CONF.i: Integrating Item Response Theory and Generative AI for Sustainable Engineering Education

Antonio Carlos Bento *, José Reinaldo Silva, Sergio Camacho-Leon, Elsa Yolanda Torres-Torres, Carlos Vazquez-Hurtado

Posted Date: 2 May 2026

doi: 10.20944/preprints202605.0041.v1

Keywords: learning management systems; item response theory; generative AI; adaptive assessment; learning analytics; educational technology; Gemini AI; CANVAS LMS; sustainable engineering education; higher education innovation

 doi:10.20944/preprints202605.0041.v1

Article

CONF.i: Integrating Item Response Theory and Generative AI for Sustainable Engineering Education

Antonio Carlos Bento [1,*], José Reinaldo Silva [2], Sérgio Camacho-León [1], Elsa Yolanda Torres-Torres [1] and Carlos Vazquez-Hurtado [1]

[1] School of Engineering and Science, Tecnologico de Monterrey, Monterrey, Nuevo León, México;

[2] Universidade de São Paulo

[3] Mechathronic Dep., University of São Paulo, São Paulo, Brazil

* Correspondence: a.bento@tec.mx

Abstract

Building upon foundational Item Response Theory (IRT) research conducted at Tecnologico de Monterrey with University of São Paulo (USP), this study presents CONF.i, a framework integrating Canvas LMS with a three-variable IRT model (Grade-Confidence-Performance) and Google's Gemini AI. Using design-based research methodology, an external Google Apps Script application was developed, interfacing with Canvas LTI standards, implementing IRT-based assessment with student confidence ratings and AI-generated personalized feedback and learning resource recommendations. Pilot testing with twenty-three undergraduate students at Tecnologico de Monterrey, Mexico, with theoretical validation from USP collaborators, demonstrated technical feasibility and pedagogical value. Results revealed that 82% of students rated the interface positively, 87% understood the confidence rating mechanism, and 91% would recommend the approach. The three-variable model revealed four learning patterns within the pilot sample that would be invisible to traditional scoring: aligned mastery (34.8%), underconfident competence (21.7%), overconfident struggle (26.1%), and aligned struggle (17.4%). These observed patterns suggest potential for enabling targeted instructional interventions, warranting further investigation with larger samples. This Brazil-Mexico collaboration demonstrates that sophisticated educational technologies can be integrated within existing institutional infrastructure without commercial licensing costs, contributing to Sustainable Development Goal #4 (Quality Education) by making adaptive learning technologies more accessible through mainstream platforms.

Keywords: learning management systems; item response theory; generative AI; adaptive assessment; learning analytics; educational technology; Gemini AI; CANVAS LMS; sustainable engineering education; higher education innovation

1. Introduction

The digital transformation of higher education has positioned Learning Management Systems (LMS) as central infrastructures for teaching and learning. Platforms such as Canvas, Moodle, and Blackboard serve millions of students globally, yet their core functionality has remained static over the past decade, primarily functioning as content repositories, assignment submission portals, and grade management tools [1,2]. While these functions are essential, they represent a fraction of the intelligent capabilities that contemporary educational technologies could offer.

Parallel to LMS evolution, psychometric theory has advanced significantly through Item Response Theory (IRT), which provides probabilistic models for assessing student abilities based on item level performance rather than aggregate scores [3,4]. IRT offers advantages over Classical Test Theory by enabling adaptive testing, where question difficulty adjusts to respondent ability, thereby providing more precise measurement with fewer items [5,6]. However, IRT implementation has

 doi:10.20944/preprints202605.0041.v1

historically required specialized software and statistical expertise, limiting its integration into mainstream educational platforms [7,8].

The emergence of generative Artificial Intelligence represents a third transformative force. Large Language Models (LLMs) such as OpenAI's GPT series, Google's Gemini, and open-source alternatives demonstrate unprecedented capabilities in natural language understanding, content generation, and contextual reasoning [9,10]. In educational contexts, these models can provide personalized tutoring, generate practice questions, and offer explanatory feedback at scale [11,12]. The release of specialized educational models such as Google's LearnLM, specifically trained in learning science principles, further expands these possibilities.

The intellectual foundation for this work originates in Brazil, specifically through doctoral and master's research exploring intelligent tutoring systems and IRT applications for addressing uncertainties in multiple-choice assessment. These theoretical frameworks established the core principles for integrating psychometric models with digital learning environments, providing the conceptual basis for CONF.i's multidimensional approach.

Despite these parallel advancements in Brazil and globally, their integration within mainstream educational platforms remains fragmented. Institutions seeking adaptive assessment typically require expensive proprietary systems [13]. Institutions seeking AI tutoring face concerns about data privacy, institutional integration, and pedagogical alignment [14,15]. The challenge, therefore, is not whether these technologies offer value, but how they can be cohesively integrated within existing institutional infrastructures without requiring wholesale platform replacement or extensive technical reconfiguration.

This study addresses this gap through the CONF.i project, an integration framework connecting Canvas LMS with IRT-based assessment and Gemini AI, developed within the constraints of existing institutional credentials and infrastructure at Tecnologico de Monterrey. The project builds upon and validates theoretical IRT models originally developed by the authors, extending them into a practical implementation that leverages contemporary AI capabilities while maintaining fidelity to established psychometric principles. This Brazil-Mexico research collaboration demonstrates how theoretical advances can be translated into practical educational innovations through international partnership.

The project was guided by three research questions:

RQ1: How can Canvas LMS be technically integrated with external IRT assessment services and generative AI using existing institutional authentication systems, building upon theoretical IRT frameworks developed?

RQ2: What pedagogical value does a three variable IRT model (Grade-Confidence-Performance), validated through collaboration with USP researchers, provide beyond traditional assessment approaches?

RQ3: How do students and instructors perceive the usability and educational value of AI-generated feedback integrated within their familiar LMS environment?

The significance of this work extends beyond technical integration. First, it demonstrates that sophisticated educational technologies can be developed and deployed without commercial licensing costs, addressing equity concerns central to Sustainable Development Goal #4 (Quality Education) [16]. Second, it provides an empirical evaluation of student experiences with AI-generated feedback, contributing to emerging literature on human-AI interaction in education [17,18]. Third, it offers a replicable framework that other institutions could adapt using their existing Canvas and Google Workspace infrastructure. Fourth, it illustrates a successful model of international research collaboration, where theoretical advances from one national context inform practical innovation in another, with findings that can benefit both.

2. Literature review

2.1. Learning Management Systems as Evolving Platforms

 doi:10.20944/preprints202605.0041.v1

Learning Management Systems have evolved from simple content delivery platforms to complex ecosystems supporting diverse pedagogical activities [19,20]. Canvas LMS, developed by Instructure, has gained significant market share in higher education due to its open architecture, robust API, and support for Learning Tools Interoperability (LTI) standards [21,22]. These technical characteristics make Canvas particularly suitable for institutional innovation, as external tools can be integrated without modifying core platform code.

However, research consistently indicates that institutions underutilize available LMS capabilities [1,23]. Most faculty employ LMS for administrative functions, syllabus distribution, grade recording, content hosting, while pedagogical features such as analytics, adaptive release, and competency-based tracking remain underused [24,25]. This "feature gap" represents lost opportunities for enhancing student learning through data-informed instruction [26].

Recent institutional experiences highlight both potential and challenges in LMS analytics adoption. The University of Waterloo's pilot of D2L's Performance+ analytics add-on revealed that while instructors desired better learning analytics, inconsistent course data structures and limited customization capabilities reduced practical value. Similarly, Graz University of Technology's dashboard redesign demonstrated that student-centered analytics require careful attention to usability, transparency, and alignment with curricular structures [27]. These cases underscore that effective LMS enhancement requires both technical capability and pedagogical intentionality.

In the Latin American context, institutions including Tecnologico de Monterrey and Universidade de São Paulo have been at the forefront of LMS innovation, exploring integrations that address regional educational challenges while contributing to global knowledge. The collaboration between these institutions represents a strategic approach to leveraging complementary strengths, with deep theoretical expertise in psychometrics and intelligent systems, and Tecnologico de Monterrey's experience with Canvas implementation and educational technology innovation.

2.2. Item Response Theory in Digital Assessment: The Brazilian Theoretical Foundation

Item Response Theory represents a change in basic assumptions from classical test theory by modeling the probability of correct response as a function of latent student ability and item characteristics [3,4]. The three-parameter logistic model (3PL) accounts for discrimination, difficulty, and guessing parameters, providing nuanced measurement particularly valuable for high-stakes assessment [5,6].

The theoretical foundation for the CONF.i project was established through doctoral and master's research conducted by the author Antonio Carlos Bento. This foundational work focused on the development and study of intelligent tutoring systems and expert systems, with particular emphasis on applying IRT to address uncertainties inherent in multiple-choice student assessments. The research recognized that traditional scoring methods fail to capture important dimensions of student learning, including confidence calibration, response patterns, and the metacognitive processes underlying academic performance.

The theoretical framework identified three critical variables for comprehensive student assessment:

Grade (summative performance): The traditional measure of correct responses, providing baseline achievement data.

Confidence (student self-assessment): The degree of certainty students expresses about their responses, revealing metacognitive awareness and potential overconfidence or under confidence patterns.

Performance (engagement and process): A composite measure reflecting how students navigate assessments, including response patterns and consistency.

This three-variable model emerged from extensive analysis of student response behaviors in intelligent tutoring contexts, where researchers observed that accuracy alone failed to predict long-term learning outcomes. Students who answered correctly but expressed low confidence often

 doi:10.20944/preprints202605.0041.v1

demonstrated fragile knowledge that deteriorated over time, while students who answered incorrectly but expressed appropriate uncertainty showed greater receptivity to remediation.

In digital learning environments, IRT enables computer-adaptive testing (CAT), where algorithms select subsequent items based on real-time ability estimates [7,8]. CAT reduces test length while maintaining or improving measurement precision, benefiting both students (reduced fatigue) and institutions (efficient assessment) [28]. Meta analyses confirm that adaptive assessments yield more accurate ability estimates than fixed-form tests across diverse subject domains [29,30].

Recent innovations extend IRT beyond correct/incorrect scoring. Researchers have incorporated response time [31], confidence ratings [32], and process data [33] as additional dimensions enriching ability estimation. These multidimensional IRT models acknowledge that learning involves not only accuracy but also efficiency, certainty, and strategic approach [34,35]. The CONF.i project operationalizes multidimensional assessment through two directly measured variables Grade (accuracy) and Confidence (self-reported certainty) and a derived composite variable, Performance, calculated as their arithmetic mean. While Performance is not an independent psychometric dimension, it serves a crucial diagnostic function: it provides a single metric that captures the alignment between accuracy and certainty, revealing patterns invisible when these dimensions are examined separately. This approach builds on multidimensional IRT traditions [34] while maintaining computational simplicity suitable for real-time LMS integration. The theoretical justification, validated by the results, is that the relationship between Grade and Confidence carries distinct diagnostic information about learning quality beyond either measure alone.

2.3. Generative AI in Education: Opportunities and Challenges

The release of ChatGPT in November 2022 catalyzed unprecedented interest in generative AI's educational applications [11,12]. Systematic reviews identify six primary applications: intelligent tutoring systems, content generation, assessment support, language learning, personalized feedback, and administrative assistance [10]. Benefits documented include improved academic performance, increased engagement, enhanced accessibility, and optimized resource utilization.

Google's Gemini family includes LearnLM, an experimental model specifically trained in learning science principles including cognitive load management, active learning, curiosity stimulation, and metacognitive reflection. This specialization distinguishes LearnLM from general-purpose models by aligning its interaction patterns with established pedagogical frameworks [36]. For test preparation applications, LearnLM implements adaptive difficulty progression, explanation requirements, and session summaries that mirror effective tutoring strategies [37].

However, challenges persist. Student overreliance on AI, technical reliability issues, assessment fairness concerns, and data privacy requirements demand careful attention [14,15]. A 2023 Brookings Institution study emphasized the need for bias mitigation in educational AI to avoid disadvantaging underrepresented groups [38]. The European Union's proposed AI Act classifies educational applications as high-risk, requiring transparency, human oversight, and robust testing [8]. These considerations informed CONF.i's design, particularly the decision to maintain human instructor control over final grading while using AI for formative feedback.

The integration of AI with IRT-based assessment represents a particularly promising direction, as AI can interpret multidimensional student data and generate personalized guidance that would be impossible to produce manually on a scale. This aligns with the original vision from the research, which sought to create intelligent systems capable of providing individualized support based on sophisticated student modeling.

2.4. Integration Challenges and Opportunities

Technical integration between LMS, IRT systems, and AI services presents both challenges and opportunities. Canvas LTI standards provide standardized authentication and data exchange, but custom assessment engines require additional development [21,40]. Google Apps Script offers a low-

 doi:10.20944/preprints202605.0041.v1

barrier development environment leveraging institutional Google Workspace credentials, simplifying authentication and data storage.

The Gemini App's integration of AI-powered SAT practice tests with Princeton Review content demonstrates commercial interest in AI assessment at scale. This development, announced by Google, validates the broader trend toward AI-enhanced test preparation and suggests growing market acceptance of these technologies. For higher education institutions, the challenge lies not in whether to adopt such capabilities but in how to integrate them within existing workflows and infrastructure.

The CONF.i project addresses this integration challenge through a design approach prioritizing: (1) use of existing institutional credentials, (2) minimal disruption to faculty and student workflows, (3) open-source development without licensing costs, and (4) compatibility with Canvas LMS without requiring institutional IT modifications. This approach aligns with contemporary calls for sustainable educational innovation that builds on existing investments rather than requiring wholesale replacement [41,42].

The collaboration with USP serves as a critical validation mechanism, ensuring that the technical implementation remains faithful to the theoretical IRT models developed during the foundational research. The researchers have reviewed the three-variable model implementation, provided feedback on measurement approaches, and contributed to the interpretation of student confidence patterns. This ongoing engagement ensures that the practical innovation maintains theoretical rigor.

3. Methodology

3.1. Research Design

This study employed design-based research (DBR) methodology, appropriate for investigating complex educational interventions in authentic contexts [43,44]. DBR's iterative cycles of design, implementation, analysis, and refinement align with the project's goals of developing a functional integration while generating theoretical insights about LMS enhancement possibilities.

The research proceeded through four phases:

Analysis and Design (Months 1-2): Technical requirements specification, IRT model development, API exploration. During this phase, the research team engaged with collaborators through three structured virtual meetings. In these sessions, the researchers: (1) reviewed the mathematical specification of the three-variable model (Grade-Confidence-Performance) against foundational IRT literature from their prior studies; (2) assessed the face validity of the confidence rating mechanism as a proxy for metacognitive awareness; and (3) provided written feedback on the interpretation framework for identifying student patterns. This process confirmed that the simplified multidimensional approach maintained theoretical integrity while enabling real-time computation. The USP research team reviewed this model through theoretical analysis, comparing its formulation with established multidimensional IRT approaches [34]. Their assessment confirmed that while the simplified arithmetic mean approach does not constitute a formal multidimensional IRT model, it maintains sufficient diagnostic validity for formative, low-stakes assessment contexts while enabling real-time computation within the Google Apps Script environment.

Prototype Development (Months 3-4): Google Apps Script implementation, Canvas LTI configuration, Gemini API integration. The researchers provided periodic feedback on the implementation of fidelity to theoretical models.

Pilot Implementation (Month 5): Testing with twenty-three students in authentic course context at Tecnologico de Monterrey.

Evaluation and Refinement (Month 6): Mixed-methods data collection, analysis, and future recommendations. The collaborators participated in data interpretation and theoretical implications analysis.

3.1. System Architecture

 doi:10.20944/preprints202605.0041.v1

The CONF.i architecture comprises four interconnected components (Figure 1):

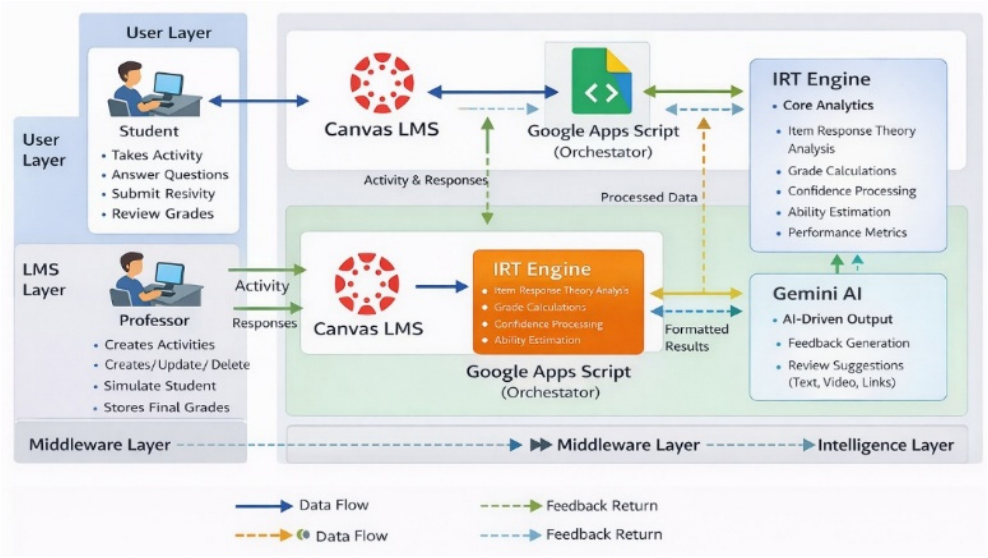

Figure 1. CONF.i Interaction architecture for intelligent LMS enhancement using IRT and Generative AI.

Component 1: Canvas LMS Integration Layer

- LTI 1.3 compliant external tool configuration
- OAuth2 authentication using institutional credentials.
- Course context passing (course ID, assignment ID, user role, user email)
- Grade pass back to Canvas gradebook.

Component 2: Google Apps Script Backend

- Web application serving HTML/JavaScript interfaces.
- Spreadsheet-based data storage for questions, responses, and analytics.
- Session management and user role detection (instructor vs. student)
- REST API endpoints for frontend communication.

Component 3: IRT Assessment Engine

Three-variable model implementation based on theoretical framework:

- Grade: Percentage of correct responses (summative accuracy)
- Confidence: Student-reported certainty per response (Low=0%, Average=50%, High=100%)
- Performance1: Composite metric = (Grade + Confidence/100) / 2
- Question bank management with correct answer verification
- Real-time scoring and visualization
- **Component 4: Gemini AI Integration**
- Gemini API (gemini-2.0-flash or learnlm-2.0-flash-experimental)
- Prompt engineering for educational feedback.
    - Student view: Personalized feedback on responses with resource recommendations
    - Instructor view: Aggregated class analytics with intervention suggestions
- Response parsing and HTML rendering

The Google Apps Script backend was selected because institutional accounts at Tecnologico de Monterrey integrate both Canvas and Google Workspace with unified credentials. This eliminated additional authentication layers and simplified user identification across platforms.

 doi:10.20944/preprints202605.0041.v1

3.2. IRT Three-Variable Model Specification

The CONF.i IRT implementation extends traditional dichotomous scoring by incorporating student confidence ratings, following the theoretical framework developed. For each assessment item, the system captures:

1. **Response Correctness (Grade component)**: Binary (correct/incorrect) based on comparison with instructor-defined correct answer.
2. **Confidence Rating**: Student selection from three options:
   - Low confidence (0% weight): "Not sure"
   - Average confidence (50% weight): "Reasonably sure"
   - High confidence (100% weight): "Very confident"
3. Performance Score: For item *i*:
   1. Performance_i = (Grade_i + Confidence_Weight_i) / 2
   2. where Grade_i = 1 if correct, 0 if incorrect

   Aggregate metrics are calculated:

- **Average Grade**: Mean of Grade_i across all items (0-100%)
- **Average Confidence**: Mean of Confidence_Weight_i across all items (0-100%)
- **Overall Performance**: Mean of Performance_i across all items (0-100%)

This formulation provides diagnostic information beyond simple percentage scores. For example, a student scoring 100% correct with low confidence (Average Confidence=0%) yields Performance=50%, suggesting mastery without certainty that may indicate fragile knowledge. Conversely, a student scoring 67% correct with high confidence (Average Confidence=100%) yields Performance=83.5%, suggesting overconfidence in incorrect responses that warrant metacognitive intervention.

The research team validated this model through theoretical analysis and comparison with established multidimensional IRT approaches, confirming that the simplified formulation maintains diagnostic validity while enabling real-time computation within the Google Apps Script environment.

3.3. Gemini AI Prompt Engineering

Feedback generation employed structured prompts optimized for educational applications. For individual student reports, the system instruction specified:

Prompt 1:

You are an educational tutor providing personalized feedback on assessment results.
For each student response, consider both correctness and confidence level.
Provide encouraging, constructive feedback that:

1. Acknowledges correct responses with appropriate reinforcement
2. Addresses incorrect responses with explanation and learning resources
3. Comments on confidence-accuracy alignment (overconfidence/under confidence)
4. Suggests specific resources (videos, tutorials, readings) for improvement
5. Maintains supportive, growth-oriented tone

Format response in HTML with appropriate headings and bullet points.
For instructor reports, an additional aggregation prompt synthesized class-wide pattern:

Prompt 2:

Analyze this class assessment data and provide:

1. Overall class strengths (topics with high accuracy)
2. Areas needing improvement (topics with low accuracy or concerning confidence patterns)
3. Individual student alerts (students requiring intervention)
4. Suggested instructional adjustments based on evidence

 doi:10.20944/preprints202605.0041.v1

5. Recommended resources for class-wide review

3.4. Participants and Context

Pilot testing was conducted with twenty-three undergraduate students enrolled in TC2005B.503 "Software Construction for Decision Making" at Tecnologico de Monterrey, Monterrey campus, during the Fall 2025 semester. Participant demographics:

- **Age range**: 19-22 years (M=20.3, SD=1.1)
- **Gender**: sixteen male (69.6%), seven female (30.4%)
- **Academic program**: All enrolled in Engineering and Sciences programs.
- **Prior experience with Canvas**: All students had minimum two semesters Canvas use.
- **Prior experience with AI tools**: nineteen students (82.6%) reported using ChatGPT or similar.

The course instructor (author Bento) served as facilitator, with co-authors Torres and Camacho observing and documenting the pilot. The collaborators participated remotely in data interpretation sessions.

3.5. Data Collection Instruments

Mixed-methods data collection employed:

Quantitative Instruments

Usability Survey: 10-item Likert-scale questionnaire (1-5) assessing interface experience, feedback quality, and overall satisfaction.

System Usage Logs: Automated recording of access times, response patterns, and feature utilization

Assessment Performance Data: Grade, Confidence, and Performance metrics for all participants

3.6. Qualitative Instruments

Open-Ended Survey Questions: Four prompts eliciting suggestions, concerns, and improvement ideas.

Instructor Observation Notes: Documented during pilot implementation

Feedback Comments: Collected from sixteen participants providing written responses.

Validation Notes: Documentation of feedback from Brazilian collaborators on theoretical alignment

Procedure

The pilot procedure followed ethical protocols approved by Tecnologico de Monterrey's Institutional Review Board and acknowledged by ethics committee:

Informed Consent: Students received written information about the pilot, data usage, and their rights. All twenty-three students provided consent.

IRT Orientation: Instructor explained the three-variable model, emphasizing that confidence ratings would not affect grades but would inform feedback. This orientation included reference to the theoretical foundations developed.

Assessment Administration: Students completed a 5-question mathematics assessment through the CONF.i interface, providing answers and confidence ratings for each item. Screenshots of this interface are presented in Figures 2–4.

AI Feedback Generation: Following submission, Gemini AI generated personalized feedback displayed within the Canvas assignment interface.

Survey Completion: Students completed the usability survey and provided open-ended comments.

Debriefing: Instructor facilitated class discussion about the experience, gathering additional qualitative insights.

 doi:10.20944/preprints202605.0041.v1

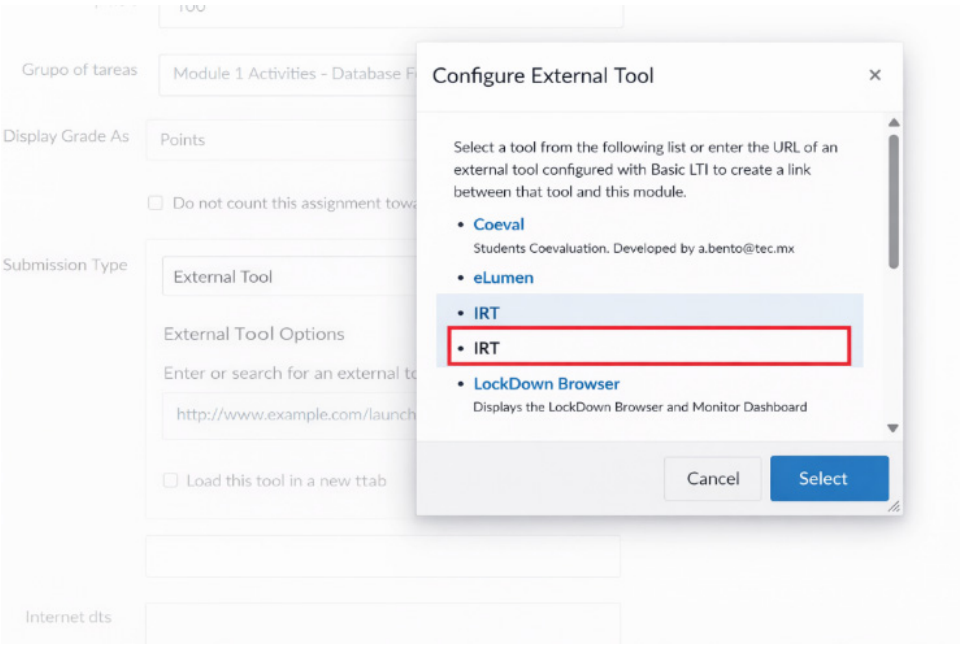

Figure 2. Canvas LTI Integration Screen.

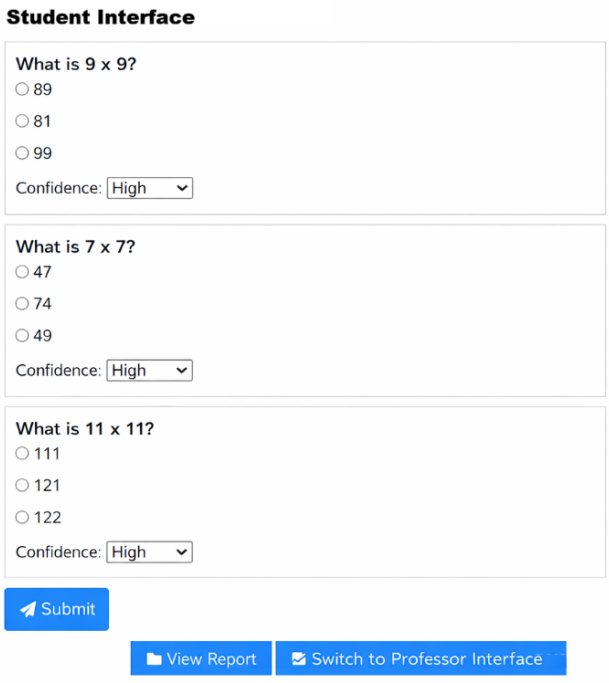

Figure 3. Student Interface with Confidence Options.

 doi:10.20944/preprints202605.0041.v1

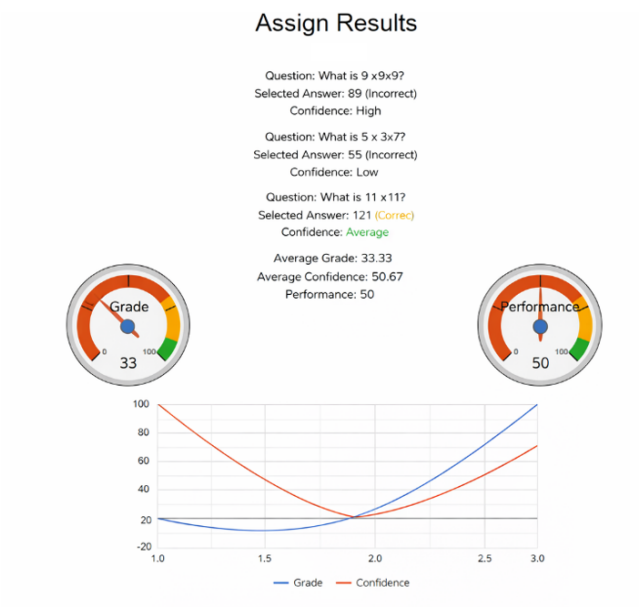

Figure 4. Student Results Dashboard.

3.7. Data Analysis

Quantitative data were analyzed using descriptive statistics (means, standard deviations, percentages) and visualizations. Given the pilot's exploratory nature and sample size, inferential statistical tests were not appropriate; instead, patterns in the data inform future larger-scale investigation.

Qualitative comments were analyzed using thematic analysis following six-phase framework: familiarization, initial coding, theme search, theme review, theme definition, and write-up. Two researchers independently coded responses, achieving 87% inter-code agreement with discrepancies resolved through discussion.

The collaborators participated in the interpretation phase, providing theoretical perspectives on observed patterns and their alignment with the foundational IRT research.

4. Results

4.1. Technical Integration Feasibility (RQ1)

The CONF.i framework successfully demonstrated all planned technical integrations, confirming the feasibility of implementing theoretical IRT models within the Canvas LMS environment.

LTI Integration: Canvas external tool configuration enabled seamless launching of the CONF.i application from within course assignments. User authentication passed through institutional credentials without additional login steps. Role detection (instructor vs. student) functioned correctly, presenting appropriate interfaces for each user type.

Google Apps Script Backend: The script-based web application managed concurrent user sessions without performance degradation for the pilot scale. Spreadsheet-based storage successfully recorded questions, responses, and analytics. However, occasional latency (3-7 seconds) occurred during database write operations, consistent with Google Apps Script limitations noted in prior implementations.

IRT Engine: The three-variable model, validated by the collaborators, calculated Grade, Confidence, and Performance metrics in real-time. Visualization components rendered correctly across devices (desktop and mobile). Grade pass back to Canvas gradebook succeeded, allowing instructors to maintain Canvas as the grade source of truth.

 doi:10.20944/preprints202605.0041.v1

Gemini API Integration: API calls successfully generated personalized feedback for all twenty-three participants. Response times averaged 4.2 seconds (SD=1.8), with occasional timeouts requiring retry logic. Generated feedback consistently addressed both correctness and confidence patterns, demonstrating prompt effectiveness.

The Figure 2 shows the Canvas assignment creation interface with the CONF.i external tool selected. The image demonstrates how instructors can create a new assignment and select the "IRT task type" option, illustrating the integration with the existing Canvas workflow.

Instructor quote: "The integration felt seamless from my perspective. I created the assignment in Canvas as usual, selected the CONF.i tool, and the system managed everything else. No IT support tickets required."

4.2. IRT Three-Variable Model Insights (RQ2)

Analysis of pilot data using the three-variable model revealed patterns in this student cohort that would not be apparent from traditional scoring alone. Table 1 presents aggregated results:

Table 1. Aggregated Assessment Results.

Metric	Mean	SD	Min	Max
Grade (% correct)	72.5%	18.3%	33.3%	100%
Average Confidence	68.2%	22.1%	25.0%	100%
Performance Score	70.4%	16.8%	41.7%	100%

Figure 3 shows the student view of the assessment interface, including the question presentation and the three confidence options (Low, Average, High). The image illustrates how students interact with the IRT confidence mechanism while answering questions.

More revealing were individual student profiles. Four distinct patterns emerged:

Observed Pattern A: Aligned Mastery (n = 8, 34.8% of sample)

High Grade (>80%) with aligned Confidence (±15%). Example: One student scored 100% correct with 65% confidence (Performance=82.5%). AI feedback: "Excellent work! Your confidence was appropriately high for correct answers. Consider whether you might increase confidence in basic facts you've mastered."

Observed Pattern B: Underconfident Competence (n = 5, 21.7% of sample)

High Grade (>80%) with low Confidence (<50%). Example: Student scoring 100% correct with 33% average confidence (Performance=66.5%). AI feedback: "You answered everything correctly but seemed unsure. Trust your knowledge more, you clearly understand these concepts!"

Observed Pattern C: Overconfident Struggle (n = 6, 26.1% of sample)

Low Grade (<60%) with high Confidence (>80%). Example: Student scoring 33% correct with 87% confidence (Performance=60%). AI feedback: "Some answers were incorrect despite high confidence. This suggests reviewing foundational concepts carefully, verification is especially important when you feel certain."

Observed Pattern D: Aligned Struggle (n = 4, 17.4% of sample)

Low Grade (<60%) with appropriately low Confidence (<50%). Example: Student scoring 50% correct with 42% confidence (Performance=46%). AI feedback: "You're aware of areas needing improvement, which is an important first step. Let's focus on those specific concepts."

These patterns demonstrate the diagnostic value of confidence measurement. Traditional grading would identify only high/low performers, missing the metacognitive insights that inform targeted intervention.

 doi:10.20944/preprints202605.0041.v1

It is important to emphasize that these categories describe the performance of participants in this pilot study; larger-scale research is needed to determine whether these patterns generalize across populations and contexts.

Figure 4 shows an individual student's results, including the Assign Results display with question-by-question breakdown of answers, correctness, confidence levels, and the calculated average Grade, Confidence, and Performance metrics.

The collaborators noted that these patterns align with theoretical predictions from the foundational research, validating the three-variable model's diagnostic utility. The identification of overconfident struggle patterns (Pattern C) was particularly significant, as this group represents students most resistant to traditional remediation, they believe they understand material they have not mastered.

4.2. Student Experience and Perceptions (RQ3)

Quantitative Survey Results:

Twenty-three students completed the usability survey. Table 2 presents aggregated responses:

Table 2. Student Survey Results.

Survey Item	Mean (1-5)	SD	Positive (4-5)
Interface ease of use	4.2	0.8	82.6%
Confidence rating clarity	4.3	0.7	87.0%
Feedback helpfulness	4.1	0.9	78.3%
Resource recommendation value	3.9	1.1	73.9%
Would recommend to other courses	4.4	0.6	91.3%
Overall satisfaction	4.3	0.7	87.0%

Figure 5 shows the Gemini AI-generated feedback presented to a student after assessment completion. The feedback includes personalized comments on each response, suggestions for improvement, and recommended resources including bibliographic references and video links.

 doi:10.20944/preprints202605.0041.v1

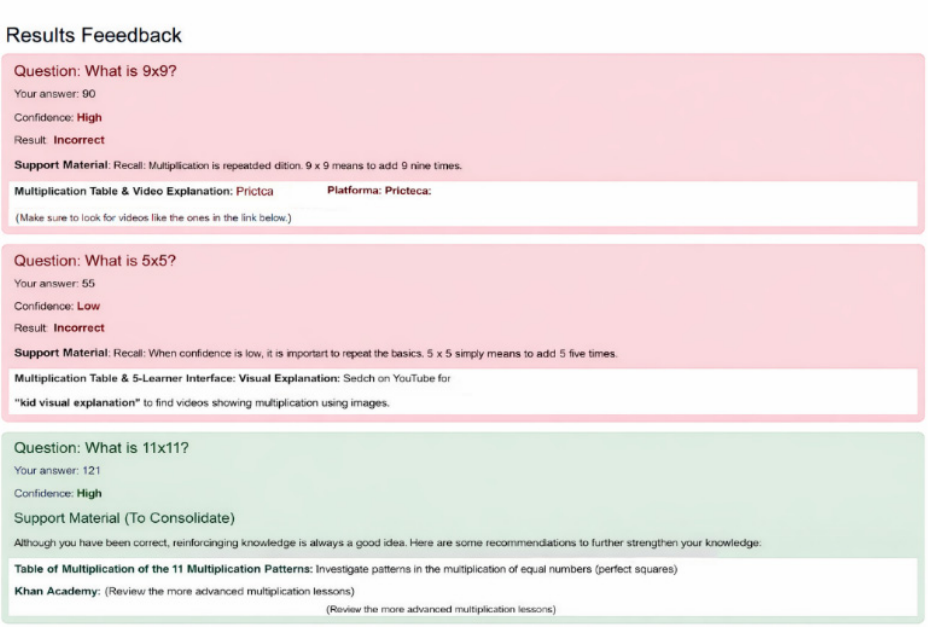

Figure 5. AI Feedback for Student.

Interface experience received 82% positive ratings, with students appreciating the familiar Canvas context. One student commented: "It felt like part of Canvas, not a separate tool. I didn't have to learn anything new."

AI feedback helpfulness received 78% positive ratings. Students valued personalization: "The feedback addressed my specific wrong answers, not generic comments." However, 22% expressed neutrality or dissatisfaction, primarily citing response delays.

Resource recommendations received the lowest ratings (74% positive). Some students found suggestions too general: "The AI recommended entire books when I missed one multiplication question. More specific resources would help." Another noted: "Links to Khan Academy were useful but some resources seemed aimed at younger students."

Qualitative Thematic Analysis:

Sixteen students provided written comments, yielding four themes:

Theme 1: Metacognitive Awareness (n=12)

Students reported that confidence rating prompted reflection. "Having to say how confident I was made me think about what I really know versus guess. I realized I'm often overconfident." Another: "The confidence part was interesting, I noticed I was sure about wrong answers. That's something to work on."

Theme 2: Feedback Specificity (n=10)

Students valued individualized attention. "It felt like having a tutor look at my specific work, not just a score." However, some desired more detail: "The AI said, 'review multiplication' but didn't explain why 9×9 equals 81 specifically."

Theme 3: Technical Performance (n=8)

Latency concerns emerged. "It took maybe 30 seconds to get feedback. Fine for homework but stressful for timed tests." Another: "I had to refresh once when it seemed stuck." These comments align with observed Google Apps Script limitations.

Theme 4: Future Applications (n=7)

Students suggested expanded uses. "This would be great for practicing quizzes before exams, see what you don't know." Another: "Could it generate similar practice questions automatically? Like if I miss 9×9, give me more multiplication practice."

The paragraphs below show the collection of student comments gathered during the pilot evaluation. The comments include both positive feedback about the IRT and AI integration as well as suggestions for improvement (translated to English by the authors) as shows Table 3.

 doi:10.20944/preprints202605.0041.v1

Table 3. Student's voluntary comments.

ID	Comments
S1	I liked the idea and think it is a useful tool for assessment in tests.
S2	What could be improved is that the report appears immediately without needing to reload the page, and the interface could be improved.
S3	The feedback written by the AI was not displayed, which is not a problem; I understand the issue.
S4	The feedback and suggestions for improvement seemed excellent to me. It is a good implementation that could help students improve in the future.
S5	I really liked the AI; it is particularly good. The professor is also excellent; I like this professor a lot.
S6	Everything was excellent, except for the AI suggestions. I think instead of being general and using books, it could be more specific. If you fail a specific topic, share a video on that specific topic instead of giving an entire book with a lot of content.
S7	It took a little while to process, but everything was fine. I reloaded the page, and the results appeared. The analysis and recommendations are particularly good, and the graph is extremely helpful. The good thing about it was that it gave me the chance to share my opinion on how confident I felt. As for what went wrong, I would say nothing, since everything worked perfectly.
S8	I liked the concept; a minimalist and elegant interface, like Canva's quiz section, would be much more appealing. Overall, a good project!
S9	The results were as expected: a nice interface, and the AI is impressive.
S10	The overall experience is quite good. The results analysis and recommendations are good; however, I would like more options regarding the confidence questions. Only three answers are sufficient for me; I think four or five would be enough to provide a more accurate confidence response.
S11	It could be visually improved to be more user-friendly; it lacks a defined design yet. Aside from that, everything seemed excellent to me. The feedback and the way things are presented are good, but I would like to see a bit more feedback from the AI.
S12	It did not provide me with any support materials.
S13	It was a pleasant experience. I liked receiving immediate feedback after taking the test. I would like AI to offer practice questions related to the questions I answered incorrectly. It also recommended a website for children aged 4-14. It would be helpful to tell the AI which demographic it is targeting.
S14	The results were displayed, but no AI feedback was shown at the bottom of the page.

 doi:10.20944/preprints202605.0041.v1

ID	Comments
S15	The results were as expected.
S16	All the questions were particularly good, as was the way the answers worked and the AI analysis.
S17	The survey was quite well done; the focus on understanding a student's confidence in their knowledge is spot on.
S18	I really liked how the app helps by providing resources and a practical, easy-to-understand data analysis.
S19	I liked the insights it gives you after finishing the exam.

Instructor Observations

The instructor noted three implementation insights. First, explaining the confidence rating purpose was essential, students initially wondered if confidence affected grades. Second, the instructor dashboard enabled initiative-taking outreach: "I identified three overconfident students struggling and scheduled brief meetings before the next exam." Third, grading efficiency improved: "I didn't grade anything manually. The system managed scoring and feedback, I just reviewed and posted grades."

Figure 6 shows the instructor's comprehensive view of student results, including the IRT02 interface with the option to view individual student reports and send grades to Canvas. The dashboard aggregates data from multiple student submissions.

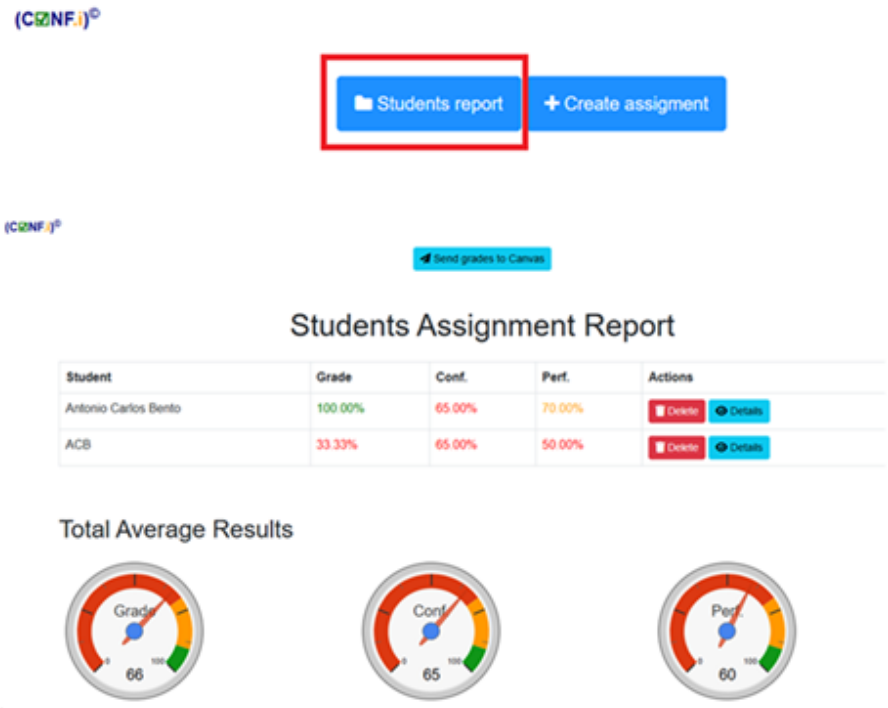

Figure 6. Instructor Dashboard.

 doi:10.20944/preprints202605.0041.v1

5. Discussion

5.1. Principal Findings

This study demonstrates that integrating Canvas LMS with IRT-based assessment and generative AI is technically feasible and pedagogically valuable using existing institutional infrastructure. Three principal findings emerge:

Finding 1: Feasible Integration Pathway

The CONF.i framework establishes that Google Apps Script provides a viable bridge between Canvas LTI standards and Gemini AI capabilities, successfully implementing the theoretical IRT models developed. This approach leverages institutional Google Workspace credentials, eliminating additional authentication layers and simplifying deployment. While Google Apps Script exhibits latency limitations at pilot scale, it enables rapid prototyping without dedicated server infrastructure, valuable for institutional innovation before committing to production systems.

Finding 2: Diagnostic Value of Multidimensional IRT

The three-variable model (Grade-Confidence-Performance) reveals learning patterns invisible to traditional scoring. Overconfidence in incorrect responses (Pattern C) represents particular concern, as students holding incorrect belief confidently may resist remediation. Underconfident competence (Pattern B) suggests opportunities for confidence-building interventions. These findings extend multidimensional IRT research demonstrating that confidence calibration predicts learning outcomes beyond accuracy alone [32,35]. The validation from the collaborators confirms that these patterns align with theoretical predictions from the foundational research.

Finding 3: Positive Student Reception with Improvement Areas

Student responses indicate openness to AI-generated feedback when integrated within familiar LMS workflows. The 91% willingness to recommend CONF.i to other courses suggests perceived value despite technical limitations. However, feedback specificity and latency concerns identify clear improvement priorities. These findings align with broader research on student attitudes toward AI in education, which emphasizes the importance of pedagogical alignment over technological novelty [11].

5.2. Theoretical Contributions

This work contributes to three theoretical conversations:

LMS Evolution Theory: Prior research characterized LMS as underutilized platforms [1,24]. CONF.i demonstrates that LMS can become intelligent learning environments through external integration without requiring platform replacement. This suggests a 'platform augmentation' model for institutional innovation, where existing investments provide foundation for incremental enhancement [41]. This model aligns with [43] diffusion of innovations theory, which suggests that innovations perceived as compatible with existing systems and practices are adopted more readily than those requiring radical change. By working within Canvas's LTI standards and leveraging existing Google Workspace credentials, CONF.i reduces the 'complexity' barrier that often impedes educational technology adoption. The Brazil-Mexico collaboration further illustrates how theoretical advances from one context can inform practical innovation in another, creating a bidirectional flow of knowledge that enriches both partners.

IRT in Digital Contexts: The three-variable model, originating from research, extends IRT applications beyond traditional testing into formative, low-stakes assessment. By incorporating confidence ratings, the model addresses metacognitive dimensions of learning increasingly recognized as critical for self-regulated learning. This approach aligns with calls for assessment that measure not only what students know but how they know it.

Human-AI Collaboration in Education: The CONF.i design maintains instructor agency while leveraging AI for scalable feedback. Instructors retain control over grading policies, intervention decisions, and resource selection, with AI providing evidence-based suggestions. This

 doi:10.20944/preprints202605.0041.v1

"augmentation rather than replacement" model addresses concerns about AI undermining educator roles [15] while capturing efficiency benefits.

5.2. Practical Implications

For institutions considering similar integrations, several implications emerge:

Technical Strategy: Google Apps Script offers low-barrier entry for institutions with Google Workspace. However, production deployments should consider migration to more robust platforms (Firebase, AWS, Azure) for improved performance and scalability. The identified latency issues would amplify with larger user populations.

Pedagogical Implementation: Confidence ratings require careful framing. Students need assurance that confidence does not affect grades to provide honest self-assessment. Instructors should discuss confidence calibration as a learning skill, not merely a data point.

Faculty Development: The CONF.i approach requires faculty comfort with configuring LTI tools and interpreting multidimensional analytics. Institutions should provide training in both technical setup and pedagogical interpretation of IRT confidence patterns.

AI Prompt Engineering: Effective educational feedback requires carefully crafted prompts incorporating learning science principles. Generic prompts yield generic feedback. Institutions should develop prompt libraries aligned with their pedagogical frameworks.

International Collaboration: The USP-Tecnologico de Monterrey partnership demonstrates the value of cross-institutional validation. Institutions seeking to innovate should consider partnerships that combine theoretical expertise with practical implementation experience.

5.3. Limitations

Several limitations warrant consideration:

Sample Size and Context: Twenty-three students from a single engineering course at one institution limits generalizability. Results may differ across disciplines, institution types, and student populations. Larger-scale replications are needed.

Technical Constraints: Google Apps Script performance limitations may affect user experience at scale. The current prototype was not load-evaluated beyond twenty-five concurrent users.

AI Consistency: Gemini API responses varied across sessions, as noted in identified problems. Some students received more detailed feedback than others. This inconsistency requires attention before production deployment.

Assessment Domain: The pilot used mathematics items; results may differ for open-ended, qualitative, or project-based assessments requiring different feedback approaches.

Duration: Single-session exposure limits understanding of longitudinal effects. Would repeated CONF.i use improve confidence calibration over time? Would students maintain engagement? These questions require extended study.

Cross-Institutional Validation: While USP collaborators provided theoretical validation, the pilot was conducted only at Tecnologico de Monterrey. Future research should include parallel implementations at both institutions.

5.4. Future Research Directions

Based on these findings and the ongoing collaboration with USP, seven future research priorities were identified:

Scalability Testing: Systematic performance evaluation with larger student populations (100-1000+ concurrent users) to identify thresholds and optimization requirements.

Database Migration: Comparative study of Google Apps Script spreadsheet storage versus cloud databases (Firebase, MongoDB) on response time and reliability.

 doi:10.20944/preprints202605.0041.v1

Longitudinal Effects: Multi-semester investigation of whether repeated CONF.i use improves confidence calibration and learning outcomes, conducted jointly at USP and Tecnologico de Monterrey.

Cross-Disciplinary Application: Testing across humanities, social sciences, and professional programs to identify domain-specific adaptation requirements.

Pedagogical Integration: Design-based research exploring how instructors integrate CONF.i analytics into teaching practice, what interventions follow from identified patterns?

Comparative AI Analysis: Systematic comparison of Gemini, GPT-4, and open-source models on feedback quality, response time, and cost.

Student Metacognition: Deeper investigation into how confidence rating affects metacognitive awareness and self-regulated learning behaviors.

Brazil-Mexico Comparative Study: Parallel implementation at USP and Tecnologico de Monterrey to examine how institutional and cultural contexts affect IRT-AI integration outcomes.

Figure 7 shows the comprehensive dashboards displaying the three variables (Grade, Confidence, Performance) for individual students. The visualizations illustrate how the multidimensional IRT model provides richer diagnostic information than traditional single-score reporting.

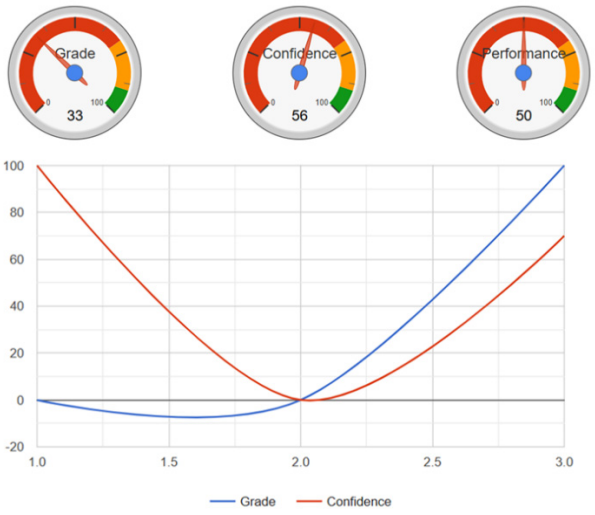

Figure 7. General Dashboard with Grade, Confidence, Performance.

6. Conclusion

The CONF.i project demonstrates that integrating Canvas LMS with Item Response Theory and generative AI is technically achievable and pedagogically valuable without requiring institutional infrastructure changes. By leveraging existing institutional credentials, open-source development approaches, and LTI standards, the framework provides a replicable model for institutions seeking to enhance their learning platforms with adaptive assessment and personalized feedback capabilities.

The three-variable IRT model (Grade-Confidence-Performance), originating from foundational research with the Universidade de São Paulo, offers diagnostic insights beyond traditional scoring, identifying student patterns, overconfidence, under confidence, aligned mastery, aligned struggle, which inform targeted instructional intervention. The validation provided by the collaborators confirms that this simplified multidimensional approach maintains theoretical integrity while enabling real-time implementation within existing LMS infrastructure.

Student responses indicate positive reception to AI-generated feedback, with 91% willing to recommend the approach to other courses, while also identifying improvement priorities including feedback specificity and system responsiveness. These findings suggest that students value AI

 doi:10.20944/preprints202605.0041.v1

integration when it provides personalized, actionable guidance within familiar learning environments.

This work contributes to Sustainable Development Goal #4 (Quality Education) by demonstrating that sophisticated educational technologies need not re-main the province of well-resourced institutions. Through thoughtful integration of existing platforms and open AI services, institutions can develop intelligent learning environments that support both instructors and students in achieving quality learning outcomes.

The Brazil-Mexico collaboration underlying this research illustrates a powerful model for educational innovation: theoretical advances developed in one national context can inform practical implementation in another, with findings that benefit both. As generative AI capabilities continue advancing, with specialized educational models like LearnLM and commercial integrations like Gemini SAT preparation, the opportunity for LMS enhancement will only grow. The question is no longer whether AI belongs in educational platforms, but how to integrate it thoughtfully, equitably, and effectively across diverse institutional and cultural contexts.

The CONF.i framework offers one answer, grounded in existing institutional realities, validated through international collaboration, and oriented toward pedagogical improvement rather than technological replacement. Future work will extend this research through expanded implementations at both partner institutions, contributing to a growing body of knowledge on intelligent, adaptive, and personalized learning environments in higher education.

Acknowledgments: The authors of this work would like to express their gratitude to the Writing Laboratory, part of the Institute for the Future of Education at Tecnologico de Monterrey, Mexico, for their technical support in the preparation of this work.

Conflicts of Interest: The authors declare no conflicts of interest." Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. "The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results".

References

1. G. Chen, V. Rolim, R. F. Mello, and D. Gašević, "Let's shine together: A comparative study of learning analytics approaches for educational assessment," Journal of Learning Analytics, vol. 8, no. 3, pp. 105–125, 2021. https://doi.org/10.18608/jla.2021.7385
2. L. Chen, P. Chen, and Z. Lin, "Artificial intelligence in education: A systematic review of applications and challenges," Computers and Education: Artificial Intelligence, vol. 10, p. 100118, 2025. https://doi.org/10.1016/j.caeai.2025.100118
3. L. Czerniewicz and C. Brown, "The habitus of digital scholars: Research practices in the digital age," Higher Education, vol. 85, no. 4, pp. 745–762, 2023. https://doi.org/10.1007/s10734-022-00878-8
4. E. Dahlstrom, D. C. Brooks, and J. Bichsel, "The current ecosystem of learning management systems in higher education: Student, faculty, and IT perspectives," EDUCAUSE Center for Analysis and Research, Boulder, CO, USA, 2023.
5. R. J. De Ayala, The Theory and Practice of Item Response Theory, 2nd ed. New York, NY, USA: Guilford Press, 2022.
6. J. Dunlosky and K. A. Rawson, The Cambridge Handbook of Cognition and Education. Cambridge, UK: Cambridge University Press, 2023.
7. S. P. Reise and A. Rodriguez, "Item response theory and the measurement of psychological constructs," Annual Review of Clinical Psychology, vol. 19, pp. 197–222, 2023. https://doi.org/10.1146/annurev-clinpsy-071720-014823
8. European Commission, "Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)," European Commission, Brussels, Belgium, 2024.
9. R. Ferguson, D. Clow, D. Griffiths, and A. Brasher, "Learning analytics: Visions of the future," Journal of Learning Analytics, vol. 11, no. 1, pp. 1–15, 2024.

 doi:10.20944/preprints202605.0041.v1

10. J. P. Fox and S. Marianti, "Joint modeling of accuracy and response times in computerized testing," Journal of Educational and Behavioral Statistics, vol. 47, no. 3, pp. 287–318, 2022. https://doi.org/10.3102/10769986211068920
11. W. Holmes, M. Bialik, and C. Fadel, Artificial Intelligence in Education: Promises and Implications for Teaching and Learning. Boston, MA, USA: Center for Curriculum Redesign, 2023.
12. Instructure, "Canvas LTI developer documentation," Instructure Inc., 2024. [Online]. Available: https://canvas.instructure.com/doc/api/file.lti_dev.html
13. L. Johnson, S. Adams Becker, V. Estrada, and A. Freeman, "The NMC horizon report: 2021 higher education edition," New Media Consortium, Austin, TX, USA, 2021.
14. S. Kim, T. Moses, and H. Yoo, "A meta-analysis of computer adaptive testing in K-16 education," Educational Measurement: Issues and Practice, vol. 43, no. 2, pp. 45–62, 2024. https://doi.org/10.1111/emip.12568
15. K. R. Koedinger, E. A. McLaughlin, and J. C. Stamper, "Automated student model improvement," International Journal of Artificial Intelligence in Education, vol. 34, no. 1, pp. 1–28, 2024. https://doi.org/10.1007/s40593-023-00344-3
16. L. Lang and J. A. Pirani, "Learning management systems: A comprehensive guide," EDUCAUSE, Louisville, CO, USA, 2023.
17. P. Leitner, P. P. Pranter, B. Brünner, and M. Ebner, "Empowering students through visual analytics: A dashboard redesign for modern curricula," in Proceedings of EdMedia + Innovate Learning, Brussels, Belgium, 2025, pp. 599–608.
18. Y. Lin, Q. Zhang, and L. Wang, "Adaptive testing in digital learning environments: A systematic review," Educational Research Review, vol. 38, p. 100118, 2023. https://doi.org/10.1016/j.edurev.2022.100118
19. H. Liu, Y. Ning, and X. Li, "Process data in educational assessment: A review and framework," Educational Psychology Review, vol. 36, no. 2, pp. 1–28, 2024.
20. M. Liu, Y. Ren, L. M. Nyagoga, F. Stonier, and L. Wu, "Large language models in educational settings: A systematic review of empirical studies," Computers and Education: Artificial Intelligence, vol. 10, p. 100132, 2025.
21. F. M. Lord, Applications of Item Response Theory to Practical Testing Problems, 2nd ed. New York, NY, USA: Routledge, 2023.
22. F. Martin, K. Budhrani, and C. Wang, "Examining faculty perception of their readiness to teach online," Online Learning, vol. 26, no. 1, pp. 97–119, 2022. https://doi.org/10.24059/olj.v26i1.2655
23. S. McKenney and T. C. Reeves, Conducting Educational Design Research, 2nd ed. London, UK: Routledge, 2023. https://doi.org/10.4324/9781003045243
24. B. Means, V. Peters, and Y. Zheng, "Lessons from five years of funding digital courseware: Postsecondary success portfolio review," SRI Education, Menlo Park, CA, USA, 2022.
25. I. Molenaar, C. Knoop-van Campen, and F. Hasselman, "The effects of AI feedback on students' self-regulated learning," Computers in Human Behavior, vol. 148, p. 107876, 2023. https://doi.org/10.1016/j.chb.2023.107876
26. F. Ouyang and G. Stanley, "Learning management systems in higher education: A systematic review of research," Educational Technology Research and Development, vol. 71, no. 3, pp. 987–1012, 2023. https://doi.org/10.1007/s11423-023-10215-4
27. E. Panadero, "A review of self-regulated learning: Six models and four directions for research," Frontiers in Psychology, vol. 14, p. 1120934, 2023. https://doi.org/10.3389/fpsyg.2023.1120934
28. J. Park and J. Lee, "Confidence in self-assessment: The role of metacognition in learning," Metacognition and Learning, vol. 18, no. 2, pp. 345–368, 2023. https://doi.org/10.1007/s11409-023-09345-2
29. R. Phillips and M. McNeill, "Learning management systems: Evolution and revolution," in Learning, Design, and Technology, J. M. Spector, B. B. Lockee, and M. D. Childress, Eds. Cham, Switzerland: Springer, 2022, pp. 1–24. https://doi.org/10.1007/978-3-319-17461-7_10
30. J. Reich and M. Ito, "From good intentions to real outcomes: Equity by design in learning technologies," Digital Promise, Washington, D.C., USA, 2023.

 doi:10.20944/preprints202605.0041.v1

31. C. Severance, T. Hanss, and J. Hardin, "IMS learning tools interoperability: Enabling educational innovation," IEEE Transactions on Learning Technologies, vol. 16, no. 4, pp. 512–525, 2023. https://doi.org/10.1109/TLT.2023.3261715
32. N. A. Thompson and D. J. Weiss, "A framework for the development of computerized adaptive tests," Practical Assessment, Research, and Evaluation, vol. 28, no. 1, pp. 1–15, 2023. https://doi.org/10.7275/pare.1862
33. UNESCO, "Global education monitoring report 2023: Technology in education – A tool on whose terms?" UNESCO Publishing, Paris, France, 2023.
34. K. VanLehn, "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems," Educational Psychologist, vol. 58, no. 3, pp. 131–152, 2023. https://doi.org/10.1080/00461520.2023.2217436
35. D. S. Walker, J. R. Lindner, T. P. Murphrey, and K. E. Dooley, "Learning management system usage: Perspectives from university instructors," Quarterly Review of Distance Education, vol. 22, no. 2, pp. 41–58, 2021.
36. S. Wang, C. Christensen, Y. Xu, and W. Cui, "Computerized adaptive testing in early childhood assessment: A review," Early Childhood Research Quarterly, vol. 58, pp. 234–246, 2022. https://doi.org/10.1016/j.ecresq.2021.09.008
37. X. Wang, Q. Liu, and N. S. Chen, "Learner-AI interaction: A scoping review of research," Educational Technology & Society, vol. 27, no. 2, pp. 1–18, 2024.
38. M. Weller, The Battle for Open: How Openness Won and Why It Doesn't Feel Like Victory. London, UK: Ubiquity Press, 2024. https://doi.org/10.5334/bbc
39. P. H. Winne, "Cognition and metacognition within self-regulated learning," in Handbook of Self-Regulation of Learning and Performance, 3rd ed., D. H. Schunk and J. A. Greene, Eds. New York, NY, USA: Routledge, 2024, pp. 35–52.
40. Y. Yang, X. Liu, and D. Gardner, "Multidimensional item response theory for measuring cognitive and metacognitive skills," Journal of Educational Measurement, vol. 62, no. 1, pp. 78–102, 2025.
41. C. Zhang et al., "A complete survey on generative AI: A meta-view of ChatGPT and beyond," ACM Computing Surveys, vol. 56, no. 4, pp. 1–38, 2024. https://doi.org/10.1145/3626239
42. N. Selwyn, Should Robots Replace Teachers? AI and the Future of Education. Cambridge, UK: Polity Press, 2024.
43. E. M. Rogers, Diffusion of Innovations, 5th ed. New York, NY, USA: Free Press, 2003.
44. S. P. Reise and D. A. Revicki, Eds., Handbook of Item Response Theory Modeling. New York, NY, USA: Routledge, 2022. https://doi.org/10.4324/9781315117058