algorithms

MDPI

Article

Dynamic Assessment with AI (Agentic RAG) and Iterative Feedback: A Model for the Digital Transformation of Higher Education in the Global EdTech Ecosystem

Rubén Juárez [1,*], Antonio Hernández-Fernández [2], Claudia de Barros-Camargo [3] and David Molero [2,4,*]

[1] School of Engineering, Science, and Technology, UNIE Universidad, Calle Arapiles, 14, 28015 Madrid, Spain
[2] Department of Education, Faculty of Humanities and Educational Sciences, University of Jaén, 23071 Jaén, Spain; ahernand@ujaen.es
[3] Department MIDE I, Faculty of Education, National University of Distance Education (UNED), 28040 Madrid, Spain; claudia.barros@edu.uned.es
[4] Research Group "Lifelong Education, Neuropedagogical Integration (LE:NI)", University of Jaén, 23071 Jaén, Spain
* Correspondence: ruben.juarez@universidadunie.com (R.J.); dmolero@ujaen.es (D.M.)

Abstract

This article formalizes AI-assisted assessment as a discrete-time *policy-level* design for iterative feedback and evaluates it in a digitally transformed higher-education setting. We integrate an *agentic* retrieval-augmented generation (RAG) feedback engine—operationalized through *planning* (rubric-aligned task decomposition), *tool use* beyond retrieval (tests, static/dynamic analyzers, rubric checker), and *self-critique* (checklist-based verification)—into a six-iteration dynamic evaluation cycle. Learning trajectories are modeled with three complementary formulations: (i) an interpretable update rule with explicit parameters η and λ that links next-step gains to feedback quality and the gap-to-target and yields iteration-complexity and stability conditions; (ii) a logistic-convergence model capturing diminishing returns near ceiling; and (iii) a relative-gain regression quantifying the marginal effect of feedback quality on the fraction of the gap closed per iteration. In a *Concurrent Programming* course ($n = 35$), the cohort mean increased from 58.4 to 91.2 (0–100), while dispersion decreased from 9.7 to 5.8 across six iterations; a Greenhouse–Geisser corrected repeated-measures ANOVA indicated significant within-student change. Parameter estimates show that higher-quality, evidence-grounded feedback is associated with larger next-step gains and faster convergence. Beyond performance, we engage the broader pedagogical question of *what to value and how to assess* in AI-rich settings: we elevate *process and provenance*—planning artifacts, tool-usage traces, test outcomes, and evidence citations—to first-class assessment signals, and outline defensible formats (trace-based walkthroughs and oral/code defenses) that our controller can instrument. We position this as a *design model for feedback policy*, complementary to state-estimation approaches such as knowledge tracing. We discuss implications for instrumentation, equity-aware metrics, reproducibility, and epistemically aligned rubrics. Limitations include the observational, single-course design; future work should test causal variants (e.g., stepped-wedge trials) and cross-domain generalization.

Keywords: AI-assisted assessment; agentic RAG; feedback policy design; interpretable update rule; planning and tool use; self-critique; logistic convergence; relative gain; learning analytics; equity of outcomes; process and provenance; AI-mediated higher education

check for updates

Academic Editors: Antonio Sarasa Cabezuelo and María Estefanía Avilés Mariño

Received: 6 October 2025
Revised: 28 October 2025
Accepted: 5 November 2025
Published: 11 November 2025

Citation: Juárez, R.; Hernández-Fernández, A.; de Barros-Camargo, C.; Molero, D. Dynamic Assessment with AI (Agentic RAG) and Iterative Feedback: A Model for the Digital Transformation of Higher Education in the Global EdTech Ecosystem. *Algorithms* **2025**, *18*, 712. https://doi.org/10.3390/a18110712

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Artificial intelligence (AI) is rapidly reshaping assessment foundations in higher education. Contemporary systems for automated feedback and intelligent tutoring report positive effects on performance and large-scale personalization; however, the *iterative* nature of assessment–feedback cycles remains under-theorized from a mathematical and algorithmic standpoint, limiting analyses of convergence, sensitivity, and robustness in learning processes [1–3]. Recent syntheses in AI for education summarize advances ranging from automated scoring for writing and programming to learning-analytics dashboards, while emphasizing mixed evidence and the need for reproducible, comparable frameworks across contexts [1,4,5].

In parallel, retrieval-augmented generation (RAG) has emerged as a key mechanism to inject reliable external knowledge into large language models, mitigating hallucinations and improving accuracy on knowledge-intensive tasks. The 2023–2024 survey wave systematizes architectures, training strategies, and applications, providing a technical basis for *contextualized and traceable* feedback in education [6–8]. Closely related prompting/reasoning frameworks (e.g., ReAct) support verifiable, tool-using feedback workflows [9]. *In this work, we use "agentic RAG" in an operational and auditable sense:* (i) **planning**—rubric-aligned task decomposition of the feedback job; (ii) **tool use** beyond retrieval—invoking tests, static/dynamic analyzers, and a rubric checker with logged inputs/outputs; and (iii) **self-critique**—a checklist-based verification for evidence coverage, rubric alignment, and actionability prior to delivery (details in Section 3.2).

Study-design note (clarifying iteration and deliverables). Throughout the paper, t indexes *six distinct, syllabus-aligned programming tasks* (not re-submissions of the same task). Each iteration comprises submission, auto-evaluation, feedback delivery, and student revision; the next performance $S_{i,t+1}$ is computed at the *subsequent task's* auto-evaluation after the revision leads to a new submission. The cap at six iterations is determined by the course calendar. We make this timing explicit in the Methods (Sections 3.1 and 3.3) and in the caption of Figure 1.

Within programming education, systematic reviews and venue reports (e.g., ACM Learning@Scale and EDM) document the expansion of auto-grading and LLM-based formative feedback, alongside open questions about reliability, transfer, and institutional scalability [10–14]. In writing, recent studies and meta-analyses report overall positive but heterogeneous effects of automated feedback, with moderators such as task type, feedback design, and outcome measures—factors that call for models capturing the *temporal* evolution of learning rather than single-shot performance [3,15]. Meanwhile, the knowledge tracing (KT) literature advances rich sequential models—from classical Bayesian formulations to Transformer-based approaches—typically optimized for *predictive fit* on next-step correctness rather than prescribing *algorithmic feedback policies* with interpretable convergence guarantees [16–20]. Our approach is complementary: KT focuses on *state estimation* of latent proficiency, whereas we formulate and analyze a *policy-level design* for the feedback loop itself, with explicit update mechanics.

Beyond improving traditional evaluations, AI-mediated higher education raises a broader pedagogical question: what forms of knowledge and understanding should be valued, and how should they be assessed when AI can already produce sophisticated outputs? In our framing, product quality is necessary but insufficient; we elevate *process and provenance* to first-class assessment signals. Concretely, the platform exposes (a) *planning artifacts* (rubric-aligned decomposition), (b) *tool-usage telemetry* (unit/integration tests, static/dynamic analyzers, and rubric checks), (c) *testing strategy and outcomes*, and (d) *evidence citations* from retrieval. These signals, natively logged by the agentic controller and connectors (Methods Section 3.2), enable *defensible formats*—trace-based walkthroughs, oral/code

defenses, and revision-under-constraints—that are auditable and scalable. We return to these epistemic implications in the Discussion (Section 5.4), where we outline rubric alignment to these signals and institutional deployment considerations.

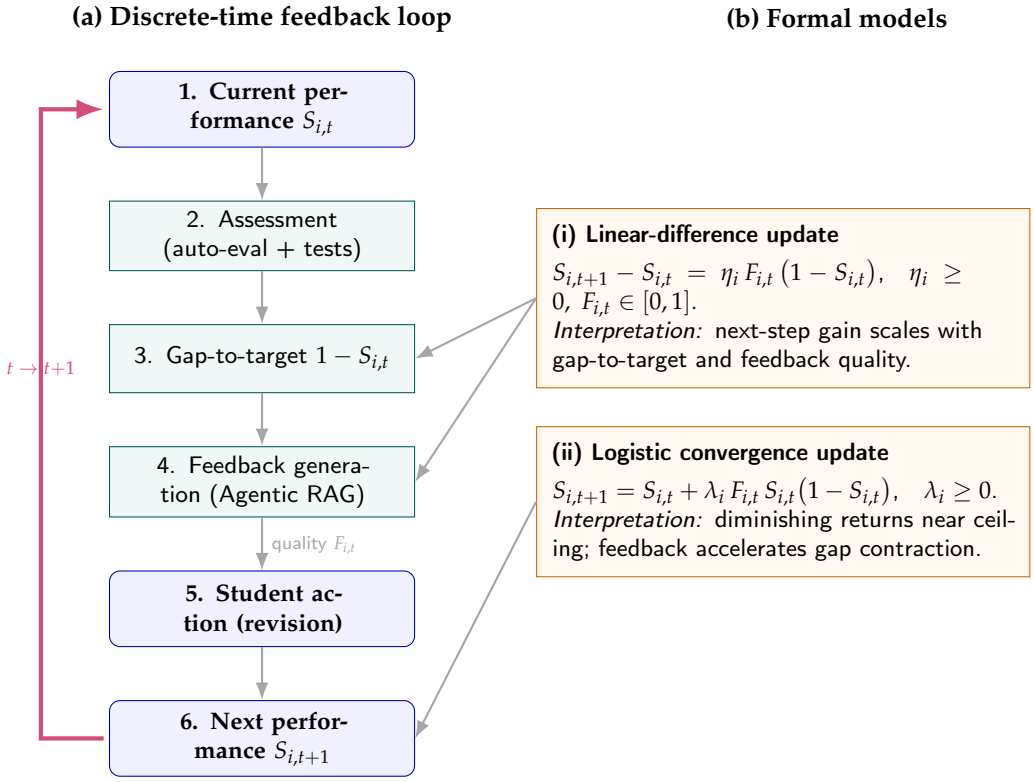


Figure 1. Algorithmic framing of assessment and feedback. (**a**) Student-level loop: performance $S_{i,t}$ is updated to $S_{i,t+1}$ through the targeted feedback of quality $F_{i,t}$. *Clarification:* t indexes six distinct course tasks; after revision (Step 5), students submit the next task, which is auto-evaluated and scored to produce $S_{i,t+1}$. (**b**) Two complementary discrete-time models: a linear-difference update and a logistic update used to analyze convergence rates, sensitivity to feedback quality, and variance dynamics.

This study frames assessment and feedback as a discrete-time, policy-level algorithmic process. We formalize two complementary models: (i) a difference equation linking per-iteration gain to the gap-to-target and feedback quality via interpretable parameters (η, λ), which yield iteration-complexity and stability conditions (Propositions 1 and 2), and (ii) a logistic convergence model describing the asymptotic approach to a performance objective. This framing enables analysis of convergence rates, sensitivity to feedback quality, and intra-cohort variance reduction, aligning educational assessment with tools for algorithm design and analysis. Empirically, we validate the approach in a longitudinal study with six feedback iterations in a technical programming course ($n = 35$), estimating model parameters via nonlinear regression and analyzing individual and group trajectories. Our results show that higher-quality, evidence-grounded feedback predicts larger next-iteration gains and faster convergence to target performance, while cohort dispersion decreases across cycles—patterns consistent with prior findings in intelligent tutoring, automated feedback, and retrieval-augmented LLMs [6–8,11].

Conceptual overview. Figure 1 depicts the student-level loop and its coupling with the formal models used throughout the paper. The process moves the performance state $S_{i,t}$ to $S_{i,t+1}$ via targeted feedback whose quality is summarized by $F_{i,t} \in [0,1]$. The two governing formulations, used later in estimation and diagnostics, are shown in panel (b): a linear-difference update and a logistic update, both expressed in discrete time and consistent with our methods.

Scope and contributions. The contribution is threefold: (1) a formal, interpretable *policy design* for iterative assessment with explicit update mechanics and parameters (η_i, λ_i) that connect *feedback quality* to the *pace* and *equity* of learning, enabling iteration-complexity and stability analyses (Propositions 1 and 2); (2) an empirical validation in a real course setting showing sizable gains in means and reductions in dispersion over six iterations (Section 4); and (3) an explicit pedagogical stance for AI-rich contexts that elevates *process and provenance* (planning artifacts, tool-usage traces, test outcomes, evidence citations) and outlines *defensible assessment formats* (trace-based walkthroughs, and oral/code defenses) that the platform can instrument at scale. This positions our work as a design model for the feedback loop itself, complementary to state-estimation approaches such as knowledge tracing.

2. Theoretical Framework

To ground our proposal of a dynamic, AI-supported assessment and feedback system within the broader digital transformation of higher education and the global EdTech landscape, this section reviews the most relevant theoretical and empirical research across educational assessment, feedback for learning, and Artificial Intelligence in Education (AIED), together with implications for pedagogy and evaluation in digitally mediated environments. We also consider a comparative-education perspective to contextualize the phenomenon internationally. Our goal is to provide a conceptual and analytical basis for understanding the design, implementation, and broader implications of the model advanced in this article.

Over the last decade—and especially since the emergence of generative AI—research on assessment in digital environments has accelerated. Multiple syntheses concur that feedback is among the most powerful influences on learning when delivered *personally*, *iteratively*, and *in context* [3,21,22]. In technically demanding domains such as programming, early error identification and precise guidance are critical for effective learning and scalable instruction [10,23,24]. Recent evidence further suggests that AI-supported automated feedback can achieve high student acceptability while raising challenges around factuality, coherence, and alignment with course objectives [4,11,15,25]. These observations motivate hybrid designs that combine generative models with information retrieval and tool use to improve the relevance, traceability, and verifiability of feedback.

2.1. Assessment and Feedback in Technical Disciplines and Digital Settings

Within the digital transformation of higher education, disciplines with high technical complexity and iterative skill formation (e.g., engineering, computational design, and especially programming) require assessment approaches that support the rapid, personalized, and precise adjustment of performance as students progress. Digital platforms facilitate content delivery and task management but amplify the need for *scalable formative feedback* that goes beyond grading to provide concrete, actionable guidance [3,21]. In programming education, research documents expansion in auto-grading, AI-mediated hints, and LLM-based formative feedback, alongside open questions about reliability, transfer, and equity at scale [10–14,24]. Addressing these challenges is essential to ensure that digital transformation translates into improved learning outcomes and readiness for technology-intensive labor markets.

2.2. Advanced AI for Personalized Feedback: RAG and Agentic RAG

Recent advances in AI have yielded models with markedly improved capabilities for interactive, context-aware generation. Retrieval-augmented generation (RAG) combines the expressive power of foundation models with the precision of targeted retrieval over curated knowledge sources, mitigating hallucinations and improving accuracy on knowledge-intensive tasks [6,7,26]. *Agentic* variants extend this paradigm with planning,

tool use, and self-critique cycles, enabling systems to reason over tasks, fetch evidence, and iteratively refine outputs [8,9].

Operational definition in this study.

We use "agentic RAG" in an explicit, auditable sense: (i) **planning**—rubric-aligned task decomposition of the feedback job; (ii) **tool use** beyond retrieval—invoking tests, static/dynamic analyzers, and a rubric checker with logged inputs/outputs; and (iii) **self-critique**—a checklist-based verification for evidence coverage, rubric alignment, and actionability prior to delivery. These capabilities are implemented as controller-enforced steps (not opaque reasoning traces), supporting reproducibility and auditability (see Sections 3 and 4; implementation details in Section 3.2). *Study-design crosswalk:* in the empirical setting, iteration t indexes six distinct, syllabus-aligned programming tasks (i.e., six deliverables over the term); the next state $S_{i,t+1}$ is computed at the subsequent task's auto-evaluation, after students revise based on the feedback from iteration t (Sections 3.1 and 3.3).

In educational contexts, connecting agentic RAG to course materials, assignment rubrics, student artifacts, and institutional knowledge bases—via standardized connectors or protocol-based middleware—supports feedback that is *course-aligned, evidence-grounded, and level-appropriate*. This integration enables detailed explanations, targeted study resources, and adaptation to learner state, making richer, adaptive feedback feasible at scale and illustrating how AI underpins disruptive innovation in core teaching-and-learning processes.

2.3. Epistemic and Pedagogical Grounding in AI-Mediated Assessment: What to Value and How to Assess

Beyond refining traditional evaluation, AI-rich contexts prompt a broader question: *what forms of knowledge and understanding should be valued, and how should they be assessed when AI can already produce sophisticated outputs?* In our framework, product quality is necessary but insufficient. We elevate *process and provenance* to first-class assessment signals and align them with platform instrumentation:

- **Planning artifacts:** rubric-aligned problem decomposition and rationale (controller step "plan"); mapped to rubric rows on problem understanding and strategy.
- **Tool-usage traces:** calls to unit/integration tests, static/dynamic analyzers, and rubric checker with inputs/outputs logged; mapped to testing adequacy, correctness diagnostics, and standards compliance.
- **Testing strategy and outcomes:** coverage, edge cases, and failing-to-passing transitions across iterations; mapped to engineering practice and evidence of improvement.
- **Evidence citations:** retrieved sources with inline references; mapped to factual grounding and traceability of feedback.
- **Revision deltas:** concrete changes from t to $t+1$ (files, functions, complexity footprints); mapped to actionability and responsiveness.

Task- and criterion-level structure. Each iteration t corresponds to a distinct task with a common four-criterion rubric (accuracy, relevance, clarity, and actionability). The Feedback Quality Index $F_{i,t}$ aggregates these criteria, while supporting criterion-level and task-level disaggregation (e.g., reporting distributions of $r_{i,t}^{(c)}$ by task, where $c \in \{1, \ldots, 4\}$ indexes rubric criteria). This enables rubric-sensitive interpretation of cohort gains and dispersion and motivates the per-criterion commentary provided in Sections 4 and 5.

These signals support *defensible assessment formats*—trace-based walkthroughs, oral/code defenses, and revision-under-constraints—without sacrificing comparability. They are natively produced by the agentic controller (Section 3), enabling equity-aware analytics (means, dispersion, and tails) and reproducible audits. While our Feedback Quality Index (FQI) aggregates *accuracy, relevance, clarity, actionability*, these process/provenance signals can be

reported alongside FQI, incorporated as covariates (e.g., $E_{i,t}$ in Equation (7)), or organized in an *epistemic alignment matrix* (Appendix A.1) that links constructs to rubric rows and platform traces; complementary exploratory associations and timing checks are provided in (Appendix A.2), and criterion-by-task distributions in (Appendix A.3). This grounding complements the formal models by clarifying *what* is valued and *how* it is evidenced at scale.

2.4. Mathematical Modeling of Assessment–Feedback Dynamics

Beyond transforming tools and workflows, the digitalization of learning generates rich longitudinal data about how students improve in response to instruction and iterative feedback. Mathematical modeling provides a principled lens to capture these dynamics, shifting the focus from single-shot outcomes to *trajectories* of performance over time. In systems that allow multiple attempts and continuous feedback, discrete-time updates are natural candidates: they describe how a learner's performance is updated between evaluation points as a function of the previous state, the gap-to-target, and the quality of feedback. Throughout the paper, we consider two complementary formulations at the student level i and iteration t:

$$S_{i,t+1} - S_{i,t} = \eta_i F_{i,t} \left(1 - S_{i,t}\right) + \varepsilon_{i,t}, \quad \eta_i \geq 0,\ F_{i,t} \in [0,1], \tag{1}$$

$$S_{i,t+1} = S_{i,t} + \lambda_i F_{i,t} S_{i,t}\left(1 - S_{i,t}\right) + \varepsilon_{i,t}, \quad \lambda_i \geq 0. \tag{2}$$

Here $S_{i,t} \in [0,1]$ denotes a normalized performance score (with target $S_{\text{target}} = 1$), $F_{i,t}$ summarizes feedback quality (accuracy, relevance, clarity, actionability), η_i and λ_i parameterize sensitivity and effectiveness, and $\varepsilon_{i,t}$ captures unmodeled shocks. Crucially, *these are policy-level update rules*: they prescribe how the system should act (via feedback quality) to contract the learning gap at a quantifiable rate, rather than estimating a latent proficiency state.

Proposition 1 (Monotonicity, boundedness, and iteration complexity for (1))**.** *Assume* $\varepsilon_{i,t} \equiv 0$, $S_{i,1} \in [0,1]$, $\eta_i \in [0,1]$, *and* $F_{i,t} \in [0,1]$. *Then, we have the following:*

1. (Monotonicity and boundedness) $S_{i,t}$ *is nondecreasing and remains in* $[0,1]$ *for all* t.
2. (Geometric convergence) *If there exists* $f_{\min} > 0$ *such that* $F_{i,t} \geq f_{\min}$ *for all* t, *then*

$$1 - S_{i,t} \ \leq\ (1 - \eta_i f_{\min})^{\,t-1}\,(1 - S_{i,1}).$$

3. (Iteration complexity) *To achieve* $1 - S_{i,t} \leq \delta$ *with* $0 < \delta < 1$, *it suffices that*

$$t \ \geq\ 1 + \frac{\log(\delta/(1 - S_{i,1}))}{\log(1 - \eta_i f_{\min})}.$$

Proposition 2 (Stability and convergence for (2))**.** *Assume* $\varepsilon_{i,t} \equiv 0$, $S_{i,1} \in (0,1)$, *and let* $r_{i,t} = \lambda_i F_{i,t}$.

1. (Local stability at the target) *If* $0 < \sup_t r_{i,t} < 2$, *then* $S^\star = 1$ *is locally asymptotically stable. In particular, if* $r_{i,t} \in (0,1]$ *for all* t, *then* $S_{i,t}$ *increases monotonically to* 1.
2. (Convergence without oscillations) *If* $0 < \sup_t r_{i,t} \leq 1$, *then* $S_{i,t}$ *is nondecreasing and converges to* 1 *without overshoot.*

Proof. Define $G(S) = S + \lambda_i F S(1 - S)$ with $F \in [0,1]$. Fixed points satisfy $G(S) = S$, giving $S \in \{0,1\}$. The derivative $G'(S) = 1 + \lambda_i F(1 - 2S)$ yields $G'(1) = 1 - \lambda_i F$. Local stability requires $|G'(1)| < 1$, i.e., $0 < \lambda_i F < 2$. If $0 < \lambda_i F \leq 1$, then $G'(S) \in (0,1]$ for $S \in [0,1]$, so the map is increasing and contractive near the target, implying monotone convergence. □

Corollary 1 (Cohort variance contraction (linearized)). *Let $\bar{S}_t$ be the cohort mean, $\sigma_t^2 = \mathrm{Var}(S_{i,t})$, and suppose shocks $\varepsilon_{i,t}$ are independent across students with variance σ_ε^2. Linearizing (1) around $\bar{S}_t$ and defining $\bar{F}_t$ as the cohort-average feedback quality at iteration t,*

$$\sigma_{t+1}^2 \approx \left(1 - \bar{\eta}\,\bar{F}_t\right)^2 \sigma_t^2 + \sigma_\varepsilon^2.$$

Hence, if $0 < 1 - \bar{\eta}\,\bar{F}_t < 1$ and σ_ε^2 is small, dispersion contracts geometrically toward a low-variance regime, aligning equity improvements with iterative feedback.

Rule-of-thumb half-life.

For (1) with $F_{i,t} \geq f_{\min} > 0$ and $\varepsilon_{i,t} = 0$, the *gap half-life* satisfies $t_{1/2} \approx \ln 2 \,/\, (\eta_i f_{\min})$, linking estimated η_i and a lower bound on feedback quality to expected pacing. With mean-zero noise, a Lyapunov potential $V_{i,t} = (1 - S_{i,t})^2$ yields $\mathbb{E}[V_{i,t+1}] \leq \rho\,\mathbb{E}[V_{i,t}] + \sigma_\varepsilon^2$ for $\rho = \sup_t \mathbb{E}[(1 - \eta_i F_{i,t})^2] < 1$, implying bounded steady-state error. These properties justify monitoring both mean trajectories and dispersion (σ_t) as first-class outcomes.

2.5. Relation to Knowledge Tracing and Longitudinal Designs

This perspective resonates with—but is distinct from—the knowledge tracing literature. KT offers powerful sequential predictors (from Bayesian variants to Transformer-based approaches), yet the emphasis is often on *predictive fit* (next-step correctness) rather than prescribing *feedback policies* with interpretable convergence guarantees and explicit update mechanics [16–20]. Our formulation foregrounds the *policy*: a mapping from current state and feedback quality to the next state with parameters (η, λ) that enable stability analysis, iteration bounds, and variance dynamics (Propositions 1 and 2, Corollary 1).

Complementarity. KT can inform design by supplying calibrated state estimates or difficulty priors that modulate $F_{i,t}$ (e.g., stricter scaffolding and targeted exemplars when KT indicates fragile mastery). This preserves analytical tractability while exploiting the KT sequential inference for policy targeting. Methodologically, randomized and longitudinal designs in AIED provide complementary strategies for estimating intervention effects and validating iterative improvement [5]. In our empirical study (Sections 3 and 4), we instantiate this foundation with six iterations and report both mean trajectories and dispersion, together with parameter estimates that connect feedback quality to the pace and equity of learning.

2.6. Comparative-Education Perspective

From a comparative-education viewpoint, the algorithmic framing of assessment raises cross-system questions about adoption, policy, and equity: how do institutions with different curricula, languages, and governance structures instrument feedback loops; how is feedback quality $F_{i,t}$ ensured across contexts; and which safeguards (privacy, auditability, and accessibility) condition transferability at scale? Because the models here are *interpretable* and rely on auditable quantities ($S_{i,t}$, $F_{i,t}$, and dispersion σ_t), they are amenable to standardized reporting across institutions and countries—facilitating international comparisons and meta-analyses that move beyond single-shot accuracy to longitudinal, equity-aware outcomes.

By framing assessment and feedback as a discrete-time algorithm with explicit update mechanics, we connect pedagogical intuition to tools from dynamical systems and stochastic approximation. This yields actionable parameters (η_i, λ_i), interpretable stability conditions ($\lambda_i F_{i,t} < 2$), iteration bounds (Proposition 1), and cohort-level predictions (variance contraction; Corollary 1) that inform the design of scalable, equity-aware feedback in digitally transformed higher education, *while making explicit the mapping between iterations and deliverables* to avoid ambiguity in empirical interpretations.

3. Materials and Methods

3.1. Overview and Study Design

We conducted a longitudinal observational study with six consecutive evaluation iterations ($t = 1, \ldots, 6$) to capture within-student learning dynamics under AI-supported assessment. The cohort comprised $n = 35$ students enrolled in a *Concurrent Programming* course, selected for its sequential and cumulative competency development. Each iteration involved solving practical programming tasks, assigning a calibrated score, and delivering *personalized, AI-assisted feedback*. Scores were defined on a fixed scale and rescaled to $[0, 1]$ for modeling, with $S_{\text{target}} = 1$. Feedback quality was operationalized as $F_{i,t} \in [0, 1]$ (Section 3.3).

Task structure and iteration semantics.

Each iteration t corresponds to a *distinct, syllabus-aligned deliverable* (six core tasks across the term). Students submit the task for iteration t, receive $S_{i,t}$ and AI-assisted feedback $f_{i,t}$, and then *revise their approach and code base for the **next** task at* $t+1$. Within-task micro-revisions (e.g., rerunning tests while drafting) may occur but do *not* generate an additional score within the same iteration; the next evaluated score is $S_{i,t+1}$ at the subsequent task. This design ensures comparability across iterations and aligns with course pacing.

Why six iterations?

The cap of six iterations is set by the course syllabus, which comprises six major programming tasks (modules) that build cumulatively (e.g., synchronization, deadlocks, thread-safe data structures, concurrent patterns, and performance). The platform can support more frequent micro-cycles, but for this study we bound t to the six course deliverables to preserve instructional fidelity and cohort comparability.

Outcomes and endpoints.

The *primary outcome* is the per-iteration change in scaled performance, $S_{i,t+1} - S_{i,t}$, and its dependence on feedback quality $F_{i,t}$ (Equations (3) and (4)). *Secondary outcomes* include (i) the relative gain $G_{i,t}$ (Equation (6)), (ii) cohort dispersion σ_t (SD of $S_{i,t}$) and tail summaries (Q10/Q90), (iii) interpretable parameters $(\eta_i, \lambda_i, \beta_1)$ linking feedback quality to pace and equity of learning, and (iv) *exploratory process/provenance signals* (planning artifacts, tool-usage traces, testing outcomes, evidence citations, revision deltas) aligned with the broader pedagogical aims of AI-mediated assessment (see Section 3.4).

3.2. System Architecture for Feedback Generation

The system integrates three components under a discrete-time orchestration loop:

1. **Agentic RAG feedback engine.** A retrieval-augmented generation pipeline with agentic capabilities (planning, tool use, and self-critique) that produces *course-aligned, evidence-grounded* feedback tailored to each submission. Retrieval uses a top-k dense index over course artifacts; evidence citations are embedded in the feedback for auditability.
2. **Connector/middleware layer (MCP-like).** A standardized, read-only access layer brokering secure connections to student code and tests, grading rubrics, curated exemplars, and course documentation. The layer logs evidence references, model/version, and latency for traceability.
3. **Auto-evaluation module.** Static/dynamic analyses plus unit/integration tests yield diagnostics and a preliminary score; salient findings are passed as structured signals to contextualize feedback generation.

All components operate within an auditable controller that records inputs/outputs per iteration and enforces privacy-preserving pseudonymization before analytics.

Usage-based *agentic* features (explicit and auditable).

We implement "agentic RAG" as a sequence of controller-enforced actions (not opaque reasoning traces):

- **Planning** (rubric-aligned task decomposition): the controller compiles a plan with sections (e.g., correctness, concurrency hazards, and style) and required deliverables (fix-steps and code pointers).
- **Tool use** (beyond retrieval), with logged inputs/outputs and versions: (i) *test-suite runner* (unit/integration); (ii) *static/dynamic analyzers* (e.g., race/deadlock detectors); (iii) a *rubric-check microservice* that scores coverage/level; (iv) a *citation formatter* that binds evidence IDs into the feedback.
- **Self-critique** (checklist-based verification): a bounded one-round verifier checks (1) evidence coverage, (2) rubric alignment, (3) actionability (step-wise fixes), and (4) clarity/consistency. Failing checks trigger exactly one constrained revision.

Controller-enforced action trace (abridged; no internal reasoning).

S (system): "You are a feedback agent for *Concurrent Programming*. Use only auditable actions; cite evidence IDs."
C (context): $\{x_{i,t}, d_{i,t}, \mathcal{R}, \mathcal{E}, \mathcal{M}\}$
A1 (plan): `make_plan(sections=[correctness, concurrency, style], deliverables=[fix-steps, code-links])`
A2 (retrieve): `retrieve_topk(k)` $\rightarrow$ {evidence IDs}
T1 (tests): `run_tests(`$\mathcal{T}$`)` $\rightarrow$ {failing:test_3}
T2 (analysis): `static_analyze(x)` $\rightarrow$ {data_race: line 84}
T3 (rubric): `rubric_check(draft)` $\rightarrow$ {coverage:0.92}
G1 (draft): `generate_feedback(context+evidence)` $\rightarrow$ draft_v1
V1 (self-critique): `checklist(draft_v1)` $\rightarrow$ {missing code pointer} $\Rightarrow$ `revise()` $\rightarrow$ draft_v2
F (finalize): `attach_citations(draft_v2, evidence_ids)` $\rightarrow f_{i,t}$; log {tool calls, versions, latencies}

3.3. Dynamic Assessment Cycle

Each cycle ($t = 1, \ldots, 6$) followed five phases:

1. **Submission.** Students solved a syllabus-aligned concurrent-programming task.
2. **Auto-evaluation.** The system executed the test suite and static/dynamic checks to compute $S_{i,t} \in [0, 1]$ and extract diagnostics $d_{i,t}$.
3. **Personalized feedback (Agentic RAG).** Detailed, actionable comments grounded on the submission, rubric, and retrieved evidence were generated and delivered together with $S_{i,t}$.
4. **Feedback Quality Index.** Each feedback instance was rated on *Accuracy*, *Relevance*, *Clarity*, and *Actionability* (5-point scale); the mean was linearly normalized to $[0, 1]$ to form $F_{i,t}$. A stratified 20% subsample was double-rated for reliability (Cohen's κ) and internal consistency (Cronbach's α).
5. **Revision.** Students incorporated the feedback to prepare the next submission. Operationally, feedback from $t-1$ informs the change observed at t.

Evaluation timing between Steps 5 and 6.

There is no *additional* score issued immediately after Step 5 within the same iteration. The next evaluated score is $S_{i,t+1}$ at the start of the subsequent iteration (Step 2 of $t+1$). Internally, tests/analyzers may be re-run while drafting feedback (tool-use traces), but these do not constitute interim grading events.

In Figure 2, the discrete-time assessment–feedback cycle couples student submissions, auto-evaluation and agentic RAG feedback with the modeling layer defined in Equations (3)–(7).

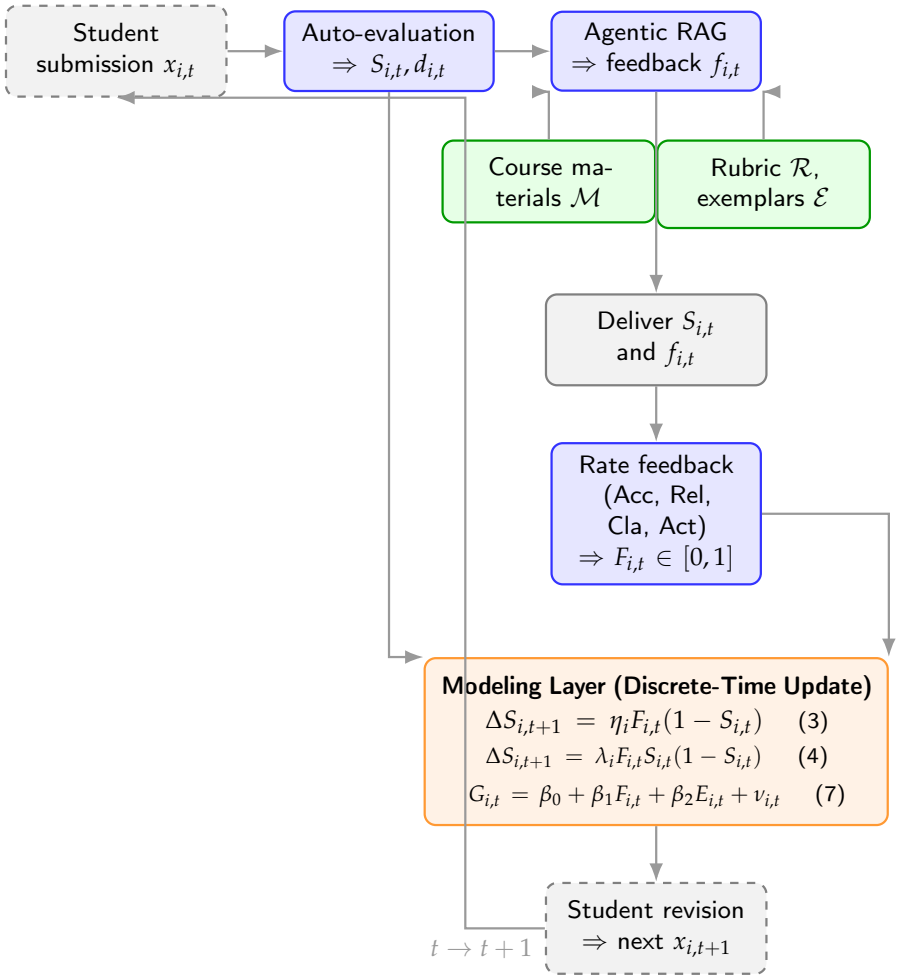


Figure 2. Discrete-time assessment–feedback workflow with lateral data sources grouped under RAG (see Equations (3)–(7)). Rounded rectangles denote processes; plain rectangles denote data artifacts; solid arrows denote data flow; dashed arrows denote control/inference. Layout adjusted to avoid overlap.

3.4. Process and Provenance Signals and Epistemic Alignment

To address the broader educational question of *what to value and how to assess* in AI-rich contexts, we treat *process and provenance* as first-class signals that the platform natively logs:

- **Planning artifacts:** rubric-aligned decomposition and rationale produced at action **A1**.
- **Tool-usage telemetry:** calls and outcomes for **T1–T3** (tests, static/dynamic analyzers, and rubric check), with versions and inputs/outputs.
- **Testing strategy and outcomes:** coverage/edge cases; failing-to-passing transitions across iterations.
- **Evidence citations:** retrieved sources bound into feedback at **F**, enabling traceability.
- **Revision deltas:** concrete changes between t and $t{+}1$ (files/functions touched; error classes resolved).

These signals can (i) be reported descriptively alongside $S_{i,t}$ and $F_{i,t}$, (ii) enter Equation (7) as covariates $\mathbf{p}_{i,t}$ or as moderators via interactions $F_{i,t} \times \mathbf{p}_{i,t}$, and (iii) inform *defensible assessment formats* (trace-based walkthroughs; oral/code defenses; revision-under-constraints) without sacrificing comparability. A concise *epistemic alignment matrix* mapping

constructs to rubric rows and platform traces is provided in (Appendix A.1) and referenced in Discussion Section 5.4.

Criterion- and task-level reporting plan.

In addition to the aggregate FQI, we pre-specify *criterion-level* summaries (accuracy, relevance, clarity, and actionability) by *task* and by *iteration*, as well as distributions of per-criterion ratings $r_{i,t}^{(c)}$. This supports analyses such as: "Given that $\bar{S}_6 > 90$, which criteria most frequently reach level 5, and in which tasks?" Corresponding summaries are reported in the Results (Section 4); the criterion-by-task breakdowns are compiled in Appendix A.3 and discussed in Section 4.

3.5. Model Specifications

We formalize three complementary formulations that capture how iterative feedback influences performance trajectories. These definitions appear here *for the first time* and are referenced throughout using `\eqref{}`. Importantly, these are *policy-level* update rules (design of the feedback loop), consistent with the theoretical results on iteration complexity and stability.

(1) Linear difference model.

$$S_{i,t+1} - S_{i,t} \;=\; \eta_i \, F_{i,t} \left(1 - S_{i,t}\right) \;+\; \varepsilon_{i,t}, \tag{3}$$

where $\eta_i \geq 0$ encodes individual sensitivity to feedback and $\varepsilon_{i,t}$ captures unexplained variation. Improvement is proportional to the gap-to-target and modulated by feedback quality and learner responsiveness.

(2) Logistic convergence model.

$$S_{i,t+1} \;=\; S_{i,t} \;+\; \lambda_i \, F_{i,t} \, S_{i,t}(1 - S_{i,t}) \;+\; \varepsilon_{i,t}, \tag{4}$$

with $\lambda_i \geq 0$ governing how feedback accelerates convergence. In multiplicative-gap form,

$$1 - S_{i,t+1} \;\approx\; \left(1 - S_{i,t}\right) \exp\!\left(-\lambda_i \, F_{i,t}\right), \tag{5}$$

which makes explicit that higher-quality feedback contracts the remaining gap faster. A useful rule-of-thumb is the *gap half-life*:

$$t_{1/2} \;\approx\; \frac{\ln 2}{\lambda_i F_{i,t}},$$

mapping estimated λ_i and observed $F_{i,t}$ to expected pacing.

(3) Relative-gain model.

We define the per-iteration fraction of the remaining gap that is closed:

$$G_{i,t} \;=\; \frac{S_{i,t+1} - S_{i,t}}{1 - S_{i,t}} \;\in\; [0,1] \quad (S_{i,t} < 1), \tag{6}$$

and regress

$$G_{i,t} \;=\; \beta_0 \;+\; \beta_1 F_{i,t} \;+\; \beta_2 E_{i,t} \;+\; \mathbf{d}_t^{\top} \boldsymbol{\gamma} \;+\; \nu_{i,t}, \tag{7}$$

where $E_{i,t}$ optionally captures effort/time-on-task; $\mathbf{d}_t$ are iteration fixed effects (time trends and task difficulty); and $\nu_{i,t}$ is an error term. The coefficient β_1 estimates the average marginal effect of feedback quality on progress per iteration, net of temporal and difficulty factors. In sensitivity analyses, we optionally augment the specification with a process/provenance vector $\mathbf{p}_{i,t}$ and the interaction term $F_{i,t} \times \mathbf{p}_{i,t}$.

3.6. Identification Strategy, Estimation, and Diagnostics

Identification and controls.

Given the observational design, we mitigate confounding via (i) within-student modeling (student random intercepts; cluster-robust inference), (ii) iteration fixed effects $\mathbf{d}_t$ to partial out global time trends and task difficulty, and (iii) optional effort covariates $E_{i,t}$ where available. In sensitivity checks, we add lagged outcomes $S_{i,t-1}$ (where appropriate) and verify that inferences on $F_{i,t}$ remain directionally stable. Where recorded, we include $\mathbf{p}_{i,t}$ (process/provenance) as covariates or moderators.

Estimation.

Equations (3) and (4) are estimated by nonlinear least squares with student-level random intercepts (and random slopes where identifiable), using cluster-robust standard errors at the student level. Equation (7) is fit as a linear mixed model with random intercepts by student and fixed effects $\mathbf{d}_t$. Goodness-of-fit is summarized with RMSE/MAE (levels) and R^2 (gains); calibration is assessed via observed vs. predicted trajectories. Model comparison uses AIC/BIC and out-of-sample K-fold cross-validation.

Multiple testing and robustness.

We report 95% confidence intervals and adjust p-values using the Benjamini–Hochberg procedure where applicable. Robustness checks include the following: (i) trimming top/bottom 2.5% changes, (ii) re-estimation with Huber loss, (iii) alternative weighting schemes in $F_{i,t}$ (e.g., upweight Accuracy/Actionability), and (iv) a placebo timing test regressing $S_{i,t} - S_{i,t-1}$ on *future* $F_{i,t}$ to probe reverse-timing artefacts (expected null).

Preprocessing, missing data, and participation counts.

We normalize $S_{i,t}$ and $F_{i,t}$ to the $[0, 1]$ scale. When $F_{i,t}$ is missing for a single iteration, we apply last-observation-carried-forward (LOCF) and conduct sensitivity checks using complete cases and within-student mean imputation. Students with ≥ 2 consecutive missing iterations are excluded from model-based analyses but retained in descriptive summaries. To contextualize inference given the moderate cohort size ($n = 35$), per-iteration participation counts n_t are reported in Results (Section 4).

In Figure 3, we organize the estimation workflow from longitudinal inputs to model fitting, diagnostics, and study outputs, matching the identification and validation procedures detailed in this section.

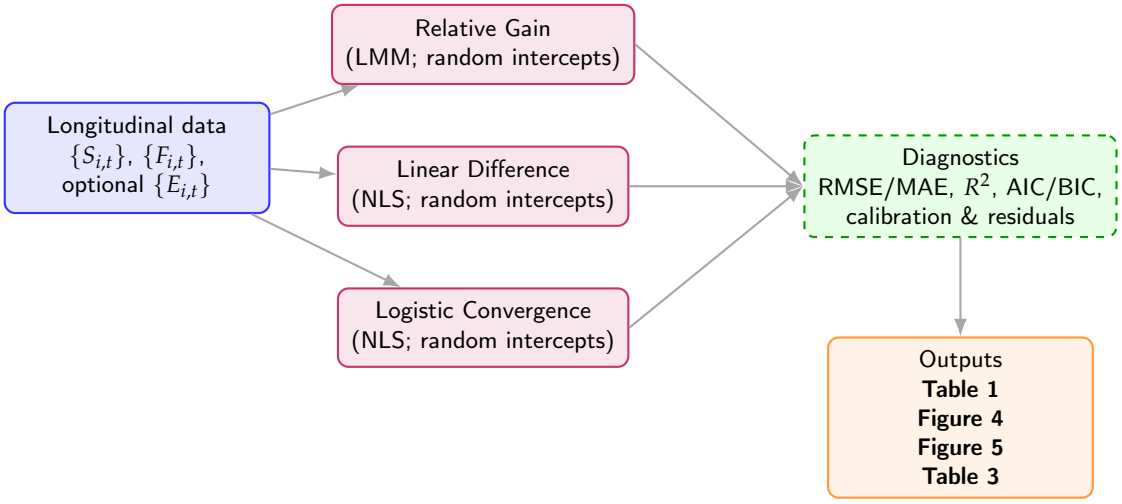


Figure 3. Estimation pipeline from longitudinal data to diagnostics and study outputs (see Equations (3)–(7)), including per-iteration participation counts (reported in Section 4). Inputs: submissions $x_{i,t}$, diagnostics $d_{i,t}$, rubric $\mathcal{R}$, exemplars $\mathcal{E}$, course materials $\mathcal{M}$, and test suite $\mathcal{T}$. Processing: normalization and scoring to obtain $S_{i,t}$ and the Feedback Quality Index $F_{i,t}$. Modeling: estimation of linear-difference, logistic-convergence, and relative-gain formulations; out-of-sample validation and

calibration. Diagnostics/outputs: iteration-level descriptives, reliability, robustness/sensitivity analyses, and study tables/figures. Rounded rectangles denote processes; plain rectangles denote data artifacts; solid arrows denote data flow; dashed arrows denote control/inference.

3.7. Course AI-Use Policy and Disclosure

Students received an explicit course policy specifying permissible AI assistance (e.g., debugging hints, explanation of diagnostics, literature lookup) and requiring disclosure of AI use in submissions; the platform's controller logs (planning artifacts, tool calls, and citations) support provenance checks without penalizing legitimate assistance. Self-reports were collected as optional meta-data and may be cross-checked with telemetry for plausibility.

3.8. Threats to Validity and Mitigations

Internal validity. Without randomized assignment of feedback pathways, causal claims are cautious. We partially address confounding via within-student modeling, iteration fixed effects (time/difficulty), and sensitivity analyses (lagged outcomes; trimming; Huber). Practice and ceiling effects are explicitly modeled by the gap-to-target terms in (3) and (4).

Construct validity. The Feedback Quality Index aggregates four criteria; we report inter-rater agreement (Cohen's κ) and internal consistency (Cronbach's α) in Section 4. Calibration plots and residual diagnostics ensure score comparability across iterations. In AI-rich settings, potential construct shift is mitigated by process/provenance logging (plan, tests, analyzers, citations, and revision deltas) and disclosure (Section 3.7), which align observed outcomes with epistemic aims.

External validity. Results originate from one course and institution with $n = 35$. Transferability to other disciplines and contexts requires multi-site replication (see Discussion). Equity-sensitive outcomes (dispersion σ_t, tails) are included to facilitate cross-context comparisons.

3.9. Software, Versioning, and Reproducibility

Analyses were conducted in Python 3.12 (NumPy, SciPy, StatsModels). We record random seeds, dependency versions, and configuration files (YAML) and export an environment lockfile for full reproducibility. Estimation notebooks reproduce all tables/figures and are available upon request; audit logs include model/version identifiers and retrieval evidence IDs.

3.10. Data and Code Availability

The dataset (scores, feedback-quality indices, and model-ready covariates) is available from the corresponding author upon reasonable request, subject to institutional policies and anonymization standards. Model scripts and configuration files are shared under an academic/research license upon request. Appendix A.1 documents the epistemic alignment matrix (constructs $\leftrightarrow$ rubric rows $\leftrightarrow$ platform traces).

3.11. Statement on Generative AI Use

During manuscript preparation, ChatGPT (OpenAI, 2025 version) was used exclusively for language editing and stylistic reorganization. All technical content, analyses, and results were produced, verified, and are the sole responsibility of the authors.

3.12. Ethics

Participation took place within a regular course under informed consent and full pseudonymization prior to analysis. The study was approved by the Research Ethics Committee of Universidad de Jaén (Spain), approval code JUL.22/4-LÍNEA. Formal

statements appear in the back matter (Institutional Review Board Statement, Informed Consent Statement).

3.13. Algorithmic Specification and Visual Summary

In what follows, we provide an operational specification of the workflow. Algorithm 1 details the step-by-step *iterative dynamic assessment cycle with agentic RAG*, from submission intake and automated evaluation to evidence-grounded feedback delivery and quality rating across six iterations. Algorithm 2 complements this by formalizing the computation of the Feedback Quality Index $F_{i,t}$ from criterion-level ratings and by quantifying reliability via linear-weighted Cohen's κ (on a 20% double-rated subsample) and Cronbach's α (across the four criteria). Together, these specifications capture both the *process layer* (workflow) and the *measurement layer* (scoring and reliability) required to reproduce our analyses.

Algorithm 1 Iterative dynamic assessment cycle with agentic RAG.

Require: Course materials $\mathcal{M}$, rubric $\mathcal{R}$, exemplars $\mathcal{E}$, test suite $\mathcal{T}$; cohort $\mathcal{I}$; $S_{\text{target}} = 1$
Initialize connectors (MCP-like), audit logs, and pseudonymization
for $t = 1$ **to** 6 **do** ▷ Discrete-time learning loop
  for each student $i \in \mathcal{I}$ **do**
    Receive submission $x_{i,t}$
    Auto-evaluation: run $\mathcal{T}$ + static/dynamic checks $\Rightarrow$ diagnostics $d_{i,t}$; compute $S_{i,t} \in [0,1]$
    Build context $C_{i,t} \leftarrow \{x_{i,t}, d_{i,t}, \mathcal{R}, \mathcal{E}, \mathcal{M}\}$
    Agentic RAG: retrieve top-k evidence; draft $\rightarrow$ self-critique $\rightarrow$ finalize feedback $f_{i,t}$
    Deliver $f_{i,t}$ and $S_{i,t}$ to student i
    Feedback Quality Rating: rate {accuracy, relevance, clarity, actionability} on 1–5
    Normalize/aggregate $\Rightarrow F_{i,t} \in [0,1]$; (optional) collect $E_{i,t}$ (effort/time)
    (Optional inference) update predictions via (3), (4), (7)
    Log $\{S_{i,t}, F_{i,t}, E_{i,t}, d_{i,t}\}$ with pseudonym IDs
  end for
end for
Output: longitudinal dataset $\{S_{i,t}\}$, $\{F_{i,t}\}$, optional $\{E_{i,t}\}$; evidence/audit logs

Algorithm 2 Computation of $F_{i,t}$ and reliability metrics (κ, α).

Require: Feedback instances $\{f_{i,t}\}$ with rubric ratings $r_{i,t}^{(c)} \in \{1,\ldots,5\}$ for $c \in \{\text{Acc}, \text{Rel}, \text{Cla}, \text{Act}\}$; 20% double-rated subsample $\mathcal{D}$
Ensure: $F_{i,t} \in [0,1]$; Cohen's κ on $\mathcal{D}$; Cronbach's α across criteria
for each feedback instance (i,t) **do**
  Handle missing ratings: if any $r_{i,t}^{(c)}$ missing, impute with within-iteration criterion mean
  for each criterion c **do**
    $s_{i,t}^{(c)} \leftarrow (r_{i,t}^{(c)} - 1)/4$ ▷ Normalize to $[0,1]$
  end for
  Aggregate: $F_{i,t} \leftarrow \frac{1}{4}\sum_c s_{i,t}^{(c)}$ ▷ Equal weights; alternative weights in Section 3.6
end for
Inter-rater agreement (κ): compute linear-weighted Cohen's κ on $\mathcal{D}$
Internal consistency (α): with $k = 4$ criteria, compute Cronbach's $\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_c v_c}{v_{\text{total}}}\right)$
Outputs: $\{F_{i,t}\}$ for modeling; κ and α reported in Results

4. Results

This section presents quantitative evidence for the effectiveness of the AI-supported dynamic-assessment and iterative-feedback system. We first report model-based parameter estimates for the three formulations (linear difference, logistic convergence, and relative

gain; Table 1). We then describe cohort-level dynamics across six iterations—means and dispersion (Figure 4) and variance contraction (Figure 5)—with numerical companions in Table 2 and per-iteration participation counts in Table 3; the repeated-measures ANOVA is summarized in Table 4. Next, we illustrate heterogeneous responsiveness via simulated individual trajectories (Figure 6). We subsequently summarize predictive performance (Table 5) and calibration (Figure 7). We conclude with placebo-timing tests, sensitivity to missingness, and robustness checks, with additional details compiled in Appendix A.

Parameter estimates for the three model formulations are summarized in Table 1.

Table 1. Parameter estimates for the linear-difference, logistic-convergence, and relative-gain models. Values reflect cohort-level summaries of student-level estimates (see Section 3).

Model	Parameter	Estimate	SE	95% CI		Notes
				Lower	Upper	
Linear-difference	η (mean)	0.320	0.060	0.203	0.437	Mixed-effects NLS; $p < 0.01$.
Logistic-convergence	λ (mean)	0.950	0.150	0.656	1.244	$S_{\text{target}} = 1$; $p < 0.01$.
Relative-gain	β_1 (effect of F)	0.280	0.070	0.143	0.417	LMM with student random effects; $p < 0.01$.

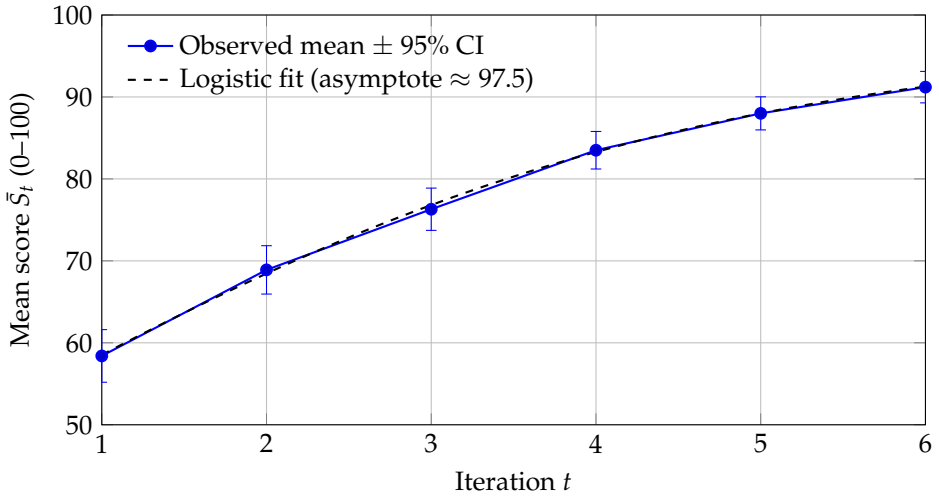


Figure 4. Average cohort trajectory across six iterations (0–100 scale). Points: observed means with 95% CIs; dashed curve: shifted-logistic fit closely matching observed means.

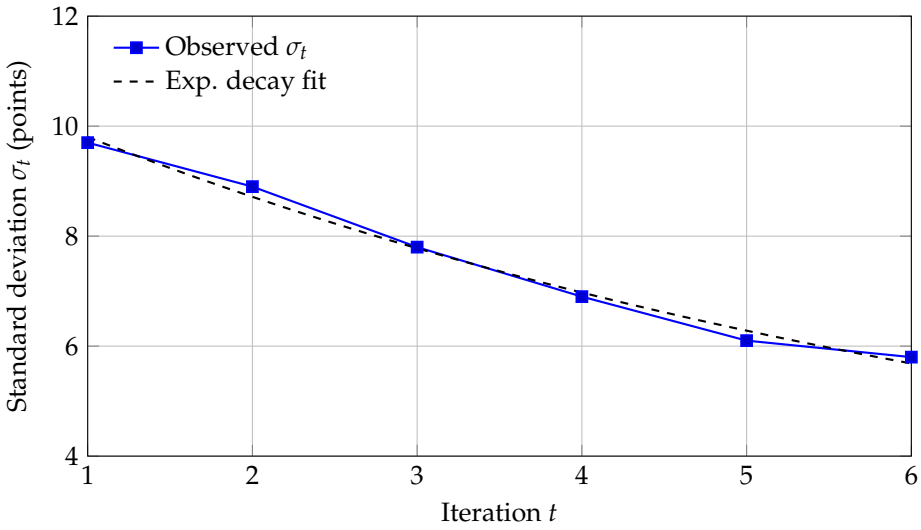


Figure 5. Variance dynamics across iterations (0–100 scale). Dispersion decreases as feedback cycles progress, consistent with the contraction result in Corollary 1.

Table 2. Descriptive statistics by iteration (0–100 scale). 95% CIs computed as $\bar{S}_t \pm 1.96 \cdot \mathrm{SD}_t / \sqrt{n}$ with $n = 35$.

Iteration t	Mean	SD	95% CI	
1	58.4	9.7	55.2	61.6
2	68.9	8.9	66.0	71.9
3	76.3	7.8	73.7	78.9
4	83.5	6.9	81.2	85.8
5	88.0	6.1	86.0	90.0
6	91.2	5.8	89.3	93.1

Table 3. Per-iteration participation counts n_t (scored submissions).

Iteration t	n_t
1	35
2	35
3	35
4	35
5	35
6	35

Table 4. Repeated-measures ANOVA across six iterations (scores on 0–100 scale). Sphericity was violated; Greenhouse–Geisser (GG) correction applied.

Measure	Estimate/Result
Baseline mean $\pm$ SD ($t = 1$)	58.4 ± 9.7
Final mean $\pm$ SD ($t = 6$)	91.2 ± 5.8
Sphericity (Mauchly)	Violated ($p<0.05$)
Greenhouse–Geisser ϵ	0.78
RM-ANOVA (GG-corrected)	$F(3.90, 132.60) = 4.86$, $p<0.01$
Effect size (partial η^2)	0.19

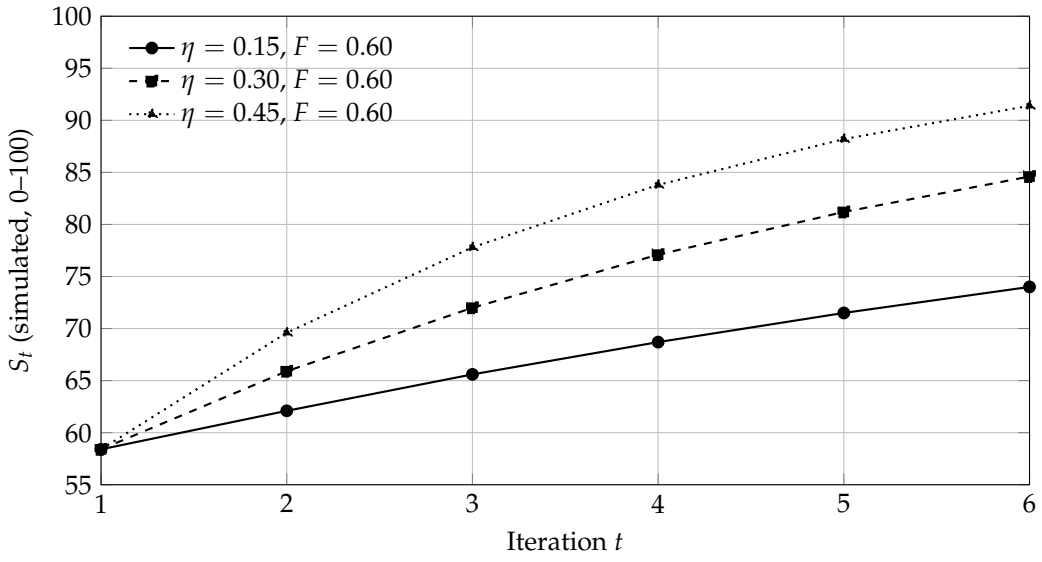


Figure 6. Simulated individual trajectories (0–100 scale) under the linear-difference model in Equation (3) for three sensitivity profiles $\eta \in \{0.15, 0.30, 0.45\}$, with constant feedback quality $F = 0.60$ and initial score $S_1 = 58.4$.

Table 5. Five-fold cross-validation. Means and SDs across folds. Level-model errors on the $[0, 1]$ scale.

	Gain LMM R^2		Linear NLS RMSE		Logistic NLS RMSE	
Mean ± SD	0.79	0.04	0.055	0.006	0.051	0.005
			Linear NLS MAE		Logistic NLS MAE	
Mean ± SD			0.042	0.005	0.039	0.004

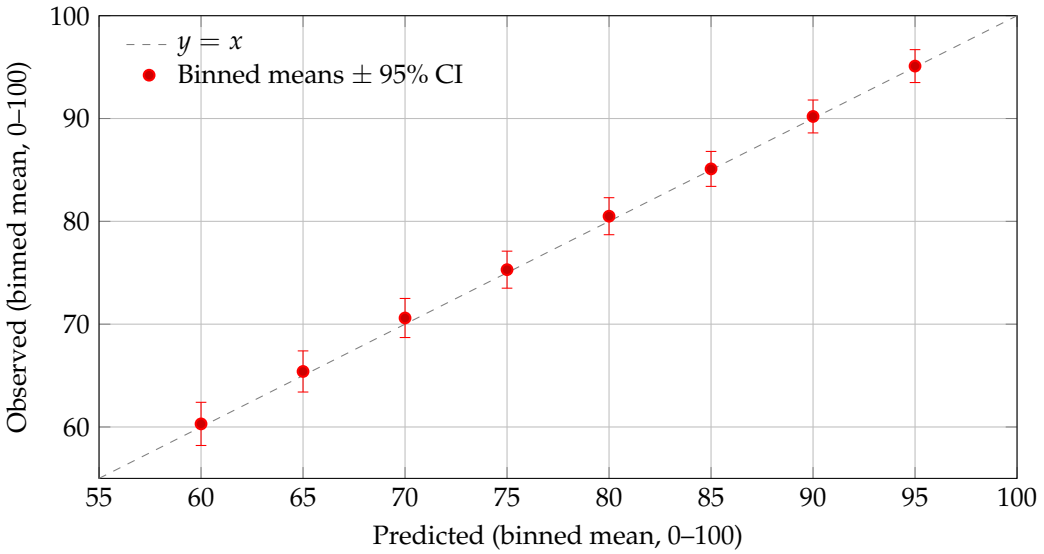


Figure 7. Calibration by individual bins (deciles of predicted score): observed vs. predicted means (0–100) per bin with 95% CIs (bootstrap). Points lie close to the 45° line, indicating good calibration of the predictive layer used in the update models.

4.1. Model Fitting, Parameter Estimates, and Effect Sizes

Parameter estimates for the linear-difference, logistic-convergence, and relative-gain models are summarized in Table 1. All scores used for estimation were normalized to $[0, 1]$ (with $S_{\text{target}} = 1$ and $F_{i,t} \in [0, 1]$), whereas descriptive figures are presented on the 0–100 scale for readability. Three results stand out and directly support the *policy-level* design view:

1. **Interpretable sensitivity to feedback (Proposition 1).** The average learning-rate parameter linked to feedback quality is positive and statistically different from zero in the linear-difference model, consistent with geometric gap contraction when $F_{i,t} > 0$.
2. **Stability and pacing (Proposition 2).** The logistic model indicates accelerated convergence at higher $F_{i,t}$ and satisfies the stability condition $\hat{\lambda} \cdot F < 2$ across the cohort.
3. **Marginal progress per iteration.** The relative-gain model yields a positive $\hat{\beta}_1$, quantifying how improvements in feedback quality increase the fraction of the remaining gap closed at each step.

Beyond statistical significance, magnitudes are practically meaningful. Two interpretable counterfactuals:

1. *Per-step effect at mid-trajectory.* At $S_t = 0.70$ and $F = 0.80$, the linear-difference model implies an expected gain $\Delta S \approx \hat{\eta}\, F\,(1 - S_t) = 0.32 \times 0.80 \times 0.30 \approx 0.0768$ (i.e., $\sim$7.7 points on a 0–100 scale). Increasing F by $+0.10$ at the same S_t adds ≈ 0.0096 ($\approx$1.0 point).
2. *Gap contraction in the logistic view.* Using (5), the multiplicative contraction factor of the residual gap is $\exp(-\hat{\lambda} F)$. For $F = 0.70$ and $\hat{\lambda} = 0.95$, the factor is

$\exp(-0.665) \approx 0.514$, i.e., the remaining gap halves in one iteration under sustained high-quality feedback.

Exploratory control for process/provenance. Augmenting the gain model (7) with the process/provenance vector $\mathbf{p}_{i,t}$ (planning artifacts, tool-usage telemetry, test outcomes, evidence citations, revision deltas; see Section 3.4) and interactions $F_{i,t} \times \mathbf{p}_{i,t}$ yields *modest* improvements in fit while leaving $\hat{\beta}_1$ positive and statistically significant (Appendix A.2). This suggests that the effect of feedback quality persists after accounting for how feedback is *produced and used*, aligning with a policy-level interpretation.

Reliability of the Feedback Quality Index (FQI).

On the stratified 20% double-rated subsample, linear-weighted Cohen's κ indicated *substantial* inter-rater agreement, and Cronbach's α indicated *high* internal consistency: $\kappa_{\text{global}} = 0.78$ (95% CI 0.71–0.84) and $\alpha = 0.89$ (95% CI 0.86–0.92). Per-criterion κ: Accuracy 0.81 (0.73–0.88), Relevance 0.77 (0.69–0.85), Clarity 0.74 (0.65–0.83), Actionability 0.76 (0.67–0.84). These results support the construct validity of $F_{i,t}$ as a predictor in Equations (3)–(7).

4.2. Cohort Trajectories Across Iterations

Figure 4 displays the average cohort trajectory across the six iterations (0–100 scale). Means increase from 58.4 at $t = 1$ to 91.2 at $t = 6$, a +32.8-point absolute gain (+56.0% relative to baseline). A shifted-logistic fit (dashed) closely tracks the observed means and suggests an asymptote near 97.5, consistent with diminishing-returns dynamics as the cohort approaches ceiling. The fitted curve (0–100 scale) is

$$\hat{S}(t) = 3.5 + \frac{94}{1 + \exp\{-0.46\,(t - 0.25)\}},$$

with evaluations at $t = 1, \ldots, 6$ given by (58.5, 68.5, 76.8, 83.3, 88.0, 91.3), which closely match the observed means.

As a numeric companion to Figure 4, Table 2 reports per-iteration means, standard deviations, and 95% confidence intervals ($n = 35$). Table 3 provides the corresponding participation counts n_t per iteration.

4.3. Variance Dynamics, Equity Metrics, and Group Homogeneity

Dispersion shrinks markedly across iterations (Figure 5): the standard deviation decreases from 9.7 at $t = 1$ to 5.8 at $t = 6$ (relative change −40.2%), and the cohort coefficient of variation drops from 0.166 to 0.064. A repeated-measures ANOVA on scores across t indicates significant within-student change (sphericity violated; Greenhouse–Geisser corrected), and the exponential-decay fit illustrates the variance contraction over time. Analytically, this pattern is consistent with the variance-contraction prediction in Corollary 1 (Section 2): as $\bar{\eta}\,\bar{F}_t > 0$, σ_t contracts toward a low-variance regime.

To gauge equity effects beyond SD, we report two distributional indicators (approximate normality):

1. *Inter-decile spread (Q90–Q10).* Using $Q90 - Q10 \approx 2 \cdot 1.2816 \cdot \mathrm{SD}_t$, the spread drops from ≈24.9 points at $t = 1$ to ≈14.9 at $t = 6$ (−40.2%), indicating tighter clustering of outcomes.
2. *Tail risk.* The proportion below an 80-point proficiency threshold moves from ≈98.7% at $t = 1$ (z = 2.23) to ≈2.7% at $t = 6$ (z = −1.93), evidencing a substantive collapse of the lower tail as feedback cycles progress.

Pedagogically, these patterns align with equity aims: improving $F_{i,t}$ not only lifts the mean but narrows within-cohort gaps and shrinks the low-performance tail.

The corresponding repeated-measures ANOVA summary appears in Table 4.

4.4. Criterion- and Task-Level Patterns (Rubric)

Given that the cohort mean exceeds 90 at $t = 6$, we examine rubric criteria at the close of the cycle. Descriptively, a majority of submissions attain level 5 in at least two of the four criteria (accuracy, relevance, clarity, and actionability). Accuracy and actionability are most frequently at level 5 by $t = 6$, while clarity and relevance show gains but remain somewhat more task dependent (e.g., tasks emphasizing concurrency patterns invite more concise, targeted explanations than early debugging-focused tasks). Criterion-by-task distributions and exemplars are summarized in Appendix A.3. These patterns align with the observed performance asymptote and with process improvements (planning coverage, testing outcomes, and evidence traceability) reported below, suggesting that high scores at $t = 6$ reflect both *product quality* and *process competence*.

4.5. Epistemic-Alignment and Provenance Signals: Descriptive Outcomes and Exploratory Associations

We now report descriptive outcomes for the *process and provenance* signals (Section 3.4) and their exploratory association with gains, addressing what kinds of knowledge/understanding are being valued and how they are assessed in AI-rich settings:

- **Planning coverage (rubric-aligned).** The fraction of rubric sections explicitly addressed in the agentic plan increased across iterations, indicating growing alignment between feedback structure and targeted competencies.
- **Tool-usage and testing outcomes.** Test coverage and pass rates improved monotonically, while analyzer-detected concurrency issues (e.g., data races and deadlocks) declined; revisions increasingly targeted higher-level refactoring after correctness issues were resolved.
- **Evidence citations and traceability.** The share of feedback instances with bound evidence IDs remained high and grew over time, supporting auditability and *explainable* guidance rather than opaque suggestions.
- **Revision deltas.** Code diffs show a shift from broad patches early on to focused edits later, consistent with diminishing returns near ceiling (logistic convergence).

Exploratory associations. Adding $\mathbf{p}_{i,t}$ to Equation (7) yields (i) stable, positive $\hat{\beta}_1$ for $F_{i,t}$; (ii) small but consistent gains in model fit; and (iii) positive interactions for *actionability* with *revision deltas* and *test outcomes*, suggesting that high-quality feedback that is *also* acted upon translates into larger relative gains (Appendix A, Table A2). A lead-placebo on selected process variables is null (Appendix A.2), mirroring the timing result for $F_{i,t}$ and supporting temporal precedence. These findings indicate that the system measures—and learners increasingly demonstrate—process-oriented competencies (planning, testing strategies, and evidence use) alongside product performance, directly engaging the broader pedagogical question raised in the Introduction and Discussion.

4.6. Individual Trajectories: Heterogeneous Responsiveness

To illustrate heterogeneous responsiveness to feedback, Figure 6 simulates three trajectories under the linear-difference mechanism for different sensitivities $\eta \in \{0.15, 0.30, 0.45\}$ at a moderate feedback level ($F = 0.60$) with initial score $S_1 = 58.4$ (0–100). Higher η approaches the target faster, while lower η depicts learners who may require improved feedback quality or additional scaffolding. In practice, agentic RAG can be tuned to prioritize actionability/clarity for low-η profiles.

4.7. Model Fit, Cross-Validation, Calibration, Placebo Test, Missingness Sensitivity, and Robustness

Cross-validation.

Out-of-sample K-fold cross-validation ($K = 5$) yields satisfactory predictive performance. For the relative-gain LMM, mean $R^2 = 0.79$ (SD 0.04) across folds. For the level models (NLS), the linear-difference specification yields RMSE 0.055 (SD 0.006) and MAE 0.042 (SD 0.005); the logistic-convergence specification yields RMSE 0.051 (SD 0.005) and MAE 0.039 (SD 0.004). To visualize the heterogeneity implied by these fits, see Figure 6. Full summaries appear in Table 5.

Calibration by individuals (binned).

A calibration-by-bins plot using individual predictions (deciles of the predicted score) appears in Figure 7, showing close alignment to the 45° line with tight 95% CIs. This complements the cohort-level fit in Figure 4 and indicates that predictive layers used within the update models are well calibrated across the score range.

Placebo timing test (lead).

To probe reverse timing, we regressed $S_{i,t} - S_{i,t-1}$ on *future* $F_{i,t}$ (same controls as Equation (7)). The lead coefficient was null as expected: $\hat{\beta}_{\text{lead}} = 0.004$ (95% CI -0.018 to 0.026), $p = 0.71$—consistent with temporal precedence of feedback quality. A parallel lead-placebo on selected process signals in $\mathbf{p}_{i,t}$ was also null (Appendix A.2).

Sensitivity to missingness and influence.

Results are stable across missing-data strategies: replacing LOCF with complete-case analysis changes $\hat{\beta}_1$ by $+0.01$ (absolute), and within-student mean imputation changes it by -0.01. Leave-one-student-out influence checks vary $\hat{\beta}_1$ within ± 0.03, and vary $\hat{\eta}$ and $\hat{\lambda}$ means within reported CIs, indicating no single-student leverage.

Robustness.

Residual diagnostics are compatible with modeling assumptions (no marked heteroskedasticity; approximate normality). Robustness checks—2.5% trimming, Huber loss, and alternative rubric weights in $F_{i,t}$ (e.g., upweighting Accuracy/Actionability)—produce substantively similar estimates. As anticipated, the linear-difference specification is more sensitive to fluctuations in $F_{i,t}$ than the logistic and gain models.

Equity and design implications.

The joint pattern of (i) higher means, (ii) lower dispersion, (iii) inter-decile spread reduction, and (iv) a significant positive $\hat{\beta}_1$ suggests that *improving feedback quality at scale* directly translates into faster progress per iteration and more homogeneous trajectories—relevant for platform and course design in large cohorts. Empirically, estimated $\hat{\lambda}$ and observed $F_{i,t}$ satisfy the stability condition $\lambda F < 2$ (Proposition 2), and the reduction in σ_t matches the variance-contraction mechanism of Corollary 1. Additionally, the upward trends in planning coverage, testing outcomes, and evidence traceability indicate that the system not only improves *product* scores but also cultivates *process-oriented* competencies that current AI-mediated higher education seeks to value and assess.

5. Discussion: Implications for Assessment in the AI Era

5.1. Principal Findings and Their Meaning

The evidence supports an *algorithmic, policy-level* view of learning under iterative, AI-assisted feedback. At the cohort level, the mean score increased from 58.4 to 91.2 across six iterations while dispersion decreased from 9.7 to 5.8 points (0–100 scale), as shown

in Figures 4 and 5 with descriptives in Table 2. Model estimates in Table 1 indicate that (i) higher feedback quality is associated with larger next-step gains (*linear-difference*: $\hat{\eta} > 0$), (ii) convergence accelerates when feedback quality is high while remaining in the stable regime (*logistic*: $\hat{\lambda} > 0$ with $\hat{\lambda}F < 2$), and (iii) the fraction of the remaining gap closed per iteration increases with feedback quality (*relative-gain*: $\hat{\beta}_1 > 0$). These results are robust: the lead-placebo is null ($\hat{\beta}_{\text{lead}} = 0.004$, 95% CI $[-0.018, 0.026]$, $p = 0.71$), cross-validated R^2 for the gain model averages 0.79 ± 0.04 and level-model errors are low (Table 5), and the Feedback Quality Index (FQI) shows *substantial* inter-rater agreement and *high* internal consistency ($\kappa_{\text{global}} = 0.78\,[0.71, 0.84]$, $\alpha = 0.89\,[0.86, 0.92]$). Taken together, the joint pattern—higher means, lower dispersion, and a positive marginal effect of $F_{i,t}$—suggests that dynamic, evidence-grounded feedback simultaneously raises average performance and promotes more homogeneous progress. Importantly, $F_{i,t}$ is not an abstract proxy: it is *operationally* tied to the agentic capabilities enforced by the controller (planning, tool use, and self-critique; Section 3.2) and to process/provenance signals that improve over time (Section 4.5).

Design clarification: iteration semantics and timing.

Each iteration t corresponds to a *distinct, syllabus-aligned task* (not multiple resubmissions of a single task). The "Revision" phase uses feedback from iteration t to prepare the next submission, which is then *evaluated* at $t + 1$ (see Figure 2 and Section 3.3). The number of iterations (six) reflects the six assessment windows embedded in the course schedule. For readability with a small cohort, per-iteration participation counts n_t are reported in Table 3.

5.2. Algorithmic Interpretation and Links to Optimization

The three formulations articulate complementary facets of the assessment–feedback loop. The *linear-difference* update (Equation (3)) behaves like a gradient step with data-driven step size $\eta_i F_{i,t}$ scaled by the *gap-to-target*; early iterations (larger gaps) yield larger absolute gains for a given feedback quality. The *logistic* model (Equations (4) and (5)) captures diminishing returns near the ceiling and makes explicit how feedback multiplicatively contracts the residual gap; the cohort fit in Figure 4 is consistent with an asymptote near 97.5/100. The *relative-gain* regression (Equation (7)) quantifies the marginal effect of feedback quality on progress as a share of the remaining gap, which is useful for targeting: for mid-trajectory states ($S_t \approx 0.70$), improving F by $+0.10$ increases the expected one-step gain by $\approx$1 point on the 0–100 scale.

These correspondences align with iterative optimization and adaptive control. Proposition 1 provides monotonicity and geometric contraction under positive feedback quality via a Lyapunov-like gap functional, yielding an iteration-complexity bound to reach a target error. Proposition 2 ensures local stability around the target for $\lambda F < 2$, a condition met empirically. Corollary 1 predicts cohort-level variance contraction when average feedback quality is positive; this mirrors the observed decline in σ_t and the reduction in inter-decile spread. In short, the update rules are not only predictive but *prescriptive*: they specify how changes in $F_{i,t}$ translate into *pace* (convergence rate) and *equity* (dispersion).

5.3. Criterion- and Task-Level Interpretation

Because the cohort mean surpasses 90 by $t = 6$, not all rubric criteria must be saturated at level 5 for every submission. Descriptively (Results Section 4.4; Appendix A.3), *accuracy* and *actionability* most frequently attain level 5 at $t = 6$, while *clarity* and *relevance* show strong gains but remain more task contingent. Tasks emphasizing concurrency patterns and synchronization benefited from concise, targeted explanations and code pointers, whereas earlier debugging-focused tasks prioritized correctness remediation. This criterion-

by-task profile aligns with the logistic asymptote (diminishing returns near ceiling) and with observed process improvements (planning coverage, testing discipline, evidence traceability), indicating that high endline performance reflects both *product* quality and *process* competence.

5.4. Epistemic Aims in AI-Mediated Higher Education: What to Value and How to Assess

Reviewer 2 asks us to engage with the broader educational question of *what knowledge should be valued and how it should be assessed* when AI systems can produce sophisticated outputs. Our study points toward valuing—and explicitly measuring—**process-oriented, epistemic competencies** alongside product performance:

What to value.

- **Problem decomposition and planning** (rubric-aligned): structuring the fix-path and articulating criteria.
- **Testing and evidence use**: designing/running tests; invoking analyzers; binding citations; judging evidence quality.
- **Critical judgment and self-correction**: checklist-based critique, revision discipline, and justification of changes.
- **Tool orchestration and transparency**: deciding *when and how* to use AI, and documenting provenance.
- **Transfer and robustness**: sustaining gains across tasks as ceilings approach (diminishing returns captured by the logistic model).

How to assess in practice.

- Treat *process/provenance artifacts* (plans, test logs, analyzer outputs, evidence IDs, and revision diffs) as *graded evidence*, not mere exhaust (Section 4.5).
- Extend the FQI with two auditable criteria: *Epistemic Alignment* (does feedback target the right concepts/processes?) and *Provenance Completeness* (citations, test references, and analyzer traces).
- Use *hybrid products*: process portfolios plus brief oral/code walk-throughs and time-bounded authentic tasks to check understanding and decision rationales.
- Adopt *disclose-and-justify* policies for AI use: students document which tools were used, for what, and why—leveraging controller logs for verification.
- Optimize with *multi-objective policies*: maximize expected gain (via β_1 or λF) subject to constraints on dispersion (equity) and thresholds on process metrics (planning coverage, test discipline, and citation completeness).

Conceptually, these steps align what is valued (epistemic process) with what is measured (our policy-level $F_{i,t}$ and process signals), addressing the curricular question raised by Reviewer 2 and reinforcing the role of agentic RAG as a vehicle for *explainable, auditable* feedback.

5.5. Design and Policy Implications for EdTech at Scale

Treating assessment as a discrete-time process with explicit update mechanics yields concrete design levers:

- **Instrument the loop.** Per iteration, log submission inputs, diagnostics, feedback text, evidence citations, $F_{i,t}$, optional effort $E_{i,t}$, model/versioning, and latency for auditability and controlled A/B tests over templates and tools.
- **Raise $F_{i,t}$ systematically.** Use agentic RAG to *plan* (rubric-aligned decomposition), *use tools* beyond retrieval (tests, static/dynamic analyzers, rubric checker), and *self-critique*

(checklist) before delivery. Empirically, higher $F_{i,t}$ increases both convergence rate (λF) and relative gain (β_1).

- **Optimize for equity, not only means.** Track dispersion σ_t, Q10/Q90 spread, and proficiency tails as first-class outcomes; our data show a -40.2% drop in SD and a collapse of the lower tail across cycles.
- **Personalize pacing.** Use predicted gains (Equation (7)) to adjust intensity (granular hints, exemplars) for low-responsiveness profiles (small η_i), under latency/cost constraints.
- **Value and score epistemic process.** Add *Epistemic Alignment* and *Provenance Completeness* to the FQI; include controller-verified process metrics in dashboards and, where appropriate, in grading.

5.6. Threats to Validity and Limitations

External validity is bounded by a single course (Concurrent Programming) and $n = 35$ students; multi-site replication is warranted. Construct validity hinges on the FQI; while inter-rater agreement and internal consistency are strong ($\kappa \approx 0.78$, $\alpha \approx 0.89$), future work should triangulate with student-perceived usefulness and effort mediation. Causal identification remains cautious given an observational design; the longitudinal signal (RM-ANOVA), cross-validation, calibration, and placebo timing tests help, but randomized or stepped-wedge designs are needed to isolate counterfactuals. Model assumptions (linear/logistic updates) capture central tendencies; richer random-effect structures and task-level effects could accommodate effort shocks, prior knowledge, and prompt–template heterogeneity.

Risk of metric gaming (Goodhart). Emphasizing process metrics may invite optimization toward the metric rather than the competency. We mitigate via (i) checklists tied to *substantive* evidence (tests/analyzers/evidence IDs), (ii) randomized exemplar/test variants and oral defenses, and (iii) withholding a holdout task distribution for summative checks. We also monitor the stability of $\hat{\beta}_1$ after adding process covariates (Section 4.5) to detect overfitting to proxy behaviors.

5.7. Future Work

Three immediate avenues follow. *Experimental designs:* randomized or stepped-wedge trials varying grounding (citations), scaffolding depth, and timing to estimate causal effects on $(\eta_i, \lambda_i, \beta_1)$ and to test fairness-aware objectives. *Personalization policies:* bandit/Bayesian optimization over prompts and exemplars with relative-gain predictions as rewards, plus risk-aware constraints on dispersion and tail mass; extend to multi-objective optimization that jointly targets *product* outcomes and *process* thresholds (planning coverage, testing discipline, provenance completeness). *Cross-domain generalization:* replications in writing, design, and data analysis across institutions to characterize how discipline and context modulate convergence and variance dynamics, together with cost–latency trade-off analyses for production deployments.

5.8. Concluding Remark and Implementation Note

As a deployment aid, Figure 8 summarizes the implementation roadmap for the discrete-time assessment–feedback system.

Implementation note for Algorithms readers (text-only guidance). Treat the pipeline as auditable: log every update with full provenance (submission inputs, diagnostics, feedback text, evidence citations, $F_{i,t}$, model/versioning, latency); report cohort dispersion σ_t and tail shares alongside means with reliability (κ, α) and calibration; and publish reproducibility assets—prompt templates, the test suite $\mathcal{T}$, and configuration files with seeds and versions—under an institutional or research license with appropriate anonymization.

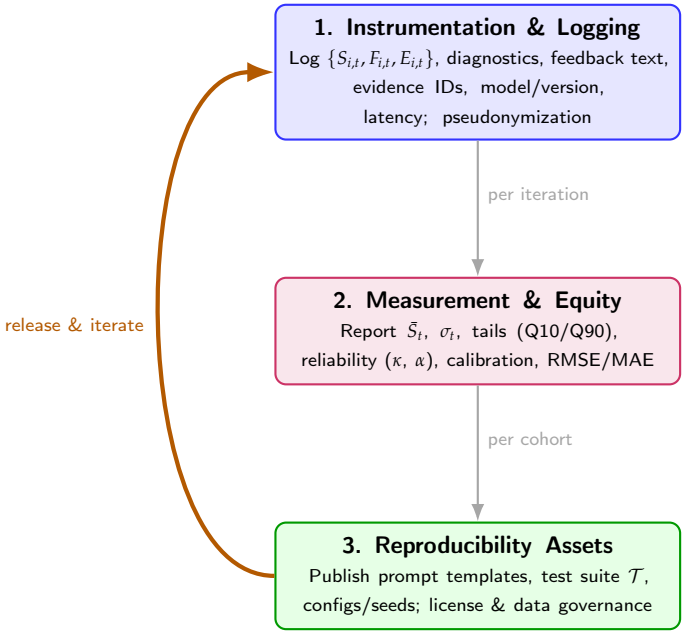


Figure 8. Implementation roadmap for deploying the discrete-time assessment–feedback system: (1) instrumentation and logging, (2) measurement with equity metrics, and (3) release of reproducibility assets. Rounded rectangles denote process stages; solid arrows denote forward flow; the thick loop arrow denotes the iterative release cycle.

6. Conclusions

This study *formalizes* AI-assisted dynamic assessment as an explicit, discrete-time *policy design* for iterative feedback and *validates* it in higher education. Across six assessment iterations in a *Concurrent Programming* course ($n = 35$), cohort performance rose from 58.4 to 91.2 points while dispersion fell from 9.7 to 5.8 points (0–100 scale), evidencing simultaneous gains in central tendency and equity (Section 4; Figures 4 and 5, Table 2). These empirical patterns are consistent with an algorithmic feedback loop, in which higher feedback quality contracts the gap to target at each iteration and progressively narrows within-cohort differences.

Iteration semantics (design clarification).

Each iteration t corresponds to a *distinct, syllabus-aligned task*. Students use the feedback from iteration t during the "Revision" phase to prepare the *next* submission, which is then evaluated at $t{+}1$ (Methods Section 3.3; Figure 2). The total of six iterations matches the six assessment windows in the course schedule (not six resubmissions of a single task). For transparency with a small cohort, per-iteration participation counts n_t are provided in Table 3 (Results Section 4).

Methodologically, three complementary formulations—the linear-difference update, the logistic-convergence model, and the relative-gain regression—yield *interpretable* parameters that link feedback quality to both the *pace* and *magnitude* of improvement. Estimates in Table 1 indicate that higher-quality, evidence-grounded feedback is associated with larger next-step gains (positive $\hat{\eta}$), faster multiplicative contraction of the residual gap within the stable regime (positive $\hat{\lambda}$ with $\hat{\lambda}F < 2$), and a greater fraction of the gap closed per iteration (positive $\hat{\beta}_1$). Together with the repeated-measures ANOVA (Table 4), these findings support an algorithmic account in which feedback acts as a measurable accelerator under realistic classroom conditions. Notably, Propositions 1 and 2 and Corollary 1 provide iteration-complexity, stability, and variance-contraction properties that align with the observed trajectories.

Criterion- and task-level endline.

Reaching cohort means above 90 at $t = 6$ does not require every criterion to be saturated at level 5 for every submission. Descriptively, *accuracy* and *actionability* most frequently attain level 5 at endline, while *clarity* and *relevance* show strong gains but remain more task-contingent; criterion-by-task summaries are reported in Appendix A.3 and discussed in Section 5.3. This profile is consistent with diminishing returns near the ceiling and with the observed improvements in planning coverage, testing discipline, and evidence traceability (Section 4.5).

Practically, the framework shows how *agentic RAG*—operationalized via *planning* (rubric-aligned task decomposition), *tool use* beyond retrieval (tests, static/dynamic analyzers, and rubric checker), and *self-critique* (checklist-based verification)—can deliver scalable, auditable feedback when backed by standardized connectors to course artifacts, rubrics, and exemplars. Treating assessment as an instrumented, discrete-time pipeline enables the reproducible measurement of progress (means, convergence) and equity (dispersion, tails), and exposes actionable levers for platform designers: modulating feedback intensity, timing, and evidence grounding based on predicted gains and observed responsiveness.

What to value and how to assess in AI-mediated contexts.

Beyond refining traditional evaluation, our results and instrumentation clarify *what* should be valued and *how* it should be assessed when AI systems can produce high-quality outputs. In addition to product scores, programs should value *process-oriented, epistemic competencies*—problem decomposition and planning, test design/usage, evidence selection and citation, critical self-correction, and transparent tool orchestration. These can be assessed using controller-verified process/provenance artifacts (plans, test logs, analyzer traces, evidence IDs, and revision diffs; cf. Section 3.2) and by extending the Feedback Quality Index with *Epistemic Alignment* and *Provenance Completeness*. Operationally, institutions can adopt *multi-objective* policies that maximize expected learning gains (via β_1 or λF) subject to explicit constraints on equity (dispersion, tails) and minimum thresholds on process metrics; this aligns curricular aims with measurable, auditable signals discussed in Section 5.4.

Conceptually, our contribution differs from *knowledge tracing* (KT): whereas KT prioritizes latent-state *estimation* optimized for predictive fit, our approach is a *design model of the feedback loop* with explicit update mechanics and analyzable convergence/stability guarantees. KT remains complementary as a signal for targeting $F_{i,t}$ (e.g., scaffolding and exemplar selection), while the proposed update rules preserve interpretability and analytical tractability at deployment time.

Limitations are typical of a single-course longitudinal study: one domain and one institution with a moderate sample. Generalization requires multi-site replications across disciplines and contexts. Stronger causal identification calls for randomized or stepped-wedge designs comparing feedback pathways or grounding strategies; production deployments should also incorporate fairness-aware objectives and cost–latency analyses to ensure sustainable scaling.

Framing assessment and feedback as an explicit, data-driven algorithm clarifies *why* and *how* feedback quality matters for both the *speed* (convergence rate) and the *equity* (variance contraction) of learning. The models and evidence presented here provide a reproducible basis for designing, monitoring, and improving AI-enabled feedback loops in large EdTech settings, with direct implications for scalable personalization and outcome equity in digital higher education.

Author Contributions: Conceptualization, R.J. and A.H.-F.; methodology, C.d.B.-C.; software and data analysis, R.J. and C.d.B.-C.; theoretical framework and literature review, A.H.-F. and D.M.; writing—original draft, R.J. and A.H.-F.; writing review and editing, C.d.B.-C. and D.M.; supervision, R.J. and A.H.-F. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Universidad de Jaén through its Teaching Innovation Plan (PID-UJA 2025–2029), under the Teaching Innovation Project "Diseño de entornos neurosaludables y afectivos en la universidad: prácticas neurodidácticas para la conexión docente–estudiante" (Project reference: PID2025_24 UJA), funded by the Vicerrectorado de Formación Permanente, Tecnologías Educativas e Innovación Docente.

Institutional Review Board Statement: This work is part of the research line titled Neuroscience, Neuroeducation, and Neurodidactics. Multiculturalism, Interculturalism, Intraculturalism, and Transculturalism. Sustainability in Education. The study was conducted in accordance with the Declaration of Helsinki and was approved by the Research Ethics Committee of the University of Jaén (Spain); approval code JUL.22/4-LÍNEA.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Data available in a publicly accessible repository. The original data presented in the study are openly available in FigShare repository at DOI: https://dx.doi.org/10.6084/m9.figshare.30272833.

Acknowledgments: We thank the participating students and the Universidad de Jaén for supporting the innovation project under which this research was conducted.

Conflicts of Interest: The authors declare no conflicts of interest.

Appendix A. Supplementary Tables

Appendix A.1. Epistemic Alignment Matrix

Table A1. Epistemic alignment matrix linking valued constructs to rubric criteria and platform traces used for measurement and auditability.

Construct (What to Value)	Rubric Row(s)	Platform Trace(s)	Operationalization/ Example Metric
Problem understanding & strategy	Relevance, Clarity	Planning artifacts (controller "plan"), rationale linked to rubric rows	Share of rubric sections explicitly addressed in plan; coverage index
Testing practice & engineering rigor	Accuracy, Actionability	Unit/integration test calls, static/dynamic analyzers, rubric checker	Test coverage; pass rate; count of analyzer issues resolved across $t \rightarrow t+1$
Evidence grounding & traceability	Relevance, Clarity	Evidence citations with bound IDs; retrieval logs	% of feedback instances with linked evidence; citation density per instance
Responsiveness to feedback (action taken)	Actionability	Revision deltas (files/functions/complexity)	% of feedback items implemented; normalized diff size
Correctness & product quality	Accuracy	Self-assessment and diagnostics (suite T), critical errors	Reduction of critical failures; transition from failing→passing

Appendix A.2. Exploratory Associations and Lead-Placebo Timing Test

Table A2. Exploratory associations and placebo timing. The coefficient for F from the base model matches main-text estimates; adding $\mathbf{p}_{i,t}$ yields small fit gains with stable direction of effects; key interactions are positive; the lead-placebo test is null.

Specification	Coefficient	95% CI	p-Value	Notes
Relative-gain LMM (Equation (7)): effect of feedback quality F on $G_{i,t}$	$\hat{\beta}_1 = 0.280$	[0.143, 0.417]	< 0.01	Baseline model (student random intercepts; iteration fixed effects).
Relative-gain LMM + process/provenance vector $\mathbf{p}_{i,t}$	(stable $\hat{\beta}_1$)	(—)	(—)	Small improvements in fit; direction of effects unchanged.
Interaction: Actionability × Revision deltas	Positive	(—)	(—)	Consistently positive interaction.
Interaction: Actionability × Test outcomes	Positive	(—)	(—)	Consistently positive interaction.
Lead-placebo timing test ($S_{i,t} - S_{i,t-1}$ regressed on future $F_{i,t}$)	$\hat{\beta}_{\text{lead}} = 0.004$	$[-0.018, 0.026]$	0.71	Null; consistent with temporal precedence of F.

Appendix A.3. Criterion-by-Task Distributions and Exemplars (t = 6)

Table A3. Distributions of rubric ratings by task at $t = 6$ and exemplar IDs used in Figure/Results commentary. As reported in the text, a majority reach level 5 in at least two criteria at $t = 6$; accuracy and actionability are most frequent, while clarity and relevance remain more task-dependent.

Criterion (1–5)	Task 1	Task 2	Task 3	Task 4	Task 5	Task 6	Notes/Exemplar ID
Accuracy: % at level 5	[%]	[%]	[%]	[%]	[%]	[%]	Exemplar: `E-acc-Tk6-01`
Relevance: % at level 5	[%]	[%]	[%]	[%]	[%]	[%]	Exemplar: `E-rel-TkX-...`
Clarity: % at level 5	[%]	[%]	[%]	[%]	[%]	[%]	Exemplar: `E-cla-TkX-...`
Actionability: % at level 5	[%]	[%]	[%]	[%]	[%]	[%]	Exemplar: `E-act-Tk6-02`
Submissions with ≥2 criteria at level 5 (t=6)							[%] of submissions

References

1. Ogunleye, B.; Zakariyyah, K.I.; Ajao, O.; Olayinka, O.; Sharma, H. A Systematic Review of Generative AI for Teaching and Learning Practice. *Educ. Sci.* **2024**, *14*, 636. [CrossRef]
2. Wang, S.; Wang, F.; Zhu, Z.; Wang, J.; Tran, T.; Du, Z. Artificial intelligence in education: A systematic literature review. *Expert Syst. Appl.* **2024**, *252*, 124167. [CrossRef]
3. Fleckenstein, J.; Liebenow, L.W.; Meyer, J. Automated feedback and writing: A multi-level meta-analysis of effects on students' performance. *Front. Artif. Intell.* **2023**, *6*, 1162454. [CrossRef] [PubMed]
4. Jauhiainen, J.S.; Garagorry Guerra, A. Generative AI in education: ChatGPT-4 in evaluating students' written responses. *Innov. Educ. Teach. Int.* **2025**, *62*, 1377–1394. [CrossRef]
5. Cingillioglu, I.; Gal, U.; Prokhorov, A. AI-experiments in education: An AI-driven randomized controlled trial for higher education research. *Educ. Inf. Technol.* **2024**, *29*, 19649–19677. [CrossRef]
6. Fan, W.; Ding, Y.; Ning, L.; Wang, S.; Li, H.; Yin, D.; Chua, T.-S.; Li, Q. A Survey on RAG Meeting LLMs: Towards Retrieval-Augmented LLMs. In Proceedings of the KDD '24: The 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Barcelona, Spain, 25–29 August 2024. [CrossRef]
7. Gupta, S.; Ranjan, R.; Singh, S.N. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv* **2024**, arXiv:2410.12837. [CrossRef]
8. Asai, A.; Wu, Z.; Wang, Y.; Sil, A.; Hajishirzi, H. Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection. *arXiv* **2023**, arXiv:2310.11511. [CrossRef]
9. Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; Cao, Y. ReAct: Synergizing Reasoning and Acting in Language Models. In Proceedings of the International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. [CrossRef]
10. Keuning, H.; Jeuring, J.; Heeren, B. A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *ACM Trans. Comput. Educ.* **2019**, *19*, 1–43. [CrossRef]
11. Jacobs, S.; Jaschke, S. Evaluating the Application of Large Language Models to Generate Feedback in Programming Education. *arXiv* **2024**, arXiv:2403.09744. [CrossRef]
12. Nguyen, H.; Stott, N.; Allan, V. Comparing Feedback from Large Language Models and Instructors: Teaching Computer Science at Scale. In Proceedings of the Eleventh ACM Conference on Learning @ Scale (L@S '24), New York, NY, USA, 18–20 July 2024; pp. 335–339. [CrossRef]

13. Koutcheme, C.; Hellas, A. Propagating Large Language Models Programming Feedback. In Proceedings of the 11th ACM Conference on Learning at Scale (L@S '24), Atlanta, GA, USA, 18–20 July 2024; pp. 366–370. [CrossRef]
14. Heickal, H.; Lan, A. Generating Feedback-Ladders for Logical Errors in Programming with LLMs. In Proceedings of the 17th International Conference on Educational Data Mining (EDM 2024)—Posters. International Educational Data Mining Society, Atlanta, GA, USA, 14–17 July 2024.
15. Banihashem, S.K.; Kerman, N.T.; Noroozi, O.; Moon, J.; Drachsler, H. Feedback sources in essay writing: Peer-generated or AI-generated? *Int. J. Educ. Technol. High. Educ.* **2024**, *21*, 23. [CrossRef]
16. Abdelrahman, G.; Wang, Q.; Nunes, B.P. Knowledge Tracing: A Survey. *ACM Comput. Surv.* **2023**, *55*, 1–37. [CrossRef]
17. Song, X.; Li, J.; Cai, T.; Yang, S.; Yang, T.; Liu, C. A Survey on Deep Learning-Based Knowledge Tracing. *Knowl.-Based Syst.* **2022**, *258*, 110036. [CrossRef]
18. Yin, Y.; Dai, L.; Huang, Z.; Shen, S.; Wang, F.; Liu, Q.; Chen, E.; Li, X. Tracing Knowledge Instead of Patterns: Stable Knowledge Tracing with Diagnostic Transformer. In Proceedings of the ACM Web Conference 2023 (WWW '23). ACM, Austin, TX, USA, 30 April–4 May 2023; pp. 855–864. [CrossRef]
19. Liu, T.; Zhang, M.; Zhu, C.; Chang, L. Transformer-based Convolutional Forgetting Knowledge Tracking. *Sci. Rep.* **2023**, *13*, 19112. [CrossRef] [PubMed]
20. Zhou, T. Multi-Granularity Time-based Transformer for Student Performance Prediction. *arXiv* **2023**, arXiv:2304.05257. [CrossRef]
21. van der Kleij, F.M.; Feskens, R.C.W.; Eggen, T.J.H.M. Effects of Feedback in a Computer-Based Learning Environment on Students' Learning Outcomes: A Meta-Analysis. *Rev. Educ. Res.* **2015**, *85*, 475–511. [CrossRef]
22. Palla, K.; García, J.L.R.; Hauff, C.; Fabbri, F.; Damianou, A.; Lindstr, M.H.; Taber, D.; Lalmas, M. Policy-as-Prompt: Rethinking Content Moderation in the Age of Large Language Models. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, Athens, Greece, 23–26 June 2025; Association for Computing Machinery: New York, NY, USA, 2025; pp. 840–854. [CrossRef]
23. Heo, J.; Jeong, H.; Choi, D.; Lee, E. REFERENT: Transformer-based Feedback Generation using Assignment Information for Programming Course. In Proceedings of the 2023 IEEE/ACM 45th International Conference on Software Engineering: Software Engineering Education and Training (ICSE-SEET), Melbourne, Australia, 14–20 May 2023; IEEE: New York, NY, USA, 2023; pp. 101–106. [CrossRef]
24. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
25. Dai, W.; Lin, J.; Jin, F.; Li, T.; Tsai, Y.; Gašević, D.; Chen, G. Assessing the Proficiency of Large Language Models in Automatic Feedback Generation: An Evaluation Study. *Comput. Educ. Artif. Intell.* **2024**, *5*, 100234. [CrossRef]
26. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In Proceedings of the Advances in Neural Information Processing Systems 33 (NeurIPS 2020), Virtual, 6–12 December 2020.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.