with a controlled, analytical lens to anticipate and diagnose learner difficulties before engaging directly with real student work. This dataset offers a unique window into how teachers can operationalise AI for assessment automation within the IAT framework.

Building on the IAT framework and the empirical material generated through the bilingual teacher training, the present study pursues three research questions:

1. What patterns can be identified in the constructed-response questions designed by teachers within the IAT framework?
2. What are the most common conceptual errors in AI-simulated student responses, and how can they be categorised?
3. What structure do AI-generated feedback messages follow, and how can they help teachers provide effective feedback?

By addressing these questions, this study aims to propose an operational framework for the full assessment cycle using AI—from question design to automated personalised feedback—within the IAT model. The article demonstrates how teachers can use AI to automate time-consuming parts of assessment, thereby freeing time for human interaction and

individual guidance. In doing so, it contributes to the growing body of research on AI-assisted pedagogy (Dogan, 2025; Chiu, 2024) and extends the IAT model into the domain of assessment automation, offering both theoretical and practical insights for educators, researchers, and policymakers concerned with the future of assessment in an AI-mediated educational landscape.

The article proceeds as follows. The next section reviews the theoretical foundations underpinning the study, drawing on formative assessment theory, feedback models, cognitive load theory, self-regulated learning, and the AI-TPACK framework. The methodology section then describes the design-based research approach, the dataset, and the analytical procedures employed. Subsequently, the findings present patterns in question design, conceptual errors, and feedback structures. The discussion interprets these findings in light of the theoretical framework and considers their implications for teacher practice and professional development. Finally, the conclusion summarises the contributions, acknowledges limitations, and outlines directions for future research.

2. Theoretical Framework

2.1. Formative Assessment and the Feedback Cycle

The theoretical foundation of this study rests on well-established principles of formative assessment and effective feedback. Black and Wiliam (1998) demonstrated that formative assessment—the process of gathering evidence about student learning to inform subsequent instruction—can significantly improve learning outcomes when implemented effectively. Their seminal review established that assessment *for* learning, as opposed to assessment *of* learning, shifts the focus from measuring achievement to supporting growth. This distinction is central to understanding how AI-assisted assessment automation might enhance, rather than replace, teachers' formative practices.

Building on this foundation, Pellegrino (2018) conceptualised assessment as a process of reasoning from evidence, comprising three interconnected components: (a) theories and research about content-based cognition that identify the knowledge and skills of interest; (b) tasks and observations that can provide evidence of student mastery; and (c) interpretive techniques that capture important variation in knowledge and skill. This framework underscores that effective assessment design is not merely a technical exercise but a form of conceptual research requiring deliberate alignment between learning goals, assessment tasks, and interpretation. The IAT model's emphasis on 'Smart Question Design' directly operationalises this principle by positioning teachers as designers who construct tasks intentionally calibrated to elicit evidence of specific learning outcomes.

Central to formative assessment is the quality of feedback provided to learners. Hattie and Timperley (2007) proposed a model distinguishing three levels of feedback: task-level feedback (how well a task is performed), process-level feedback (the strategies used to perform the task), and self-regulation-level feedback (students' capacity to monitor and regulate their own learning). Their meta-analysis revealed that feedback is among the most powerful influences on achievement, but its impact depends critically on its timing, specificity, and alignment with learners' needs. Shute (2008) further elaborated that formative feedback should be timely, specific, and focused on the task rather than the learner, avoiding praise that might undermine self-efficacy. These principles provide a benchmark against which AI-generated feedback can be evaluated.

Nicol and Macfarlane-Dick (2006) explicitly connected formative assessment to self-regulated learning, proposing seven principles of good feedback practice that support student self-regulation. These include clarifying goals and criteria, facilitating self-assessment, delivering high-quality information about learning, encouraging dialogue, and promoting positive motivational beliefs. This integration is particularly relevant to the present study because AI-generated feedback, when structured appropriately, can support self-regulation by providing students with actionable information about their performance while encouraging them to take ownership of their learning. The present dataset's feedback messages—

structured with a positive opening, a guiding question, and an encouraging closure—reflect an attempt to operationalise these principles within an AI-mediated context.

2.2. Self-Regulated Learning and Assessment

The concept of self-regulated learning (SRL) provides a complementary lens through which to understand the role of feedback in student development. Zimmerman (1990) defined SRL as the processes whereby learners personally activate and sustain cognitions, affects, and behaviours oriented towards the attainment of personal goals. This definition emphasises the proactive role of learners in setting goals, monitoring their effectiveness, and adapting their functioning—processes that are fundamentally supported by high-quality feedback. Zimmerman and Schunk (2011) further elaborated that self-regulation includes self-initiated forms of social learning, such as seeking help from peers, coaches, and teachers, challenging the notion that self-regulation is purely individualistic.

Panadero (2017) reviewed six major models of SRL, identifying common phases (forethought, performance, and self-reflection) while also highlighting differential effects of SRL models based on students' developmental stages. This review underscores that SRL is not a monolithic construct but a family of related processes that interact with context, task demands, and individual differences. The relevance of SRL to the present study lies in the observation that feedback—whether from teachers, peers, or AI systems—mediates the relationship between task performance and self-regulatory development. When feedback provides specific, actionable information about how to improve, it supports the self-reflection phase of the SRL cycle, enabling students to adjust their strategies and set more effective goals.

More recently, Chiu (2024) explored how generative AI tools such as ChatGPT can foster SRL by satisfying students' needs for autonomy, competence, and relatedness, drawing on self-determination theory. His classification tool identified 20 learning activities using ChatGPT that support the three phases of SRL—forethought, performance, and self-reflection—while also addressing motivational dimensions. This work suggests that AI, when designed with pedagogical intentionality, can scaffold self-regulatory processes rather than undermining them. The present study extends this line of inquiry by examining how AI-generated feedback messages might similarly support self-regulation through their structure and content.

2.3. Cognitive Load Theory and Pedagogical Stimuli

Cognitive load theory (CLT) offers a third theoretical lens, particularly relevant to understanding both the design of assessment tasks and the generation of feedback. While Sweller's original formulation (Sweller, 1988) is foundational, more recent work has elaborated its implications for instructional design. Mayer (2024) reviewed the development of the cognitive theory of multimedia learning, emphasising three core principles: dual channels (separate verbal and visual processing), limited capacity (processing capacity is severely constrained), and active processing (meaningful learning requires selecting, organising, and integrating information). These principles have direct implications for assessment design: tasks that impose excessive extraneous cognitive load may fail to measure the intended constructs, while tasks that appropriately challenge working memory can promote deeper learning.

Kriegelstein et al. (2025) recently demonstrated that anchoring effects can bias cognitive load assessments, suggesting that students' first impressions of task complexity influence their subsequent judgments. This finding has implications for feedback design: if students misperceive their cognitive load, feedback that addresses this misperception may be more effective than feedback that assumes accurate self-assessment.

Crucially, the concept of pedagogical stimuli—introduced in this study as a methodological innovation—directly engages with CLT to reduce teachers' cognitive load during error analysis. By providing AI-simulated student responses at varying levels of quality (excellent, average, and weak), the workflow enables teachers to anticipate and diagnose conceptual errors without the cognitive burden of collecting, reading, and categorising large volumes of authentic student work. This pre-diagnostic phase reduces extraneous cognitive load by offering a structured, representative sample of potential

misconceptions, allowing teachers to focus their cognitive resources on designing targeted interventions rather than on the initial detection of patterns. In this way, AI-simulated responses function as cognitive scaffolds that streamline the analytical process, freeing working memory for higher-order pedagogical reasoning. The present study's use of AI-simulated responses as pedagogical stimuli thus operationalises CLT principles by distributing cognitive effort between the AI system (pattern detection and categorisation) and the teacher (interpretation and instructional decision-making).

2.4. AI-Assisted Assessment and Automation

The integration of AI into assessment represents a significant shift in how evaluation can be conducted. Luckin and Holmes (2016) argued that AI in education should augment human expertise rather than replace it, enabling more personalised, flexible, and inclusive learning experiences. Cukurova (2024) extended this argument, proposing three conceptualisations of AI: the externalisation of human cognition (AI as a tool that performs cognitive functions previously done by humans), the internalisation of AI models (AI that influences human mental models), and the extension of human cognition via tightly coupled human-AI hybrid systems. The present study aligns most closely with the third conceptualisation, viewing AI as a collaborative partner that extends teachers' capacity to design, evaluate, and respond to student work.

Molenaar (2022) introduced a detect-diagnose-act framework to describe core functions of AI in education, along with a six-level model of automation that specifies the transition of control between AI, learners, and teachers. This framework is particularly relevant to operationalising the IAT model, as it provides a language for discussing the appropriate distribution of tasks between human and machine intelligence. In the context of assessment, lower levels of automation might involve AI generating draft feedback for teacher review, while higher levels might involve fully automated feedback generation with minimal human oversight. The present study examines a middle ground where AI generates feedback drafts that teachers can adapt and personalise, preserving human agency while leveraging AI's capacity for rapid analysis.

Samuel, Soh, and Jung (2025) demonstrated that ChatGPT can enhance reflective assignment design while reducing faculty preparation time, suggesting that AI can support both quality and efficiency in assessment. Similarly, Fernández-Sánchez, Lorenzo-Castiñeiras, and Sánchez-Bello (2025) found that AI tools improved rubric precision, efficiency, and alignment with curricular objectives. These studies indicate that AI's role in assessment extends beyond automation to encompass design and quality assurance. Topping et al. (2025) reviewed research on AI-enhanced peer assessment, finding that AI generally improved peer assessment quality while also highlighting under-researched areas such as automated assignment, calibration, and teamwork effectiveness.

However, Selwyn (2019) cautioned against uncritical adoption of AI in education, emphasising that technological integration must be presented as a choice and that the social, emotional, and relational dimensions of teaching cannot be replicated by machines. Corbin et al. (2025) similarly argued that the research landscape on assessment after AI is fragmented, and they called for a collective research agenda to guide future inquiry. Their principles—why, who, what, how, and where we assess—provide a structure for thinking systematically about assessment in the age of AI. The present study responds to this call by examining a concrete instantiation of AI-assisted assessment within a structured professional development framework.

2.5. The IAT Model and AI-TPACK Framework

The Integrated AI Triad (IAT) model (Talebzadeh, 2025a) provides the overarching framework for this study. Positioned within the broader IAT model, the third session of the bilingual teacher training programme operationalises the concept of the 'Bilingual Professional' by introducing a five-step workflow for assessment automation. The IAT model's emphasis on human agency aligns with the work of scholars who caution against the replacement of teachers by technology

(Selwyn, 2019) while also embracing the augmentation perspective advocated by Luckin and Holmes (2016) and Cukurova (2024).

The AI-TPACK framework (Koehler & Mishra, 2009; An et al., 2026; Mishra & Bozkurt, 2026) further informs this study's theoretical orientation. TPACK posits that effective teaching with technology requires an integration of technological, pedagogical, and content knowledge. An et al. (2026) extended this model to distinguish between human teachers and AI machine teachers, separating TPK and TPCK into distinct dimensions. Mishra and Bozkurt (2026) argued that in the era of generative AI, deep content knowledge becomes more vital than ever to verify and curate the output of algorithmic tools, which they characterise as 'smart drunk interns'—highly capable but fundamentally unreliable. This perspective underscores the importance of teachers maintaining pedagogical oversight even as they leverage AI for efficiency.

The IAT model operationalises these theoretical insights by providing concrete training modules that build teachers' capacity to design smart tasks, create personalised content, automate assessment, and develop analytical rubrics. The present study examines the third of these modules—assessment automation—in detail, analysing the patterns and structures that emerge when teachers apply AI within their professional practice. By doing so, it contributes to the growing body of research that seeks to understand how AI can be integrated into teaching in ways that enhance rather than diminish human expertise.

3. Methodology

3.1. Research Context and Participants

This study is situated within a broader professional development programme titled *Creative Teaching with AI*, conducted by a teacher education institution in Iran. The programme was designed as a design-based research (DBR) intervention (McKenney & Reeves, 2018), iteratively developed and refined through multiple cycles of implementation, feedback, and revision. The DBR framing allowed the research team to simultaneously advance theoretical understanding of AI-integrated assessment while generating practical solutions for teacher professional development.

The programme was structured as a series of four weekly workshops delivered across four cohorts, comprising approximately 138 participating teachers ($N = 138$). Based on the workshop records, the participants were distributed as follows: 64 novice teachers—recent graduates of teacher training universities with no prior classroom teaching experience at the time of the workshops—and 74 experienced teachers with teaching experience ranging from 2 to 32 years ($M = 14.2$, $SD = 8.1$), organised into four cohorts designated as N1, N2, E1, and E2. Each cohort completed four weekly sessions, resulting in a total of 16 workshop sessions across the entire programme. The workshops were conducted online and included a combination of synchronous instruction, guided practice, and independent project work.

The present study focuses specifically on Session 3: Assessment Automation, which introduced a five-step workflow for AI-assisted assessment: (1) content generation, (2) objective and qualitative test design, (3) generating hypothetical student responses, (4) diagnosing conceptual errors, and (5) producing constructive feedback. The selection of this session as the focus of analysis was deliberate: it represents the most comprehensive instantiation of the assessment cycle within the Integrated AI Triad (IAT) model (Talebzadeh, 2025a), encompassing the full arc from question design to feedback generation.

All participating teachers were provided with a bilingual (Persian-English) guide containing five ready-to-use prompt templates corresponding to each step of the workflow. Teachers were then asked to apply these templates to construct their own assessment materials, selecting a topic of their choice from their respective subject areas and grade levels. The resulting dataset—representing the practical output of this professional development activity—serves as the primary data source for the present study.

3.2. Dataset Design and Structure

The dataset analysed in this study is a curated collection of assessment artifacts generated by the participating teachers during the third workshop session. It is organised into four interconnected sections, each corresponding to a specific phase of the five-step assessment workflow introduced in the session:

Section 1: Smart Question Design. This section contains approximately 70 constructed-response questions designed by teachers across a range of subject areas and grade levels. Subject areas include science (physical sciences, life sciences, earth sciences), mathematics, Persian language and literature, social studies, and religious education. Grade levels span from primary (Grades 1–6) to secondary (Grades 7–12). Each question is accompanied by metadata indicating the subject, grade level, and, in some cases, the specific learning objectives it was designed to assess.

Section 2: AI-Generated Hypothetical Student Responses. For each teacher-designed question, the AI generated three levels of hypothetical student responses: one excellent response, one or several average responses, and several weak responses containing conceptual errors. In total, this section comprises over 330 responses across all questions. These AI-simulated responses function as pedagogical stimuli—a methodological innovation introduced in this study—providing teachers with a controlled, analytical lens to anticipate and diagnose learner difficulties before engaging directly with real student work.

Section 3: AI-Driven Diagnosis of Common Conceptual Errors. For each set of responses, the AI identified the most common conceptual mistakes and classified the responses into strong, average, and weak categories. This section includes both the classification output and the AI's explanatory reasoning for each diagnosis.

Section 4: Feedback and Error Analysis. For each weak response, the AI generated a constructive feedback message structured in three parts: (a) a positive opening that acknowledges what the student did well; (b) a guiding question designed to help the student discover their own error rather than receiving the correct answer directly; and (c) an encouraging closure that motivates further effort. This section also implicitly embeds evaluative criteria such as scientific accuracy, coherence, logical reasoning, and appropriate use of technical vocabulary.

The dataset thus represents a complete operationalisation of the assessment cycle, from initial question design through to personalised feedback, within a structured professional development context.

3.3. Analytical Approach

This study employed a qualitative content analysis approach with a directed (deductive-inductive) orientation. The analysis was guided by the theoretical framework established in the preceding sections, including formative assessment theory, feedback models, cognitive load theory, and self-regulated learning, while remaining open to emergent patterns not anticipated by existing theory (Braun & Clarke, 2020).

The analysis proceeded in three phases, corresponding to the three research questions:

Phase 1: Analysis of Question Design Patterns. The first phase focused on the 70 constructed-response questions in Section 1 of the dataset. An initial open coding process identified emergent patterns in question design, including subject area distribution, grade level variation, question structure, and cognitive demand. Each question was then classified according to Krathwohl's (2002) revised Bloom's taxonomy to determine the predominant cognitive level being assessed. Additionally, the analysis examined question features such as the use of scenarios, authentic contexts, scaffolding, and interdisciplinary connections. Two researchers independently coded a subset of 20 questions to establish inter-coder reliability ($\kappa = 0.84$), with discrepancies resolved through discussion.

Phase 2: Analysis of Conceptual Errors. The second phase examined the AI-diagnosed conceptual errors in Section 3 of the dataset. A thematic analysis approach was used to identify recurring error types across different subject areas and grade levels. The coding process was informed by the conceptual error categories identified by the AI in the dataset, while also allowing for the emergence of new categories not explicitly labelled by the AI. The resulting taxonomy of conceptual errors was then refined through iterative review of the data, with particular attention to the relationship between error type and the cognitive demands of the corresponding question.

Phase 3: Analysis of Feedback Structure. The third phase analysed the AI-generated feedback messages in Section 4 of the dataset. The analysis focused on the structural and linguistic features of the feedback messages, including the presence and nature of the three-part structure (positive opening, guiding question, encouraging closure), the types of questions used to guide students, and the implicit evaluative criteria embedded in the feedback. The analysis also examined the alignment between the feedback messages and the principles of effective feedback identified in the theoretical framework (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006).

To support the analytical process, ChatGPT was used as an assistive tool for initial coding and pattern identification, following the protocol described by Naeem, Smith, and Thomas (2025). The use of AI in this context was limited to assisting with data organisation and initial thematic identification; all final coding decisions and interpretations were made by the researchers. This approach aligns with the methodological guidance provided by Braun and Clarke (2020) for maintaining researcher reflexivity and interpretive control while leveraging AI for efficiency.

3.4. Trustworthiness and Rigour

Several strategies were employed to ensure the trustworthiness and rigour of the analysis (Lincoln & Guba, 1985). Triangulation was achieved through the use of multiple data sources (the four sections of the dataset), multiple analytical phases, and the involvement of two researchers in the coding process. Member checking was conducted by sharing preliminary findings with a subset of the participating teachers, who confirmed that the identified patterns resonated with their professional experiences. Audit trails were maintained throughout the analytical process, documenting all coding decisions, interpretive judgments, and revisions to ensure transparency and reproducibility. Researcher reflexivity was practised through ongoing critical reflection on the researchers' own assumptions and biases, particularly regarding the role of AI in assessment. Finally, the use of a directed content analysis approach with explicit theoretical grounding ensured that the analysis remained anchored in established educational theory while remaining open to emergent findings.

3.5. Ethical Considerations

This study was conducted in accordance with standard ethical guidelines for educational research. All data were anonymised to protect the identities of participating teachers and their students. The AI-generated responses used in the dataset are hypothetical and do not represent actual student work; they were generated solely for pedagogical purposes within the professional development context. The use of AI tools in the analytical process was transparent and consistent with the researchers' institution policies on AI-assisted research. No personally identifiable information was collected or analysed during the study.

4. Findings

This section presents the findings of the analysis, organised according to the three research questions: patterns in question design, common conceptual errors, and the structure of AI-generated feedback.

4.1. Patterns in Teacher-Designed Constructed-Response Questions

Analysis of the 70 constructed-response questions revealed several distinctive patterns in how teachers operationalised the IAT framework for question design.

Subject and Grade Distribution. The questions spanned a wide range of subjects, with science (physical sciences, life sciences, earth sciences) comprising the largest category (approximately 45%), followed by Persian language and literature (20%), mathematics (15%), social studies (12%), and religious education (8%). Grade levels ranged from Grade 1 to Grade 12, with a notable concentration in Grades 4–6 (approximately 50% of all questions).

Cognitive Demand Patterns. Classification using Krathwohl's (2002) revised Bloom's taxonomy revealed that the majority of questions targeted higher-order thinking skills. Approximately 35% of questions required **analysis** (e.g., explaining causal mechanisms, comparing and contrasting), 25% required **application** (e.g., applying concepts to novel scenarios), and 20% required **evaluation** (e.g., justifying decisions, critiquing perspectives). Only 10% focused on **remembering**, and 10% on **understanding**. This distribution suggests that teachers prioritised questions that elicit deeper reasoning rather than simple recall.

Question Structure and Scaffolding. Three predominant structural patterns emerged:

1. **Scenario-Based Questions with Guided Sub-Questions.** Many questions embedded the target concept within a realistic scenario and provided explicit sub-questions to scaffold student thinking. For example, a Grade 4 science question presented a family choosing to wear warm clothes instead of using an electric heater, followed by four directed prompts: identifying the energy type, its source, environmental impacts, and the role of individual behaviour in conservation. This structure reduces cognitive load by breaking a complex task into manageable components while maintaining conceptual coherence.
2. **Creative and Imaginative Framing.** Several questions used imaginative scenarios to engage younger learners. A Grade 1 science question asked students to imagine a "magic magnet" and select items for a school bag, explaining why each would be attracted. Such framing promotes engagement while still requiring scientific reasoning about magnetic properties. The accompanying learning objectives explicitly noted the dual goals of "strengthening imagination and creativity alongside scientific learning."
3. **Analytical and Interdisciplinary Prompts.** For older grades, questions often required integration of multiple concepts or perspectives. A Grade 12 Persian literature question asked students to analyse the betrayal of Shaghad in the story of Rostam from both ethical and social perspectives, using textual evidence and a contemporary example. This design demands synthesis of literary analysis, ethical reasoning, and real-world application.

Subject-Specific Characteristics. Science questions frequently required explanation of causal mechanisms (e.g., "Why does a ball stop moving after being pushed?"), often accompanied by requests for examples or illustrations. Persian language questions emphasised interpretative and analytical skills, particularly in higher grades. Mathematics questions, though fewer, included both procedural and conceptual items, with some requiring step-by-step explanation of reasoning.

Design Quality and Limitations. While most questions demonstrated thoughtful design, some were overly broad or lacked sufficient scaffolding. For instance, questions such as "What is a polymer?" or "What is the difference between homogeneous and heterogeneous mixtures?" without additional context or guiding prompts may elicit shallow responses that fail to reveal deep understanding. Conversely, the most effective questions provided clear expectations, contextual grounding, and multiple entry points for students at different levels.

Table 1
Summary of Patterns in Teacher-Designed Constructed-Response Questions

Pattern Dimension	Key Characteristics	Representative Example	Frequency / Distribution
Subject Distribution	Dominance of science (45%); followed by Persian language (20%), mathematics (15%), social studies (12%), religious education (8%)	—	Science: 45% Persian: 20% Mathematics: 15% Social Studies: 12% Religious Education: 8%
Grade Level Concentration	Strongest representation in Grades 4–6 (50%), reflecting transition to higher-order thinking	—	Grades 4–6: 50% Grades 1–3: 20% Grades 7–9: 18% Grades 10–12: 12%
Cognitive Demand (Bloom's Taxonomy)	Predominantly higher-order: analysis (35%), application (25%), evaluation (20%); lower-order: remembering (10%), understanding (10%)	—	Analysis: 35% Application: 25% Evaluation: 20% Understanding: 10% Remembering: 10%
Scenario-Based with Guided Sub-Questions	Real-world context + explicit scaffolded prompts (2–4 sub-questions); reduces cognitive load	<i>Energy conservation scenario:</i> family wearing warm clothes instead of electric heater, with 4 guided prompts	Most frequent (40% of all questions)
Creative / Imaginative Framing	Fantasy-based scenarios; engages younger learners; promotes creativity alongside scientific reasoning	<i>Grade 1 Science:</i> "Imagine you have a magic magnet... what items would you choose for your school bag and why?"	Common in primary grades (1–3)
Analytical / Interdisciplinary Prompts	Integration of multiple concepts; requires synthesis, textual evidence, and real-world application	<i>Grade 12 Persian:</i> Analyse betrayal of Shaghad in the story of Rostam from ethical and social perspectives, with textual evidence and contemporary example	Concentrated in secondary grades (10–12)
Subject-Specific Characteristics	Science: causal mechanisms with examples Persian: interpretative and analytical skills Mathematics: procedural + conceptual reasoning	—	Science: mechanism-focused Persian: interpretation-focused Mathematics: reasoning-focused
Design Limitations	Some questions overly broad (e.g., "What is a polymer?") lacking scaffolding; may elicit shallow responses	<i>Overly broad:</i> "What is a polymer?" <i>Scaffolded:</i> "Why is the book called the 'kind friend'? ... refer to its roles in knowledge, culture, history, and great figures"	15% of questions lacked adequate scaffolding

4.2. Common Conceptual Errors in Simulated Student Responses

Analysis of the AI-generated weak responses, combined with the AI-driven diagnoses in Section 3 of the dataset, revealed recurring patterns of conceptual misunderstanding across subject areas. These errors were systematically categorised into six types.

Error Type 1: Causal Confusion (Reversing Cause and Effect). This was particularly common in science questions involving dynamic processes. For example, in response to the question "Why do earthquakes occur?" a weak response stated: "Earthquake comes because the earth shakes. When the earth shakes, it cracks." The AI diagnosis noted that the student had reversed cause and effect—treating the shaking as the cause rather than the consequence of tectonic forces. Similarly, in a question about the water cycle, a weak response focused on temperature and darkness rather than explaining how solar radiation drives evaporation and precipitation.

Error Type 2: Over-Simplification of Complex Processes. Many weak responses reduced multi-step processes to single-stage explanations. In response to the question "How do underground waters form?" a weak response stated simply: "Underground waters are made from rain and are in the ground." The diagnosis highlighted the absence of any reference to infiltration, percolation through soil and rock layers, or aquifer formation—key stages that distinguish a scientific from a surface-level understanding.

Error Type 3: Misconceptions About Foundational Concepts. Several responses revealed fundamental misunderstandings of core scientific ideas. In a question about electrical conductors and insulators, a weak response claimed: "A conductor is something that produces electricity, like a battery." The diagnosis identified confusion between a conductor (which allows current to flow) and a source of electricity. Similarly, in a question about mass measurement, a weak response suggested using a ruler to measure length and width, conflating mass with volume or size.

Error Type 4: Failure to Recognise Causal Relationships. In questions requiring explanation of why phenomena occur, weak responses often described what happens without explaining why. For example, in response to "Why does a ball stop moving?" a weak response stated: "The ball is light, so it cannot move much and stops by itself." The diagnosis noted the absence of any reference to friction or opposing forces—the core scientific explanation. In another instance, a weak response to a question about energy conservation claimed that wearing warm clothes "produces energy" in the home, revealing confusion between generating energy and reducing consumption.

Error Type 5: Narrowing or Misinterpreting Abstract Concepts. In humanities and social science questions, weak responses often reduced multidimensional concepts to single dimensions. In response to "Why is the book called the 'kind friend'?" a weak response stated: "Because it plays with us and entertains us." The diagnosis noted the failure to recognise the book's educational, cultural, and informational roles. Similarly, in response to "What is identity?" a weak response equated identity solely with nationality, ignoring individual, familial, and social dimensions.

Error Type 6: Confusing Physical and Chemical Changes. This was a recurrent error in chemistry-related questions. Weak responses often labelled phase changes (e.g., melting ice, boiling water) as chemical changes because the appearance of the substance changed, while describing burning wood as a physical change because "only its colour turned black." The diagnosis highlighted the fundamental misunderstanding: physical changes alter state or form without producing new substances, whereas chemical changes produce new substances with different properties.

Cross-Subject Patterns. The analysis revealed that error types were not uniformly distributed across subjects. Causal confusion and over-simplification were most frequent in science questions, particularly in topics involving dynamic systems (earthquakes, water cycles, forces). Misconceptions about foundational concepts appeared across all subjects but were most pronounced in topics requiring abstract reasoning (mass vs. weight, conductors vs. sources). In humanities subjects, narrowing of abstract concepts was the dominant pattern.

Relationship Between Error Type and Cognitive Demand. Errors were more frequent and more severe in questions requiring higher-order thinking (analysis, evaluation). For remembering-level questions, errors were typically omissions or inaccuracies rather than conceptual misunderstandings. This suggests that while teachers designed questions targeting higher-order thinking, some students (as simulated) lacked the prerequisite conceptual foundations to engage effectively with such demands.

Table 2
Taxonomy of Common Conceptual Errors in Simulated Student Responses

Error Type	Description	Representative Example (Weak Response)	AI Diagnosis	Most Affected Subject Areas	Frequency
1. Causal Confusion	Reversing cause and effect; treating consequence as cause	"Earthquake comes because the earth shakes. When the earth shakes, it cracks."	Student treated shaking as cause rather than consequence of tectonic forces; failed to explain why earth shakes	Science (earthquakes, water cycle)	Very High
2. Over-Simplification	Reducing multi-step processes to single-stage explanations	"Underground waters are made from rain and are in the ground."	Absence of key stages: infiltration, percolation through soil/rock layers, aquifer formation	Science (geology, water cycle)	High
3. Foundational Misconceptions	Fundamental misunderstanding of core scientific concepts	"A conductor is something that produces electricity, like a battery."	Confusion between conductor (allows current flow) and source of electricity; conflating mass with volume	Science (physics, measurement)	Medium-High
4. Failure to Recognise Causal Relationships	Describing <i>what</i> happens without explaining <i>why</i>	"The ball is light, so it cannot move much and stops by itself."	No reference to friction or opposing forces; confused energy generation with energy reduction	Science (forces, energy)	High
5. Narrowing Abstract Concepts	Reducing multidimensional concepts to single dimensions	"Why is the book called the 'kind friend'? Because it plays with us and entertains us."	Failure to recognise educational, cultural, informational roles of book; equating identity solely with nationality	Humanities (literature, social studies)	Medium
6. Confusing Physical vs. Chemical Changes	Labelling phase changes as chemical; describing burning as physical	"Melting ice and boiling water are chemical changes... burning wood is physical because only its colour turned black."	Fundamental misunderstanding: physical changes alter state without new substances; chemical changes produce new substances	Science (chemistry)	Medium-High
Cross-Subject Pattern	—	—	Causal confusion + over-simplification frequent in dynamic systems; foundational misconceptions in abstract reasoning;	—	Varies

Error Type	Description	Representative Example (Weak Response)	AI Diagnosis	Most Affected Subject Areas	Frequency
			narrowing of concepts in humanities		

4.3. Structure of AI-Generated Feedback Messages

Analysis of the feedback messages in Section 4 of the dataset revealed a consistent three-part structure that aligns with principles of effective feedback identified in the theoretical framework.

Part 1: Positive Opening. Every feedback message began with affirmation that acknowledged what the student had done well or expressed appreciation for their effort. Examples included:

"Well done that you were able to imagine the earth cracking like a wall; this shows you pay good attention and use examples from around you."

"It's great that you tried to answer and gave an example—this shows you're thinking and trying to connect concepts to real life."

This opening serves to maintain student motivation and self-efficacy, consistent with Shute's (2008) recommendation that feedback should be non-threatening and focused on the task rather than the learner.

Part 2: Guiding Question. The core of each feedback message was a question designed to prompt the student to reconsider their response and discover the error independently, rather than being told the correct answer directly. The guiding questions varied in form but consistently aimed to stimulate reflection:

- **Probing causal reasoning:** "What do you think causes the earth to start shaking in the first place? If there is no force from inside the earth, does the earth shake on its own?"
- **Challenging assumptions:** "When you mix oil and water, do they mix completely and become uniform, or do they separate after a while and form two layers?"
- **Directing attention to missing steps:** "When it rains, do all the drops turn into underground water immediately, or might they have a longer path to reach the depths of the earth?"
- **Prompting re-examination of definitions:** "If heat caused attraction, could hot objects like a cup of tea or a heater attract things from a distance like a magnet?"

Part 3: Encouraging Closure. Each message concluded with a statement that reinforced the student's capacity to improve and expressed confidence in their ability to discover the correct understanding. Examples included:

"Try to think again and see what exactly a magnet works with. I'm waiting to read your new discoveries!"

"I'm sure this time you will discover the correct answer yourself."

"You can discover this hidden layer because you have a curious mind."

Structural Consistency Across Subjects. The three-part structure was remarkably consistent across subjects and grade levels. Whether the feedback addressed a science misconception, a literary interpretation error, or a mathematical calculation mistake, the same underlying pattern was maintained. This consistency suggests that the AI was prompted with a structured template that enforced a specific pedagogical approach.

Implicit Evaluative Criteria. While not explicitly stated, the feedback messages implicitly embedded evaluative criteria that aligned with the quality dimensions identified in the theoretical framework. The guiding questions directed students toward:

- **Scientific accuracy:** Questioning whether explanations are complete and based on established principles.
- **Coherence:** Prompting students to check whether their reasoning is logically connected.
- **Use of evidence:** Encouraging students to support claims with examples or observations.
- **Depth of understanding:** Challenging surface-level explanations and probing for underlying mechanisms.

Alignment with Theoretical Principles. The observed feedback structure aligns closely with Hattie and Timperley's (2007) model, particularly the focus on process-level and self-regulation-level feedback. By guiding students to discover errors themselves, the feedback promotes self-monitoring and metacognitive awareness. The positive opening and encouraging closure address affective dimensions, supporting motivation and self-efficacy. The use of questions rather than direct answers is consistent with Nicol and Macfarlane-Dick's (2006) principle that feedback should encourage dialogue and promote self-assessment.

Variations in Question Quality. While the overall structure was consistent, the quality and specificity of the guiding questions varied. Some questions were highly targeted, directing attention to specific omissions or contradictions in the student's response. Others were more general, potentially less effective in prompting focused reflection. For example, a general question like "What do you think?" is less likely to generate productive revision than a specific question that highlights a precise point of confusion.

Summary. The AI-generated feedback messages operationalised a structured, theory-informed approach to formative feedback. The three-part pattern—positive opening, guiding question, encouraging closure—provides a replicable template that teachers can adapt for their own practice. The focus on student discovery rather than correction aligns with research on self-regulated learning and feedback effectiveness.

Table 3
Structure of AI-Generated Feedback Messages

Structural Component	Purpose	Key Characteristics	Representative Example	Alignment with Theory
1. Positive Opening	Acknowledge student effort / correct elements; maintain motivation and self-efficacy	Affirmative, non-threatening, task-focused; specific recognition of what the student did well	"Well done that you were able to imagine the earth cracking like a wall; this shows you pay good attention and use examples from around you."	Shute (2008): Feedback should be non-threatening, task-focused; supports self-efficacy
2. Guiding Question	Prompt student to reconsider response; facilitate independent error discovery	Open-ended, reflective, Socratic; directs attention to omissions or contradictions; avoids providing direct answers	<i>Probing causal reasoning:</i> "What do you think causes the earth to start shaking in the first place?" <i>Challenging assumptions:</i> "When you mix oil and water, do they mix completely or separate?" <i>Directing to missing steps:</i> "When it rains, do all drops turn into underground water immediately?"	Hattie & Timperley (2007): Process-level and self-regulation-level feedback; promotes self-monitoring Nicol & Macfarlane-Dick (2006): Feedback should encourage dialogue and promote self-assessment
3. Encouraging Closure	Reinforce capacity to improve; express confidence in student's ability	Positive, forward-looking, motivational; emphasises growth mindset	"Try to think again and see what exactly a magnet works with. I'm waiting to read your new discoveries!" "I'm sure this time you will discover the correct answer yourself."	Shute (2008): Feedback should support motivation and self-efficacy Nicol & Macfarlane-Dick (2006): Promotes positive motivational beliefs
Implicit Evaluative Criteria	Indirectly guide students toward quality standards	Embodied in the framing of guiding questions; not explicitly stated	<i>Scientific accuracy:</i> "Is your explanation complete and based on established principles?" <i>Coherence:</i> "Is your reasoning logically connected?" <i>Use of evidence:</i> "Can you support your claim with examples?"	Pellegrino (2018): Assessment reasoning from evidence Black & Wiliam (1998): Formative assessment as process of gathering and using evidence
Structural Consistency	Uniform pattern across subjects and grade levels	Reliable replication of three-part template across all feedback messages	Applied consistently across science, mathematics, Persian language, social studies, and religious education	Molenaar (2022): Levels of automation in hybrid human-AI systems

Summary of Key Findings

1. **Question Design Patterns.** Teachers predominantly designed scenario-based questions with guided sub-questions (40%), targeting higher-order thinking (analysis, application, evaluation = 80%). Science was the most frequent subject (45%), with Grades 4–6 heavily represented (50%). Approximately 15% of questions lacked adequate scaffolding.
2. **Conceptual Error Taxonomy.** Six error types were identified, with causal confusion and over-simplification being most frequent (very high to high frequency). Foundational misconceptions and confusion of physical vs. chemical changes were medium-high frequency. Errors were more severe in higher-order thinking questions, suggesting gaps in prerequisite conceptual foundations.
3. **Feedback Structure.** AI-generated feedback consistently followed a three-part structure: positive opening → guiding question → encouraging closure. This structure aligns with theoretical principles of effective feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Shute, 2008) and implicitly embeds evaluative criteria such as scientific accuracy, coherence, and use of evidence.

5. Discussion and Conclusion

The findings of this study offer a detailed empirical account of how teachers can operationalise AI-assisted assessment automation within the Integrated AI Triad (IAT) framework. In this section, I move beyond summarising the patterns identified in the dataset to interpret their broader significance, introduce new conceptual distinctions, and articulate the theoretical and practical contributions of the study. I then consider the limitations of the research and outline directions for future inquiry.

5.1. The Assessment Automation Cycle: A Five-Step Operational Framework

The dataset analysed in this study represents a complete instantiation of the assessment cycle—from question design through to personalised feedback—within a structured professional development context. Synthesising the three strands of findings, I propose a five-step operational framework that captures the workflow enacted by teachers and the AI system:

1. **Content Generation:** Teachers (with AI assistance) create foundational instructional texts or identify existing content as the basis for assessment.
2. **Smart Question Design:** Teachers construct constructed-response questions that target higher-order thinking, often using scenario-based framing and guided sub-questions to scaffold student responses.
3. **Generating Hypothetical Student Responses:** The AI system produces simulated responses at multiple performance levels (excellent, average, weak), functioning as **pedagogical stimuli** that allow teachers to anticipate and diagnose potential learner difficulties before engaging with authentic student work.
4. **Diagnosing Conceptual Errors:** The AI system analyses the simulated responses, identifies patterns of error, and classifies responses according to quality and error type—providing teachers with a pre-diagnostic lens that reduces the cognitive load of error detection.
5. **Producing Constructive Feedback:** The AI system generates draft feedback messages structured as positive opening → guiding question → encouraging closure, which teachers can review, personalise, and deliver to students.

This framework differs from existing models of AI-assisted assessment in several respects. Unlike automated essay scoring systems that focus primarily on grade prediction, the IAT workflow preserves teacher agency by positioning AI as a partner in a collaborative process rather than a replacement for human judgment (Cukurova, 2024; Molenaar, 2022). Unlike generic AI tutoring systems that provide direct answers, the feedback generation step is explicitly designed to promote student discovery through Socratic questioning—a pedagogical choice consistent with formative assessment principles (Nicol & Macfarlane-Dick, 2006).

The framework also addresses a critical gap identified by Corbin et al. (2025): the need for a shared sense of what kinds of studies are most needed in assessment after AI. By providing a concrete, replicable workflow grounded in empirical material, this study offers a reference point for researchers and practitioners seeking to implement AI-assisted assessment in ways that maintain human oversight and pedagogical integrity.

5.2. Pedagogical Stimuli: A Methodological Innovation

A central contribution of this study is the introduction and operationalisation of the concept of pedagogical stimuli—AI-simulated student responses that serve as catalysts for teacher reflection, error diagnosis, and feedback design. While the use of simulated data in educational research is not new, the specific framing of AI-generated responses as *stimuli* for pedagogical reasoning represents a departure from typical applications.

In conventional assessment practice, teachers encounter student errors organically—often after they have already been committed, in the context of grading or classroom interaction. This reactive stance limits the opportunity for proactive instructional design: teachers cannot easily anticipate the full range of misconceptions that may emerge, nor can they prepare targeted interventions in advance. The pedagogical stimuli approach reverses this logic by providing teachers with a controlled, analytical lens to explore potential difficulties *before* they appear in authentic student work.

The cognitive load implications of this shift are significant. Krieglstein et al. (2025) demonstrated that anchoring effects can bias cognitive load assessments, suggesting that teachers' initial impressions of student difficulty may influence their subsequent judgments. By exposing teachers to a representative sample of potential errors in a low-stakes, simulated environment, the pedagogical stimuli workflow reduces extraneous cognitive load (Sweller, 1988) associated with on-the-spot error detection, allowing teachers to focus their cognitive resources on interpretation and instructional decision-making rather than on initial pattern identification.

From a self-regulated learning perspective (Zimmerman, 1990), the pedagogical stimuli approach also models the kind of reflective practice that teachers are expected to cultivate in their students. By engaging in systematic error analysis and feedback design, teachers externalise the metacognitive processes that underlie effective formative assessment—a practice that may, over time, become internalised as part of their professional repertoire (Panadero, 2017).

5.3. Redefining Teacher Agency in AI-Assisted Assessment

The findings of this study complicate simplistic narratives about AI replacing or diminishing teacher agency (Selwyn, 2019). Rather than reducing the need for teacher expertise, the IAT workflow appears to **re-allocate** teacher effort: automating the more time-consuming and repetitive aspects of assessment (content generation, multiple-choice test design, draft feedback production) while amplifying the need for human judgment in question design, error interpretation, and feedback personalisation.

This observation aligns with Priestley, Biesta, and Robinson's (2026) ecological model of teacher agency, which conceptualises agency as emerging from the interplay of individual capacity, relational dynamics, and structural conditions. In the context of the IAT workflow, the AI system functions as a structural enabler—providing teachers with tools and data that extend their analytical reach—while the teacher's professional judgment remains the decisive factor in interpreting and acting upon the AI's outputs. The positive opening → guiding question → encouraging closure structure of the AI-generated feedback, for instance, is a pedagogical choice that reflects a specific theory of formative feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006) rather than a neutral technical output.

This distribution of labour suggests that teacher agency in AI-assisted assessment is not a fixed property but a dynamic achievement that depends on the quality of the human-AI partnership. When AI systems are designed to supplement rather than supplant teacher judgment, and when teachers are provided with the training and support to engage critically with AI outputs, agency can be enhanced rather than eroded. Cukurova's (2025) UNESCO position paper on protecting

teacher agency in the age of AI is particularly relevant here: the findings of this study provide empirical grounding for the claim that well-designed professional development can equip teachers to maintain pedagogical oversight even as they leverage AI for efficiency.

5.4. Implications for Practice and Professional Development

The practical implications of this study extend to three key audiences: teachers, teacher educators, and school leaders.

For **teachers**, the five-step IAT workflow offers a concrete, replicable template for integrating AI into assessment practice. The prompt templates provided in the professional development materials (Talebzadeh, 2025b) lower the barrier to entry, allowing teachers to experiment with AI-assisted assessment without requiring advanced technical skills. The consistent three-part structure of the AI-generated feedback also provides a pedagogical scaffold that teachers can adapt to their own contexts, maintaining the core principles of formative feedback while leveraging AI for efficiency.

For **teacher educators**, the pedagogical stimuli approach suggests a new direction for pre-service and in-service training. By engaging teachers in the systematic analysis of AI-simulated student responses, professional development programmes can cultivate diagnostic reasoning skills in a low-stakes, reflective environment—building capacity for error detection and feedback design that transfers to authentic classroom practice. The AI-TPACK framework (An et al., 2026; Mishra & Bozkurt, 2026) provides a useful lens for conceptualising this training: teachers must develop not only technical knowledge of AI tools but also pedagogical knowledge of how to interpret and respond to AI-generated insights.

For **school leaders**, the IAT model offers a systemic framework for AI integration that extends beyond individual classrooms to encompass administrative and family engagement. The findings of this study, focused on the teacher pillar of the IAT triad, suggest that leadership support for AI-assisted assessment should include sustained professional development, access to reliable AI tools, and the establishment of ethical guidelines for AI use—including protocols for reviewing and adapting AI-generated feedback before it reaches students (Corbin et al., 2025; Gallant, Davis, & Khan, 2026).

5.5. Limitations

Several limitations of this study warrant consideration. First, the student responses analysed in the dataset are AI-simulated rather than authentic. While this design choice was intentional—pedagogical stimuli are meant to serve as analytical tools rather than replacements for real student work—it limits the generalisability of the findings to authentic classroom contexts. Future research should examine how teachers apply the IAT workflow to authentic student responses and how the patterns of error identified in simulated data compare to those observed in real student work.

Second, the study is situated within a specific professional development context in Iran, with a particular cohort of 138 teachers. While the dataset is rich and diverse, the findings may not generalise to other educational contexts, cultural settings, or teacher populations. Replication studies in different contexts would strengthen the external validity of the proposed framework.

Third, the analysis relies on AI-generated diagnoses and classifications in Sections 3 and 4 of the dataset. While the use of AI as an assistive tool in qualitative analysis is consistent with emerging methodological practice (Naeem, Smith, & Thomas, 2025), the interpretations are ultimately those of the researchers. The possibility of AI-generated artefacts influencing the pattern identification process cannot be entirely ruled out.

Fourth, the study focuses exclusively on constructed-response questions, leaving other assessment formats (e.g., performance tasks, project-based assessments, portfolios) largely unexplored. The applicability of the IAT workflow to these formats remains an open question.

5.6. Future Research Directions

The findings of this study open several avenues for future inquiry. First, longitudinal research is needed to examine the impact of AI-assisted assessment workflows on student learning outcomes. Does the use of pedagogical stimuli and AI-generated feedback lead to measurable improvements in student understanding and self-regulation? Second, comparative studies could examine the relative effectiveness of AI-generated feedback versus teacher-generated feedback, or the conditions under which hybrid (human-AI) feedback is most beneficial. Third, research is needed on the scalability of the IAT workflow—how can it be adapted for different subjects, grade levels, and institutional contexts? Fourth, the ethical implications of AI-generated feedback warrant further investigation, particularly regarding issues of privacy, bias, and the potential for over-reliance on automated systems (Ritzhaupt, Dawson, & Crompton, 2026). Finally, future research should explore the integration of the IAT model across all three pillars—teachers, managers, and families—examining how systemic AI integration can create coherent, supportive learning ecosystems.

5.7. Conclusion

This study set out to examine the operationalisation of AI-assisted assessment within the Integrated AI Triad (IAT) model, focusing on the patterns in teacher-designed questions, common conceptual errors in simulated student responses, and the structure of AI-generated feedback. The findings reveal that teachers can design constructed-response questions that target higher-order thinking, particularly when supported by scenario-based framing and guided sub-questions. The AI-simulated student responses, functioning as pedagogical stimuli, expose a taxonomy of six recurrent error types—causal confusion, over-simplification, foundational misconceptions, failure to recognise causal relationships, narrowing of abstract concepts, and confusion of physical and chemical changes. The AI-generated feedback consistently follows a three-part structure—positive opening, guiding question, encouraging closure—that aligns with theoretical principles of effective formative feedback.

Synthesising these findings, the study proposes a five-step operational framework for assessment automation that preserves teacher agency while leveraging AI for efficiency. This framework addresses a critical gap in the literature on assessment after AI, offering a concrete, replicable workflow that can inform practice, professional development, and future research. The introduction and operationalisation of pedagogical stimuli as a methodological innovation extends the conceptual toolkit of AI-assisted assessment, providing teachers with a proactive lens for anticipating and diagnosing learner difficulties.

Ultimately, this study contributes to the growing body of research on AI-assisted pedagogy by demonstrating how teachers can operationalise the IAT model in authentic professional development contexts. The findings suggest that AI, when thoughtfully integrated into the assessment cycle, can augment rather than diminish teacher expertise—freeing time for human interaction and individual guidance while maintaining the pedagogical integrity that lies at the heart of effective teaching. The challenge moving forward is not whether AI should be used in assessment, but how it can be used to support the human capacities—judgment, empathy, and creativity—that remain essential to education.

Declarations

Ethics Approval Statement: This study is an independent, practitioner-led research initiative based on a professional development course designed and delivered solely by the author. Since the research involved standard educational practices and posed no more than minimal risk to participants, formal review by a named institutional ethics committee was not applicable. However, all procedures were conducted in full compliance with the ethical principles outlined in the Declaration of Helsinki, with particular attention to participant confidentiality, anonymity, and voluntary participation.

Participant Consent Statement: Written informed consent was obtained from all individual participants (teachers, school administrators, and educational staff) prior to their enrolment in the study. Each participant was clearly informed about the research objectives, data collection methods, the scope of the intervention, and their unconditional right to withdraw at any time without negative consequences.

References

1. An X, Bai J, Li Y, Zhou Y, Chai CS, Hong J-C (2026) Development of an Artificial Intelligence Technology Pedagogical Content Knowledge (AI-TPACK) Scale for Human and AI Concurrently as Teachers. *Asia-Pacific Educ Researcher* 35(3):483–493
2. Black P, Wiliam D (1998) Assessment and classroom learning. *Assess Education: Principles Policy Pract* 5(1):7–74
3. Braun V, Clarke V (2020) One size fits all? What counts as quality practice in (reflexive) thematic analysis. *Qualitative Res Psychol* 18(3):328–352
4. Chiu TKF (2024) A classification tool to foster self-regulated learning with generative artificial intelligence by applying self-determination theory: A case of ChatGPT. *Education Tech Research Dev* 73(1):1–16
5. Corbin T, Bearman M, Boud D, Crawford N, Dawson P, Fawns T, Walton J (2025) Assessment after Artificial Intelligence: The research we should be doing. *J Univ Teach Learn Pract* 22(7):1–12
6. Cukurova M (2024) The interplay of learning, analytics and artificial intelligence in education: A vision for hybrid intelligence. *Br J Edu Technol* 55(6):1–18
7. Cukurova M (2025) Promoting and protecting teacher agency in the age of artificial intelligence. UNESCO Publishing
8. Dogan S (2025) A human-AI hybrid co-design model: Planning effective instruction with ChatGPT. *Asian J Distance Educ* 20(1):1–18
9. Fernández-Sánchez A, Lorenzo-Castiñeiras JJ, Sánchez-Bello A (2025) Navigating the future of pedagogy: The integration of AI tools in developing educational assessment rubrics. *Eur J Educ*, 60(1), e12826
10. Gallant TB, Davis M, Khan ZR (2026) Academic Integrity in the Age of AI. *Elements in Generative AI in Education*. Cambridge University Press
11. Hattie J, Timperley H (2007) The power of feedback. *Rev Educ Res* 77(1):81–112
12. Koehler MJ, Mishra P (2009) What is technological pedagogical content knowledge (TPACK)? *Contemp Issues Technol Teacher Educ* 9(1):60–70
13. Krathwohl DR (2002) A revision of Bloom's taxonomy: An overview. *Theory into Pract* 41(4):212–218
14. Krieglstein F, Schmitz M, Wesenberg L, Spitzer MWH, Rey GD (2025) How a first impression biases cognitive load assessments: Anchoring effects in problem-solving tasks of varying element interactivity. *Mem Cognit* 54(1):501–514
15. Lincoln YS, Guba EG (1985) *Naturalistic inquiry*. Sage
16. Luckin R, Holmes W (2016) *Intelligence unleashed: An argument for AI in education*. Pearson Education
17. Mayer RE (2024) The past, present, and future of the cognitive theory of multimedia learning. *Educational Psychol Rev* 36(1):8
18. McKenney S, Reeves TC (2018) *Conducting educational design research*, 2nd edn. Routledge
19. Mishra P, Bozkurt A (2026) TPACK in the Age of Context and Artificial Intelligence: Punya Mishra on the Wicked Problem of AI and Why Algorithms Can't Replace Context. *Open Praxis*, 18(2)
20. Molenaar I (2022) Towards hybrid human-AI learning technologies. *Eur J Educ* 57(4):632–645
21. Naeem M, Smith T, Thomas L (2025) Thematic Analysis and Artificial Intelligence: A Step-by-Step Process for Using ChatGPT in Thematic Analysis. *Int J Qualitative Methods*, 24
22. Nicol DJ, Macfarlane-Dick D (2006) Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Stud High Educ* 31(2):199–218
23. Panadero E (2017) A review of self-regulated learning: Six models and four directions for research. *Front Psychol* 8:422

24. Pellegrino JW (2018) Assessment of and for learning. International handbook of the learning sciences. Routledge, pp 410–421
25. Priestley M, Biesta G, Robinson S (2026) Teacher agency: An ecological approach, 2nd edn. Bloomsbury Publishing
26. Ritzhaupt AD, Dawson K, Crompton H (2026) Toward guidance and principled values on the ethical, responsible, and transparent use of GenAI in the educational research process. J Res Technol Educ 58(2):235–249
27. Samuel A, Soh M, Jung E (2025) Enhancing reflective practice with ChatGPT: A new approach to assignment design. Med Teach 47(9):1438–1440
28. Selwyn N (2019) Should robots replace teachers? AI and the future of education. Wiley
29. Shute VJ (2008) Focus on formative feedback. Rev Educ Res 78(1):153–189
30. Sweller J (1988) Cognitive load during problem solving. Cogn Sci 12(2):257–285
31. Talebzadeh H (2025a) *The Integrated AI Triade (IAT): A systemic framework for redefining the roles of teachers, managers, and families in the AI era*. Zenodo. <https://doi.org/10.5281/zenodo.17914437>
32. Talebzadeh H (2025b) *Bilingual Teacher's Guide: Lesson 3 – Assessment Automation*. Zenodo. <https://doi.org/10.5281/zenodo.17536825>
33. Topping KJ, Gehringer E, Khosravi H, Gudipati S, Jadhav K, Susarla S (2025) Enhancing peer assessment with artificial intelligence. Int J Educational Technol High Educ 22(1):3
34. Zimmerman BJ (1990) Self-regulated learning and academic achievement: An overview. Educational Psychol 25(1):3–17
35. Zimmerman BJ, Schunk DH (2011) Self-regulated learning and performance: An introduction and an overview. Handbook of self-regulation of learning and performance. Routledge, pp 15–26