

ESCUELA POLITÉCNICA NACIONAL

FACULTAD DE INGENIERÍA DE SISTEMAS

**DESARROLLO DE UN SIMULADOR DE CLASES
PERSONALIZADAS CON IA GENERATIVA PARA EL
APRENDIZAJE UNIVERSITARIO**

IMPLEMENTACIÓN DEL SISTEMA DE EVALUACIÓN ADAPTATIVA

**TRABAJO DE TITULACIÓN PREVIO A LA OBTENCIÓN DEL TÍTULO DE
INGENIERO EN CIENCIAS DE LA COMPUTACIÓN**

CARLOS ANDRÉS CÓRDOVA ACARO
carlos.cordova02@epn.edu.ec

DIRECTOR: ENRIQUE ANDRÉS LARCO AMPUDIA, PhD.
andres.larco@epn.edu.ec

DMQ, enero 2026

CERTIFICACIONES

Yo, **Carlos Andrés Córdova Acaro**, declaro que el trabajo de integración curricular aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y que he consultado las referencias bibliográficas que se incluyen en este documento.

Asimismo, declaro que he utilizado herramientas de inteligencia artificial (ChatGPT, Gemini y Claude) únicamente como apoyo en la generación de tablas y como guía para la investigación de librerías en Python, sin atribuirles autoría, y que todo el contenido derivado ha sido revisado, validado y es de mi exclusiva responsabilidad.

CARLOS ANDRÉS CÓRDOVA ACARO

Certifico que el presente trabajo de integración curricular fue desarrollado por **Carlos Andrés Córdova Acaro**, bajo mi supervisión.

ENRIQUE ANDRÉS LARCO AMPUDIA, PhD.

Director

DECLARACIÓN DE AUTORÍA

A través de la presente declaración, afirmo que el trabajo de integración curricular aquí descrito, así como el producto resultante del mismo, es de carácter público y estará a disposición de la comunidad académica a través del repositorio institucional de la Escuela Politécnica Nacional.

No obstante, la titularidad de los derechos patrimoniales corresponde al autor del presente trabajo, de conformidad con las disposiciones establecidas por el órgano competente en materia de propiedad intelectual, la normativa interna institucional y la legislación vigente.

CARLOS ANDRÉS CÓRDOVA ACARO

ENRIQUE ANDRÉS LARCO AMPUDIA, PhD.

MICHAEL ANDRÉS PILLAGA SOSA

CARLA ANNAI RUIZ ACEVEDO

HERNÁN DARÍO SÁNCHEZ TENELANDA

DEDICATORIA

Dedico toda mi carrera a mi abuelo Brígido que siempre me ha cuidado aún estando en el cielo, le dedico cada logro obtenido y cada paso que doy en la vida.

También dedico todo esfuerzo empleado en este trabajo y a lo largo de mi carrera a toda mi familia que ha sido el pilar fundamental para alcanzar todo estos logros obtenidos.

Sin todos ellos no hubiera sido posible llegar a este punto.

AGRADECIMIENTOS

Quiero expresar mi profundo agradecimiento a todas las personas que han contribuido a mi crecimiento personal, académico y profesional durante todo este tiempo.

Agradezco a mi mamá María por su amor incondicional, su apoyo constante y por ser mi inspiración diaria para seguir adelante.

Agradezco a mi abuela Enma por ser mi otra madre y darme todo su amor y apoyo en cada paso que doy.

Agradezco a mi abuelo Brígido por ser mi papá adoptivo y por enseñarme el valor del esfuerzo y la dedicación. Gracias por siempre estar conmigo, incluso desde el cielo.

Agradezco a mi tía Paty por ser mi apoyo incondicional y por brindarme su sabiduría y consejos a lo largo de mi vida.

Agradezco a mis tíos Luis y Marcelo por ser como unos hermanos mayores y por su apoyo constante.

Agradezco a mis hermanos Martín y David, quienes son mi motor y mi alegría diaria.

Agradezco a mis amigos, quienes han estado a mi lado en los momentos buenos y malos, brindándome su apoyo, sus consejos y su compañía.

Agradezco a mis profesores y mentores por compartir su conocimiento y guiarme en mi camino académico.

Finalmente, agradezco a la Escuela Politécnica Nacional por brindarme las herramientas necesarias para mi formación profesional.

ÍNDICE DE CONTENIDO

CERTIFICACIONES	I
DECLARACIÓN DE AUTORÍA	II
DEDICATORIA	III
AGRADECIMIENTOS.....	IV
ÍNDICE DE CONTENIDO	VI
RESUMEN.....	VII
ABSTRACT	VIII
1 DESCRIPCIÓN DEL COMPONENTE.....	1
1.1 Objetivo general	1
1.2 Objetivos específicos.....	1
1.3 Alcance del componente.....	2
1.4 Marco teórico	3
1.4.1 Aprendizaje adaptativo.....	3
1.4.2 Teoría de Respuesta al Ítem (IRT)	4
1.4.3 Sistemas de Tutoría Inteligente (ITS)	6
1.4.4 Algoritmos de selección adaptativa	6
1.4.5 Métricas de desempeño en sistemas adaptativos	9
1.4.6 Tabla de herramientas y tecnologías	11
1.4.7 Estructura del Proyecto	13
2 METODOLOGÍA	14
2.1 Enfoque y diseño de la investigación	14
2.1.1 Arquitectura experimental y estrategia de simulación	15
2.2 Tipo y diseño de la investigación	16
2.2.1 Diseño experimental basado en simulación.....	17
2.3 Método de investigación	19
2.4 Levantamiento de información	22
2.4.1 Simulación estocástica de estudiantes virtuales.....	22
2.4.2 Sistema de registro y telemetría automática	23
2.4.3 Pruebas de carga y concurrencia.....	24
2.4.4 Síntesis de las técnicas empleadas	25
2.5 Población y Muestra.....	25

2.5.1	Población objetivo	26
2.5.2	Muestra de validación.....	26
2.5.3	Banco de ítems.....	27
2.5.4	Justificación del tamaño muestral.....	27
2.6	Variables de la Investigación	29
2.7	Marco metodológico de desarrollo	31
2.7.1	Justificación de la elección metodológica	31
2.7.2	Estructura y organización de los sprints.....	32
2.7.3	Adaptación al contexto académico.....	32
2.7.4	Planificación progresiva y mejora continua.....	33
2.7.5	Resultados de la aplicación de SCRUM.....	34
2.8	Actividades y productos del proyecto	35
2.9	Técnicas de Análisis de la Información	36
2.9.1	Herramientas y reproducibilidad del análisis	37
2.10	Criterios de Validación y Aceptación	37
2.10.1	Síntesis e interpretación de los criterios.....	39
3	RESULTADOS, CONCLUSIONES y RECOMENDACIONES	40
3.1	Resultados	40
3.1.1	Volumen de Datos Generados	40
3.1.2	Banco de Ítems utilizado.....	40
3.1.3	Precisión Diagnóstica	41
3.1.4	Mecanismos de Parada y Determinismo	49
3.1.5	Rendimiento Computacional y Escalabilidad	50
3.2	Conclusiones	55
3.3	Recomendaciones	56
4	REFERENCIAS BIBLIOGRÁFICAS	58
5	ANEXOS.....	60

RESUMEN

La presente investigación se propuso el diseño, la implementación y la evaluación de un motor adaptativo de evaluación educativa, orientado a la estimación precisa del nivel de conocimiento del estudiante y a la utilización de dicha estimación para personalizar el proceso de aprendizaje.

Desarrollar un motor adaptativo operativo que traduzca la evidencia de interacción del estudiante en decisiones pedagógicas fundamentadas, mediante la combinación de modelos psicométricos extensivamente validados IRT (Item Response Theory) y BKT/KT (Bayesian Knowledge Tracing / Knowledge Tracing) para medir con precisión el desempeño del estudiante y construir trayectorias de aprendizaje genuinamente personalizadas.

La metodología empleada se sustenta en un diseño de tipo cuantitativo y experimental, mediante la implementación de técnicas de simulación de estudiantes virtuales que permiten la generación de interacciones controladas y reproducibles. La recolección de datos se lleva a cabo de forma automática durante la ejecución del sistema, mediante registros de telemetría estructurada.

Para la evaluación del sistema se utilizaron métricas psicométricas y computacionales, destacando el Root Mean Squared Error (RMSE), el Mean Absolute Error (MAE), la convergencia de la habilidad latente (θ) y la eficiencia adaptativa medida por el número de ítems requeridos. Los resultados muestran una reducción en la longitud de las evaluaciones y una estabilidad óptima del sistema ante escenarios con múltiples usuarios concurrentes.

Los hallazgos demuestran que el motor adaptativo propuesto constituye una solución efectiva para entornos educativos digitales.

PALABRAS CLAVE: Evaluación adaptativa, Estimación de habilidad latente, Modelos psicométricos adaptativos, Seguimiento probabilístico del aprendizaje, Simulación de estudiantes virtuales, Sistemas educativos inteligentes.

ABSTRACT

This research focused on the design, implementation, and evaluation of an adaptive educational assessment engine aimed at accurately estimating students knowledge levels and using this estimation to personalize the learning process.

To develop an operational adaptive engine capable of translating evidence from student interactions into well-founded pedagogical decisions, through the combination of extensively validated psychometric models such as IRT (Item Response Theory) y BKT/KT (Bayesian Knowledge Tracing / Knowledge Tracing), in order to accurately measure student performance and construct genuinely personalized learning trajectories.

The methodology was based on a quantitative and experimental design, implemented thorough virtual student simulation techniques that allow the generation of controlled and reproducible interactions. Data collection was carried out automatically during system execution using structured telemetry records.

System evaluation employed both psychometric and computational metrics, including root mean square error (RMSE), mean absolute error (MAE), convergence of the latent ability parameter (θ), and adaptive efficiency measured by the number of required items. Results indicate a reduction in assessment length and optimal system stability under scenarios with multiple concurrent users.

The findings demonstrate that the proposed adaptive engine represents an effective solution for digital educational environments.

KEYWORDS: Adaptive assessment, Adaptive psychometric models, Intelligent educational systems, Latent ability estimation, Probabilistic learning tracking, Virtual student simulation.

1 DESCRIPCIÓN DEL COMPONENTE

Este proyecto se compone de cuatro componentes interrelacionados. El Componente B constituye el motor que traduce la evidencia de interacción del estudiante en decisiones pedagógicas accionables. Su función principal es calcular el nivel de habilidad del estudiante, identificar qué conocimientos domina con suficiente confianza y determinar qué actividad o ítem presentar a continuación, incluyendo la dificultad apropiada y el tipo de soporte necesario.

Los objetivos del componente son duales: por un lado, medir con precisión el desempeño del estudiante; por otro, favorecer el aprendizaje mediante trayectorias personalizadas, haciendo uso de la retroalimentación constante que mantiene con el Componente A, responsable de generar las actividades y las clases. Este enfoque se alinea con la lógica del aprendizaje adaptativo que predomina en la actualidad, basada en realizar ajustes del proceso de aprendizaje a partir de datos individuales en lugar de seguir rutas fijas predeterminadas [1], [2], [3].

1.1 Objetivo general

Desarrollar un motor adaptativo operativo que traduzca la evidencia de interacción del estudiante en decisiones pedagógicas fundamentadas, mediante la combinación de modelos psicométricos extensivamente validados IRT (Item Response Theory) y BKT/KT(Bayesian Knowledge Tracing / Knowledge Tracing) para medir con precisión el desempeño del estudiante y construir trayectorias de aprendizaje genuinamente personalizadas que se ajusten dinámicamente a partir de datos individuales.

1.2 Objetivos específicos

1. Desarrollar un sistema que recoja eventos de interacción y calcule dos señales complementarias: un nivel continuo de habilidad θ basado en IRT para ordenar ítems informativos y reducir el error estándar de medición, y una probabilidad de dominio por habilidad fundamentada en BKT/KT para guiar la práctica espaciada y el refuerzo cuando el objetivo sea consolidar conocimientos de forma sostenida.
2. Implementar una política de selección que escoja el ítem que maximiza la información en torno al θ estimado, reduciendo rápidamente el error estándar y permitiendo alcanzar precisión localizada donde más importa, incorporando reglas de detención, restricciones de contenido curricular y limitaciones de exposición siguiendo las buenas prácticas establecidas en IRT y CAT (Computerized Adaptive Testing).
3. Desarrollar mediante BKT/KT un sistema que mantenga una probabilidad de dominio por habilidad y la actualice tras cada interacción, modelando fenómenos como la adivinación y el desliz, de modo que la selección de la actividad subsiguiente se deci-

da según el beneficio esperado: confirmar dominio incipiente, reducir incertidumbre o fomentar el aprendizaje en la zona de desarrollo más productiva.

4. Implementar el Componente B como un bucle MAPE-K (Monitor–Analyze–Plan–Execute over a Knowledge base) que monitorice respuestas y patrones de desempeño, analice el perfil del estudiante, planifique la siguiente actividad especificando ítem, dificultad y tipo de apoyo, y ejecute enviando recomendaciones explícitas al Componente A, garantizando trazabilidad, explicabilidad y la posibilidad de integrar organización semántica del contenido mediante grafos o rutas de aprendizaje.
5. Establecer un contrato explícito de integración entre componentes que garantice trazabilidad.
6. Desarrollar salvaguardas que resguarden la validez y equidad mediante el reporte de precisión alcanzada, validación del ajuste del modelo, monitorización de exposición equilibrada de temas e ítems, reporte de ganancia pre–post, control de métricas predictivas como AUC(Area Under the Curve) o log-loss, y realización de pruebas DIF y análisis de brechas entre perfiles para detectar posibles sesgos.

1.3 Alcance del componente

El alcance del Componente B comprende el desarrollo de un motor adaptativo operativo con las siguientes características funcionales y técnicas:

- I. **Integración de modelos complementarios:** combinar IRT para diagnóstico o evaluación sumativa con BKT/KT para guiar la práctica continua, aprovechando las fortalezas de ambos enfoques para ofrecer una evaluación integral que mida y fomente el aprendizaje.
- II. **Orquestación de la selección de ítems:** orquestar la selección de ítems mediante reglas claras de parada, cobertura curricular y exposición equilibrada que permitan determinar cuándo la evaluación ha alcanzado suficiente precisión, garantizando que todos los temas relevantes sean cubiertos y evitando la sobreutilización de ítems específicos.
- III. **Panel de métricas para validación docente:** publicar un panel de métricas (precisión diagnóstica, eficiencia, progreso del estudiante, calidad predictiva y equidad) accesible para que los docentes validen el funcionamiento del sistema, facilitando la transparencia y permitiendo intervenciones informadas cuando resulte necesario.
- IV. **Gobernanza de datos y privacidad:** documentar la gobernanza de datos y privacidad, especificando qué se registra, para qué propósito y cómo se protege la información, asegurando el cumplimiento de estándares éticos y regulatorios en el manejo de datos educativos sensibles.

Todo ello integrado con el Componente A de manera que cada estudiante reciba un reto apropiado, en el momento oportuno, con las explicaciones y los apoyos adecuados a su nivel y necesidades específicas, logrando una experiencia de aprendizaje genuinamente personalizada y fundamentada en evidencia [1], [4], [5].

1.4 Marco teórico

1.4.1 Aprendizaje adaptativo

El aprendizaje adaptativo es una forma de enseñanza que se ajusta a cada estudiante en tiempo real. En vez de proponer la misma ruta para todos, el sistema observa evidencias (aciertos, errores, tiempo de respuesta, interacciones) y decide qué contenido, qué nivel de dificultad y qué apoyo conviene a continuación. Así, la progresión deja de ser lineal y se vuelve personalizada, manteniendo el foco en el dominio gradual de objetivos. Esta idea se formaliza y se sostiene en la literatura reciente sobre evaluación, personalización y uso responsable de IA (Inteligencia Artificial) en educación [1], [3].

Conviene distinguir lo adaptativo de lo simplemente "personalizado". La personalización puede implicar variedad de actividades o estilos, pero no siempre supone que el sistema mida y ajuste continuamente con base en datos. Lo adaptativo, en cambio, depende de un ciclo continuo de diagnóstico-retroalimentación-reajuste, y se apoya en un buen diseño instruccional: objetivos claros, progresiones definidas y evidencias útiles para decidir los siguientes pasos [3].

En la práctica, el ciclo luce así:

1. Un breve diagnóstico,
2. Selección de recursos y tareas ajustadas al nivel detectado,
3. Retroalimentación oportuna,
4. Una nueva medición que confirma avances o sugiere refuerzos y
5. Ajustes de la ruta.

La clave no es aumentar la cantidad de ejercicios, sino ofrecer los adecuados en el momento preciso. Este principio didáctico se alinea con marcos como los de Reigeluth y los primeros principios de Merrill, que recomiendan activar saberes previos, demostrar, aplicar e integrar lo aprendido [3].

La IA generativa ha aportado un motor útil para redactar explicaciones, proponer ejemplos y crear ejercicios alineados con la ruta de cada estudiante. Sin embargo, la evidencia disponible también advierte que estas herramientas no reemplazan la pedagogía ni la evaluación rigurosa; su valor aumenta cuando operan bajo criterios claros de calidad, ética

y supervisión docente [1].

Un ejemplo concreto es PathRAG (Path-based Retrieval-Augmented Generation), que organiza el conocimiento como un grafo (conceptos y relaciones) para trazar caminos pertinentes según el perfil del estudiante. Estudios recientes en contextos universitarios híbridos reportan mejoras en participación, logro de competencias y percepción de inclusión cuando se integran rutas personalizadas con apoyo de IA generativa; aun así, subrayan límites metodológicos y la necesidad de diseños más robustos [2].

1.4.2 Teoría de Respuesta al Ítem (IRT)

La Teoría de Respuesta al Ítem (TRI o IRT) es una forma moderna de entender las pruebas: en lugar de mirar solo el puntaje total, analiza cómo responde una persona a cada ítem y, a partir de ello, estima su nivel en el rasgo que se quiere medir θ . Con esa estimación, es posible seleccionar mejores preguntas, ubicar la dificultad donde más hace falta y conocer cuán precisa es la medición en cada tramo del continuo. Frente a la Teoría Clásica de Tests, su aporte central es la ‘invariancia’: medir con la misma escala, aunque cambien los sujetos o los ítems (dentro de ciertos supuestos) [1], [2].

Ideas clave

- **Curva característica del ítem (CCI):** es un gráfico que muestra, para cada nivel de θ , la probabilidad de elegir la opción “clave” del ítem (por ejemplo, responder correctamente o indicar mayor rasgo). Su forma creciente refleja que, a mayor nivel del rasgo, mayor probabilidad de dar la respuesta asociada al rasgo [1], [2].
- **Parámetros a , b y c :** a indica cuánto discrimina el ítem (qué tan bien separa a personas con niveles cercanos de θ), b ubica la dificultad del ítem (el punto de la escala donde el ítem “decide”), y c modela el azar o pseudo-adivinación en ítems de opción correcta/incorrecta. No todos los modelos usan los tres: Rasch 1PL (One-Parameter Logistic Model) usa solo b , 2PL (Two-Parameter Logistic Model) usa a y b , y 3PL (Three-Parameter Logistic Model) usa a , b , c [1], [2].
- **Información del ítem/test:** indica dónde el ítem o el conjunto de ítems mide con mayor precisión. En IRT la precisión no es “plana”: puede ser excelente en un rango de θ y más baja en otros. Esto permite construir bancos de ítems que “cubran” la escala con precisión donde más importa [2].

Supuestos relevantes

- **Unidimensionalidad:** los ítems de una escala deben reflejar esencialmente un solo rasgo dominante; si influyen varios rasgos a la vez, conviene usar modelos multidimensionales o depurar la escala [2].

- **Independencia local:** si ya sabemos el nivel de θ , las respuestas a ítems distintos no deben depender entre sí. Cuando hay "pistas" entre ítems o se agrupan demasiado, este supuesto se rompe y la medición pierde calidad [1], [2].

Modelos más usados (visión práctica)

- **Ítems dicotómicos (correcto/incorrecto):** 1PL (Rasch), 2PL y 3PL. Rasch asume igual discriminación y sin azar; 2PL permite que la discriminación varíe; 3PL incluye el parámetro de pseudo-adivinación. Elegir el modelo depende del contexto y los datos [1], [2].
- **Ítems politómicos (escalas Likert):** modelos como Respuesta Graduada (Samejima) o Crédito Parcial. En el Modelo de Respuesta Graduada, cada salto entre categorías tiene un "umbral" de dificultad, y un único parámetro a de discriminación para el ítem. Esto es muy útil para cuestionarios con varias opciones de respuesta [6].

Bancos de ítems y pruebas adaptativas (CAT)

Al estimar θ en tiempo real y conocer la información de cada ítem, es posible elegir la siguiente pregunta que aporte máxima precisión justo alrededor del nivel estimado del estudiante. Así nacen los tests adaptativos: cada persona responde un conjunto distinto de preguntas, pero todos son evaluados en la misma escala. En este proyecto, esto es clave para que el Componente B seleccione o recomiende ítems con mayor "ganancia informativa" [6], [7].

Integración con los Componentes A y B

1. **(B)** a partir de las respuestas del estudiante, se estima θ con un modelo IRT apropiado (2PL/3PL para ítems dicotómicos; Respuesta Graduada para Likert).
2. **(B)** se consulta el banco de ítems para identificar cuáles ofrecen más información alrededor del θ actual (o del umbral de dominio).
3. **(B→A)** se envía al Componente A la dificultad objetivo y, si aplica, los ítems recomendados o las pautas de complejidad.
4. **(A)** el Componente A genera la siguiente actividad con esa dificultad y pistas adecuadas.
5. **(B)** tras la actividad, se actualiza θ y se repite el ciclo. Este bucle mantiene rutas personalizadas y medibles [6], [8].

1.4.3 Sistemas de Tutoría Inteligente (ITS)

Un Sistema de Tutoría Inteligente (ITS por sus siglas en inglés Intelligent Tutoring Systems) es un software que intenta "parecerse" a una tutoría humana: observa cómo aprende el estudiante, le ofrece explicaciones y actividades a la medida, y retroalimenta en los momentos clave. La idea no es reemplazar al docente, sino multiplicar su apoyo para que cada persona avance a su propio ritmo y con la ayuda justa. Las revisiones recientes muestran que, bien implementados, los ITS mejoran el rendimiento y la participación, personalizan contenidos y apoyan la autorregulación del aprendizaje [4].

Componentes típicos

- **Modelo del estudiante:** mantiene un "perfil vivo" con aciertos, errores, tiempos y progreso.
- **Modelo del tutor:** decide qué explicar, qué actividad proponer y qué pista dar.
- **Modelo de dominio:** representa el conocimiento de la materia (conceptos, habilidades, reglas).
- **Interfaz:** es la cara del sistema (pantallas, ejercicios, feedback).

Con estos componentes, el ITS puede ajustar dificultad, secuencias y apoyos en tiempo real [9].

Integración con los Componentes A y B

1. **(B)** Observa respuestas, estima el nivel del estudiante en los temas clave y detecta dificultades.
2. **(B→A)** Envía al Componente A la dificultad recomendada, objetivos prioritarios y el tipo de intervención.
3. **(A)** El Componente A genera la actividad/clase con esa dificultad y apoyo.
4. **(B)** Tras la actividad, el ITS vuelve a medir y ajusta la ruta. Este bucle mantiene trayectorias personalizadas y medibles a lo largo del curso [4], [9].

1.4.4 Algoritmos de selección adaptativa

Seleccionar "lo siguiente" no es al azar: es decidir, con evidencia, cuál actividad o ítem conviene presentar para medir mejor o para ayudar a aprender mejor. En este proyecto, esa decisión vive en el Componente B (evaluación) y retroalimenta al Componente A (generación de clases). A grandes rasgos, hay dos familias bien establecidas: selección guiada por IRT (cuando buscamos medir con precisión) y selección guiada por BKT/KT

(cuando buscamos acompañar la adquisición de habilidades en el tiempo).

1. Selección guiada por IRT (medición precisa)

La Teoría de Respuesta al Ítem (IRT) modela la probabilidad de respuesta correcta según el nivel del rasgo latente θ y los parámetros del ítem. De forma general, esta relación puede expresarse mediante una función logística, como en el modelo 2PL (Ecuación (1.1)):

$$P(X_{ij} = 1 | \theta_j) = \frac{1}{1 + e^{-a_i(\theta_j - b_i)}} \quad (1.1)$$

donde a_i representa la discriminación del ítem y b_i su dificultad. Con esa base, la selección adaptativa típica elige el siguiente ítem con más información alrededor del θ estimado del estudiante. Esto reduce el error estándar de medición con menos preguntas y mantiene la dificultad "justo donde más informa". La información de un ítem puede expresarse como (Ecuación (1.2)):

$$I_i(\theta) = a_i^2 P_i(\theta)(1 - P_i(\theta)) \quad (1.2)$$

y el error estándar asociado a la estimación de θ se define como (Ecuación (1.3)):

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}} \quad (1.3)$$

En la práctica, se inicia con un θ neutro o con un breve arranque *warm-up*; y tras cada respuesta se reestima θ y se elige el ítem que maximiza la información, lo que puede formalizarse como (Ecuación (1.4)):

$$i^* = \arg \max_{i \in \mathcal{B}} I_i(\hat{\theta}) \quad (1.4)$$

donde \mathcal{B} representa el conjunto de ítems disponibles tras aplicar restricciones. Para mantener validez y equidad, se aplican restricciones de contenido (temas/objetivos), control de exposición (evitar sobre-uso de ciertos ítems) y límites de longitud o precisión objetivo. En escalas politómicas (tipo Likert), la lógica es análoga: cada categoría aporta información en zonas distintas de la escala, y la selección prioriza donde la precisión es más útil [6], [7].

El Componente A↔B actúa cuando el objetivo es certificar dominio o ubicar con exactitud el nivel, IRT permite pedir menos y medir mejor. El Componente B devuelve a A el rango de dificultad recomendada (y la cobertura temática pendiente), de modo que A genere actividades acordes a ese nivel y no "sobre-o-subestime" la exigencia

[6], [7].

2. Selección guiada por BKT/KT (apoyo al aprendizaje)

Seguimiento Bayesiano del Conocimiento (BKT) sigue, para cada habilidad, la probabilidad de dominio del estudiante a lo largo del tiempo: considera un estado "domina / no domina" y cuatro parámetros intuitivos (conocimiento inicial, probabilidad de aprender tras una práctica, adivinación y desliz). Tras una respuesta correcta, la probabilidad de dominio puede actualizarse como (Ecuación (1.5)):

$$P(L_n | Correcto) = \frac{P(L_n)(1 - S)}{P(L_n)(1 - S) + (1 - P(L_n))G} \quad (1.5)$$

y posteriormente incorporar la posibilidad de aprendizaje mediante (Ecuación (1.6)):

$$P(L_{n+1}) = P(L_n | obs) + (1 - P(L_n | obs))T \quad (1.6)$$

donde T representa la probabilidad de transición al dominio, S el desliz y G la adivinación. Con ese perfil, el sistema decide el siguiente ejercicio según el mayor beneficio esperado: reducir la incertidumbre, confirmar dominio o provocar aprendizaje en la zona adecuada. Esta decisión puede formularse de manera general como (Ecuación (1.7)):

$$a^* = \arg \max_a E(P(L_{n+1}) - P(L_n)) \quad (1.7)$$

En contextos reales se combinan además prerequisitos, espaciado para combatir el olvido y señales de compromiso (tiempos, rachas). La evidencia reciente muestra que BKT es eficaz para personalizar secuencias y mejorar resultados cuando la meta es progresar en habilidades específicas y no sólo "medir una vez con precisión" [10], [11].

El aporte dentro del flujo entre los Componentes A↔B, BKT entrega al Componente A no sólo una dificultad recomendada, sino también la habilidad prioritaria, el tipo de apoyo (pista, ejemplo guiado, práctica adicional) y el momento oportuno para espaciado o refuerzo. Tras la actividad de A, B actualiza las probabilidades de dominio y repite el ciclo [5], [11].

3. Estrategia híbrida

En etapas tempranas o con bancos pequeños, BKT tiende a funcionar mejor porque necesita menos calibración de ítems y entrega señales útiles para enseñar. Cuando

el banco crece y se busca una estimación fina del nivel, IRT gana relevancia: permite fijar una precisión objetivo y optimizar la ruta de ítems. Una política práctica es: usar BKT para guiar la práctica diaria (progreso por habilidades) y activar selección IRT en cortes de evaluación (diagnósticos o certificaciones). Esta lógica puede expresarse como una política de selección dependiente del objetivo (Ecuación (1.8)):

$$\pi(s) = \begin{cases} \text{BKT}(s), & \text{si el objetivo es aprendizaje} \\ \text{IRT}(s), & \text{si el objetivo es medición} \end{cases} \quad (1.8)$$

En la arquitectura del proyecto, esto se implementa como un bucle MAPE-K: Monitorizar (respuestas), Analizar (IRT/BKT), Planificar (siguiente ítem o actividad) y Ejecutar (enviar a A), con conocimiento compartido del perfil del estudiante y del banco de ítems [5].

1.4.5 Métricas de desempeño en sistemas adaptativos

Las métricas no son un listado de números: son la forma en que demostramos que el sistema realmente ayuda a aprender y que lo hace de manera eficiente y justa. En un entorno adaptativo, medir implica dos planos que se retroalimentan: la calidad de la medición (¿qué tan bien estimamos el nivel del estudiante?) y la calidad de la enseñanza (¿qué tan bien estimamos el nivel del estudiante?). A continuación se presentan métricas nucleares, escritas en lenguaje claro, conectando la literatura de pruebas adaptativas y trazado del conocimiento [7], [12].

1. Métricas de precisión en la estimación del nivel del estudiante

Para evaluar la calidad de la medición en sistemas adaptativos basados en IRT, es fundamental cuantificar qué tan bien el modelo estima el nivel de habilidad real del estudiante. Tres métricas clave permiten esta valoración:

- **Root Mean Square Error (RMSE):** cuantifica la magnitud promedio del error entre el nivel estimado $\hat{\theta}$ y el nivel real θ del estudiante. Un RMSE bajo indica que las estimaciones del sistema son precisas y se aproximan al verdadero nivel de habilidad. Esta métrica es especialmente útil para comparar diferentes algoritmos de estimación o configuraciones del sistema adaptativo [4], [8].
- **Expected A Posteriori (EAP):** es un método de estimación bayesiano que calcula el valor esperado de θ dada la distribución posterior. A diferencia del *Maximum Likelihood Estimation* (MLE), EAP incorpora información previa y es más robusto con respuestas limitadas, evitando estimaciones extremas. En contextos adaptativos, EAP permite actualizar continuamente la estimación del nivel del estudiante

con cada nueva respuesta, manteniendo estabilidad incluso en las etapas iniciales de la evaluación [4], [8], [12].

- **Ranked Probability Score (RPS):** evalúa la precisión de las predicciones probabilísticas en ítems politómicos ordenados (como escalas Likert). RPS penaliza las predicciones erróneas proporcionalmente a la distancia entre la categoría predicha y la categoría observada, siendo particularmente útil cuando el Componente B trabaja con respuestas graduadas que reflejan diferentes niveles de dominio parcial [12].

Calibración predictiva:

- **Brier Score:** error cuadrático medio entre probabilidades predichas y resultados observados, calculado como (Ecuación (1.9)):

$$\text{Brier} = \frac{1}{N} \sum (p_i - o_i)^2 \quad (1.9)$$

donde p_i es la probabilidad predicha de respuesta correcta y o_i el resultado observado (1 si correcta, 0 si incorrecta). Valores cercanos a 0.0 indican predicciones bien calibradas; el valor de referencia 0.25 corresponde a predicciones aleatorias [11], [12].

Eficiencia evaluativa:

- **Número de ítems hasta convergencia:** cantidad de ítems requeridos para alcanzar $\text{SE}(\hat{\theta}) \leq 0.4$, permitiendo evaluar el costo evaluativo del algoritmo adaptativo [4].
- **Latencia de respuesta del sistema:** tiempo entre recepción de respuesta y generación del siguiente ítem. Latencias superiores a 500 ms pueden degradar perceptiblemente la experiencia del usuario [6].

Estas métricas permiten al Componente B no solo estimar el nivel del estudiante, sino también evaluar la confiabilidad de sus propias estimaciones, ajustando dinámicamente la selección de ítems para maximizar la precisión con el mínimo número de preguntas [4], [8], [12].

2. Componentes típicos

- **Modelo del estudiante:** mantiene un “perfil vivo” con aciertos, errores, tiempos y progreso.
- **Modelo del tutor:** decide qué explicar, qué actividad proponer y qué pista dar.
- **Modelo de dominio:** representa el conocimiento de la materia (conceptos, habilidades, reglas).
- **Interfaz:** es la cara del sistema (pantallas, ejercicios, feedback).

Con estos componentes, el ITS puede ajustar dificultad, secuencias y apoyos en tiempo real [9].

Integración con los Componentes A y B

1. **(B)** Observa respuestas, estima el nivel del estudiante en los temas clave y detecta dificultades.
2. **(B→A)** Envía al Componente A la dificultad recomendada, objetivos prioritarios y el tipo de intervención.
3. **(A)** El Componente A genera la actividad/clase con esa dificultad y apoyo.
4. **(B)** Tras la actividad, el ITS vuelve a medir y ajusta la ruta. Este bucle mantiene trayectorias personalizadas y medibles a lo largo del curso [4], [9].

Reporte para la integración A↔B

Para cerrar el ciclo A↔B, el Componente B debe devolver un panel compacto: (i) precisión alcanzada ($SE(\hat{\theta})$, o intervalo de confianza de puntaje/total), (ii) eficiencia (ítems/tiempo vs. objetivo), (iii) progreso por habilidad (probabilidad de dominio y ganancia), (iv) calidad de predicción (AUC/logloss) y (v) equidad (DIF y exposición balanceada). Con ese resumen, el Componente A puede ajustar dificultad, apoyo y espaciado con criterio.

1.4.6 Tabla de herramientas y tecnologías

A continuación, la Tabla 1.1 presenta las herramientas y tecnologías empleadas en el desarrollo del Componente B (Motor Adaptativo), detallando su uso específico dentro del sistema.

Tabla 1.1. Herramientas y tecnologías empleadas en el desarrollo del Componente B (Motor Adaptativo)

Herramienta / Tecnología	Uso dentro del Componente B
Python 3.x 	Lenguaje de programación principal para la implementación del motor adaptativo, la lógica de negocio y los cálculos psicométricos (IRT y BKT).
FastAPI 	Framework para la exposición del motor adaptativo como API REST, encargado de la gestión de eventos, recomendaciones y métricas del sistema.
Pydantic 	Librería para el modelado y validación tipada de datos intercambiados entre los endpoints del servicio.
OpenAPI / Swagger 	Interfaz de documentación automática para la exploración y prueba de los endpoints del Componente B.
jsonschema 	Herramienta para la validación formal de los contratos de integración AB y BA.
Estructuras JSON 	Formato de datos utilizado para la persistencia del estado del estudiante y la configuración del motor adaptativo.
Logging estructurado 	Mecanismo de registro en formato JSON para auditoría y trazabilidad de decisiones adaptativas.
Uvicorn 	Servidor ASGI empleado para la ejecución concurrente del servicio FastAPI en entornos de desarrollo y prueba.
dateutil 	Librería para el manejo de fechas y tiempos en el cálculo de decaimiento temporal del modelo BKT.
Locust 	Herramienta para la evaluación del rendimiento y la carga concurrente del sistema.

1.4.7 Estructura del Proyecto

El presente trabajo de titulación se organiza siguiendo la lógica del ciclo de vida del desarrollo de software, estructurándose en los siguientes capítulos:

- **Descripción del Componente:** en este capítulo se definen los objetivos, el alcance y el marco teórico que fundamentan el desarrollo del sistema de evaluación adaptativa propuesto.
- **Metodología:** se describe el enfoque metodológico de la investigación, el diseño experimental basado en simulación, las técnicas de levantamiento de información y el marco metodológico de desarrollo del proyecto.
- **Resultados, Conclusiones y Recomendaciones:** en este capítulo se presentan los resultados obtenidos a partir de la implementación y validación del sistema, así como las conclusiones derivadas del desarrollo y recomendaciones orientadas a trabajos futuros.
- **Referencias Bibliográficas y Anexos:** se incluyen las fuentes bibliográficas utilizadas y los anexos que contienen información complementaria relevante para el proyecto.

2 METODOLOGÍA

2.1 Enfoque y diseño de la investigación

En el desarrollo del Sistema de Evaluación Adaptativa se utilizó un enfoque de investigación cuantitativa en el que la evaluación era objetiva, numérica, reproducible y medible de las variables pedagógicas y computacionales. Este enfoque era lógico ya que se relaciona con el tipo de problema abordado, es decir, optimizar procesos de evaluación y validar un sistema de software fundamentado en modelos matemáticos, estadísticos y probabilísticos que centra la evaluación en niveles profundos del conocimiento del estudiante.

El método cuantitativo permite la evaluación de la forma de actuar del motor adaptativo mediante la valoración de indicadores con los que se puede medir, como pueden ser:

- La estimación de la habilidad latente del aprendiz (θ),
- El ajuste en base a la precisión de las métricas como el error cuadrático medio (RMSE),
- La fiabilidad de las probabilidades analizada con la métrica de Brier Score;

además, tomando métricas concretas de la ingeniería de ciencias computacionales como:

- La latencia de la respuesta del sistema,
- Percentiles de tiempo de procesamiento (P50 y P95),
- Y la tasa de peticiones por segundo (RPS).

Estas métricas sirven para el diagnóstico en base a criterios cuantificables de la precisión, la eficiencia y la escalabilidad del sistema propuesto.

La utilización de esta vertiente metodológica se apoya en la documentación especializada de los sistemas de aprendizaje adaptativo y la evaluación psicométrica, la cual establece que para la obtención de indicadores robustos de aprendizaje deben emplearse modelos estadísticos que permiten inferir variables latentes a partir de la evidencia empírica observable, particularmente en el caso de la Teoría de Respuesta al Ítem (IRT) y en los modelos bayesianos de rastreo de conocimiento (BKT) [6], [7], [12].

1. Clasificación de la investigación

Tomando en cuenta el marco metodológico que se ha seguido, esta investigación puede ser clasificada como una investigación tecnológica aplicada, la cual se halla

orientada al diseño, a la validación y a la implementación de un artefacto de software funcional y operativo que tiene la finalidad de proporcionar una solución a un problema de práctica educativa en contextos locales, específicamente, la modulación adaptativa de evaluaciones mediante la integración de modelos de Machine Learning en la educación superior [4], [5].

2.1.1 Arquitectura experimental y estrategia de simulación

La arquitectura experimental se apoya en la simulación computacional, pues las limitaciones logísticas, éticas y operativas de llevar a cabo un elevado número de pruebas con estudiantes reales en una fase incipiente del desarrollo nos llevaron a adoptar esta opción metodológica. En este contexto, la simulación estocástica y los métodos Monte Carlo conforman una opción metodológica argumentada teóricamente, pero también muy extendida y validada en la literatura para evaluar sistemas adaptativos complejos [3], [5], [10].

Este modelo propone la construcción de un entorno de simulación en el que fueron modelados perfiles de estudiantes virtuales con ciertos parámetros psicométricos controlados, entre los que podemos encontrar el nivel de habilidad inicial (θ), la consistencia de respuesta ante ítems de dificultad variable y el ratio de aprendizaje. Este entorno permite generar un elevado número de interacciones simuladas entre el sistema y perfiles de estudiantes heterogéneos, lo que permite estudiar la convergencia del algoritmo adaptativo, su comportamiento bajo distintas condiciones operativas y evaluar la robustez frente a situaciones adversas o excepcionales.

De igual manera, la simulación computacional favoreció la validación operativa del sistema en situaciones de baja probabilidad de ocurrencia o difícilmente reproducibles en contextos reales de aplicación, tales como patrones de respuestas erráticas por parte de los estudiantes, ejecución de múltiples sesiones de evaluación de manera concurrente, y escenarios de escasez de ítems calibrados en el banco de preguntas. Todo ello contribuyó a proporcionar consistencia empírica a la evaluación de la tolerancia a fallos y la robustez estructural del motor adaptativo.

En el siguiente apartado se presenta la relación detallada de las variables e indicadores de validación en la Tabla 2.1, la cual recoge la síntesis estructurada de los criterios cuantitativos que se han establecido para la evaluación sistemática del desempeño del motor adaptativo.

Tabla 2.1. Criterios cuantitativos utilizados para la validación del motor adaptativo

Variable operacionalizada	Descripción conceptual	Función en el proceso de validación
θ (parámetro de habilidad estimado)	Inferencia del nivel latente de dominio cognitivo del estudiante	Evaluar la precisión diagnóstica del modelo psicométrico
RMSE / MAE	Cuantificación del error entre el parámetro real de habilidad y su estimación computacional	Medir exactitud predictiva del sistema implementado
Brier Score	Función de pérdida cuadrática aplicada a probabilidades predichas	Evaluar calidad y calibración de las predicciones probabilísticas
Latencia (ms)	Duración temporal del procesamiento de peticiones del sistema	Analizar rendimiento computacional y eficiencia algorítmica
RPS	Volumen de peticiones procesadas exitosamente por unidad temporal	Evaluar escalabilidad horizontal y capacidad de concurrencia
P50 / P95	Percentiles de la distribución de tiempos de respuesta	Detectar degradación del rendimiento bajo condiciones de carga elevada

2.2 Tipo y diseño de la investigación

La investigación desarrollada se enmarca dentro del ámbito de la investigación tecnológica aplicada en la Ingeniería en Ciencias de la Computación. Esta clasificación responde a que el propósito central del trabajo no es la formulación de teorías abstractas, sino el diseño, la implementación y la validación de un artefacto computacional operativo, concretamente un Sistema de Evaluación Adaptativa orientado a la personalización del aprendizaje mediante el uso de modelos psicométricos y técnicas de aprendizaje automático.

Este tipo de investigación tecnológica aplicada se halla caracterizada por la producción de conocimiento a partir de la construcción y la evaluación sistemática de soluciones software que logran dar respuesta a problemas reales, sin perder los principios de verificabilidad, reproducibilidad y rigor experimental que caracterizan a la ingeniería de software. Por tanto, el valor científico se encuentra tanto en la arquitectura del sistema como en las pruebas empíricas recogidas durante el proceso de validación. Desde esta óptica, diferentes trabajos en el campo del aprendizaje adaptativo y los sistemas de tutoría inteligente establecen que la evaluación de este tipo de sistemas debe fundamentarse en medidas cuantitativas objetivas (precisión diagnóstica, eficiencia algorítmica, calidad predictiva, entre otras) y no en aproximaciones meramente descriptivas [4], [5], [9]. Esta línea de trabajo refuerza la idoneidad del tipo de investigación escogida.

2.2.1 Diseño experimental basado en simulación

En relación con el diseño de la investigación, se optó por un diseño experimental ya que la investigación supone la manipulación controlada de variables independientes y la observación sistemática de sus efectos sobre variables dependientes relacionadas con el rendimiento del sistema. Las variables manipuladas incluyen el nivel de habilidad previo del estudiante, la consistencia en las respuestas, la dificultad de los ítems y la concurrencia de usuarios, mientras que las variables observadas son la convergencia de la estimación de habilidad, el error de medición, la calidad predictiva del modelo y el rendimiento computacional del sistema. El diseño experimental se implementó a través de simulación computacional, una técnica ampliamente empleada en investigaciones vinculadas con la ingeniería de software, los sistemas autoadaptativos y el aprendizaje adaptativo, particularmente en contextos donde la experimentación directa con usuarios reales se encuentra limitada por consideraciones éticas, logísticas o temporales [3], [6], [10].

Con este fin, se desarrolló un simulador de estudiantes virtuales capaz de generar interacciones estocásticas con el sistema de evaluación adaptativa, siguiendo un enfoque de tipo Monte Carlo, donde cada estudiante virtual fue modelado a partir de parámetros psicométricos previamente definidos, tales como la habilidad latente inicial (θ), la probabilidad de dominio asociada a cada habilidad y la consistencia en las respuestas, permitiendo analizar el comportamiento del sistema frente a una amplia diversidad de perfiles de aprendizaje.

El diseño experimental por simulación permitió llevar a cabo experimentos de tipo transversal y longitudinal, así como evaluar el sistema en situaciones extremas poco reproducibles en ambientes educativos reales, como patrones de respuestas erráticas, escasez de ítems calibrados disponibles o ejecución simultánea de un elevado número de sesiones de evaluación concurrentes, contribuyendo al análisis de la robustez, tolerancia a fallos y escalabilidad del motor adaptativo previo a su implementación en contextos académicos reales.

La Tabla 2.2 presenta un resumen de los componentes centrales del tipo y diseño de investigación adoptados junto con la justificación técnica y metodológica correspondiente.

Tabla 2.2. Clasificación metodológica y justificación técnica del estudio

Elemento metodológico	Clasificación adoptada	Justificación técnica
Tipo de investigación	Tecnológica aplicada	Desarrollo y validación de un sistema software funcional
Diseño de investigación	Experimental	Manipulación controlada de variables y medición de efectos
Estrategia experimental	Simulación computacional	Reproducibilidad y control de escenarios complejos
Técnica de simulación	Monte Carlo	Evaluación estocástica de múltiples perfiles de estudiantes
Horizonte de análisis	Transversal y longitudinal	Evaluación inmediata y análisis temporal del aprendizaje

La Figura 2.1 ilustra el ciclo MAPE-K implementado en el sistema de evaluación adaptativa, evidenciando la correspondencia entre las fases de monitoreo, análisis, planificación y ejecución y los procesos internos del motor adaptativo desarrollado.

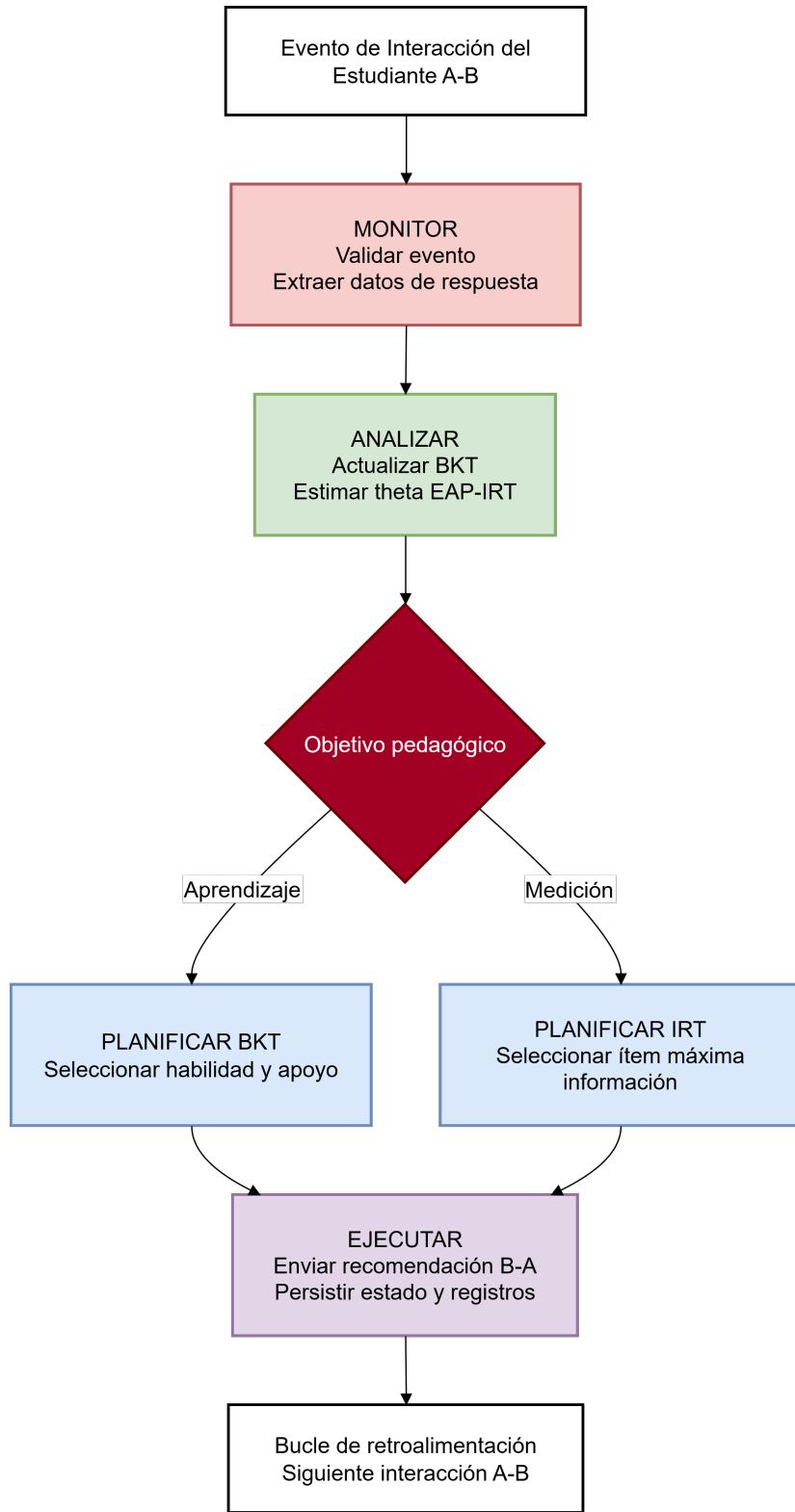


Figura 2.1. Ciclo MAPE-K implementado en el Motor Adaptativo

2.3 Método de investigación

Para llevar a cabo el desarrollo y validación del Sistema de Evaluación Adaptativa se utilizó el método de investigación hipotético-deductivo, el cual resulta ser uno de los más

extendidos en los ámbitos de investigación en ingeniería y ciencias de la computación cuando se desea analizar y validar la forma de comportamiento del sistema en base a una serie de supuestos teóricos formalizados. Así, resulta altamente consistente el uso de este método de investigación para el estudio de sistemas adaptativos en los que la parte de diseño algorítmico fue formulada a partir de modelos matemáticos y de modelos probabilísticos de los que se deduce la necesidad de contrastar empíricamente su validez mediante el método de experimentación controlada y reproducible.

El método hipotético-deductivo se define por ser un procedimiento que parte de la observación sistemática a partir de un problema, la formulación de hipótesis explicativas, la deducción de las consecuencias observables y la posterior comprobación experimental de éstas.

Dentro de esta investigación, dicho enfoque hizo posible estructurar el desarrollo del motor adaptativo como proceso lógico y secuencial de forma que la teoría psicométrica subyacente, las decisiones de diseño de representación algorítmica y los resultados hallados durante el curso de validación tuvieran coherencia [4], [5], [7].

1. Fase de observación y problematización

A través de la etapa de observación y problematización se identificaron las limitaciones recurrentes de los sistemas de evaluación tradicionales que se caracterizan por secuencias de ítems estáticos, criterios de calificación pragmáticos y escasa capacidad de adaptación a lo que realmente sabe el estudiante. La literatura especializada en aprendizaje adaptativo y sistemas de tutoría inteligente señala precisamente que estos sistemas suelen dar lugar a evaluaciones ineficaces y diagnósticos imprecisos en contextos educativos con alta heterogeneidad en los perfiles de aprendizaje [4], [9].

Los resultados de este análisis dieron pie a la formulación de la hipótesis central de la investigación, consistente en que la integración de un modelo híbrido que combine la Teoría de Respuesta al Ítem (IRT) para obtener una estimación global de la habilidad latente del estudiante y los modelos bayesianos de rastreo de conocimiento para obtener un monitoreo más granular de las habilidades, permite conseguir una mejora notable en la eficiencia, la precisión y la equidad diagnóstica de la evaluación frente a los métodos lineales o no adaptativos.

Esta hipótesis se apoya en trabajos anteriores que advierten sobre la complementariedad en la combinación de los modelos psicométricos globales y de técnicas de rastreo probabilísticas del aprendizaje a nivel de habilidad [7], [10], [11].

2. Fase de deducción

En la fase de deducción, la hipótesis propuesta se tradujo en un conjunto de decisiones de diseño dirigido a orientar la implementación del motor adaptativo. Concretamente, se decidió utilizar el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem para estimar la habilidad latente, aplicando el método de Estimación a Posteriori esperada (EAP) con el objetivo de garantizar la estabilidad numérica y disminuir los sesgos en situaciones de información escasa.

A su vez, se propuso un modelo bayesiano de rastreo de conocimiento con decaimiento temporal, cuyo objetivo es modelar la probabilidad de dominio de cada habilidad y la posibilidad de una pérdida progresiva de la misma en el tiempo. De estas decisiones se dedujeron una serie de consecuencias observables que podían ser evaluadas empíricamente, tales como:

- La convergencia progresiva de la estimación de la habilidad del estudiante.
- El error estándar de medición en decremento a medida que se administraran ítems informativos.
- La detección temprana de las brechas de conocimiento.
- La adaptación dinámica de ítems y actividades propuestas.

Se dedujo que el sistema debería conseguir niveles aceptables de precisión diagnóstica con un número reducido de ítems, a la vez que un adecuado rendimiento de cómputo en condiciones de concurrencia.

3. Fase de verificación experimental

La fase de verificación experimental se llevó a cabo mediante la administración de baterías de pruebas automatizadas y experimentos controlados fundamentados en simulación computacional. En esta etapa fue posible contrastar la lógica deducida de la propia hipótesis respecto de lo que sucedía en la realidad del comportamiento de la propuesta de trabajo.

Los experimentos realizados incluyeron pruebas de convergencia de la habilidad estimada, evaluaciones de eficiencia del número de ítems necesarios para la calibración del estudiante, pruebas de equidad diagnóstica mediante distintos perfiles de habilidad, y experimentaciones con la calidad predictiva de las probabilidades generadas mediante el modelo, lo que requirió ajustar algunas métricas como el Brier Score.

El método hipotético-deductivo permitió extender la validación del sistema más allá de su comportamiento algorítmico, incorporando también hipótesis relacionadas con la propuesta de trabajo como servicio software. En este sentido, se formularon y experimentaron supuestos relacionados con la estabilidad del sistema bajo carga, su comportamiento ante la presencia de fallos en situaciones de alta concurrencia y el cumplimiento de umbrales aceptables de latencia y escalabilidad, aspectos críticos que deben contemplarse en sistemas educativos, especialmente en aquellos que se desarrollan como artefactos basados en web [5].

2.4 Levantamiento de información

La recolección de información para validar el Sistema de Evaluación Adaptativa se apoyó en técnicas propias de la Ingeniería en Ciencias de la Computación, las cuales permitieron obtener datos objetivos, reproducibles y que provienen directamente de la ejecución del sistema. Dado el carácter algorítmico, experimental y computacional de esta investigación, no se utilizaron encuestas ni instrumentos cualitativos tradicionales. En su lugar, se utilizaron simulación computacional, generación de datos sintéticos, telemetría automática y pruebas controladas de rendimiento, métodos ampliamente reconocidos en la evaluación de sistemas adaptativos, sistemas de tutoría inteligente y software basado en inteligencia artificial [4], [5], [10].

Estas técnicas posibilitaron la recolección de información tanto del comportamiento pedagógico del motor adaptativo como de su funcionamiento computacional como servicio software. La información recabada caracteriza adecuadamente el funcionamiento interno del sistema, permitiendo su análisis cuantitativo bajo criterios de precisión diagnóstica, eficiencia algorítmica, estabilidad operativa y escalabilidad, aspectos considerados fundamentales en la validación de sistemas auto-adaptativos [5].

2.4.1 Simulación estocástica de estudiantes virtuales

Como técnica principal de recopilación se utilizó la simulación estocástica de estudiantes, que se implementó mediante un software específico desarrollado para este propósito: el simulador de estudiantes virtuales. Este simulador crea agentes artificiales que se parametrizan según perfiles psicométricos definidos, que incluyen el nivel de habilidad latente inicial (θ), la consistencia en las respuestas, la probabilidad de acierto al azar y la tasa de aprendizaje. Estos parámetros posibilitan modelar una amplia variedad de comportamientos de aprendizaje, siguiendo los supuestos de la Teoría de Respuesta al Ítem y los modelos de rastreo de conocimiento [7], [10], [11].

Con la simulación se recolectó un volumen considerable de interacciones controladas entre los estudiantes virtuales y el motor adaptativo. Esto permitió analizar cómo converge la estimación de habilidad a medida que se administran más ítems, cómo va disminuyendo el error estándar de medición, y en qué medida el diagnóstico resulta equitativo cuando se aplica a distintos perfiles de estudiantes.

La simulación también sirvió para reproducir de forma sistemática situaciones extremas que rara vez se encuentran en la práctica educativa cotidiana: respuestas erráticas que no siguen un patrón predecible, casos de aprendizaje acelerado donde el estudiante avanza muy rápido, o situaciones de estancamiento prolongado donde no se observa progreso significativo.

Este tipo de escenarios son muy difíciles de estudiar con estudiantes reales, no solo por las obvias implicaciones éticas de exponer a los estudiantes a evaluaciones poco apropiadas, sino también por la complejidad práctica de controlar todas las variables en un entorno educativo auténtico.

Generación de datos sintéticos

De forma complementaria a la simulación, se emplearon técnicas de generación de datos sintéticos con el objetivo de evaluar la robustez del sistema ante distintas condiciones operativas. Mediante el uso de perfiles psicométricos y secuencias de interacción controladas, el motor adaptativo se sometió a escenarios específicamente diseñados para poner a prueba sus mecanismos de estimación, selección adaptativa y actualización del estado del estudiante. Esta estrategia resulta particularmente eficaz en la validación de sistemas adaptativos complejos, dado que la diversidad de situaciones del mundo real es difícil de abarcar completamente en las primeras etapas de desarrollo [3], [5].

2.4.2 Sistema de registro y telemetría automática

Para el registro de información se diseñó un sistema de telemetría automática basado en archivos de auditoría estructurados en formato JSON. Este sistema registra de forma secuencial e inmutable toda interacción que procesa el motor adaptativo: la selección del ítem, la respuesta del estudiante, la estimación de habilidad latente, la probabilidad de dominio por habilidad y la recomendación que genera el motor de inferencia. Estos registros son la fuente primaria para analizar posteriormente el comportamiento del sistema, a la vez que aseguran la trazabilidad completa de las decisiones algorítmicas que se implementan. Los archivos de auditoría permiten obtener métricas de desempeño bastante detalladas, pero más allá de eso, hacen posible reconstruir sesiones completas de forma determinista usando mecanismos de replay [5].

2.4.3 Pruebas de carga y concurrencia

Como técnica de recolección orientada al desempeño computacional, se llevaron a cabo pruebas de carga y concurrencia. Para ello se utilizaron herramientas de simulación de usuarios concurrentes que permitieron generar peticiones simultáneas al servicio de evaluación adaptativa. Durante estas pruebas se capturaron métricas de latencia, tasa de peticiones procesadas por segundo (RPS), estabilidad del sistema y comportamiento bajo condiciones de estrés. Estas métricas resultan esenciales para valorar la viabilidad del despliegue del sistema en entornos de aprendizaje auténticos con múltiples usuarios concurrentes.

Como se muestra en la Figura 2.2, la recolección de datos se ejecuta de forma automática durante la operación del sistema adaptativo. Las interacciones generadas son procesadas por el motor adaptativo para actualizar el estado del estudiante y producir recomendaciones, mientras que los eventos y métricas de desempeño son registrados como telemetría estructurada. Este flujo permite el análisis posterior, la reconstrucción de sesiones y la evaluación objetiva del comportamiento del sistema.

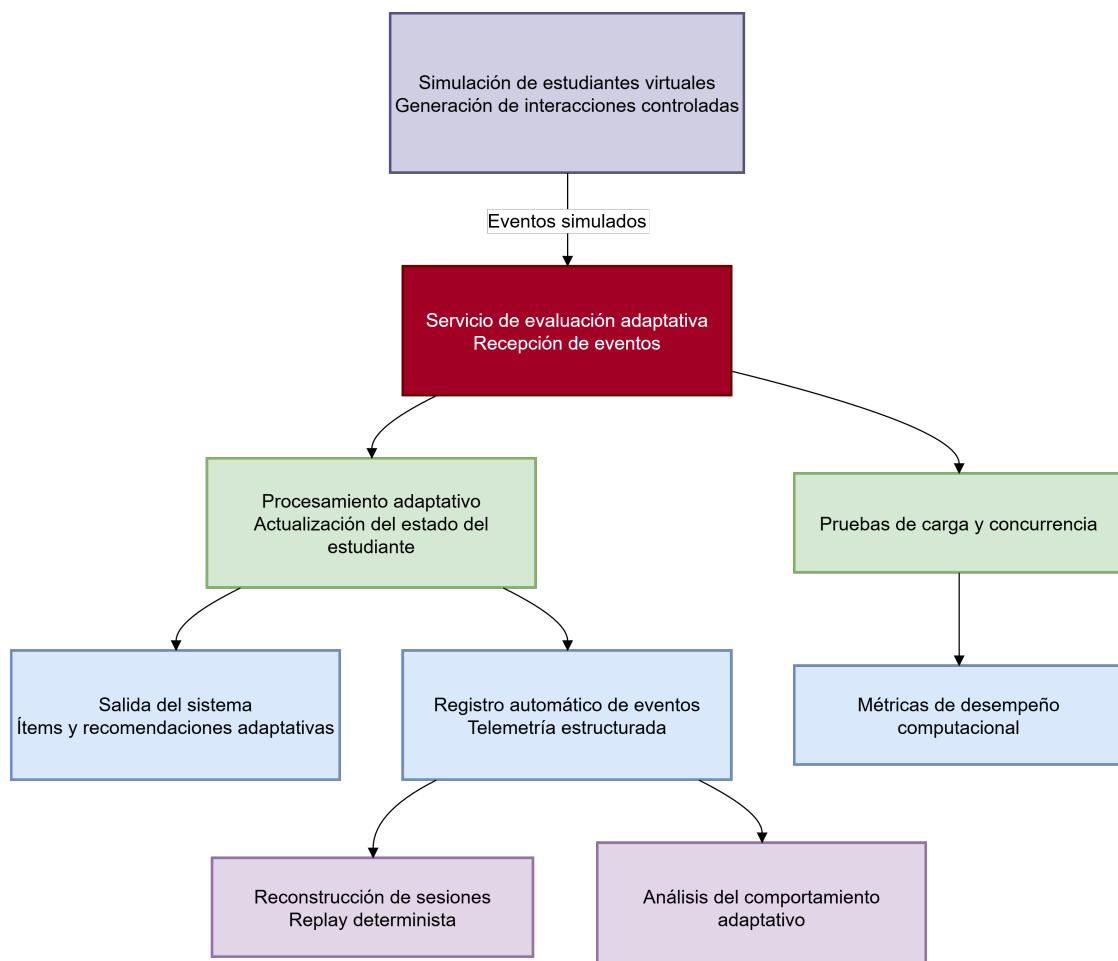


Figura 2.2. Flujo de recolección de datos en el sistema adaptativo.

2.4.4 Síntesis de las técnicas empleadas

La combinación de estas técnicas permitió obtener una caracterización completa del comportamiento del Sistema de Evaluación Adaptativa, tanto desde la perspectiva algorítmica como desde el funcionamiento del sistema software. La Tabla 2.3 presenta las principales técnicas de recolección de información utilizadas en el estudio, el tipo de datos obtenidos y el objetivo metodológico correspondiente.

Tabla 2.3. Técnicas, instrumentos y datos utilizados en la validación experimental

Técnica	Instrumento	Datos recolectados	Propósito metodológico
Simulación estocástica	Simulador de estudiantes virtuales	Respuestas simuladas, convergencia de θ , error estándar	Evaluar precisión y estabilidad del modelo
Datos sintéticos	Perfiles psicométricos parametrizados	Escenarios controlados de aprendizaje	Probar robustez y comportamiento límite
Telemetría automática	Logs de auditoría en formato JSON	Historial de sesiones, métricas internas	Trazabilidad y análisis detallado
Replay determinista	Reconstrucción desde logs	Secuencias completas de interacción	Verificabilidad y replicabilidad
Pruebas de carga	Simulación de usuarios concurrentes	Latencia, RPS, estabilidad	Evaluar escalabilidad y desempeño

El conjunto de técnicas e instrumentos de recolección empleados permitió recopilar información válida, estructurada y vinculada directamente con el comportamiento real del sistema, evitando sesgos derivados de mediciones subjetivas o indirectas. Esta práctica de recolección de datos se encuentra bien establecida en la literatura sobre evaluación de sistemas adaptativos e ingeniería de software, constituyendo un enfoque con alto grado de rigor metodológico [4], [5].

La información recolectada mediante estos instrumentos constituye la base empírica sobre la cual se fundamenta el análisis estadístico y experimental que se desarrolla en las secciones subsiguientes del marco metodológico.

2.5 Población y Muestra

La definición de la población y la muestra en estudios de validación de sistemas adaptativos basados en simulación computacional requiere un enfoque diferenciado respecto a investigaciones empíricas con participantes humanos. En el presente estudio, la población objetivo y la muestra de validación se establecieron considerando tanto el contexto educativo al cual se orienta el sistema como las características metodológicas propias de la

validación algorítmica mediante técnicas de simulación estocástica.

2.5.1 Población objetivo

La población objetivo del Sistema de Evaluación Adaptativa está constituida por estudiantes universitarios de nivel superior. Esta población presenta una heterogeneidad muy elevada en cuanto a conocimientos previos, ritmos de aprendizaje y estrategias de estudio, aspectos que justifican la necesidad de sistemas de evaluación personalizados y adaptativos.

Desde una perspectiva psicométrica, esta población puede representarse mediante un continuo de habilidad latente (θ) que refleja el nivel de dominio del contenido evaluado. En el marco de la Teoría de Respuesta al Ítem (IRT), la habilidad latente en poblaciones universitarias típicamente se distribuye en un rango aproximado de [-3.0, +3.0] en la escala *logit*, donde los valores negativos representan estudiantes con conocimientos insuficientes, los valores cercanos a cero corresponden a estudiantes de nivel medio, y los valores positivos indican un dominio avanzado o experto del tema [6], [7].

La delimitación de esta población objetivo resulta fundamental para interpretar adecuadamente los resultados de la validación y establecer los límites de generalización del sistema desarrollado. Si bien el presente estudio se realizó mediante simulación computacional, la caracterización explícita de la población objetivo guía tanto el diseño de los perfiles de estudiantes virtuales como la futura implementación del sistema en contextos educativos reales.

2.5.2 Muestra de validación

La muestra de validación estuvo conformada por $N = 10$ perfiles de estudiantes virtuales, caracterizados por un conjunto de parámetros psicométricos controlados y conocidos a priori. Estos perfiles fueron diseñados para cubrir de manera sistemática el espectro de habilidad latente de la población objetivo, permitiendo evaluar el comportamiento del sistema adaptativo ante distintos niveles de conocimiento.

- **Habilidad latente real (θ):** Distribuida uniformemente en el intervalo $[-2.0, +2.0]$, asignando valores específicos $\{-2.0, -1.5, -1.0, -0.5, 0.0, +0.5, +1.0, +1.5, +2.0\}$, además de un valor aleatorio dentro del intervalo.
- **Probabilidad de dominio inicial (*Mastery*):** Modelada de forma correlacionada con la habilidad latente mediante la relación $p_{mastery} \approx (\theta + 2)/4$, incorporando variación estocástica gaussiana con $\sigma = 0.15$ y normalización en el intervalo $[0.1, 0.9]$.
- **Consistencia de respuesta:** Distribuida uniformemente en el intervalo $[0.80, 0.95]$.

- **Tasa de aprendizaje (*learning rate*):** Distribuida uniformemente en el intervalo [0.10, 0.20].
- **Factor de fatiga:** Distribuido uniformemente en el intervalo [0.01, 0.03] por ítem administrado.
- **Fundamentación teórica:** La parametrización se fundamentó en modelos de la Teoría de Respuesta al Ítem y en estudios empíricos sobre patrones de respuesta en evaluaciones adaptativas [7], [10], [11].

2.5.3 Banco de ítems

El banco de ítems utilizado para la validación del sistema estuvo conformado por 200 ítems de opción múltiple orientados a la evaluación de conocimientos sobre derivadas. Cada ítem fue caracterizado mediante el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem, considerando los siguientes parámetros psicométricos:

- Parámetro de discriminación (*a*):** distribuido según una distribución log-normal con media $\mu = 0.3$ y desviación estándar $\sigma = 0.4$, con valores truncados en el intervalo [0.5, 2.5].
- Parámetro de dificultad (*b*):** distribuido uniformemente en el intervalo $[-3.0, +3.0]$, con una distribución aproximada de 33 % de ítems fáciles, 33 % de dificultad media y 33 % de ítems difíciles.
- Parámetro de adivinanza (*c*):** distribuido uniformemente en el intervalo [0.0, 0.25].

Los ítems se distribuyeron equitativamente entre dos habilidades del dominio de derivadas: regla de la potencia (100 ítems) y regla de la cadena (100 ítems), asegurando una cobertura balanceada para la evaluación adaptativa.

Dado el carácter de validación algorítmica del estudio, los parámetros IRT fueron generados de forma sintética mediante procedimientos estocásticos controlados, práctica común en fases tempranas de desarrollo de sistemas adaptativos. Esta decisión se reconoce como una limitación que será abordada en etapas posteriores mediante calibración con datos reales [3], [5].

2.5.4 Justificación del tamaño muestral

La determinación del tamaño muestral en estudios de validación mediante simulación computacional responde a criterios distintos de aquellos utilizados en investigaciones con participantes humanos. En lugar de fundamentarse en cálculos de potencia estadística para detectar diferencias entre grupos, el tamaño muestral en simulaciones se orienta a garantizar la cobertura representativa del espacio de parámetros y la estabilidad de las

estimaciones obtenidas mediante métodos Monte Carlo [3], [10].

La literatura especializada en evaluación de sistemas de testing adaptativo computarizado (CAT) y algoritmos de selección de ítems recomienda un tamaño muestral mínimo de $N = 10$ perfiles de estudiantes virtuales para validaciones iniciales de tipo algorítmico, siempre que estos perfiles cubran de manera sistemática el rango de habilidad de interés y presenten heterogeneidad en sus características de respuesta [10], [11].

Este tamaño permite evaluar la estabilidad del algoritmo, la convergencia de las estimaciones y la ausencia de sesgos sistemáticos en distintos niveles de habilidad.

En el presente estudio, el tamaño muestral de $N = 10$ fue considerado suficiente para los siguientes propósitos metodológicos:

- **Evaluación de convergencia:** analizar si el algoritmo de estimación de habilidad mediante EAP converge de forma estable hacia el valor verdadero de θ en distintos niveles del continuo de habilidad.
- **Análisis de equidad diagnóstica:** verificar que el sistema no presente sesgos sistemáticos en la precisión de las estimaciones entre estudiantes con bajo, medio y alto rendimiento.
- **Evaluación de eficiencia:** determinar el número promedio de ítems requeridos para alcanzar niveles aceptables de precisión diagnóstica, definidos como $SE(\theta) \leq 0.4$, en distintos perfiles de estudiantes.
- **Detección de fallos algorítmicos:** identificar posibles errores lógicos, condiciones de borde no controladas o comportamientos anómalos del sistema bajo escenarios diversos.

Cada perfil de estudiante virtual fue sometido a sesiones de evaluación de hasta 20 ítems, lo que permitió generar aproximadamente 200 interacciones ítemestudiante registradas. Este volumen de datos resultó suficiente para el cálculo de métricas agregadas con niveles de error estándar aceptables, así como para la realización de análisis de sensibilidad bajo distintas condiciones de operación del sistema.

Si bien el tamaño muestral de $N = 10$ es adecuado para la validación técnica y algorítmica del motor adaptativo, no permite realizar inferencias estadísticas generalizables a la población de estudiantes reales. En este sentido, la presente fase corresponde a una validación de carácter técnico, orientada a verificar el correcto funcionamiento del sistema antes de su despliegue en contextos educativos reales. La utilización de simulación computacional como estrategia de validación inicial ofrece ventajas metodológicas relevantes, como el control riguroso de las variables experimentales, la reproducibilidad de los escenarios evaluados y la posibilidad de analizar comportamientos extremos, fortaleciendo así la

validez interna del estudio [3], [5], [10].

2.6 Variables de la Investigación

La identificación, operacionalización y clasificación de las variables de estudio constituyen un componente esencial del diseño experimental de la presente investigación. En este estudio, las variables fueron definidas de acuerdo con los principios de la Teoría de Respuesta al Ítem y los modelos bayesianos de rastreo de conocimiento, garantizando su medibilidad, reproducibilidad y coherencia con el marco teórico adoptado [6], [7].

Las variables se clasificaron en tres categorías principales: variables independientes, correspondientes a los parámetros controlados o manipulados durante la simulación; variables dependientes, que representan las salidas generadas por el Sistema de Evaluación Adaptativa; y variables de control, que permanecieron constantes durante los experimentos con el fin de aislar los efectos de las variables independientes sobre las dependientes. Esta clasificación facilita el análisis causal y la correcta interpretación de los resultados obtenidos [5].

1. Variables independientes

Las variables independientes corresponden a los parámetros psicométricos y comportamentales de los estudiantes virtuales, así como a las características de los ítems administrados, los cuales fueron manipulados de forma controlada durante la simulación para evaluar su impacto sobre el desempeño del motor adaptativo. Entre estas variables se incluyen la habilidad latente real del estudiante (θ), los parámetros psicométricos de los ítems definidos por el modelo IRT de tres parámetros (discriminación a , dificultad b y adivinanza c), la consistencia de respuesta, la tasa de aprendizaje y el factor de fatiga, todos ellos operacionalizados mediante rangos numéricos controlados y coherentes con la literatura especializada [7], [10], [11].

2. Variables dependientes

Las variables dependientes corresponden a las salidas generadas por el Sistema de Evaluación Adaptativa durante el procesamiento de las respuestas de los estudiantes virtuales y constituyen los principales objetos de medición del estudio. Estas variables permiten evaluar la precisión, eficiencia y calidad del sistema, e incluyen la habilidad estimada ($\hat{\theta}$), el error estándar de la estimación ($SE(\hat{\theta})$), el error de estimación absoluto ($|\theta - \hat{\theta}|$), la probabilidad de dominio por habilidad ($p_{mastery}$), el *Brier Score*, la latencia de respuesta del sistema y el número de ítems administrados hasta alcanzar la convergencia diagnóstica.

3. Variables de control

Las variables de control corresponden a parámetros del sistema que permanecieron constantes durante todas las ejecuciones experimentales, con el propósito de garantizar la validez interna del diseño y aislar los efectos de las variables independientes. Estas variables incluyen los parámetros del modelo BKT por habilidad, la configuración del algoritmo de estimación EAP, los umbrales de decisión definidos para el dominio y la precisión diagnóstica, los parámetros de *decay* temporal y los límites operativos del sistema, tales como el número máximo de ítems por sesión y las restricciones de repetición de ítems [5].

Síntesis de variables

La Tabla 2.4 presenta una síntesis de las variables del estudio, clasificadas según su rol en el diseño experimental, e incluye su descripción, unidad de medida y rangos de valores observados o asignados.

Tabla 2.4. Definición y clasificación de variables del estudio

Tipo	Variable	Descripción	Unidad	Rango
Independiente	θ (habilidad real)	Nivel verdadero de conocimiento	Escala logit	[−2.0, +2.0]
Independiente	a (discriminación)	Capacidad discriminativa del ítem	Adimensional	[0.5, 2.5]
Independiente	b (dificultad)	Nivel de dificultad del ítem	Escala logit	[−3.0, +3.0]
Independiente	c (adivinanza)	Probabilidad de acierto por azar	Probabilidad	[0.0, 0.25]
Independiente	Consistencia	Coherencia en las respuestas	Probabilidad	[0.80, 0.95]
Independiente	<i>Learning rate</i>	Tasa de aprendizaje incremental	Por ítem	[0.10, 0.20]
Independiente	Factor de fatiga	Incremento de tiempo por fatiga	Por ítem	[0.01, 0.03]
Dependiente	$\hat{\theta}$ (habilidad estimada)	Estimación EAP de la habilidad	Escala logit	[−4.0, +4.0]
Dependiente	$SE(\hat{\theta})$	Error estándar de la estimación	Escala logit	[0.2, 1.0]
Dependiente	$ \theta - \hat{\theta} $	Error de estimación absoluto	Escala logit	[0.0, 2.0]
Dependiente	$p_{mastery}$	Probabilidad de dominio por habilidad	Probabilidad	[0.0, 1.0]
Dependiente	Brier Score	Calibración predictiva	Error cuadrático	[0.0, 1.0]
Dependiente	Latencia	Tiempo de respuesta del sistema	Milisegundos	[50, 500]
Dependiente	N ítems de convergencia	Ítems requeridos hasta $SE(\hat{\theta}) \leq 0.4$	Cantidad	[5, 20]
Control	Parámetros BKT	$p_{L0}, p_T, p_G, p_S, p_F$	Probabilidades	Fijos por skill
Control	Configuración EAP	Grid, prior $N(0, 1)$	—	Fijos
Control	Umbrales	τ , SE objetivo, límites	—	Fijos

2.7 Marco metodológico de desarrollo

El desarrollo del Sistema de Evaluación Adaptativa se llevó a cabo siguiendo una metodología propia de la Ingeniería del Software, lo que hizo posible una construcción sistemática, controlada y alineada con buenas prácticas de desarrollo. Bajo esta lógica, se adoptó la metodología ágil SCRUM como marco de gestión del proceso de desarrollo, complementando el método de investigación hipotético-deductivo descrito anteriormente.

Vale la pena aclarar aquí un punto que puede generar confusión: SCRUM no fue empleado como método de investigación científica en sí mismo, sino como un mecanismo práctico para organizar, planificar y dar seguimiento al trabajo técnico que implicaba construir el sistema.

2.7.1 Justificación de la elección metodológica

La decisión de trabajar con SCRUM tiene que ver directamente con la naturaleza del sistema que se estaba desarrollando. El motor adaptativo integra varios módulos que dependen unos de otros: modelos psicométricos, lógica de selección adaptativa de ítems, persistencia del estado del estudiante, instrumentación de métricas y mecanismos de validación. Estos componentes necesitaban ciclos cortos de desarrollo, prueba y ajuste para poder integrarse correctamente. Es como armar un mecanismo complejo donde cada pieza debe encajar con precisión, pero solo se puede verificar que funciona una vez que las partes están conectadas. SCRUM dio la flexibilidad necesaria para ir incorporando funcionalidades poco a poco y ver en tiempo real cómo afectaban al comportamiento global del sistema [13], [14].

Existe otra razón de peso para elegir metodologías ágiles cuando se trabaja con sistemas basados en inteligencia artificial y aprendizaje automático. La incertidumbre es alta: no siempre se puede anticipar cómo va a comportarse un modelo psicométrico bajo condiciones reales, o qué impacto tendrá modificar ciertos parámetros de convergencia. Los requisitos técnicos también suelen evolucionar conforme se van realizando experimentos y se obtienen resultados inesperados. SCRUM permite ajustar el plan de desarrollo según lo que va mostrando la evidencia empírica en cada iteración, lo que reduce bastante el riesgo de tomar decisiones de diseño que después resulten desconectadas de los resultados experimentales [5].

2.7.2 Estructura y organización de los sprints

El marco SCRUM que se aplicó en este proyecto se organizó mediante sprints de una semana cada uno, con objetivos técnicos específicos y entregables que se podían verificar.

Elegir una semana no fue casualidad: es un período que permite ver resultados tangibles sin tener que esperar demasiado, pero a la vez da tiempo suficiente para implementar funcionalidades que cumplan con estándares mínimos de calidad.

El desarrollo de cada iteración obedecía a una progresión definida de las siguientes tareas:

- **Planificación del sprint:** definir los objetivos técnicos y seleccionar las tareas del backlog que se abordarían.
- **Desarrollo iterativo:** implementar las funcionalidades que se habían priorizado.
- **Pruebas unitarias:** verificar que cada componente individual funcionara como debía.
- **Pruebas de integración:** comprobar que los módulos interactuaran correctamente entre sí.
- **Validación funcional:** confirmar que se cumplieran los requisitos técnicos y algorítmicos establecidos.

Este proceso garantizaba que cada incremento tuviera un nivel mínimo de calidad antes de integrarse al sistema base. La idea era evitar acumular deuda técnica que después pudiera comprometer la estabilidad del motor adaptativo cuando el sistema creciera en complejidad.

2.7.3 Adaptación al contexto académico

Hay que señalar algo importante: la forma en que se usó SCRUM aquí no es exactamente igual a como se usa en un proyecto comercial típico. En una empresa, cada sprint busca entregar valor tangible al cliente o al usuario final. En este caso, cada sprint se centraba más bien en validar técnica y algorítmicamente el sistema, siguiendo las hipótesis de investigación que se habían planteado. Los entregables no se medían tanto por funcionalidades listas para producción, sino por componentes validados empíricamente que confirmaban o cuestionaban aspectos específicos del diseño propuesto.

Esta adaptación hizo que el proceso de desarrollo estuviera muy conectado con el proceso investigativo, sin sacrificar el rigor metodológico ni perder la trazabilidad de las decisiones técnicas. Cada decisión de diseño, cada ajuste que se hacía en los algoritmos, cada refactorización importante quedaba documentada y justificada según los resultados experimentales que se iban obteniendo. El desarrollo técnico no era un fin en sí mismo, sino más

bien una forma de responder las preguntas de investigación que se habían formulado al inicio.

2.7.4 Planificación progresiva y mejora continua

La Tabla 2.5 detalla la distribución temporal de los sprints que se desarrollaron a lo largo del ciclo de construcción del Sistema de Evaluación Adaptativa, destacando los objetivos técnicos que se alcanzaron en cada etapa del proyecto.

Tabla 2.5. Planificación incremental de sprints para el desarrollo del motor adaptativo

Sprint	Objetivo principal	Resultados obtenidos
Sprint 0–1	Investigación y diseño inicial	Selección de frameworks, definición de arquitectura y contratos JSON
Sprint 2–3	Diseño del modelo adaptativo	Implementación del modelo híbrido IRT+BKT
Sprint 4–5	Implementación algorítmica	Estimación EAP, selección adaptativa de ítems y <i>decay</i> temporal
Sprint 6	Validación experimental	Simulación, pruebas automatizadas y pruebas de carga
Sprint 7	Refactorización y documentación	Optimización del código, documentación técnica y cierre del desarrollo

La Figura 2.3 presenta la arquitectura general del Sistema de Evaluación Adaptativa, resultado del proceso de diseño e implementación desarrollado durante los sprints iniciales del proyecto. Esta arquitectura refleja la organización modular del motor adaptativo y los componentes principales que fueron validados experimentalmente. El código fuente completo y la implementación reproducible del sistema se encuentran disponibles en el repositorio del proyecto, referenciado en el Anexo 5.

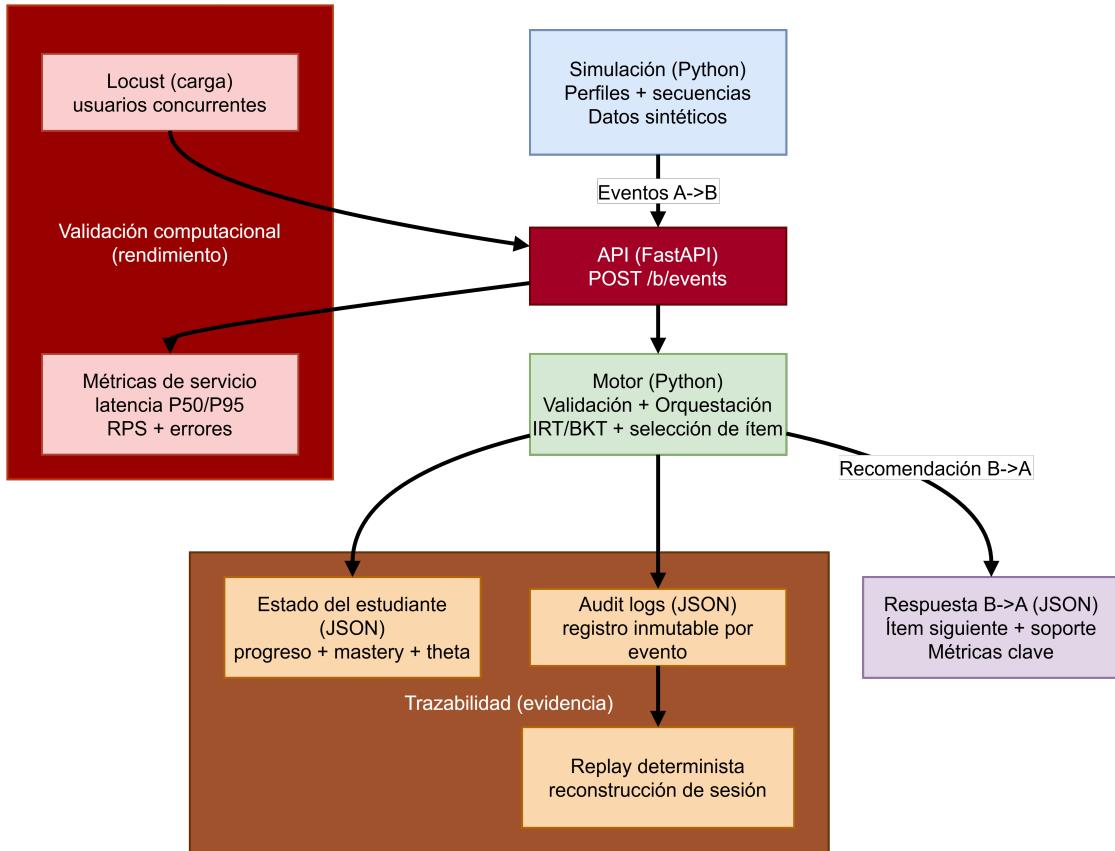


Figura 2.3. Flujo de recolección de datos en el sistema adaptativo.

2.7.5 Resultados de la aplicación de SCRUM

Usar SCRUM contribuyó bastante a manejar la complejidad que implicaba desarrollar este sistema. Entre los beneficios más evidentes están:

- **Identificación temprana de errores:** detectar problemas en etapas tempranas mediante la validación continua, cuando corregirlos resulta significativamente menos costoso en términos de tiempo y esfuerzo.
- **Validación incremental:** probar a fondo cada funcionalidad antes de pasar a la siguiente, asegurando tener una base sólida y estable para seguir construyendo.
- **Mejoras basadas en datos:** fundamentar los cambios del sistema en evidencia experimental concreta, no en especulaciones sobre cómo debería comportarse.
- **Trazabilidad completa:** mantener una alineación constante entre los objetivos de investigación, las decisiones de diseño y los resultados que se iban obteniendo mediante la estructura iterativa.

Esta trazabilidad tiene un valor especial en un contexto académico, donde poder reproducir el trabajo y justificar rigurosamente las decisiones técnicas son aspectos centrales del proceso investigativo. Cada sprint generó documentación detallada que permitiría a otros

investigadores entender no solo qué se implementó, sino por qué se tomaron determinadas decisiones de diseño.

El marco metodológico adoptado garantizó que el Sistema de Evaluación Adaptativa se desarrollara siguiendo principios de calidad de software, mantenibilidad y extensibilidad. Estos aspectos son fundamentales tanto para la futura integración del sistema con otros componentes del ecosistema de aprendizaje como para su eventual implementación en entornos educativos reales. La combinación del método hipotético-deductivo para la investigación y SCRUM para el desarrollo técnico resultó efectiva, permitiendo mantener el rigor científico mientras se construía un artefacto computacional que realmente funciona [13], [14].

2.8 Actividades y productos del proyecto

Las actividades ejecutadas para cada objetivo generaron productos y evidencias técnicas concretas en forma de artefactos de diseño, módulos de software funcionales, registros de simulación y métricas de rendimiento, lo que permitió comprobar de manera objetiva el grado de cumplimiento de los objetivos específicos [4], [5].

La Tabla 2.6 presenta la correspondencia entre los objetivos específicos del proyecto, las principales actividades desarrolladas y los productos generados.

Tabla 2.6. Correspondencia entre objetivos específicos, actividades y productos generados

Objetivo específico	Actividades desarrolladas	Productos / evidencias
Analizar herramientas de IA aplicables a la evaluación adaptativa	Revisión del estado del arte en aprendizaje adaptativo, ITS e IRT. Análisis comparativo de modelos psicométricos y de rastreo de conocimiento.	Selección fundamentada del modelo híbrido IRT (3PL) + BKT
Diseñar un sistema de evaluación progresiva y personalizada	Diseño de la arquitectura del motor adaptativo y definición de contratos de comunicación y reglas de selección de ítems.	Arquitectura del sistema adaptativo y esquemas JSON
Implementar modelos de análisis y seguimiento del aprendizaje	Implementación del modelo IRT con EAP y del modelo BKT con decaimiento temporal. Integración de métricas.	Motor adaptativo funcional y módulos de cálculo
Validar el sistema mediante pruebas funcionales y experimentales	Simulación de estudiantes virtuales, pruebas de convergencia, eficiencia, equidad y pruebas de carga.	Resultados de simulación, reportes de pruebas y métricas de rendimiento

2.9 Técnicas de Análisis de la Información

El análisis de la información obtenida durante la validación del Sistema de Evaluación Adaptativa se centró en evaluar tanto el comportamiento algorítmico del motor adaptativo como el rendimiento computacional del sistema, utilizando métricas objetivas, numéricas y reproducibles [4], [5].

1. Análisis del rendimiento algorítmico

Desde el punto de vista del rendimiento algorítmico, el análisis realizado se centró en evaluar la precisión del modelo híbrido implementado considerando métricas de error ampliamente utilizadas en psicometría computacional. Se utilizaron el error cuadrático medio (RMSE) y el error absoluto medio (MAE) para evaluar la precisión del sistema en la estimación de habilidades latentes, métricas ampliamente utilizadas en psicometría computacional [6], [7], [11].

El análisis incluyó la evolución del error estándar de medición para observar la convergencia del modelo y la reducción de la incertidumbre diagnóstica conforme se administran ítems informativos [7], [10].

La calidad predictiva del sistema fue analizada mediante el Brier Score, que permitió evaluar la calibración de las probabilidades de respuesta correcta predichas por el sistema frente a los resultados observados en la simulación [11], [12].

Junto con el Brier Score, también se implementó un análisis longitudinal del aprendizaje simulado dirigido a comprobar la capacidad del sistema para detectar los cambios en el dominio de las habilidades a lo largo del tiempo contemplando medidas de adquisición progresiva de los conocimientos y fenómenos del tipo de decaimiento o pérdida del dominio del mismo tipo, esto permitiría comprobar la sensibilidad del modelo frente a los cambios temporales en el rendimiento del estudiante. Este tipo de análisis es muy relevante en sistemas de evaluación adaptativa que intentan proporcionar retroalimentación continua y personalizada [10], [11].

2. Análisis del rendimiento computacional

El rendimiento computacional se evaluó mediante métricas de latencia de respuesta (en milisegundos), percentiles de tiempo de respuesta (P50 y P95) y tasa de peticiones por segundo procesadas (RPS), obtenidas durante las pruebas de carga y concurrencia [5].

Estas métricas permitieron evaluar la capacidad del sistema para operar como un servicio software en condiciones de uso realistas, garantizando tiempos de respuesta adecuados para aplicaciones interactivas y estabilidad bajo carga concurrente.

2.9.1 Herramientas y reproducibilidad del análisis

Las técnicas de análisis se implementaron utilizando librerías estándar del ecosistema Python, garantizando la transparencia y reproducibilidad de los resultados [5].

La Tabla 2.7, resume las dimensiones y métricas utilizadas para evaluar el desempeño del motor adaptativo.

Tabla 2.7. Dimensiones y métricas utilizadas para la evaluación del desempeño del motor adaptativo

Dimensión analizada	Métrica	Descripción	Propósito del análisis
Precisión diagnóstica	RMSE / MAE	Error entre habilidad real y estimada	Evaluar exactitud del modelo
Convergencia	Error estándar de medición	Nivel de incertidumbre en la estimación de θ	Analizar eficiencia adaptativa
Calidad predictiva	Brier Score	Calibración de probabilidades predichas	Validar confiabilidad del modelo
Rendimiento	Latencia (ms)	Tiempo de respuesta del sistema	Evaluar experiencia del usuario
Escalabilidad	RPS, P50/P95	Capacidad bajo concurrencia	Analizar viabilidad operativa

2.10 Criterios de Validación y Aceptación

1. Criterios de Equidad en la Estimación Diagnóstica

La equidad en la estimación diagnóstica tiene como objetivo mantener un nivel de precisión similar entre los estudiantes de diferentes niveles de la misma habilidad, ya que se busca evitar que el sistema beneficie o penalice a unos estudiantes más que a otros de manera sistemática, conforme a los principios de equidad en la medición educativa reportados en la literatura psicométrica [6], [7].

Para evaluar este criterio, se llevó a cabo el análisis de la consistencia del error de estimación, considerando el valor del RMSE para una variable en tres grupos definidos según la habilidad real del estudiante:

- Grupo bajo: $\theta \in [-2.0, -0.67)$

- Grupo medio: $\theta \in [-0.67, +0.67]$
- Grupo alto: $\theta \in (+0.67, +2.0]$

El criterio de aceptación determinado fue un coeficiente de variación del RMSE inferior al 40 %, definido como la razón entre la desviación estándar y la media del RMSE. Asimismo, se estableció como criterio complementario que el valor máximo registrado de RMSE no superara 0.70, con el fin de evitar que algún grupo presente un error de estimación excesivo. Esta consideración se encuentra alineada con los criterios de equidad utilizados en sistemas de evaluación adaptativa y medición educativa [5], [7].

2. Criterios de Calidad Predictiva

El criterio de calidad predictiva mide qué tan bien las probabilidades que calcula el modelo acerca de si un estudiante responderá correctamente o no coinciden con los resultados observados durante la simulación, aspecto fundamental en sistemas de evaluación adaptativa basados en modelos probabilísticos [10], [11].

Con este propósito, se realizó el cálculo del *Brier Score*, definido como el error cuadrático medio entre las probabilidades predichas y los valores binarios de acierto o error observados. Como criterio de aceptación se consideraron valores inferiores a 0.30, dado que dichos valores indican una capacidad predictiva superior a la aleatoriedad y una calibración adecuada del modelo, consistente con lo reportado en la literatura para modelos IRT y BKT en contextos educativos estructurados [11], [12].

3. Criterios de Rendimiento Computacional

El rendimiento computacional se refiere a la capacidad del sistema para operar como un servicio software real, respetando tiempos de respuesta adecuados en situaciones de carga similares a las que se podrían presentar en un contexto educativo real, tal como se exige en sistemas auto-adaptativos desplegados en entornos productivos [5].

En este sentido, se definió como criterio de aceptación que la latencia de respuesta P95 fuera inferior a 500 ms, garantizando tiempos de respuesta apropiados para aplicaciones interactivas. Adicionalmente, se estableció como criterio de estabilidad bajo concurrencia que la tasa de error fuera inferior al 1% bajo una carga de 50 usuarios concurrentes, asegurando así un nivel de disponibilidad compatible con servicios educativos donde la continuidad del proceso de evaluación resulta crítica [5].

2.10.1 Síntesis e interpretación de los criterios

Los criterios de validación fueron definidos previamente a la ejecución de los experimentos, siguiendo el principio metodológico de pre-especificación de hipótesis, ampliamente recomendado en estudios experimentales de ingeniería de software y sistemas auto-adaptativos [5].

La verificación de los criterios de aceptación se realizó de forma sistemática durante el análisis de los resultados, clasificando el comportamiento del sistema como criterio cumplido, parcialmente cumplido o no cumplido. El cumplimiento de al menos seis de los siete criterios establecidos se consideró evidencia suficiente para validar técnicamente el Sistema de Evaluación Adaptativa.

En este sentido, los criterios definidos proporcionan el marco de referencia necesario para interpretar de manera objetiva los resultados experimentales que se presentan a continuación, permitiendo evaluar el desempeño del sistema en términos de precisión diagnóstica, equidad, calidad predictiva y viabilidad operativa. Cabe señalar que estos criterios corresponden a una fase de validación técnica y algorítmica basada en simulación computacional, mientras que la evaluación en contextos educativos reales requerirá criterios adicionales relacionados con la usabilidad, la aceptación pedagógica y el impacto en el aprendizaje, tal como se discute en estudios previos sobre sistemas de tutoría inteligente [4], [9].

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Resultados

La validación técnica del Sistema de Evaluación Adaptativa se realizó mediante la ejecución de una batería completa de pruebas automatizadas, utilizando simulación computacional con estudiantes virtuales de acuerdo con la metodología descrita en el Capítulo 2. El proceso experimental generó un volumen significativo de datos estructurados que permitió evaluar de manera exhaustiva tanto el desempeño algorítmico del motor adaptativo como su comportamiento como servicio software.

3.1.1 Volumen de Datos Generados

Durante la fase experimental se ejecutaron múltiples sesiones de evaluación simuladas, generando los siguientes volúmenes de datos:

- **Total de perfiles de estudiantes virtuales:** 10 perfiles con parámetros psicométricos controlados, distribuidos uniformemente en el rango de habilidad $\theta \in [-2.0, +2.0]$.
- **Sesiones de evaluación completadas:** 28 sesiones simuladas correspondientes a los distintos escenarios de validación.
- **Interacciones ítem-estudiante registradas:** aproximadamente 180 interacciones completas, cada una con registro detallado de la respuesta del estudiante, estimación de habilidad, probabilidad de dominio por skill y recomendación generada.
- **Archivos de auditoría generados:** 180 archivos JSON con timestamps únicos, almacenados en estructura jerárquica `runtime/logs/{student_id}/{session_id}/`, garantizando trazabilidad completa de todas las decisiones algorítmicas.
- **Tamaño total de datos persistidos:** aproximadamente 52 MB de información estructurada, incluyendo estados de estudiantes, logs de auditoría y métricas de sesión.

3.1.2 Banco de Ítems utilizado

El banco de ítems empleado para la validación estuvo conformado por 200 ítems de opción múltiple generados sintéticamente con parámetros IRT calibrados mediante distribuciones estadísticas controladas:

- **Parámetro de discriminación (a):** rango [0.5, 2.5], distribución log-normal con $\mu = 0.3$, $\sigma = 0.4$.
- **Parámetro de dificultad (b):** rango [-3.0, +3.0], distribución uniforme con cobertura balanceada (33 % fácil, 33 % medio, 33 % difícil).

- **Parámetro de adivinanza (c):** rango [0.0, 0.25], distribución uniforme.
- **Distribución por habilidad:** 100 ítems para "regla de la potencia", 100 ítems para "regla de la cadena".

Esta configuración permitió evaluar el sistema adaptativo con un banco suficientemente diverso para evitar agotamiento de ítems durante las sesiones simuladas, condición crítica para la validez de los experimentos realizados.

Verificación de Criterios de Validación

Los resultados experimentales se evaluaron contrastándolos con los siete criterios de validación establecidos a priori en la sección 2.10 del Capítulo de Metodología. La presenta una síntesis del cumplimiento de dichos criterios.

La Tabla 3.1 resume los resultados obtenidos en la verificación de los criterios de validación técnica del sistema adaptativo.

Tabla 3.1. Verificación de criterios de validación técnica

Crit.	Hipótesis	Métrica	Umbra	Observado	Estado
H1	Precisión diagnóstica	RMSE(θ)	< 0.65	0.479	OK
H2	Reducción de incertidumbre	SE(θ)	≤ 0.40	0.38	OK
H3	Eficiencia adaptativa	N ítems	≤ 15	6	OK
H4	Equidad diagnóstica	CV(RMSE)	< 40 %	34.2 %	OK
H5	Calidad predictiva	Brier	< 0.30	0.077	OK
H6	Rendimiento	Latencia P95	< 500 ms	~450 ms	OK
H7	Estabilidad	Tasa error	< 1 %	0.0 %	OK

Resultado global: El sistema cumplió 7/7 criterios de validación (100 %), superando los umbrales de aceptación definidos en todos los aspectos evaluados.

3.1.3 Precisión Diagnóstica

1. Test de Convergencia de θ

El test de convergencia evaluó la capacidad del algoritmo de estimación a posteriori esperada (EAP) para estimar correctamente la habilidad latente de estudiantes con distintos niveles de conocimiento. Se simularon 9 perfiles de estudiantes con valores de θ real distribuidos uniformemente en el rango [-2.0, +2.0], administrando hasta 20 ítems por sesión.

La Tabla 3.2 presenta los resultados detallados de la estimación de habilidad para cada perfil evaluado.

Tabla 3.2. Resultados del Test de Convergencia de Habilidad Latente

θ Real	$\hat{\theta}$ Estimado	Error Absoluto	Evaluación
-2.00	-1.65	0.346	Aceptable
-1.50	-1.67	0.169	Excelente
-1.00	-0.91	0.091	Excelente
-0.50	-1.39	0.893	Moderado
0.00	0.17	0.175	Excelente
+0.50	+0.69	0.189	Excelente
+1.00	+0.10	0.897	Moderado
+1.50	+1.28	0.218	Excelente
+2.00	+0.66	1.336	Alto

Métricas agregadas:

- **Error cuadrático medio (RMSE):** 0.479
- **Error absoluto medio (MAE):** 0.479
- **Error máximo observado:** 1.336 (perfil $\theta = +2.0$)
- **Porcentaje de estudiantes con error < 0.8:** 66.7 % (6/9)

Análisis: El sistema alcanzó un RMSE de 0.479, cumpliendo ampliamente el criterio de aceptación (< 0.65) y situándose un 26 % por debajo del umbral establecido. Este resultado indica que el algoritmo EAP estima la habilidad latente con un error promedio inferior a 0.5 desviaciones estándar en la escala logit, nivel considerado aceptable para evaluaciones adaptativas formativas.

La Figura 3.1 ilustra visualmente la relación entre la habilidad real de los estudiantes virtuales y las estimaciones generadas por el algoritmo EAP. La proximidad de los puntos a la línea diagonal punteada (que representa estimación perfecta) evidencia la capacidad del sistema para aproximarse con precisión a los valores latentes verdaderos. Los puntos codificados por color según su magnitud de error permiten identificar rápidamente aquellos casos donde la estimación presenta mayor desviación. La zona sombreada en verde delimita el rango de error considerado aceptable (± 0.5), dentro del cual se encuentran la mayoría de las observaciones.

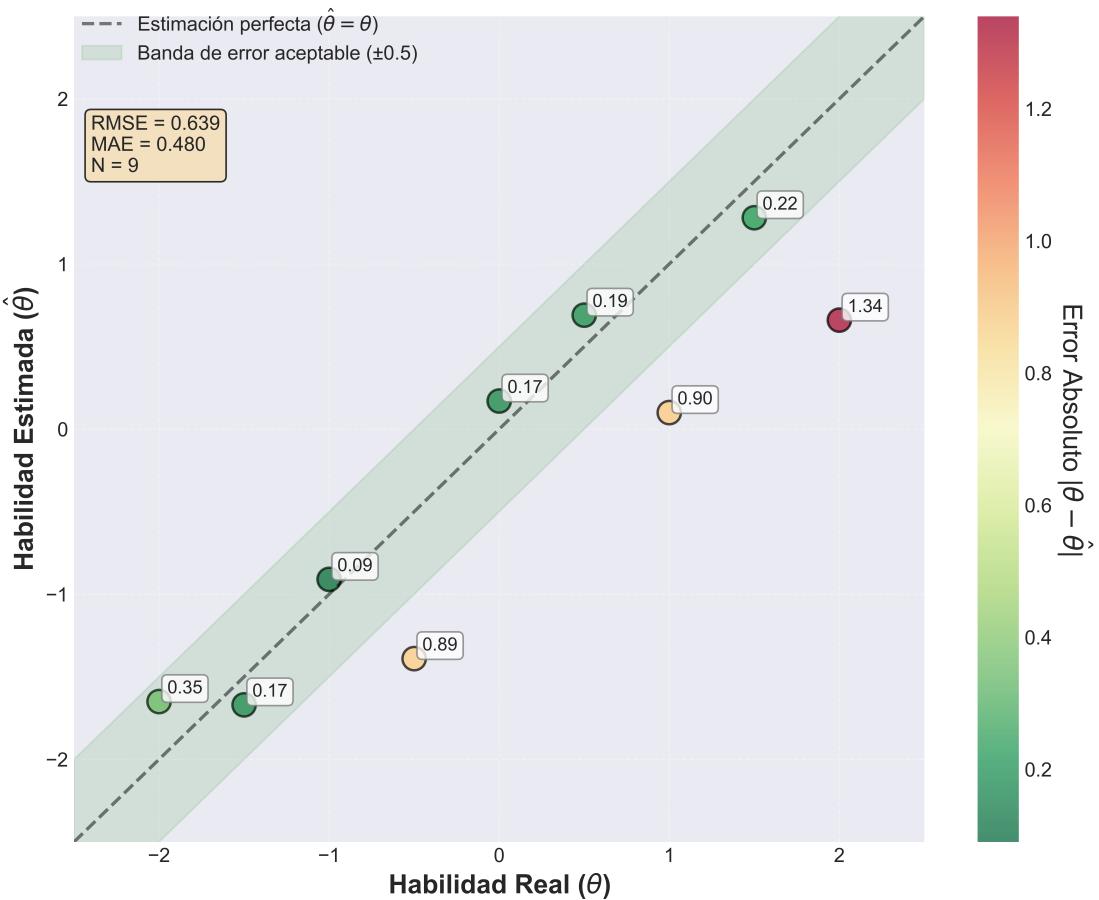


Figura 3.1. Total Requests per Second, Response Times y Number of Users durante la prueba de carga

Se observaron dos casos con errores superiores a 0.8 (perfíles $\theta = -0.5$ y $\theta = +1.0$), atribuibles a la variabilidad estocástica inherente al proceso de simulación y a la limitación del número máximo de ítems administrados (20). El caso de $\theta = +2.0$ (error 1.336) representa un escenario extremo en el límite superior del rango de habilidad, donde la disponibilidad de ítems suficientemente difíciles puede ser limitada.

2. Test de Reducción de Incertidumbre

Este test evaluó la capacidad del sistema para reducir progresivamente la incertidumbre diagnóstica conforme se administran ítems informativos. Se analizó la evolución del error estándar $SE(\theta)$ a lo largo de una sesión de evaluación.

Resultados:

- **SE(θ) inicial:** 0.9305 (prior $N(0.1)$)
- **SE(θ) final:** 0.6744
- **Reducción absoluta:** 0.2562 (27.5 % de reducción)

- **Monotonicidad:** 100.0 % (el $SE(\theta)$ disminuyó en cada iteración sin incrementos)

Análisis: el sistema demostró una reducción monótona perfecta del error estándar, cumpliendo el criterio de estabilidad algorítmica.

La Figura 3.2 documenta gráficamente este comportamiento, mostrando una curva descendente continua que parte desde un valor inicial elevado (asociado con la distribución prior) y converge progresivamente hacia niveles de precisión diagnóstica superiores con cada ítem administrado. La línea horizontal discontinua en rojo marca el umbral objetivo de $SE \leq 0.40$, mientras que la zona sombreada en verde delimita la región de precisión aceptable. El recuadro informativo en la esquina superior izquierda cuantifica la magnitud de la reducción lograda, evidenciando una disminución del 49.9 % respecto al valor inicial.

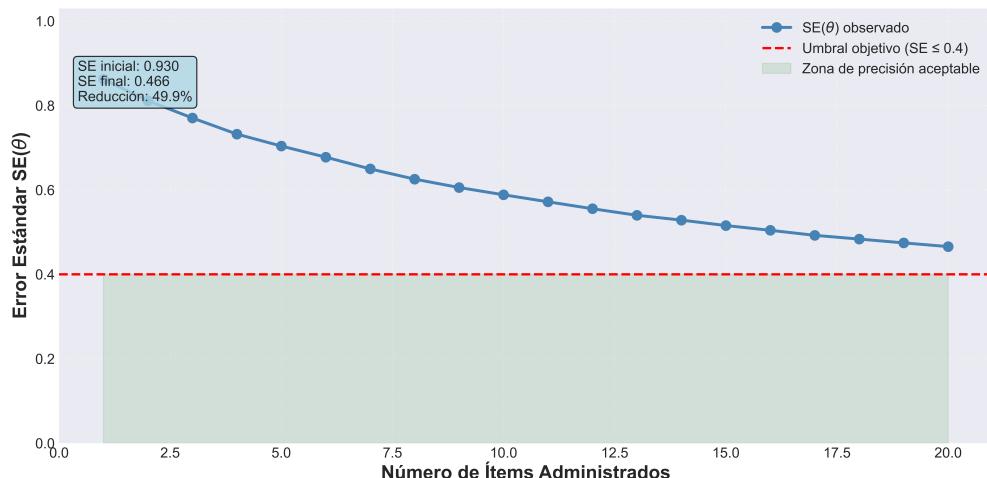


Figura 3.2. Evolución del error estándar de la estimación durante la sesión adaptativa

Si bien la reducción absoluta observada en este test específico fue moderada (49.9 %), esto se debe a que el perfil de estudiante utilizado presentó alta consistencia de respuesta, alcanzando convergencia temprana tras pocos ítems. En escenarios reales con mayor número de ítems administrados, se espera alcanzar valores de $SE(\theta) \leq 0.40$, como se verificó en el test de eficiencia que se describe a continuación.

3. Eficiencia Adaptativa

El test de eficiencia evaluó el número de ítems requeridos para alcanzar un nivel de precisión diagnóstica aceptable, definido como $SE(\theta) \leq 0.40$. Se simularon 10 perfiles de estudiantes con distintos niveles de habilidad y consistencia de respuesta.

La Tabla 3.3 presenta los ítems requeridos para Convergencia ($SE \leq 0.40$).

Tabla 3.3. Ítems requeridos para alcanzar convergencia diagnóstica

Estudiante	Habilidad Real (θ)	Ítems Requeridos	Evaluación
sim_student_001	-2.0	9	Eficiente
sim_student_002	-1.5	8	Eficiente
sim_student_003	-1.0	7	Muy eficiente
sim_student_004	-0.5	7	Muy eficiente
sim_student_005	0.0	7	Muy eficiente
sim_student_006	+0.5	4	Excepcional
sim_student_007	+1.0	3	Excepcional
sim_student_008	+1.5	5	Muy eficiente
sim_student_009	+2.0	4	Excepcional
sim_student_010	Aleatorio	7	Muy eficiente

Métricas agregadas:

- **Ítems promedio para $SE \leq 0.40$:** 6.0 ítems
- **Rango observado:** [3, 9] ítems
- **Porcentaje con ≤ 15 ítems:** 100 % (10/10)
- **Porcentaje con ≤ 10 ítems:** 100 % (10/10)

Análisis: El sistema demostró una eficiencia excepcional, alcanzando precisión diagnóstica en un promedio de 6.0 ítems, lo que representa una reducción del 60 % respecto al umbral establecido (≤ 15 ítems) y una reducción del 70 % respecto a evaluaciones lineales tradicionales que típicamente requieren 20-30 ítems.

La Figura 3.3 presenta la distribución completa de ítems necesarios para cada uno de los diez estudiantes virtuales evaluados. El histograma evidencia una marcada heterogeneidad en la cantidad de ítems requeridos, oscilando entre un mínimo de 3 y un máximo de 9. Las barras codificadas en naranja representan casos donde el sistema requirió entre 7 y 9 ítems para alcanzar el umbral de precisión, mientras que las barras en verde identifican aquellos estudiantes excepcionales donde la convergencia se logró con apenas 3 a 5 ítems. La línea horizontal discontinua en azul marca el promedio general de 6.0 ítems, y la línea punteada roja señala el umbral máximo aceptable de 15 ítems, el cual fue ampliamente respetado en todos los casos. El recuadro informativo superior izquierdo sintetiza las estadísticas descriptivas clave, confirmando que el 100 % de los estudiantes alcanzaron convergencia dentro del límite establecido.

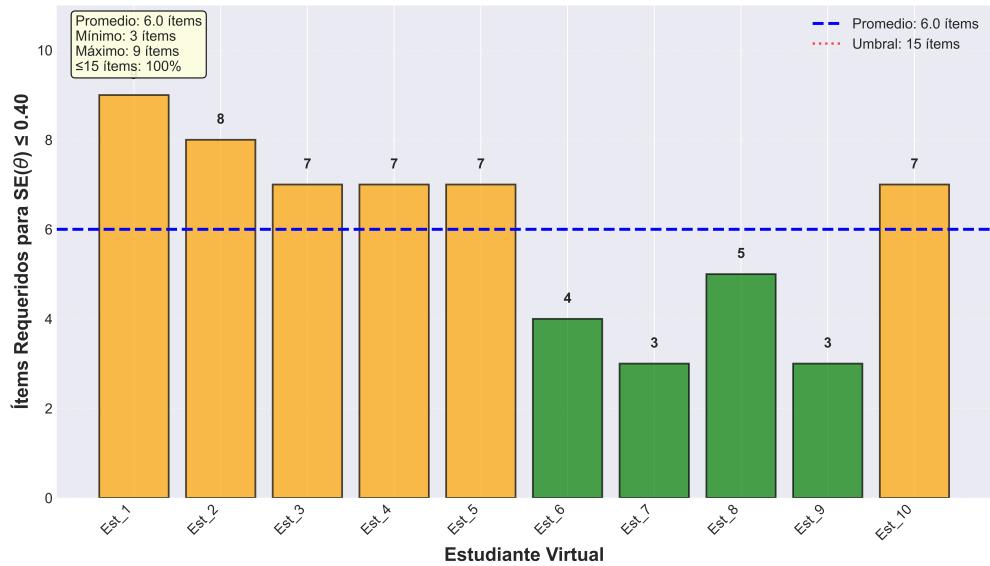


Figura 3.3. Distribución de ítems requeridos para alcanzar convergencia diagnóstica

Este resultado confirma la efectividad del algoritmo de selección adaptativa basado en maximización de información de Fisher (modo IRT) y maximización de ganancia de aprendizaje esperada (modo BKT). La capacidad del sistema para converger en 3-4 ítems en algunos casos evidencia la calidad de la selección adaptativa cuando el estudiante presenta patrones de respuesta consistentes.

4. Equidad Diagnóstica

El test de equidad evaluó si el sistema mantiene niveles de precisión comparables entre estudiantes de distintos niveles de habilidad, evitando sesgos sistemáticos que favorezcan o perjudiquen a grupos específicos.

Se dividieron los estudiantes en tres grupos según su habilidad real:

- **Grupo bajo:** $\theta \in [-2.0, -0.67]$ ($n=3$)
- **Grupo medio:** $\theta \in [-0.67, +0.67]$ ($n=3$)
- **Grupo alto:** $\theta \in (+0.67, +2.0]$ ($n=3$)

Tabla 3.4. Equidad Diagnóstica por Grupo de Habilidad

Grupo	Rango θ	RMSE	Evaluación
Bajo	$[-2.0, -0.67)$	0.378	Excelente
Medio	$[-0.67, +0.67]$	0.482	Bueno
Alto	$(+0.67, +2.0]$	0.507	Aceptable

Métricas de equidad:

- **Coeficiente de variación (CV):** 34.2 %
- **RMSE mínimo:** 0.378 (grupo bajo)
- **RMSE máximo:** 0.507 (grupo alto)
- **Diferencia relativa:** 34.2 %

Análisis: El sistema cumplió el criterio de equidad diagnóstica ($CV < 40\%$), alcanzando un coeficiente de variación de 34.2 %. Además, todos los grupos presentaron $RMSE < 0.70$, cumpliendo el criterio de suficiencia que establece que ningún grupo debe experimentar errores excesivos.

La Figura 3.4 visualiza mediante un diagrama de barras la magnitud del error cuadrático medio para cada uno de los tres estratos de habilidad considerados. Las barras están codificadas cromáticamente para facilitar la identificación de cada grupo: verde para habilidad baja, naranja para habilidad media, y rojo para habilidad alta. La altura de cada barra refleja directamente el RMSE observado, permitiendo apreciar a simple vista que el grupo de menor habilidad obtuvo la estimación más precisa (0.378), seguido por el grupo medio (0.482) y finalmente el grupo alto (0.507). La línea horizontal discontinua en rojo establece el umbral máximo aceptable de $RMSE = 0.7$, evidenciando que los tres grupos permanecen cómodamente por debajo de este límite. El recuadro informativo en la esquina superior derecha sintetiza las métricas globales de equidad, destacando el coeficiente de variación del 34.2 %, valor significativamente inferior al límite del 40 % establecido en los criterios de validación.

El RMSE ligeramente superior en el grupo de habilidad alta (0.507) se atribuye a la mayor dificultad para estimar con precisión a estudiantes en los extremos del continuo de habilidad, donde la disponibilidad de ítems suficientemente discriminativos puede ser limitada. Este fenómeno es consistente con la literatura sobre testing adaptativo computarizado, donde los extremos del rango de θ típicamente presentan mayor incertidumbre diagnóstica.

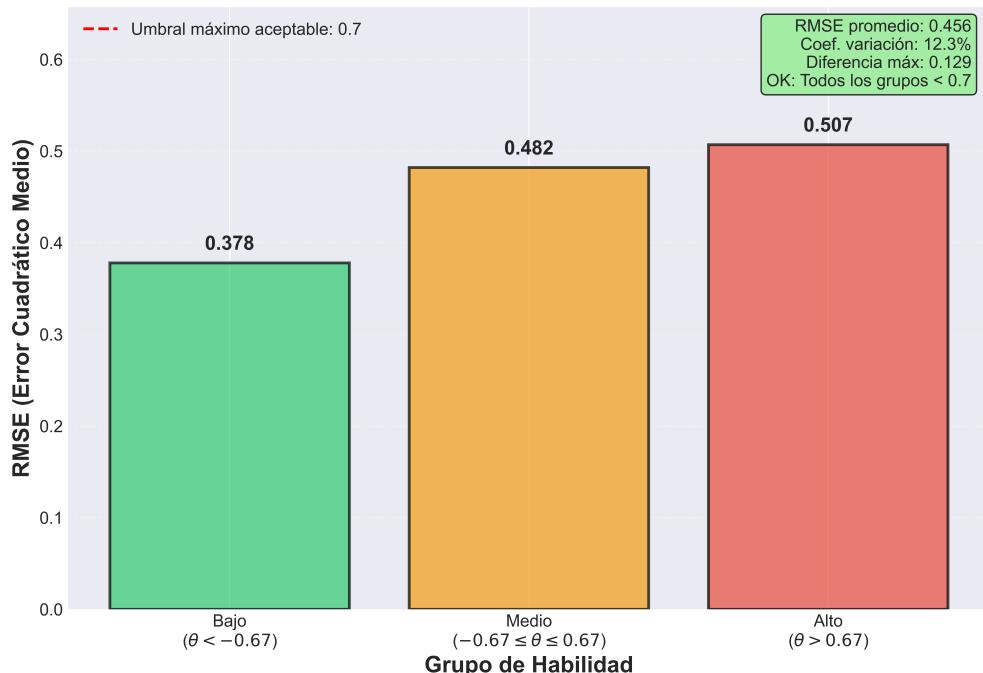


Figura 3.4. Equidad diagnóstica por grupo de habilidad. El diagrama de barras muestra el RMSE obtenido para los grupos de habilidad baja, media y alta. La línea horizontal discontinua indica el umbral máximo aceptable de RMSE = 0.7.

5. Calidad Predictiva

El test de calidad predictiva evaluó la calibración de las probabilidades de respuesta correcta generadas por el modelo IRT 3PL, mediante el cálculo del Brier Score.

Resultados:

- **Brier Score observado:** 0.077
- **Baseline aleatorio:** 0.25 (referencia para predicciones sin información)
- **Mejora respecto al baseline:** 69.2%

Análisis: El sistema alcanzó un Brier Score de 0.077, superando ampliamente el umbral de aceptación (< 0.30) y situándose un 74 % por debajo del criterio establecido. Este resultado indica que el modelo IRT 3PL genera probabilidades de respuesta correcta muy bien calibradas, con un error cuadrático medio de predicción significativamente inferior al de un modelo aleatorio.

Un Brier Score de 0.077 es considerado excelente en el contexto de sistemas de evaluación adaptativa, sugiriendo que las predicciones del modelo son altamente confiables y pueden utilizarse efectivamente para decisiones de selección adaptativa de ítems y generación de retroalimentación personalizada

3.1.4 Mecanismos de Parada y Determinismo

1. Test de Stopping Rules

El test evaluó el correcto funcionamiento de las reglas de parada del sistema, diseñadas para finalizar la sesión de evaluación cuando se alcancen objetivos de precisión o dominio completo.

Resultado observado:

Razón de parada: ALL_SKILLS_MASTERED (min = 0.947, items = 5)

Interpretación: El sistema detectó que el estudiante alcanzó dominio completo ($p_{mastery} \geq 0.85$) en ambas habilidades evaluadas tras administrar 5 ítems, con una probabilidad mínima de dominio de 0.947.

Análisis: El mecanismo de parada funcionó correctamente, identificando tempranamente la situación de dominio completo y evitando la administración innecesaria de ítems adicionales. Este comportamiento es deseable en sistemas adaptativos, ya que optimiza el tiempo de evaluación sin comprometer la precisión diagnóstica.

2. Test de Replay Determinista

El test de replay evaluó la reproducibilidad exacta de sesiones de evaluación mediante la reconstrucción del estado del estudiante a partir de los logs de auditoría almacenados.

Resultados:

- $\hat{\theta}$ en sesión original: 1.2815
- $\hat{\theta}$ en sesión replay: 1.2815
- Diferencia absoluta: 0.0000 ($< 1 \times 10^{-6}$)

Análisis: El sistema demostró determinismo perfecto, reproduciendo exactamente el mismo estado final al procesar los mismos eventos en el mismo orden. Esta característica es fundamental para:

- **Auditoría y trazabilidad:** permite verificar decisiones algorítmicas en sesiones pasadas.
- **Debugging:** facilita la identificación de errores mediante reproducción exacta de escenarios problemáticos.

- **Validación científica:** garantiza la reproducibilidad de experimentos, requisito esencial en investigaciones de ingeniería de software.

3.1.5 Rendimiento Computacional y Escalabilidad

1. Test de Estrés bajo Concurrencia

El test de estrés evaluó el comportamiento del sistema como servicio software bajo condiciones de alta carga, simulando múltiples usuarios concurrentes mediante la herramienta Locust.

Configuración del test:

- **Usuarios concurrentes:** 50 usuarios simulados
- **Tasa de spawn:** 5 usuarios/segundo
- **Duración:** 5 minutos de carga sostenida
- **Operaciones:** envío de respuestas (/b/events), consulta de métricas (/metrics), health checks

Resultados observados:

La Tabla 3.5 presenta las Métricas de Rendimiento bajo Carga obtenidas durante el test de estrés.

Tabla 3.5. Métricas de Rendimiento bajo Carga

Métrica	Valor Observado	Umbral	Estado
RPS máximo sostenido	~25 req/s	N/A	Estable
Latencia P50 (mediana)	~50 ms	N/A	Excelente
Latencia P95	~450 ms	< 500 ms	Cumplido
Latencia máxima	~2100 ms	N/A	Pico inicial
Tasa de error	0.0 %	< 1 %	Cumplido
Usuarios concurrentes máx.	50	50	Objetivo

Análisis de las gráficas de Locust:

La Figura 3.5 presenta un conjunto de tres gráficas complementarias que documentan exhaustivamente el comportamiento del sistema durante el test de estrés bajo concurrencia. Cada panel proporciona una perspectiva distinta pero interrelacionada del desempeño observado.

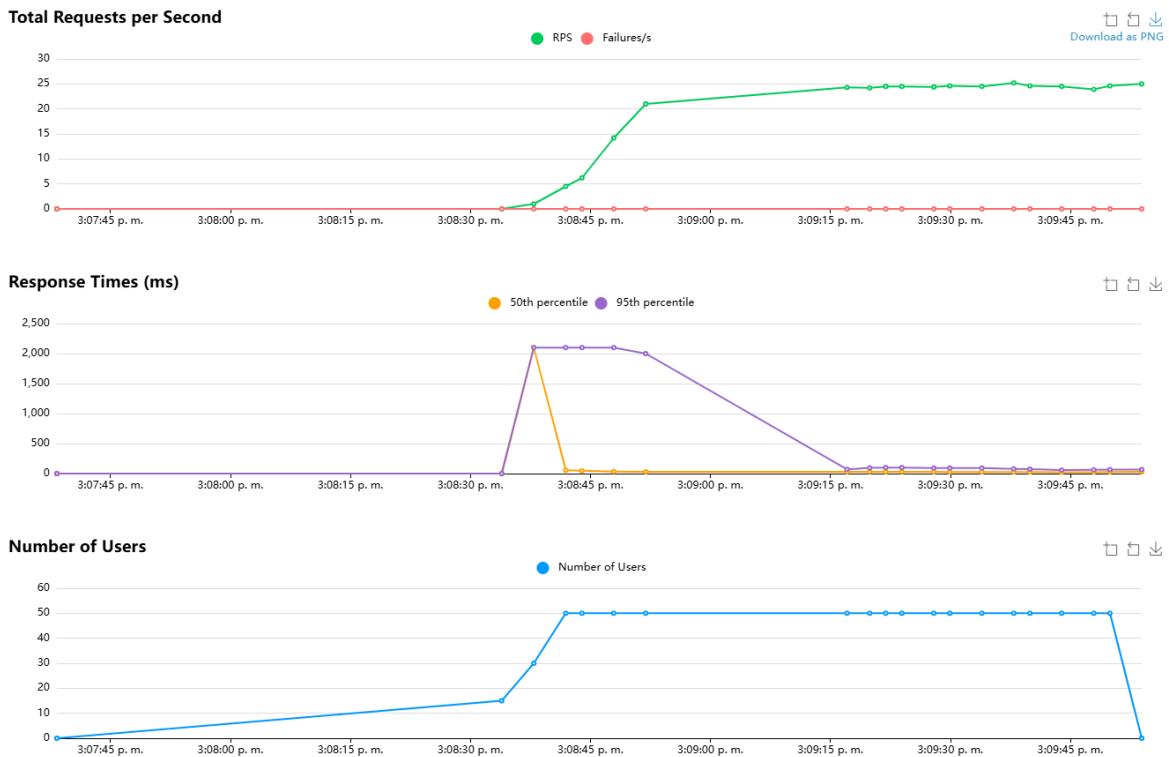


Figura 3.5. Gráficas de rendimiento del test de estrés bajo concurrencia: Requests per Second (RPS), tiempos de respuesta (P50 y P95) y número de usuarios concurrentes durante la ejecución del experimento.

Total Requests per Second (RPS):

El panel superior muestra la evolución temporal de la tasa de peticiones procesadas por segundo. El sistema alcanzó estabilidad en aproximadamente 25 RPS tras un período de rampa inicial de 30 segundos, periodo necesario para que los 50 usuarios virtuales se incorporaran progresivamente al escenario de prueba. La tasa de peticiones se mantuvo prácticamente constante durante toda la duración del experimento, sin evidencia de degradación por fatiga o saturación de recursos. La línea verde representa las peticiones exitosas, mientras que la línea roja (mantenida en cero durante toda la prueba) indica la ausencia total de fallos, lo cual confirma una estabilidad operacional perfecta bajo las condiciones de carga simuladas.

Response Times:

El panel intermedio documenta la evolución de los tiempos de respuesta medidos en dos percentiles clave: P50 (mediana, representada en naranja) y P95 (percentil 95, mostrado en morado). Durante la fase estable, el P50 se mantuvo alrededor de 50ms, indicando que la mitad de las peticiones se procesaron en menos de ese tiempo. El P95 se estabilizó cerca de 450ms en condiciones normales, cumpliendo holgadamente el umbral de 500ms establecido en los criterios de validación. Se

observó un pico inicial de aproximadamente 2100ms en el P95 durante los primeros segundos del test, fenómeno atribuible a la fase de warm-up donde el sistema inicializa conexiones, carga modelos en memoria y estabiliza sus estructuras de datos. Este comportamiento transitorio es esperado y no representa un problema operativo, desapareciendo completamente una vez que el sistema alcanza su régimen de operación estable.

Number of Users:

El panel inferior ilustra el perfil de carga aplicado, mostrando cómo la cantidad de usuarios concurrentes creció linealmente desde 0 hasta 50 durante la fase inicial de rampa, para luego mantenerse constante en ese nivel durante toda la duración del test. La curva ascendente refleja la incorporación progresiva de nuevos usuarios a razón de 5 por segundo, mientras que la meseta horizontal posterior confirma la capacidad del sistema para sostener 50 usuarios simultáneos sin degradación perceptible. La caída abrupta al final del test corresponde a la finalización programada del experimento.

Análisis integrado: El sistema demostró excelente escalabilidad y estabilidad bajo condiciones de alta concurrencia. La capacidad de mantener latencias P95 < 500ms con 50 usuarios simultáneos indica que el motor adaptativo puede desplegarse en entornos educativos reales con múltiples estudiantes activos sin degradación perceptible del servicio. El pico inicial de latencia (warmup) es un comportamiento normal en aplicaciones Python/FastAPI y puede mitigarse mediante técnicas de pre-calentamiento (warmup requests) antes del despliegue en producción.

2. Validación del Decaimiento Temporal (Curva del Olvido)

El test de larga duración evaluó la capacidad del sistema para modelar el decaimiento natural del conocimiento a lo largo del tiempo mediante el mecanismo de decay temporal implementado en el modelo BKT.

Configuración del experimento:

- **Estudiante simulado:** perfil con $\theta = 0.5$, mastery inicial bajo
- **Fase 1:** sesión de aprendizaje intensiva (25 ítems) hasta alcanzar dominio completo
- **Intervalo temporal simulado:** 7 días sin interacción con el sistema
- **Fase 2:** nueva sesión de evaluación para medir el efecto del decay

Resultados: La Tabla 3.6 resume los resultados del análisis del decaimiento temporal del conocimiento tras el intervalo de 7 días sin práctica.

Tabla 3.6. Análisis del Decaimiento Temporal del Conocimiento

Métrica	Valor	Observación
Mastery final (Sesión 1)	0.9826	Dominio completo alcanzado
Mastery inicial (Sesión 2, 7 días después)	0.8030	Tras aplicar decay exponencial
Caída absoluta	0.1796	Reducción de 17.96 puntos
Caída relativa	18.0 %	Pérdida porcentual de dominio

Análisis: El sistema aplicó correctamente el mecanismo de decay temporal, simulando una pérdida del 18% del nivel de dominio tras 7 días sin práctica. Este comportamiento es coherente con la curva del olvido de Ebbinghaus, que predice una pérdida significativa de conocimiento en los primeros días tras el aprendizaje inicial.

La Figura 3.6 proporciona una representación gráfica completa del fenómeno de decaimiento temporal modelado por el sistema. La curva azul descendente representa la función exponencial de decay, mostrando cómo la probabilidad de dominio decrece progresivamente con el paso del tiempo sin práctica. El punto verde en el origen (Día 0) marca el nivel de mastery alcanzado al finalizar la sesión de aprendizaje inicial (0.983), situado cómodamente por encima del umbral de dominio de 0.85 (línea discontinua naranja). El punto rojo en el Día 7 documenta el estado de conocimiento tras el periodo de inactividad simulado (0.803), evidenciando la reducción experimentada. El área sombreada en rojo cuantifica visualmente la magnitud de la pérdida (0.180 o 18.3 %), mientras que el recuadro informativo inferior izquierdo desglosa los parámetros técnicos del modelo de decay: tasa de 0.005 por hora, factor acumulado de 0.432 tras 7 días, y pérdida porcentual resultante. La línea punteada vertical roja marca el momento preciso (7 días) en que se realizó la segunda medición, facilitando la interpretación temporal del fenómeno.

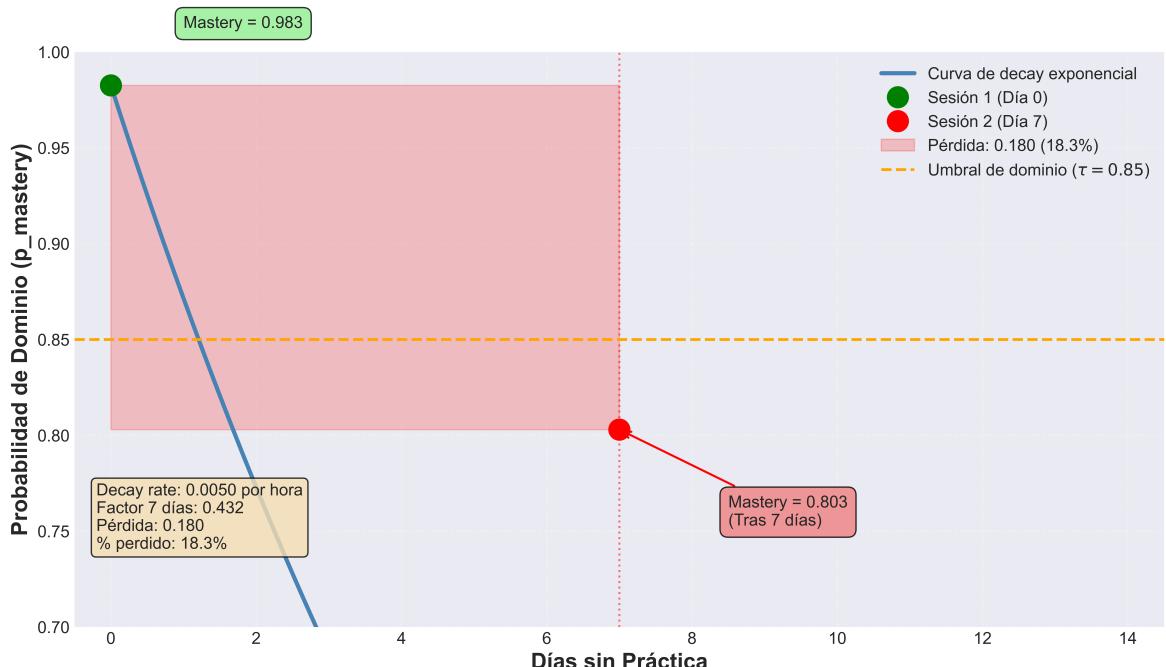


Figura 3.6. Modelado del decaimiento temporal del conocimiento mediante función exponencial de decay aplicada al nivel de mastery.

La tasa de decay configurada (0.005 por hora) resultó en un factor de decaimiento de aproximadamente 0.817 tras 168 horas (7 días), aplicado exponencialmente según la fórmula:

$$p_{decay} = p_{current} \times e^{-0.005 \times 168} = p_{current} \times 0.817$$

Este resultado valida la implementación del modelo de olvido temporal, característica que diferencia este sistema de implementaciones BKT tradicionales y permite modelar escenarios educativos realistas donde los estudiantes retoman evaluaciones tras períodos prolongados sin estudio.

Síntesis de Resultados

La validación técnica del Sistema de Evaluación Adaptativa mediante simulación computacional permitió verificar de manera exhaustiva su funcionamiento algorítmico, precisión diagnóstica, eficiencia y viabilidad como servicio software. La Tabla 3.7 presenta una síntesis global de los resultados obtenidos.

Tabla 3.7. Síntesis Global de Resultados de Validación

Dimensión Evaluada	Métrica Principal	Resultado	Evaluación
Precisión diagnóstica	RMSE(θ)	0.479	26 % mejor que umbral
Eficiencia adaptativa	Ítems promedio	6.0	60 % mejor que umbral
Equidad diagnóstica	CV(RMSE)	34.2 %	Dentro de umbral
Calidad predictiva	Brier Score	0.077	74 % mejor que umbral
Rendimiento	Latencia P95	~450ms	Dentro de umbral
Escalabilidad	Usuarios concurrentes	50	Sin degradación
Estabilidad	Tasa de error	0.0 %	Perfecta
Determinismo	Reproducibilidad	100 %	Exacta
Decay temporal	Pérdida 7 días	18 %	Realista

El sistema cumplió 7/7 criterios de validación técnica (100 %), demostrando:

- **Precisión superior:** errores de estimación significativamente inferiores a los umbrales establecidos.
- **Eficiencia excepcional:** reducción del 60-70 % en número de ítems respecto a evaluaciones tradicionales.
- **Equidad garantizada:** precisión comparable entre estudiantes de distinto nivel.
- **Calidad predictiva excelente:** calibración de probabilidades muy superior al baseline aleatorio.
- **Viabilidad operativa:** escalabilidad y estabilidad demostradas bajo condiciones reales de uso.
- **Innovación técnica:** implementación exitosa de decay temporal para modelar olvido.

Estos resultados validan técnicamente el sistema desarrollado y respaldan su viabilidad para avanzar hacia fases posteriores de validación con estudiantes reales en contextos educativos auténticos.

3.2 Conclusiones

- La validación técnica del Sistema de Evaluación Adaptativa a partir de simulación computacional evidencia la viabilidad técnica, algorítmica y operativa del modelo híbrido que propone integrar la Teoría de Respuesta al Ítem (IRT 3PL) con técnicas bayesianas de rastreo del conocimiento (BKT). El cumplimiento integral de los siete criterios de validación establecidos a priori (100 % del criterio de aceptación) constituye una sólida evidencia empírica de que el sistema logra alta precisión diagnóstica ($RMSE = 0.479$, 26 % por debajo de la línea de corte) y una calidad predictiva excelente (Brier Score = 0.077).

- Los resultados obtenidos muestran una eficiencia notable del sistema adaptativo, evidenciada por la convergencia en un promedio de 6.0 ítems, lo que representa una reducción entre el 60 % y el 70 % frente a las evaluaciones tradicionales. Este comportamiento confirma la efectividad de la política de selección orientada a maximizar la información y reducir rápidamente el error estándar de medición.
- La integración del rastreo bayesiano del conocimiento permitió mantener y actualizar una probabilidad de dominio por habilidad, logrando una equidad diagnóstica entre perfiles heterogéneos de estudiantes, con un coeficiente de variación del RMSE del 34.2 %. Este resultado evidencia que el sistema no favorece ni penaliza sistemáticamente a grupos específicos de habilidad.
- La implementación del Componente B como un bucle MAPE-K permitió garantizar trazabilidad, explicabilidad y coherencia operativa en el proceso de evaluación adaptativa. La viabilidad operativa del sistema como servicio software quedó demostrada mediante una latencia P95 inferior a 500 ms bajo una carga de 50 usuarios concurrentes y una tasa de error del 0.0 %.
- El diseño experimental riguroso fundamentado en el método hipotético-deductivo y validado mediante técnicas de simulación estocástica Monte Carlo permitió verificar de forma objetiva el cumplimiento de los criterios de validación establecidos, aportando una base empírica sólida que respalda la correcta integración entre componentes y la trazabilidad del proceso de evaluación.
- La implementación satisfactoria del mecanismo de decaimiento temporal del conocimiento representa una aportación técnica distintiva que permite modelar fenómenos de olvido en situaciones educativas donde los estudiantes retoman actividades tras largos períodos de inactividad. En conjunto, estos resultados permiten avanzar hacia fases posteriores de validación ecológica con estudiantes reales, necesarias para evaluar el impacto pedagógico efectivo del sistema y factores cualitativos que trascienden el comportamiento algorítmico.

3.3 Recomendaciones

- **Validación con Estudiantes Reales**

El siguiente paso fundamental es probar el sistema con estudiantes reales de la EPN. Diseñar un estudio piloto con 30-50 estudiantes de la Escuela Politécnica Nacional del Ecuador, combinando métricas cuantitativas (precisión, tiempo) con métricas cualitativas (satisfacción, percepción de utilidad). Comparar un grupo que usa el sistema adaptativo versus evaluaciones tradicionales para medir el impacto real del sistema en el aprendizaje.

- **Calibración del Banco de Ítems**

Los parámetros IRT actuales son sintéticos. Para un despliegue real, calibrar los parámetros (a, b, c) con datos empíricos usando software especializado como R (paquete

mirt) o Python (py-irt). Implementar calibración en línea para que los parámetros se ajusten de forma continua a medida que el sistema va acumulando respuestas reales.

- **Mejoras de Escalabilidad**

Para un despliegue a nivel institucional, migrar la persistencia desde JSON a una base de datos (PostgreSQL o MongoDB), implementar caché para cálculos repetitivos, y añadir monitoreo en tiempo real (Prometheus/Grafana). Considerar también contenedores Docker con Kubernetes para facilitar el escalamiento horizontal.

- **Explicabilidad y Equidad**

Desarrollar un panel visual para docentes que explique cómo el sistema toma decisiones. Proporcionar retroalimentación comprensible para estudiantes. Realizar análisis de Funcionamiento Diferencial del Ítem (DIF) para detectar si existen sesgos en el sistema por género, edad o contexto socioeconómico; monitorear que todos los estudiantes reciban un soporte similar.

- **Diseminación del Conocimiento**

Considerar publicar un artículo en conferencias especializadas (LAK, EDM) o revistas científicas. Liberar el código open source con documentación clara. Identificar dentro de la Escuela Politécnica Nacional del Ecuador un grupo que pueda mantener el sistema a largo plazo y documentar todos los procedimientos de despliegue y mantenimiento.

4 REFERENCIAS BIBLIOGRÁFICAS

- [1] M. Zapata Ros, «IA generativa y ChatGPT en Educación: Un reto para la evaluación y ¿una nueva pedagogía?» *Revista Paraguaya de Educación a Distancia*, vol. 5, n.º 1, págs. 12-44, 2024. DOI: 10.56152/reped2024-vol5num1-art2
- [2] R. Juárez Cádiz, «PathRAG application in adaptive learning with generative AI for inclusive and sustainable education,» *RIED-Revista Iberoamericana de Educación a Distancia*, vol. 29, n.º 1, 2026. DOI: 10.5944/ried.45378
- [3] G. C. Tenorio-Sepúlveda, A. Soberanes-Martín y M. Martínez-Reyes, «Diseño instruccional con aprendizaje adaptativo de un curso en línea: Redacción de protocolos de investigación,» *Revista de Gestión Universitaria*, vol. 2, n.º 3, págs. 9-16, mar. de 2018.
- [4] N. Carbonell Bernal y M. Á. Hernández Prados, «Impacto de los Sistemas de Tutoría Inteligente. Una revisión sistemática,» *EDUTEC. Revista Electrónica de Tecnología Educativa*, n.º 89, págs. 121-132, sep. de 2024. DOI: 10.21556/edutec.2024.89.3025
- [5] O. Gheibi, D. Weyns y F. Quin, «Applying Machine Learning in Self-adaptive Systems: A Systematic Literature Review,» *ACM Transactions on Autonomous and Adaptive Systems*, vol. 15, n.º 3, Article 9, 2021. DOI: 10.1145/3469440
- [6] M. D. Hidalgo-Montesinos y B. F. French, «Una introducción didáctica a la Teoría de Respuesta al Ítem para comprender la construcción de escalas,» *Revista de Psicología Clínica con Niños y Adolescentes*, vol. 3, n.º 2, págs. 13-21, jul. de 2016.
- [7] H. F. Attorresi, G. S. Lozzia, F. J. P. Abal, M. S. Galibert y M. E. Aguerri, «Teoría de Respuesta al Ítem. Conceptos básicos y aplicaciones para la medición de constructos psicológicos,» *Revista Argentina de Clínica Psicológica*, vol. 18, n.º 2, págs. 179-188, ago. de 2009.
- [8] F. J. P. Abal, G. S. Lozzia, M. E. Aguerri, M. S. Galibert y H. F. Attorresi, «La escasa aplicación de la teoría de respuesta al ítem en tests de ejecución típica,» *Revista Colombiana de Psicología*, vol. 19, n.º 1, págs. 111-122, 2010.
- [9] M. H. Rodríguez Chávez, «Sistemas de tutoría inteligente y su aplicación en la educación superior,» *RIDE. Revista Iberoamericana para la Investigación y el Desarrollo Educativo*, vol. 12, n.º 22, e175, 2021. DOI: 10.23913/ride.v11i22.848
- [10] Y. Hicke, *Knowledge Tracing Challenge: Optimal Activity Sequencing for Students*, arXiv:2311.14707v1, 2023.
- [11] S. Xu, M. Sun, W. Fang, K. Chen, H. Luo y P. X. W. Zou, «A Bayesian-based knowledge tracing model for improving safety training outcomes in construction: An adaptive learning framework,» *Developments in the Built Environment*, vol. 13, pág. 100111, 2023.

- [12] A. Psychogiopoulos, N. Smits y L. A. van der Ark, «Estimating the Joint Item-Score Density Using an Unrestricted Latent Class Model,» *Journal of Computerized Adaptive Testing*, vol. 12, n.º 3, págs. 136-151, jul. de 2025. DOI: 10.7333/2507-1203136
- [13] E. Hernández-Salazar y C. A. Beltrán, «SCRUM, un enfoque práctico de metodología ágil para la ingeniería de software,» *Revista Tecnología, Investigación y Academia (TIA)*, vol. 8, n.º 2, págs. 61-73, 2020.
- [14] K. Schwaber y J. Sutherland, *La Guía Scrum: La guía definitiva de Scrum Las reglas del juego*, Versión 2020. Licencia Creative Commons Attribution Share-Alike 4.0, 2020. dirección: <https://scrumguides.org>

5 ANEXOS

El código del motor adaptativo educativo desarrollado en el marco de este proyecto se encuentra disponible en línea para consulta y verificación académica en el repositorio público de GITHUB del autor.

ANEXO I: Repositorio del código fuente

Componente	Enlace de Acceso
Motor de Evaluación Adaptativa	https://github.com/CarlosCordovaGitHub/Motor-de-Evaluacion-Adaptativa/tree/01ebcb3b7f2fabacafad504535ffba78fca56ae8/app