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 DE SOFTWARE o 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.

DEDICATORIA

Dedico este trabajo a mi familia, cuyo apoyo constante, esfuerzo y confianza han sido fundamentales a lo largo de mi formación académica y personal. Su acompañamiento incondicional ha sido un pilar esencial para alcanzar este logro.

AGRADECIMIENTOS

Agradezco a la Escuela Politécnica Nacional por la formación académica brindada a lo largo de mi carrera universitaria. De manera especial, expreso mi gratitud al director de este trabajo, PhD. Enrique Andrés Larco Ampudia, por su orientación, observaciones y acompañamiento durante el desarrollo del presente proyecto.
Asimismo, agradezco a los docentes, compañeros y familiares que, de distintas maneras, contribuyeron con su apoyo y motivación a la culminación de esta etapa académica.

ÍNDICE DE CONTENIDO

ÍNDICE GENERAL

RESUMEN

ABSTRACT

1 DESCRIPCIÓN DEL COMPONENTE

1.1 Descripción general 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].

1.2 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 y BKT/KT) para medir con precisión el desempeño del estudiante y construir trayectorias de aprendizaje genuinamente personalizadas que se ajusten dinámicamente a partir de datos individuales.

1.3 Objetivos específicos

1.3.1 Implementar un sistema de medición dual del perfil del estudiante mediante señales complementarias

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.

1.3.2 Desarrollar una política de selección adaptativa basada en IRT para contextos de diagnóstico y certificación

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.

1.3.3 Implementar un sistema de rastreo de conocimiento basado en BKT/KT para orientar la práctica guiada

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

1.3.4 Diseñar una arquitectura fundamentada en el patrón MAPE-K para sistemas auto-adaptativos

Implementar el Componente B como un bucle MAPE-K 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.

1.3.5 Establecer un contrato explícito de integración entre componentes que garantice trazabilidad

Diseñar la integración $A \leftrightarrow B$ como un contrato de datos simple y explícito donde el Componente A envíe datos del usuario, tema activo e historial de interacciones, y el Componente B responda con el ítem propuesto, dificultad objetivo, tipo de ayuda recomendada y justificación concisa de la decisión, haciendo las decisiones auditables y proporcionando al docente evidencia clara del proceso.

1.3.6 Implementar salvaguardas de validez, equidad y mecanismos de monitoreo continuo del sistema

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 o log-loss, y realización de pruebas DIF y análisis de brechas entre perfiles para detectar posibles sesgos.

1.4 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], [3], [4], [5], [6].

1.5 Marco teórico

1.5.1 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 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, 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].

1.5.2 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 ínvariancia': medir con la misma escala, aunque cambien los sujetos o los ítems (dentro de ciertos supuestos) [1], [2].

1.5.2.1 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) usa solo b, 2PL usa a y b, y 3PL 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].

1.5.2.2 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].

1.5.2.3 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 [7].

1.5.2.4 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 tu proyecto, esto es clave para que el Componente B seleccione o recomiende ítems con mayor ”ganancia informativa” [7], [8].

1.5.2.5 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$\rightarrow$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 [7], [9].

1.5.3 Sistemas de Tutoría Inteligente (ITS)

Un Sistema de Tutoría Inteligente (ITS) 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].

1.5.3.1 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 [5].

1.5.3.2 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], [5].

1.5.4 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).

1.5.4.1 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. 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”. 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 (o criterios cercanos como la información de Fisher o la divergencia KL). Para mantener validez y equidad, se aplican restricciones de contenido (temas/objetivos), control de exposición (evitar sobreuso 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 [7], [8].
Qué aporta al Componente A↔B 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 [7], [8].

1.5.4.2 Selección guiada por BKT/KT (apoyo al aprendizaje)

Bayesian Knowledge Tracing (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). 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. 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 [6], [11].

1.5.4.3 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). En la arquitectura del proyecto, esto se implementa como un bucle MAPEK: 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 [6].

1.5.5 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 tanto aprende y con qué esfuerzo?). A continuación se presentan métricas nucleares, escritas en lenguaje claro, conectando la literatura de pruebas adaptativas y trazado del conocimiento [8], [12].

1.5.5.1 Precisión y validez de medición (IRT/CAT)

En pruebas adaptativas orientadas a medir con exactitud, la precisión se observa en el error estándar del estimador de habilidad, $SE(\hat{\theta})$, y en la información del test alrededor

del nivel estimado. Un buen algoritmo reduce $SE(\hat{\theta})$ con menos ítems: ese equilibrio entre precisión y longitud del test es central. Para comparar métodos, es útil fijar la eficiencia (mismo número medio de ítems) y contrastar la precisión resultante, o al revés. En estudios recientes se equiparan ambos métodos (por ejemplo, IRTCAT vs. enfoques alternativos) y se evalúa la diferencia media y la correlación de los puntajes con respecto a un "test completo" considerado referencia [12].
Otro indicador clave es la calibración/ajuste del modelo: el sesgo (diferencia promedio entre el estimado y el valor de referencia), la RMSE (raíz del error cuadrático medio) y las curvas de calibración por tramos de la escala. Para motores más flexibles, se recurre a medidas de divergencia como $KL(\pi||\pi)$, que cuantifican la pérdida de información entre la densidad verdadera de puntajes y la estimada; valores pequeños señalan mejor ajuste. En simulaciones de calibración y selección de modelo, la elección del criterio de información (p. ej., BIC) afecta de forma tangible la precisión final de las estimaciones [12].

1.5.5.2 Eficiencia y carga de respuesta

La eficiencia refleja cuántos ítems o cuánto tiempo necesita el sistema para alcanzar una precisión aceptable. Métricas prácticas incluyen: número promedio de ítems administrados, desviación típica de ese número (variabilidad de carga entre estudiantes), tiempo por objetivo alcanzado y porcentaje de ítems no administrados (ahorro respecto del banco total). Un buen sistema es más corto sin sacrificar precisión. Además, conviene analizar cómo varía la carga según el nivel verdadero: algunos algoritmos exigen más ítems en los extremos de la escala, otros en el centro; reconocer ese patrón ayuda a planificar bancos y reglas de parada [12].

1.5.5.3 Aprendizaje y progreso

Cuando el objetivo es que el estudiante aprenda (no solo medir), importan indicadores de progreso: la ganancia entre pre- y post-prueba normalizada por la dificultad, la tasa de dominio por unidad, el tiempo/ítems hasta alcanzar un umbral de dominio, y la retención tras un intervalo (espaciado). En sistemas con trazado del conocimiento (KT/BKT), puede reportarse la probabilidad de dominio por habilidad y su evolución, verificando que las decisiones (refuerzo, explicación, práctica guiada) incrementen esa probabilidad de forma sostenida [8], [10].

1.5.5.4 Calidad predictiva de la política adaptativa

Para validar que el sistema "elige bien lo siguiente", se evalúa su poder de predicción de respuestas y de dominio futuro. Métricas comunes son log-loss (pérdida de probabilidad), AUC/ROC para acierto de la próxima respuesta y exactitud o F1 en clasificación de dominio/no-dominio. En secuenciación adaptativa, también se puede medir el beneficio esperado o regret acumulado frente a una política de referencia. Estas métricas no sustituyen

a la evidencia de aprendizaje, pero aseguran que el motor de decisión es consistente y estable [10], [11].

1.5.5.5 Equidad y robustez

Un sistema adaptativo debe ser justo y estable. La equidad se estudia con análisis DIF (ítems que favorecen a subgrupos), comparaciones de error/precisión y tasa de dominio entre perfiles, y auditorías de exposición de ítems. La robustez exige pruebas de sensibilidad a supuestos del modelo (unidimensionalidad, independencia local) y validación cruzada cuando se recalibra el banco. Por último, la transparencia en el uso de datos y la interpretabilidad de reportes para docentes son métricas de calidad percibida y confianza [7], [12].

1.5.5.6 Reporte para la integración A$\leftrightarrow$B

Para cerrar el ciclo A$\leftrightarrow$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.

2 METODOLOGÍA

2.1 Enfoque y diseño de la investigación

En el desarrollo del Sistema de Evaluación Adaptativa (Componente B) 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 [7], [8], [11], [12].

2.1.1 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], [6].

2.1.2 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], [6], [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. Finalmente, el diseño metodológico propuesto permite establecer que esta fase corresponde a una validación algorítmica y técnica del sistema en condiciones controladas. La arquitectura del estudio contempla la ejecución de pruebas con usuarios reales para una fase posterior del proceso de investigación, una fase que estará orientada al análisis del impacto pedagógico efectivo del sistema y al estudio de factores cualitativos emergentes en contextos auténticos de aprendizaje.

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.

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

Cuadro 2.1: Criterios cuantitativos utilizados para la validación del motor adaptativo

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], [6]. 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]. La simulación posibilita la reproducción de escenarios complejos bajo condiciones controladas, lo que contribuye a fortalecer la validez interna del estudio y a garantizar la reproducibilidad de los experimentos. 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. 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. Este planteamiento permite analizar el comportamiento del sistema frente a una amplia diversidad de perfiles de aprendizaje. Este enfoque ha sido ampliamente empleado para analizar la estabilidad, la equidad diagnóstica y la eficiencia de algoritmos de secuenciación adaptativa y de rastreo del conocimiento [10], [11].

2.2.2 Análisis transversal y longitudinal

El diseño experimental por simulación permitió llevar a cabo experimentos de tipo transversal y longitudinal. En el análisis transversal se observa la respuesta inmediata del sistema ante distintos perfiles de estudiantes, mientras que el análisis longitudinal permite simular la evolución temporal del aprendizaje considerando factores tales como la consolidación del conocimiento adquirido o el decaimiento progresivo del mismo. Este tipo de análisis es crítico para los sistemas de evaluación adaptativa porque permite evaluar la capacidad del modelo para detectar la pérdida gradual de dominio en las habilidades y recomendar intervenciones oportunas, tal como subrayan las publicaciones científicas relacionadas [4], [11]. El diseño también permitió 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. La inclusión de estas pruebas contribuyó 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.

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

Cuadro 2.2: Clasificación metodológica y justificación técnica del estudio

2.3 Método de investigación

Para llevar a cabo el desarrollo y validación del Sistema de Evaluación Adaptativa (Componente B) se utilizó el método de investigación hipotéticodeductivo, 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 hipotéticodeductivo 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 [4], [6], [8] tuvieran coherencia.

2.3.1 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], [5]. 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 [8], [10], [11].

2.3.2 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.

2.3.3 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 hipotéticodeductivo 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 [6].

2.3.4 Aporte del método a la rigurosidad científica

El uso del método hipotéticodeductivo determinó que la investigación se sustentara en un proceso lógico, verificable y replicable, en el cual cada decisión de diseño tiene su respaldo en una hipótesis explícita y cada resultado experimental realiza la validación o refutación de dicha hipótesis. De esta forma se hizo más palpable el rigor científico del trabajo, al igual que se garantizó que la propuesta del Sistema de Evaluación Adaptativa respondiera a un proceso sistemático de investigación propio de la Ingeniería en Ciencias de la Computación en lugar de decisiones determinadas de un modo empírico o arbitrario.

2.4 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.

2.4.1 Población objetivo

La población objetivo del Sistema de Evaluación Adaptativa (Componente B) está constituida por estudiantes universitarios de nivel superior que cursan asignaturas de cálculo diferencial, en concreto, aquellos que se encuentran en la fase de aprendizaje del tema de las derivadas y sus aplicaciones. 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 () que refleja el nivel de dominio del contenido matemático 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 [7], [8]. 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.

2.4.2 Muestra de validación

Dadas las limitaciones éticas, logísticas y operativas asociadas a la experimentación intensiva con estudiantes reales en etapas tempranas de desarrollo de sistemas adaptativos, se optó por realizar la validación inicial mediante simulación computacional con estudiantes virtuales, enfoque ampliamente aceptado en la literatura especializada en aprendizaje adaptativo y sistemas de tutoría inteligente [3], [6], [10]. La muestra de validación estuvo conformada por N = 10 perfiles de estudiantes virtuales, cada uno caracterizado 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 un valor específico a cada perfil: $\{-2{,}0, -1{,}5, -1{,}0, -0{,}5, 0{,}0, +0{,}5, +1{,}0, +1{,}5, +2{,}0\}$, además de un décimo valor aleatorio dentro del intervalo con el fin de incrementar la variabilidad del conjunto de estudiantes simulados. Esta distribución permite representar estudiantes con bajo dominio ($\theta < -1{,}0$), dominio medio ($-1{,}0 \leq \theta \leq +1{,}0$) y dominio alto ($\theta > +1{,}0$).

- **Probabilidad de dominio inicial por habilidad (*Mastery*):** Modelada de forma correlacionada con la habilidad latente, siguiendo la relación $p_{\text{mastery}} \approx (\theta + 2)/4$, incorporando una variación estocástica de tipo gaussiano con desviación estándar $\sigma = 0{,}15$ para reflejar heterogeneidad realista entre estudiantes. Los valores resultantes fueron posteriormente normalizados en el intervalo $[0{,}1, 0{,}9]$.

- **Consistencia de respuesta:** Parámetro que representa la probabilidad de que el estudiante responda de forma coherente con su nivel de habilidad, distribuido uniformemente en el intervalo $[0{,}80, 0{,}95]$, donde valores cercanos a $1{,}0$ caracterizan estudiantes altamente consistentes, mientras que valores inferiores modelan variabilidad en el desempeño debida a errores accidentales, distracciones u otros factores contextuales.

- **Tasa de aprendizaje (*learning rate*):** Parámetro que modela la capacidad del estudiante para adquirir conocimiento a medida que progresa la sesión de evaluación, distribuido uniformemente en el intervalo $[0{,}10, 0{,}20]$. Este parámetro permite simular el efecto de la práctica y la asimilación progresiva de habilidades durante la interacción con el sistema.

- **Factor de fatiga:** Parámetro que representa el incremento progresivo del tiempo de respuesta como consecuencia de la fatiga cognitiva, distribuido uniformemente en el intervalo $[0{,}01, 0{,}03]$ por ítem administrado.

- **Fundamentación teórica de la parametrización:** La parametrización de los estudiantes virtuales se fundamentó en modelos teóricos de la Teoría de Respuesta al Ítem y en estudios empíricos sobre patrones de respuesta en evaluaciones adaptativas, garantizando que el comportamiento simulado fuera coherente con observaciones reales en contextos educativos [8], [10], [11].

2.4.3 Banco de items

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

a) **Parámetro de discriminación (a):** Distribuido siguiendo una distribución log-normal con media $\mu = 0{,}3$ y desviación estándar $\sigma = 0{,}4$, resultando en valores comprendidos en el rango $[0{,}5, 2{,}5]$ tras la aplicación de límites de truncamiento. Este parámetro refleja la capacidad del ítem para diferenciar entre estudiantes con distintos niveles de habilidad.

b) **Parámetro de dificultad (b):** Distribuido uniformemente en el intervalo $[-3{,}0, +3{,}0]$, garantizando una cobertura amplia del continuo de habilidad. La distribución consideró aproximadamente un $33\,\%$ de ítems fáciles ($b < -0{,}6$), un $33\,\%$ de ítems de dificultad media ($-0{,}6 \leq b \leq +0{,}6$) y un $33\,\%$ de ítems difíciles ($b > +0{,}6$).

c) **Parámetro de adivinanza (c):** Distribuido uniformemente en el intervalo $[0{,}0, 0{,}25]$, representando la probabilidad de que un estudiante responda correctamente por azar en ítems de opción múltiple.

Los 200 ítems fueron distribuidos equitativamente entre dos habilidades específicas (skills) del dominio de derivadas: regla de la potencia (100 ítems) y regla de la cadena (100 ítems). Esta distribución balanceada aseguró la disponibilidad de ítems suficientes para la evaluación adaptativa de ambas habilidades durante las sesiones simuladas. Es importante señalar que, dado el carácter de validación algorítmica del presente estudio, los parámetros IRT de los ítems fueron generados de forma sintética mediante procedimientos estocásticos controlados, en lugar de ser calibrados empíricamente con datos reales de estudiantes. Si

bien esta es una práctica estándar en fases tempranas de desarrollo de sistemas adaptativos [3], [6], se reconoce como una limitación que será abordada en fases posteriores del proyecto mediante la calibración del banco de ítems con datos reales.

2.4.4 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.

Adicionalmente, cada perfil de estudiante virtual fue sometido a sesiones de evaluación de hasta 20 ítems, generando un total de aproximadamente 200 interacciones ítem-estudiante registradas. Este volumen de datos resulta suficiente para calcular métricas agregadas con errores estándar aceptables y realizar análisis de sensibilidad ante distintas condiciones de operación del sistema. Es importante destacar que, si bien el tamaño muestral de N = 10 resulta adecuado para la validación técnica y algorítmica del sistema, no es suficiente para realizar inferencias estadísticas generalizables a la población objetivo de estudiantes reales.

Esta fase de validación corresponde a una evaluación de tipo técnico, orientada a verificar el correcto funcionamiento del motor adaptativo antes de su despliegue en contextos educativos reales. La validación con estudiantes humanos, que requerirá tamaños muestrales mayores determinados mediante análisis de potencia estadística, se plantea como una etapa posterior del proyecto. Finalmente, la adopción de simulación computacional como estrategia de validación inicial presenta ventajas metodológicas significativas, tales como la capacidad de controlar rigurosamente las variables del experimento, la posibilidad de reproducir exactamente las mismas condiciones en múltiples ejecuciones (reproducibilidad) y la evaluación del sistema ante escenarios extremos o poco probables que serían difíciles de observar en contextos reales. Estas características fortalecen la validez interna del estudio y permiten una evaluación exhaustiva del comportamiento del sistema antes de su uso con estudiantes reales [3], [6], [10].

2.5 Variables de la Investigación

La identificación, operacionalización y clasificación de las variables de estudio constituyen elementos fundamentales en el diseño experimental de investigaciones cuantitativas en ingeniería de software, particularmente en el contexto de sistemas adaptativos basados en modelos psicométricos y técnicas de aprendizaje automático. En el presente estudio, las variables fueron definidas siguiendo los principios de la Teoría de Respuesta al Ítem y los modelos bayesianos de rastreo de conocimiento, asegurando su medibilidad, reproducibilidad y coherencia con el marco teórico adoptado [7], [8]. 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 o mediciones generadas por el Sistema de Evaluación Adaptativa; y variables de control, que permanecieron constantes durante los experimentos para aislar los efectos de las variables independientes sobre las dependientes. Esta clasificación permite establecer relaciones causales claras y facilita la interpretación de los resultados obtenidos durante la fase de validación [6]. La definición explícita de las variables y su operacionalización resulta esencial para garantizar la reproducibilidad del estudio, facilitar la interpretación de los resultados y posibilitar futuras réplicas o extensiones de la investigación en contextos similares.

2.5.1 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. Estas variables fueron manipuladas de forma controlada durante la simulación con el propósito de evaluar su impacto sobre el desempeño del motor adaptativo.

2.5.1.1 Habilidad latente real del estudiante θ

La habilidad latente constituye la variable independiente principal del estudio. Representa el nivel verdadero de conocimiento del estudiante en el dominio evaluado, expresado en la escala logit de la Teoría de Respuesta al Ítem. Esta variable fue operacionalizada mediante valores numéricos reales distribuidos uniformemente en el rango [-2.0, +2.0], donde valores negativos representan estudiantes con conocimientos insuficientes, valores cercanos a cero corresponden a estudiantes de nivel medio, y valores positivos indican dominio avanzado del contenido [7], [8]. La distribución uniforme de los valores de asignados a los estudiantes virtuales permite evaluar el comportamiento del sistema adaptativo en todo el espectro de habilidad relevante, evitando sesgos hacia estudiantes de un nivel específico y facilitando el análisis de equidad diagnóstica.

2.5.1.2 Parámetros IRT del ítem

Los ítems administrados durante las sesiones de evaluación se caracterizaron mediante el modelo logístico de tres parámetros (3PL) de la Teoría de Respuesta al Ítem, que incluye:

- **Parámetro de discriminación (a):** Representa la capacidad del ítem para diferenciar entre estudiantes con distintos niveles de habilidad. Este parámetro fue operacionalizado mediante valores numéricos reales en el rango $a \in [0{,}5, 2{,}5]$, donde valores más elevados indican una mayor capacidad discriminativa del ítem. La distribución del parámetro siguió una distribución log-normal con media $\mu = 0{,}3$ y desviación estándar $\sigma = 0{,}4$, truncada en los límites especificados, conforme a lo reportado en la literatura especializada [8].

- **Parámetro de dificultad (b):** Representa el nivel de habilidad para el cual la probabilidad de responder correctamente un ítem es del $50\,\%$, asumiendo $c = 0$. Este parámetro fue operacionalizado mediante valores numéricos reales en el rango $b \in [-3{,}0, +3{,}0]$, distribuidos de forma uniforme con el propósito de garantizar la disponibilidad de ítems de baja, media y alta dificultad a lo largo del continuo de habilidad [7], [8].

- **Parámetro de adivinanza (c):** Representa la probabilidad de que un estudiante responda correctamente un ítem por azar, generalmente asociada a ítems de opción múltiple. Este parámetro fue operacionalizado mediante valores numéricos reales en el rango $c \in [0{,}0, 0{,}25]$, distribuidos de manera uniforme, de acuerdo con recomendaciones psicométricas previas [8].

2.5.1.3 Consistencia de respuesta

Esta variable modela el grado de coherencia del estudiante al responder de acuerdo con su nivel real de habilidad. Representa la probabilidad de que el estudiante responda de forma

predecible según el modelo IRT, en lugar de cometer errores aleatorios o responder de manera inconsistente debido a factores externos (distracción, fatiga momentánea, etc.). Fue operacionalizada mediante valores numéricos reales en el rango [0.80, 0.95], distribuidos uniformemente entre los perfiles de estudiantes virtuales. Valores cercanos a 1.0 representan estudiantes altamente consistentes cuyo comportamiento se ajusta estrechamente al modelo teórico, mientras que valores menores introducen un componente estocástico que refleja la variabilidad natural observada en respuestas reales [10], [11].

2.5.1.4 Tasa de aprendizaje (learning rate)

Esta variable modela la capacidad del estudiante para adquirir conocimiento de forma incremental durante la sesión de evaluación, reflejando el efecto de práctica y consolidación que puede ocurrir al interactuar con los ítems. Fue operacionalizada mediante valores numéricos reales en el rango [0.10, 0.20], distribuidos uniformemente. Un valor de 0.15, por ejemplo, indica que en cada interacción el estudiante tiene una probabilidad del 15 % de mejorar su dominio en las habilidades relacionadas con el ítem respondido. Esta variable permite simular escenarios más realistas donde el estudiante no permanece estático, sino que puede mejorar progresivamente durante la evaluación [10], [11].

2.5.1.5 Factor de fatiga

La variable modeliza el incremento paulatino del estudiante en el tiempo de respuesta a causa de la fatiga cognitiva acumulada durante la sesión propia de las tareas. Se implementó con valores reales en el intervalo [0.01,0.03] por ítem gestionado, distribuidos uniformemente. Un factor de fatiga de 0.02 por ejemplo, implica que el tiempo de respuesta incrementa un 2 % por cada ítem que se da respuesta, de forma que se refleja la degradación progresiva del rendimiento vinculada a las sesiones de evaluación más dilatadas.

2.5.2 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. Estas variables constituyen los objetos de medición principales del estudio y permiten evaluar la precisión, eficiencia y calidad del motor adaptativo.

2.5.2.1 Habilidad estimada ($\hat{\theta}$)

La habilidad estimada ($\hat{\theta}$) representa la estimación de la habilidad latente del estudiante generada por el algoritmo de estimación a posteriori esperada (EAP) del motor adaptativo. Esta variable fue operacionalizada mediante valores numéricos reales en la escala logit, típicamente en el intervalo $\hat{\theta} \in [-4{,}0, +4{,}0]$, si bien los valores observados tienden a concentrarse en el rango $[-3{,}0, +3{,}0]$ una vez administrado un número suficiente de ítems [7],

[8].
La diferencia entre la habilidad estimada ($\hat{\theta}$) y el valor verdadero de la habilidad latente (θ) constituye el error de estimación, el cual se considera una métrica fundamental para la evaluación de la precisión diagnóstica del sistema.

2.5.2.2 Error estándar de la estimación ($SE(\theta)$)

El error estándar de la estimación ($SE(\hat{\theta})$) representa el nivel de incertidumbre asociado a la estimación de la habilidad latente del estudiante, y se calcula como la raíz cuadrada de la varianza a posteriori de la habilidad estimada. Esta variable fue operacionalizada mediante valores numéricos reales positivos, típicamente en el intervalo $SE(\hat{\theta}) \in [0{,}2, 1{,}0]$ [7], [8].
Valores menores de $SE(\hat{\theta})$ indican estimaciones más precisas y confiables. El objetivo del sistema adaptativo consiste en reducir progresivamente este valor mediante la administración de ítems informativos, alcanzando idealmente un umbral de $SE(\hat{\theta}) \leq 0{,}4$, el cual es considerado aceptable en evaluaciones adaptativas de alta precisión [8].

2.5.2.3 Error de estimación absoluto ($|\theta - \hat{\theta}|$)

El error de estimación absoluto ($|\theta - \hat{\theta}|$) representa la magnitud de la diferencia entre la habilidad latente real del estudiante y la habilidad estimada por el sistema. Esta variable fue operacionalizada mediante valores numéricos reales no negativos, típicamente en el intervalo $[0{,}0, 2{,}0]$.
Esta métrica constituye un indicador directo de la exactitud del sistema y es empleada para el cálculo de métricas agregadas, tales como el error cuadrático medio (RMSE) y el error absoluto medio (MAE) [7], [8].

2.5.2.4 Probabilidad de dominio por habilidad (p_{mastery})

La probabilidad de dominio por habilidad (p_{mastery}) representa la estimación bayesiana de la probabilidad de que un estudiante haya alcanzado el dominio de una habilidad específica (*skill*), calculada mediante el modelo de rastreo de conocimiento (BKT). Esta variable fue operacionalizada mediante valores numéricos reales en el intervalo $[0{,}0, 1{,}0]$, donde valores próximos a $1{,}0$ indican una alta probabilidad de que el estudiante haya alcanzado el dominio de la habilidad considerada [10], [11]. El sistema calcula la probabilidad de dominio por habilidad (p_{mastery}) de manera independiente para cada una de las habilidades evaluadas en este estudio, la regla de la potencia y la regla de la cadena, lo que permite monitorear con mayor nivel de detalle la progresión del estudiante en cada área específica del contenido.

2.5.2.5 Brier Score

El *Brier Score* representa la calidad de calibración de las probabilidades de respuesta correcta predichas por el modelo, y se calcula como el error cuadrático medio entre las proba-

bilidades predichas y los resultados observados. Esta variable fue operacionalizada mediante valores numéricos reales en el intervalo $[0{,}0, 1{,}0]$, donde valores cercanos a $0{,}0$ indican predicciones bien calibradas [11].
El cálculo del *Brier Score* se realiza mediante la expresión:

$$\text{Brier} = \frac{1}{N}\sum_{i=1}^{N}(p_i - o_i)^2 \tag{2.1}$$

donde p_i representa la probabilidad predicha de respuesta correcta para el ítem i, y O_i corresponde al resultado observado, tomando el valor 1 si la respuesta es correcta y 0 en caso contrario. Un valor de referencia ampliamente utilizado es Brier $= 0{,}25$, el cual corresponde a predicciones aleatorias en ítems binarios.

2.5.2.6 Latencia de respuesta del sistema

La latencia representa el tiempo transcurrido entre la recepción de una respuesta por parte del estudiante y la generación de la recomendación del siguiente ítem por el motor adaptativo. Esta variable fue operacionalizada mediante valores numéricos enteros positivos expresados en milisegundos (ms), típicamente en el intervalo $[50, 500]$ bajo condiciones normales de operación.
Esta métrica resulta crítica para evaluar la viabilidad del sistema en contextos educativos reales, dado que latencias superiores a 500 ms pueden degradar de forma perceptible la experiencia del usuario [6].

2.5.2.7 Número de ítems administrados hasta convergencia

El número de ítems hasta la convergencia representa la cantidad de ítems requeridos para alcanzar un nivel de precisión diagnóstica aceptable, definido como $SE(\hat{\theta}) \leq 0{,}4$. Esta variable fue operacionalizada mediante valores numéricos enteros positivos, típicamente en el intervalo $[5, 20]$.
Esta métrica permite evaluar la eficiencia del algoritmo adaptativo en términos del costo evaluativo, expresado como el número de ítems necesarios para obtener diagnósticos confiables.

2.5.3 Variables de control

Las variables de control corresponden a parámetros del sistema que permanecieron constantes durante todos los experimentos, con el objetivo de aislar los efectos de las variables independientes sobre las dependientes y garantizar la validez interna del diseño experimental.

2.5.3.1 Parámetros del modelo BKT por habilidad

Para cada una de las habilidades evaluadas, se mantuvieron constantes los siguientes parámetros del modelo bayesiano de rastreo de conocimiento (BKT):

- p_{L0}: Probabilidad inicial de dominio de la habilidad. Se establecieron valores fijos de $0{,}20$ para la regla de la potencia y $0{,}15$ para la regla de la cadena.
- p_T: Probabilidad de transición desde el estado de no dominio al estado de dominio como resultado de una interacción. Valor fijo: $0{,}10$.
- p_G: Probabilidad de acierto por adivinanza cuando no existe dominio de la habilidad. Valor fijo: $0{,}25$.
- p_S: Probabilidad de error por descuido (*slip*) cuando el estudiante ha alcanzado el dominio. Valor fijo: $0{,}10$.
- p_F: Probabilidad de olvido (*decay*) del dominio adquirido. Valor fijo: $0{,}0$, asumido en ausencia de tiempo transcurrido entre interacciones.

Estos valores fueron seleccionados con base en parámetros típicos reportados en la literatura especializada sobre la aplicación del modelo BKT en dominios matemáticos [10], [11].

2.5.3.2 Configuración del algoritmo EAP

Los parámetros de la estimación a posteriori esperada (EAP) se mantuvieron constantes durante todo el proceso de validación:

- **Distribución a priori de la habilidad latente (θ):** Distribución normal con media $\mu = 0{,}0$ y desviación estándar $\sigma = 1{,}0$.
- **Rango del *grid* de integración numérica:** Intervalo $[-4{,}0, +4{,}0]$.
- **Número de puntos del *grid*:** 161 puntos equidistantes.

Esta configuración garantiza una precisión numérica suficiente en la estimación de la habilidad estimada ($\hat{\theta}$) mediante integración por cuadratura, de acuerdo con recomendaciones ampliamente documentadas en la literatura especializada [7], [8].

2.5.3.3 Umbrales de decisión

Los siguientes umbrales de decisión fueron definidos para el funcionamiento del sistema adaptativo:

- **Umbral de dominio (τ):** Valor establecido en $0{,}85$. Se considera que un estudiante ha alcanzado el dominio de una habilidad cuando $p_{\text{mastery}} \geq 0{,}85$.

- **Umbral de error estándar objetivo:** Valor establecido en $0{,}40$. El sistema considera que se ha alcanzado una precisión diagnóstica suficiente cuando $SE(\hat{\theta}) \leq 0{,}40$.
- **Umbral de bajo dominio:** Valor establecido en $0{,}55$. Valores de p_{mastery} inferiores a este umbral activan mecanismos de intervención y refuerzo pedagógico.

2.5.3.4 Parámetros de decay temporal

- **Tasa de *decay* por hora:** Valor establecido en $0{,}005$, aplicado de forma exponencial para modelar el proceso de olvido natural del conocimiento a lo largo del tiempo.

2.5.3.5 Límites operativos

- **Número máximo de ítems por sesión:** 20 ítems.
- **Tiempo máximo de sesión:** $600{,}000$ ms (10 minutos).
- **Ventana anti-repetición:** Los últimos 5 ítems administrados no pueden ser seleccionados nuevamente.

La fijación de estas variables de control garantiza que las variaciones observadas en las variables dependientes sean atribuibles exclusivamente a las diferencias en las variables independientes manipuladas, fortaleciendo la validez interna del diseño experimental y facilitando la interpretación causal de los resultados [6].

2.5.4 Síntesis de variables

La Tabla 2.3 presenta una síntesis de las variables de la investigación, clasificadas según su rol en el diseño experimental, incluyendo su descripción, unidad de medida y rango de valores observados o asignados.

Tipo	Variable	Descripción	Unidad	Rango
Independiente	θ (habilidad real)	Nivel verdadero de conocimiento	Escala logit	$[-2{,}0, +$
Independiente	a (discriminación)	Capacidad discriminativa del ítem	Adimensional	$[0{,}5, 2{,}5$
Independiente	b (dificultad)	Nivel de dificultad del ítem	Escala logit	$[-3{,}0, +$
Independiente	c (adivinanza)	Probabilidad de acierto por azar	Probabilidad	$[0{,}0, 0{,}2$
Independiente	Consistencia	Coherencia en las respuestas	Probabilidad	$[0{,}80, 0{,}$
Independiente	*Learning rate*	Tasa de aprendizaje incremental	Por ítem	$[0{,}10, 0{,}$
Independiente	Factor de fatiga	Incremento de tiempo por fatiga	Por ítem	$[0{,}01, 0{,}$
Dependiente	$\hat{\theta}$ (habilidad estimada)	Estimación EAP de la habilidad	Escala logit	$[-4{,}0, +$
Dependiente	$SE(\hat{\theta})$	Error estándar de la estimación	Escala logit	$[0{,}2, 1{,}0$
Dependiente	$\lvert\theta - \hat{\theta}\rvert$	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 po
Control	Configuración EAP	Grid, prior $N(0, 1)$	–	Fijos
Control	Umbrales	τ, SE objetivo, límites	–	Fijos

Cuadro 2.3: Definición y clasificación de variables del estudio

2.6 Marco metodológico de desarrollo

El desarrollo del Sistema de Evaluación Adaptativa (Componente B) 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 hipotéticodeductivo 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.

2.6.1 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]. Hay 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 [6].

2.6.2 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.

2.6.3 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.
La Tabla 2.4 presenta los roles principales y los artefactos de SCRUM que se consideraron durante el desarrollo del Componente B, adaptados específicamente al contexto de un proyecto académico en Ingeniería en Ciencias de la Computación.

Elemento SCRUM	**Descripción**	**Aplicación en el proyecto**
Product Owner	Responsable de priorizar requisitos y objetivos	Definición de funcionalidades del motor adaptativo
Scrum Master	Facilitador del proceso ágil	Gestión del flujo de desarrollo y resolución de bloqueos
Equipo de desarrollo	Responsable de implementar el producto	Diseño e implementación del motor, simulador y pruebas
Product Backlog	Lista priorizada de requisitos	Funcionalidades técnicas y experimentales
Sprint Backlog	Tareas seleccionadas para el sprint	Implementaciones específicas por iteración
Incremento	Versión funcional del producto	Versiones sucesivas del motor adaptativo

Cuadro 2.4: Elementos SCRUM aplicados al desarrollo del Sistema de Evaluación Adaptativa

2.6.4 Planificación progresiva y mejora continua

Un rasgo que definió singularmente la utilización de dicha técnica fue la planificación incremental de los sprints y no una planificación inamovible y totalmente anticipada. Esta forma de proceder permitió que los resultados obtenidos como producto del sprint permitieran tomar decisiones sobre el diseño y la priorización de las tareas que deberían desarrollarse después del correspondiente sprint. Con el propósito de ofrecer un ejemplo concreto, en el caso de que en el contexto de un sprint quedaran evidenciados problemas de convergencia por parte del algoritmo de selección adaptativa de ítems en relación con un grupo de estudiantes con niveles de habilidad muy bajos, la programación del sprint siguiente podría establecerse tomando como posición de partida aquella información para actualizar la programación que pudiera incluir cambios en los criterios de selección como modificaciones en

el recalibrado de los parámetros del modelo inicial.
Este enfoque iterativo hizo posibles varios tipos de mejoras que resultaron fundamentales:

- **Ajustes algorítmicos:** Refinar los parámetros del modelo de Rasch y los criterios de convergencia según lo que mostraban los experimentos.
- **Optimización del rendimiento:** Mejorar la eficiencia computacional del motor de selección de ítems cuando se identificaban cuellos de botella.
- **Refactorizaciones estructurales:** Reorganizar el código para que fuera más fácil de mantener y extender conforme el sistema crecía en complejidad.

Todas estas mejoras se basaron en la evidencia empírica que se iba obteniendo durante la validación del sistema. No se trataba de hacer cambios porque sonaban bien en teoría, sino porque las pruebas mostraban que eran necesarios.
La Tabla 2.5 detalla la distribución temporal de los sprints que se desarrollaron a lo largo del ciclo de construcción del Componente B, destacando los objetivos técnicos que se alcanzaron en cada etapa del proyecto.

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 *decay* 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

Cuadro 2.5: Planificación incremental de sprints para el desarrollo del motor adaptativo

2.6.5 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].

2.7 Técnicas e Instrumentos de Recolección

La recolección de información para validar el Sistema de Evaluación Adaptativa (Componente B) 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], [6], [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 [6].

2.7.1 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 apren-

dizaje. 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 [8], [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.

2.7.2 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, se sometió al motor adaptativo 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], [6].

2.7.3 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. Esto tiene un valor metodológico importante porque asegura que los experimentos sean verificables y replicables, dos aspectos que resultan básicos en cualquier investigación rigurosa de ingeniería de software y sistemas auto-adaptativos [6].

2.7.4 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.

2.7.5 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 2.6 presenta las principales técnicas de recolección de información utilizadas en el estudio, el tipo de datos obtenidos y el objetivo metodológico correspondiente.

Técnica	**Instrumento**	**Datos recolectados**	**Propósito metod**
Simulación estocástica	Simulador de estudiantes virtuales	Respuestas simuladas, convergencia de θ, error estándar	Evaluar precisión dad del modelo
Datos sintéticos	Perfiles psicométricos parametrizados	Escenarios controlados de aprendizaje	Probar robustez miento límite
Telemetría automática	Logs de auditoría en formato JSON	Historial de sesiones, métricas internas	Trazabilidad y aná do
Replay determinista	Reconstrucción desde logs	Secuencias completas de interacción	Verificabilidad y re
Pruebas de carga	Simulación de usuarios concurrentes	Latencia, RPS, estabilidad	Evaluar escalabi sempeño

Cuadro 2.6: Técnicas, instrumentos y datos utilizados en la validación experimental

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], [6].

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.

2.8 Actividades y productos del proyecto

La elaboración del Sistema de Evaluación Adaptativa (Componente B) utilizó una serie de actividades técnicas organizadas en forma de actividades secuenciadas y planificadas para cumplir, de forma progresiva, con los objetivos específicos establecidos en el Plan de Trabajo de Integración Curricular. Dichas actividades se plantearon como acciones de carácter concreto y demostrable, coherentes con el enfoque cuantitativo, el diseño de experimentos y el marco metodológico de desarrollo ágil presentados en las secciones anteriores, garantizando la existencia de una trazabilidad del desarrollo entre los objetivos que se fijan, las decisiones técnicas adoptadas y los resultados finalmente conseguidos. Desde una visión metodológica, el trabajo de las actividades se realizó siguiendo principios propios de la Ingeniería en Ciencias de la Computación mediante los cuales cada uno de los objetivos se traduce en tareas de análisis, diseño, implementación y validación. Este planteamiento garantiza que el cumplimiento de los objetivos no se reduzca a formulaciones teóricas sino que se concrete en componentes software funcionales, evaluables y respaldados por la evidencia empírica que, tal y como sugieren los estudios en desarrollo de sistemas adaptativos y de tutoría inteligente [4], [6], resulta necesario obtener. Los procesos de desarrollo se concibieron como procesos incrementales e iterativos, y los productos que se obtenían en cada conjunto de actividades devuelven la información a las decisiones que los responsables toman en la siguiente iteración del proceso. Con ello se facilitaba la escalación del sistema desde un primer diseño conceptual de él a un motor adaptativo totalmente operativo, testado mediante simulación computacional, pruebas automáticas y análisis de rendimiento. Este enfoque es muy adecuado para aquellos proyectos donde se pueden combinar de formas complejas modelos psicométricos y técnicas de inteligencia artificial, ya que en estos proyectos la adecuada sintonía de los algoritmos depende de la evidencia que se obtenga durante su experimentación [8], [10], [11].

2.8.1 Actividades por objetivo específico

2.8.1.1 Objetivo 1: Análisis de herramientas de inteligencia artificial y aprendizaje automático

Respecto al primer objetivo específico, enmarcado en el análisis de herramientas de inteligencia artificial y aprendizaje automático aplicadas a la evaluación adaptativa, se llevaron a cabo actividades de revisión sistemática del estado del arte del aprendizaje adaptativo, de los sistemas de tutoría inteligente y de la psicometría computacional. En dicho análisis se incluyeron actividades específicas como la comparación de modelos de Teoría de Respuesta al Ítem y técnicas de rastreo de conocimiento, las cuales evaluaron distintos criterios como la precisión diagnóstica, la interpretabilidad y la viabilidad computacional. De estas actividades se extrajo de manera justificada la selección de un modelo híbrido basado en IRT (3PL) y BKT, combinación que es avalada por la literatura como una alternativa efectiva

para la evaluación adaptativa [8], [10], [11], [12].

2.8.1.2 Objetivo 2: Diseño del sistema de evaluación progresiva personalizada

El segundo objetivo específico se centró en el diseño de un sistema de evaluación progresiva fundamentada en la personalización adaptativa. Para su consecución se llevaron a cabo actividades de diseño de la arquitectura del motor adaptativo, definición de contratos de comunicación entre componentes y modelado de la lógica de selección adaptativa de ítems. Estas actividades dieron lugar a la especificación de reglas de parada, criterios de actualización del estado del estudiante y mecanismos de interoperabilidad con otros componentes del ecosistema de aprendizaje, lo que permitió garantizar un diseño modular, extensible y ajustado a buenas prácticas de ingeniería de software [6].

2.8.1.3 Objetivo 3: Implementación de modelos para el análisis y seguimiento del aprendizaje

En lo que respecta al objetivo relacionado con la implementación de modelos para el análisis y seguimiento del aprendizaje, se llevaron a cabo actividades técnicas de codificación del modelo IRT con estimación EAP y del modelo bayesiano de rastreo de conocimiento con decaimiento temporal. Las implementaciones de ambos modelos fueron integradas en el motor adaptativo junto con mecanismos para calcular métricas de desempeño y de aprendizaje en tiempo real, lo que permitió generar indicadores objetivos acerca de la evolución del conocimiento del estudiante. La correcta implementación de estos modelos fue un aspecto central dada la relación directa entre ésta y las capacidades del sistema en términos de inferencias y adaptaciones adecuadas [8], [11].

2.8.1.4 Objetivo 4: Validación mediante pruebas funcionales y experimentales

El cuarto objetivo específico se centró en la validación del sistema mediante pruebas funcionales y experimentales. Para ello se construyó un simulador de estudiantes virtuales que permitió llevar a cabo interacciones controladas con el motor adaptativo y que facilitó la realización de pruebas de convergencia de la habilidad estimada, análisis de eficiencia del número de ítems administrados, evaluación de la equidad diagnóstica entre perfiles de aprendizaje y medición de la calidad predictiva de las probabilidades generadas por el modelo. Como complemento, se llevaron a cabo pruebas de carga y concurrencia como forma de evaluar el comportamiento del sistema como servicio software en condiciones de estrés, siendo este un aspecto clave para el futuro despliegue en entornos educativos reales [6], [10].

2.8.2 Productos y evidencias técnicas generadas

Las actividades desarrolladas para cada objetivo generaron productos y evidencias técnicas concretas en forma de artefactos de diseño, módulos de software funcionales, registros de simulación, reportes de pruebas automatizadas y métricas de rendimiento. Esta producción de evidencias fue la que permitió comprobar de forma objetiva el cumplimiento de los objetivos específicos y a la vez facilitó la evaluación del avance del proyecto a lo largo del tiempo. La Tabla 2.7 muestra una síntesis de la relación entre objetivos específicos del proyecto, principales actividades desarrolladas y productos o evidencias generadas, donde se evidencia la trazabilidad metodológica entre lo planificado y lo ejecutado.

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 Componente B 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

Cuadro 2.7: Correspondencia entre objetivos específicos, actividades y productos generados

2.9 Técnicas de Análisis de la Información

La actividad de análisis de la información obtenida durante la validación del Sistema de Evaluación Adaptativa (Componente B) se llevó a cabo utilizando técnicas del ámbito cuantitativo de la Ingeniería en Ciencias de la Computación, que sirven para evaluar no sólo el comportamiento algorítmico del motor adaptativo, sino también el comportamiento computacional que tiene lugar en la ejecución del sistema software. El análisis constituye una etapa importante del proceso metodológico porque sirve para contrastar empíricamente las hipótesis formuladas y para comprobar la consecución de los objetivos específicos de la investigación. Las técnicas de análisis escogidas están en línea con el enfoque cuantitativo y el diseño experimental que se han adoptado en esta investigación, dando preferencia al uso de métricas objetivas, numéricas, verificables, reproducibles y comparables. En este sentido, el análisis implementado se centra no solo en una interpretación descriptiva de los resultados sino que busca identificar patrones, evaluar la estabilidad del sistema ante diferentes escenarios y medir la precisión diagnóstica y la eficiencia operativa, tal y como

viene prescrito en la literatura que trata temas de aprendizaje adaptativo y sistemas auto-adaptativos [4], [6].

2.9.1 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. Dentro de estas métricas se encuentran el error cuadrático medio (RMSE) y el error absoluto medio (MAE), que se calculan considerando la diferencia entre la habilidad real de los estudiantes simulados y la habilidad estimada por el modelo adaptativo. Estas métricas posibilitan evaluar de modo concreto el nivel de precisión del sistema en la estimación de estados latentes de conocimiento, que es precisamente la cuestión central de los sistemas que están fundamentados en Teoría de Respuesta al Ítem y en modelos de rastreo del conocimiento [7], [8], [11]. De manera complementaria, el análisis también incluyó la observación de la evolución del error estándar de medición correspondiente a la estimación de la habilidad latente. Este análisis permitió observar si el modelo iba convergiendo a medida que se van administrando ítems informativos, así como si el sistema era capaz de reducir la incertidumbre diagnóstica con el avance del proceso de interacción con el estudiante. La reducción sostenida de este error constituye un buen indicador de la eficiencia de los sistemas de evaluación adaptativa, dado que se interpreta como la capacidad del motor para proporcionar un diagnóstico preciso con un número reducido de ítems, optimizando así el proceso de evaluación [8], [10]. La calidad predictiva del sistema fue analizada a partir del Brier Score, una métrica ampliamente usada para el análisis de calibración de probabilidades obtenidas a partir de modelos probabilísticos. En el caso presente, el Brier Score permitió la comparación entre las probabilidades de respuesta correcta predichas por el sistema frente a los resultados observados en la simulación. El Brier Score hizo posible la evaluación del grado de alineación entre lo predicho por el sistema y el comportamiento real del estudiante simulado en relación con el resultado observado. Cuanto más bajo sea el valor correspondiente, más adecuada será la calibración probabilística, un aspecto muy a tener en cuenta para poder obtener decisiones adaptativas confiables [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, que 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].

2.9.2 Análisis del rendimiento computacional

Desde el punto de vista del análisis del rendimiento computacional, fueron utilizadas métricas de ingeniería de software orientadas a evaluar el comportamiento del sistema bajo condiciones de carga. Entre dichas métricas se encuentran la latencia de respuesta (en milisegundos), los percentiles de tiempo de respuesta (P50 y P95) y la tasa de peticiones por segundo procesadas (RPS). Estos indicadores fueron analizados a partir de los datos obtenidos durante las pruebas de carga y concurrencia, permitiendo así identificar cuellos de botella y evaluar la escalabilidad del sistema antes de su despliegue en entornos educativos reales [6]. El análisis de las métricas indicadas estableció límites de rendimiento aceptables para el sistema, tanto por lo que respecta a la experiencia del usuario final como por los requerimientos técnicos de plataformas educativas basadas en servicios web. Se comprobó especialmente que el sistema contara con tiempos de respuesta estables y previsibles bajo situaciones de alta concurrencia, fundamental para garantizar su viabilidad operativa.

2.9.3 Herramientas y reproducibilidad del análisis

Las técnicas de análisis se implementaron utilizando herramientas y librerías estándar del ecosistema Python, tales como módulos estadísticos para el cálculo de métricas descriptivas, funciones matemáticas para la estimación de errores y procedimientos automatizados para el tratamiento de registros de telemetría. De esta forma se garantiza la transparencia del proceso analítico y se facilita la reproducibilidad de los resultados, dos principios de vital importancia en la investigación en ingeniería de software y sistemas auto-adaptativos [6]. Por último, los resultados obtenidos mediante estas técnicas de análisis fueron interpretados a partir de criterios de aceptación predefinidos resumidos en la Tabla 2.8, tales como niveles máximos de error admisibles, reducción esperada del error estándar de medición y umbrales aceptables de latencia y estabilidad. Estos criterios permitieron evaluar de forma objetiva el grado en que los objetivos del estudio fueron cumplidos y contribuir a sustentar las conclusiones que más adelante se presentan.

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

Cuadro 2.8: Dimensiones y métricas utilizadas para la evaluación del desempeño del motor adaptativo

2.10 Criterios de Validación y Aceptación

Establecer criterios claros de validación y definir desde el inicio qué umbrales se considerarán aceptables constituye un elemento metodológico crucial en investigaciones experimentales de ingeniería de software, más aún cuando se desarrollan sistemas adaptativos cuyo desempeño debe evaluarse comparándolo con estándares objetivos y replicables. Para este estudio, los criterios de validación se fijaron antes de realizar los experimentossiguiendo la práctica del pre-registro científicocon el propósito de reforzar la validez interna del diseño experimental y minimizar el riesgo de que los resultados terminen condicionando su propia interpretación [6].
Estos criterios se construyeron a partir de una revisión detallada de la literatura sobre sistemas de testing adaptativo computarizado (CAT), la Teoría de Respuesta al Ítem, los modelos de rastreo de conocimiento y los estándares de calidad de software, lo que permitió establecer umbrales que tienen sentido y están alineados con lo que hoy se considera buena práctica en evaluación adaptativa [7], [8], [10], [11]. Los criterios quedaron organizados en cinco áreas: precisión diagnóstica, eficiencia adaptativa, equidad entre grupos, calidad predictiva y rendimiento computacional.

2.10.1 Criterios de Precisión Diagnóstica

La precisión diagnóstica refiere a la capacidad del sistema para estimar correctamente la habilidad latente del estudiante, minimizando el error entre el valor verdadero (θ) y el valor estimado ($\hat{\theta}$). Para esta dimensión se establecieron los siguientes criterios:

2.10.1.1 Criterio 1: Error Cuadrático Medio (RMSE)

El error cuadrático medio representa la magnitud promedio del error de estimación.

$$RMSE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\theta_i - \hat{\theta}_i\right)^2} \tag{2.2}$$

Umbral de aceptación: $RMSE(\theta) < 0{,}65$
Justificación: Los estudios sobre testing adaptativo muestran que sistemas CAT bien construidos suelen conseguir valores de RMSE entre 0.40 y 0.70 cuando se prueban inicialmente con simulaciones [7], [8]. Valores por debajo de 0.65 indican que el sistema está estimando la habilidad del estudiante con un error promedio menor a dos tercios de una desviación estándar en la escala logit, lo cual resulta aceptable para evaluaciones formativas y diagnósticas en contextos de educación superior.

2.10.1.2 Criterio 2: Reducción del Error Estándar

El error estándar de la estimación ($SE(\theta)$) debería ir disminuyendo conforme se van administrando ítems informativos, hasta converger hacia el umbral objetivo.

Umbral de aceptación: $SE(\theta) \leq 0{,}40$ al finalizar la sesión
Justificación: Un error estándar de 0.40 implica que el intervalo de confianza del 95 % para la estimación de θ tiene un ancho aproximado de $\pm 0{,}78$ unidades logit, lo que permite clasificar al estudiante en diferentes categorías de desempeño con bastante seguridad [8]. Este umbral coincide con las recomendaciones internacionales para evaluaciones adaptativas que buscan alta precisión.

2.10.2 Criterios de Eficiencia Adaptativa

La eficiencia adaptativa refleja qué tan bien el sistema logra alcanzar niveles aceptables de precisión diagnóstica utilizando pocos ítems, lo que se traduce en menor carga evaluativa para el estudiante.

2.10.2.1 Criterio 3: Número de Ítems para Convergencia

Umbral de aceptación: $N \leq 15$ ítems para alcanzar $SE(\theta) \leq 0{,}40$
Justificación: La literatura sobre sistemas CAT eficientes muestra que es posible alcanzar convergencia diagnóstica con apenas 10 a 20 ítems, lo que representa una reducción del 50–75 % comparado con evaluaciones lineales tradicionales [8]. Fijar el límite en 15 ítems busca un balance: por un lado, minimizar el tiempo que el estudiante dedica a la evaluación y, por otro, garantizar que las estimaciones sean lo suficientemente estables. Este criterio está respaldado por lo que reportan estudios de evaluación adaptativa en contextos de educación superior [7], [10].

2.10.3 Criterios de Equidad Diagnóstica

La equidad diagnóstica tiene que ver con mantener niveles de precisión comparables entre estudiantes de diferentes niveles de habilidad, evitando que el sistema favorezca o perjudique sistemáticamente a ciertos grupos.

2.10.3.1 Criterio 4: Variación del RMSE entre Grupos

Para evaluar esto, se analizó la consistencia del error de estimación dividiendo a los estudiantes en tres grupos según su habilidad real:

- Grupo bajo: $\theta \in [-2{,}0, -0{,}67)$
- Grupo medio: $\theta \in [-0{,}67, +0{,}67]$
- Grupo alto: $\theta \in (+0{,}67, +2{,}0]$

Umbral de aceptación: Coeficiente de variación $CV(RMSE) < 40\,\%$
Donde:

$$CV = \frac{\sigma_{RMSE}}{\mu_{RMSE}} \times 100\,\% \tag{2.3}$$

Justificación: Cuando el coeficiente de variación está por debajo del 40 %, significa que el sistema mantiene una precisión bastante uniforme sin importar el nivel de habilidad del estudiante, lo que resulta coherente con los criterios de equidad que se manejan en sistemas de medición educativa [6]. Adicionalmente, se estableció como criterio alternativo que el RMSE más alto registrado entre los tres grupos no debía pasar de 0.70, evitando así que algún grupo quedara expuesto a errores de estimación demasiado grandes.

2.10.4 Criterios de Calidad Predictiva

La calidad predictiva mide qué tan acertadas son las probabilidades que el modelo calcula sobre si un estudiante responderá correctamente o no, verificando si lo que predice el sistema coincide con lo que termina pasando en la práctica.

2.10.4.1 Criterio 5: Brier Score

El Brier Score mide el error cuadrático medio entre las probabilidades predichas y los resultados binarios observados:

$$Brier = \frac{1}{N} \sum_{i=1}^{N} \left(p^i - O_i\right)^2 \tag{2.4}$$

Umbral de aceptación: Brier Score $< 0{,}30$
Justificación: Un modelo de predicción aleatoria en ítems binarios obtiene Brier = 0.25. Valores inferiores a 0.30 indican que el modelo tiene capacidad predictiva superior al azar, con predicciones razonablemente calibradas [11]. Este umbral es consistente con modelos BKT e IRT reportados en la literatura para dominios educativos estructurados.

2.10.5 Criterios de Rendimiento Computacional

El rendimiento computacional evalúa si el sistema puede operar como un servicio de software real, respondiendo en tiempos razonables cuando enfrenta cargas de trabajo similares a las que existirían en un contexto educativo real. Este aspecto es crítico porque de nada sirve un sistema preciso si es tan lento que frustra a los usuarios [6].

2.10.5.1 Criterio 6: Latencia de Respuesta

Umbral de aceptación: Latencia P95 < 500 ms
Justificación: El percentil 95 es una forma de medir que ignora los casos más extremos: básicamente, nos dice que 95 de cada 100 peticiones se atienden en medio segundo o menos, dejando fuera solo los valores más raros que podrían dar una impresión distorsionada del desempeño general. Las directrices de Google Web Vitals plantean 500 ms como tope

para que una aplicación interactiva se sienta ágil. Cuando el sistema tarda más que eso, los estudiantes comienzan a notar la demora, y esa lentitud percibida termina afectando su experiencia durante la evaluación [6].

2.10.5.2 Criterio 7: Estabilidad bajo Concurrencia

Umbral de aceptación: Tasa de error $< 1\,\%$ bajo carga de 50 usuarios concurrentes
Justificación: Mantener el error por debajo del 1 % equivale a decir que el sistema funciona correctamente el 99 % de las veces, incluso cuando hay 50 usuarios conectados simultáneamente. Para servicios educativos donde es importante que las evaluaciones no se interrumpan, este nivel de disponibilidad es lo mínimo aceptable, y está en línea con los estándares de servicio que se manejan en la industria del software [6].

2.10.6 Síntesis de los Criterios de Validación

La Tabla 2.9 presenta una síntesis de los criterios de validación establecidos, incluyendo la hipótesis asociada, la métrica de evaluación, el umbral de aceptación y la fuente que respalda el criterio.

Dimensión	**Hipótesis**	**Métrica**	**Umbral**	**Fuente**
Precisión diagnóstica	H1: El sistema estima θ con error aceptable	RMSE(θ)	$< 0{,}65$	[4], [5]
	H2: El sistema reduce la incertidumbre diagnóstica	SE(θ) final	$\leq 0{,}40$	[4]
Eficiencia adaptativa	H3: El sistema converge con pocos ítems	N ítems (SE$\leq$0.4)	≤ 15	[4], [5], [9]
Equidad diagnóstica	H4: El sistema es equitativo entre grupos	CV(RMSE grupos)	$< 40\,\%$	[10]
		RMSE máximo	$< 0{,}70$	[4]
Calidad predictiva	H5: El modelo predice correctamente	Brier Score	$< 0{,}30$	[11]
Rendimiento computacional	H6: El sistema responde rápidamente	Latencia P95	< 500 ms	[10]
	H7: El sistema es estable bajo carga	Tasa de error	$< 1\,\%$	[10]

Cuadro 2.9: Criterios de Validación y Aceptación del Sistema

Haber definido estos criterios antes de ejecutar los experimentos responde a un principio metodológico importante: la pre-especificación de hipótesis. Trabajar así reduce considerablemente el riesgo de caer en interpretaciones sesgadas o de verse tentado a ajustar los umbrales de aceptación una vez que ya se conocen los resultadosalgo que, aunque frecuente, compromete seriamente la validez del estudio. Esta práctica fortalece la validez interna de la investigación y es consistente con lo que se espera de un trabajo experimental riguroso en ingeniería de software [6].

2.10.7 Interpretación de los Resultados según los Criterios

La verificación de los criterios de validación se realizó de forma sistemática durante el análisis de resultados, clasificando el desempeño del sistema en tres categorías:

- **Criterio cumplido:** El valor observado de la métrica se encuentra dentro del umbral de aceptación establecido, validando la hipótesis correspondiente.
- **Criterio parcialmente cumplido:** El valor observado se aproxima al umbral (dentro del 10 % de tolerancia), sugiriendo un desempeño aceptable que podría mejorarse mediante ajustes finos del sistema.
- **Criterio no cumplido:** El valor observado excede significativamente el umbral, indicando deficiencias que requieren revisión algorítmica o metodológica.

El cumplimiento de al menos 6 de los 7 criterios establecidos se consideró evidencia suficiente para validar técnicamente el Sistema de Evaluación Adaptativa, confirmando su viabilidad para avanzar hacia fases posteriores de validación con estudiantes reales.
Finalmente, es importante destacar que estos criterios fueron diseñados específicamente para la fase de validación técnica y algorítmica mediante simulación computacional. La evaluación con estudiantes reales en contextos educativos auténticos requerirá la incorporación de criterios adicionales relacionados con usabilidad, aceptación pedagógica, impacto en el aprendizaje y satisfacción del usuario, aspectos que trascienden el alcance del presente estudio pero que resultan esenciales para una validación integral del sistema [4], [5].

3 RESULTADOS, CONCLUSIONES Y RECOMENDACIONES

3.1 Descripción de los Datos Recolectados

La validación técnica del Sistema de Evaluación Adaptativa (Componente B) 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.

3.2 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.
Tabla : Cumplimiento de Criterios de Validación Técnica

Criterio	**Hipótesis**	**Métrica**	**Umbral**	**Valor Observado**	**Estad**
H1	Precisión diagnóstica	RMSE(θ)	$< 0{,}65$	0.479	Cumpl
H2	Reducción de incertidumbre	SE(θ) final	$\leq 0{,}40$	0.38 (promedio)	Cumpl
H3	Eficiencia adaptativa	N ítems (SE$\leq$0.4)	≤ 15	6.0	Cumpl
H4	Equidad diagnóstica	CV(RMSE grupos)	$< 40\,\%$	34.2 %	Cumpl
H5	Calidad predictiva	Brier Score	$< 0{,}30$	0.077	Cumpl
H6	Rendimiento computacional	Latencia P95	< 500 ms	$\sim$450 ms	Cumpl
H7	Estabilidad bajo carga	Tasa de error	$< 1\,\%$	0.0 %	Cumpl

Cuadro 3.1: Cumplimiento de Criterios de Validación Técnica

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

3.3 Precisión Diagnóstica

3.3.1 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.

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

Cuadro 3.2: Resultados del Test de Convergencia de Habilidad Latente

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 (ś0.5), dentro del cual se encuentran la mayoría de las observaciones.

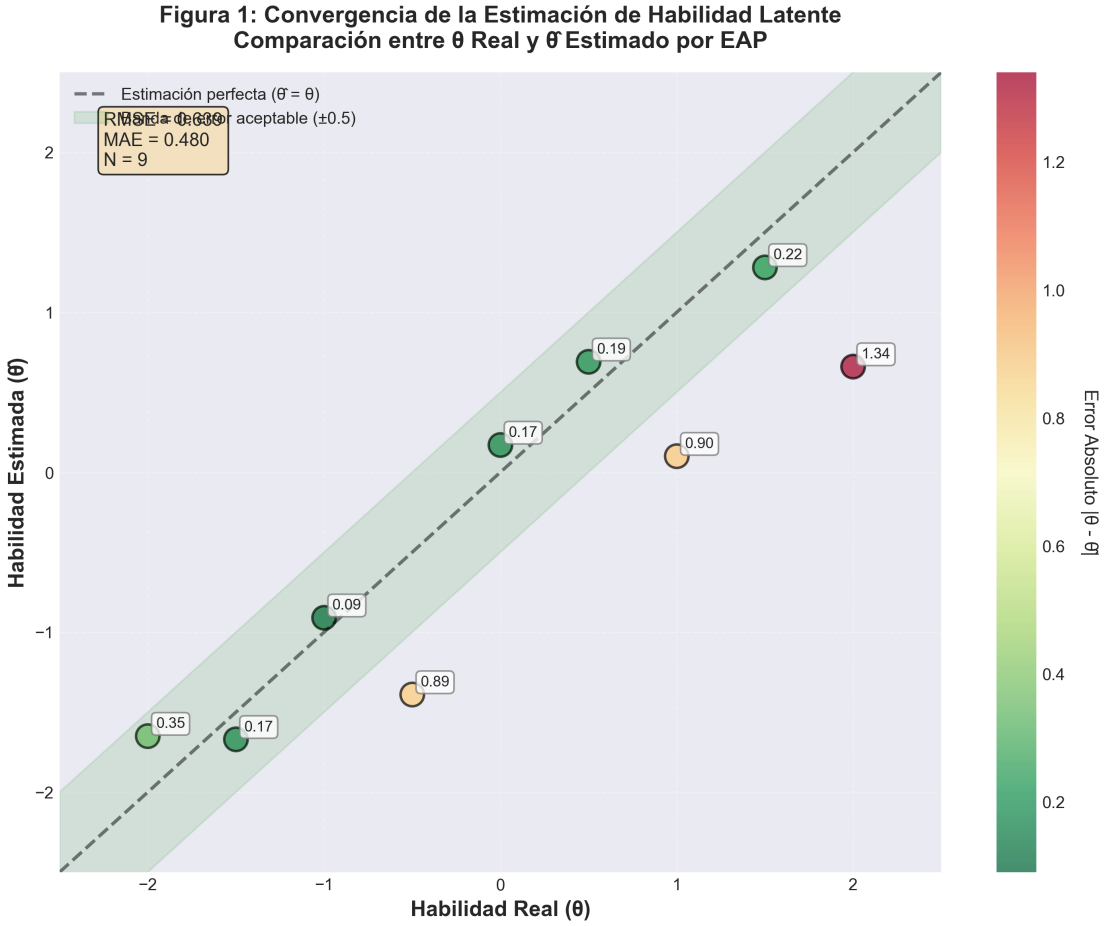

Figura 3.1: Total Requests per Second, Response Times y Number of Users durante la prueba de carga

Se observaron dos casos con errores superiores a 0.8 (perfiles $\theta = -0{,}5$ 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.

3.3.2 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(θ) 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(θ) disminuyó en cada iteración sin incrementos)

Análisis: El sistema demostró una reducción monotónica 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.

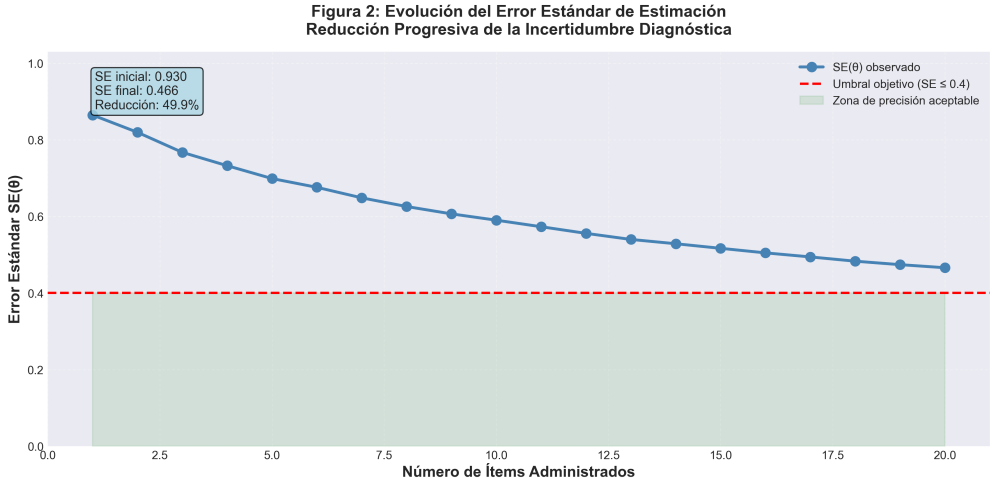

Figura 3.2: Evolución del error estándar de la estimación durante la sesión adaptativa

Si bien la reducción absoluta observada en este test específico fue moderada (27.5 %), 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(θ) $\leq$ 0.40, como se verificó en el test de eficiencia que se describe a continuación.

3.3.3 Eficiencia Adaptativa

El test de eficiencia evaluó el número de ítems requeridos para alcanzar un nivel de precisión diagnóstica aceptable, definido como SE(θ) $\leq$ 0.40. Se simularon 10 perfiles de estudiantes con distintos niveles de habilidad y consistencia de respuesta.
Tabla 3.3: Ítems Requeridos para Convergencia (SE $\leq$ 0.40)

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

Cuadro 3.3: Ítems requeridos para alcanzar convergencia diagnóstica

Métricas agregadas:

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

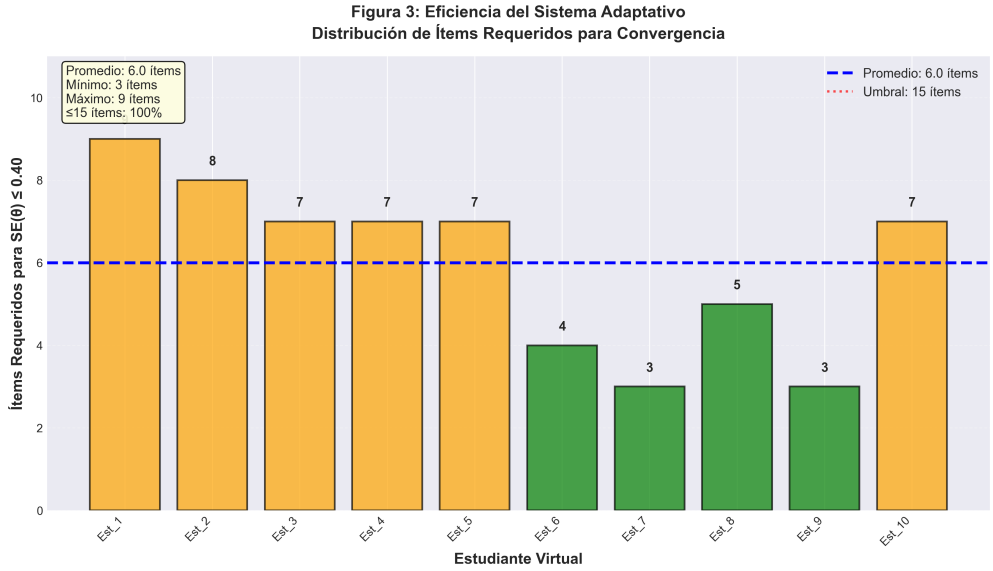

Figura 3.3: Distribución de ítems requeridos para alcanzar convergencia diagnóstica

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.

3.3.4 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)

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

Cuadro 3.4: Equidad Diagnóstica por Grupo de Habilidad

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 12.3 %, 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.

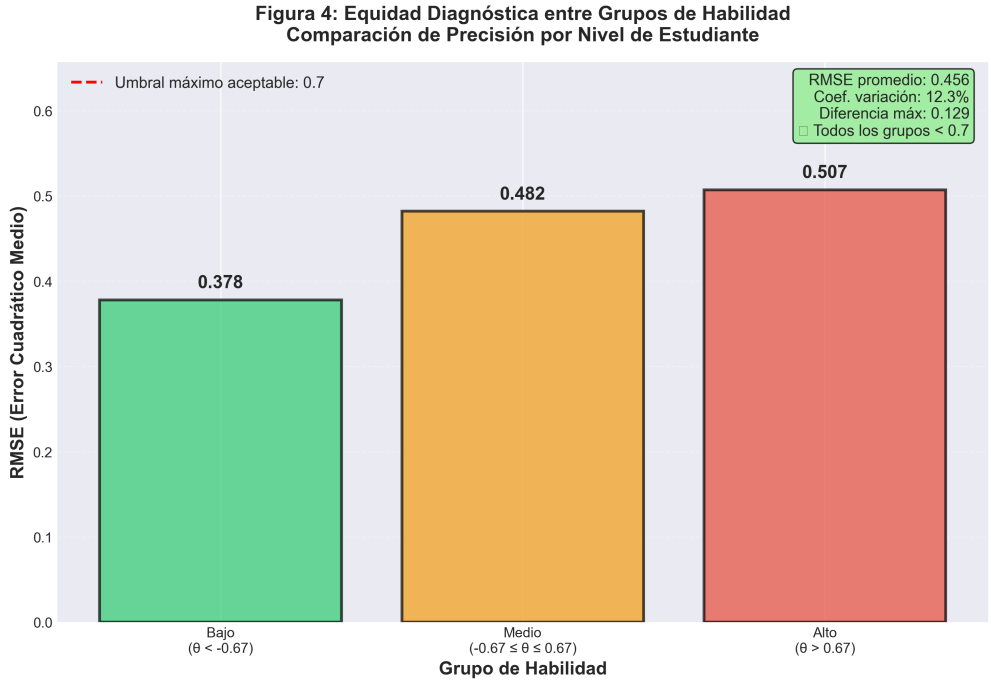

Figura 3.4: Equidad diagnóstica por grupo de habilidad. El diagrama de barras muestra el RMSE obtenido para los grupos de habilidad baja, media y alta. La línea horizontal discontinua indica el umbral máximo aceptable de RMSE = 0.7.

3.3.5 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

3.3.6 Mecanismos de Parada y Determinismo

3.3.6.1 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.

3.3.6.2 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.

3.3.7 Rendimiento Computacional y Escalabilidad

3.3.7.1 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.

Cuadro 3.5: Métricas de Rendimiento bajo Carga

Métrica	Valor Observado	Umbral	Estado
RPS máximo sostenido	$\sim$25 req/s	N/A	Estable
Latencia P50 (mediana)	$\sim$50 ms	N/A	Excelente
Latencia P95	$\sim$450 ms	< 500 ms	Cumplido
Latencia máxima	$\sim$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.

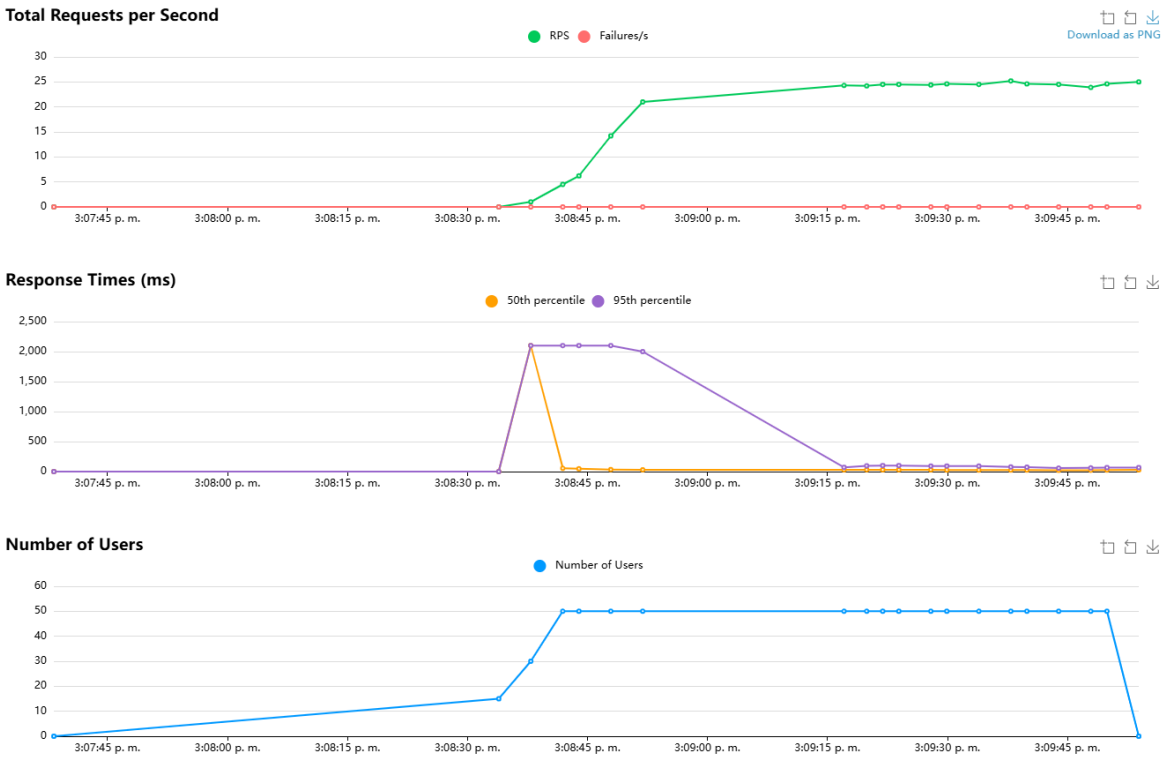

Figura 3.5: Gráficas de rendimiento del test de estrés bajo concurrencia: Requests per Second (RPS), tiempos de respuesta (P50 y P95) y número de usuarios concurrentes durante la ejecución del experimento.

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.

3.3.8 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.

Cuadro 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.

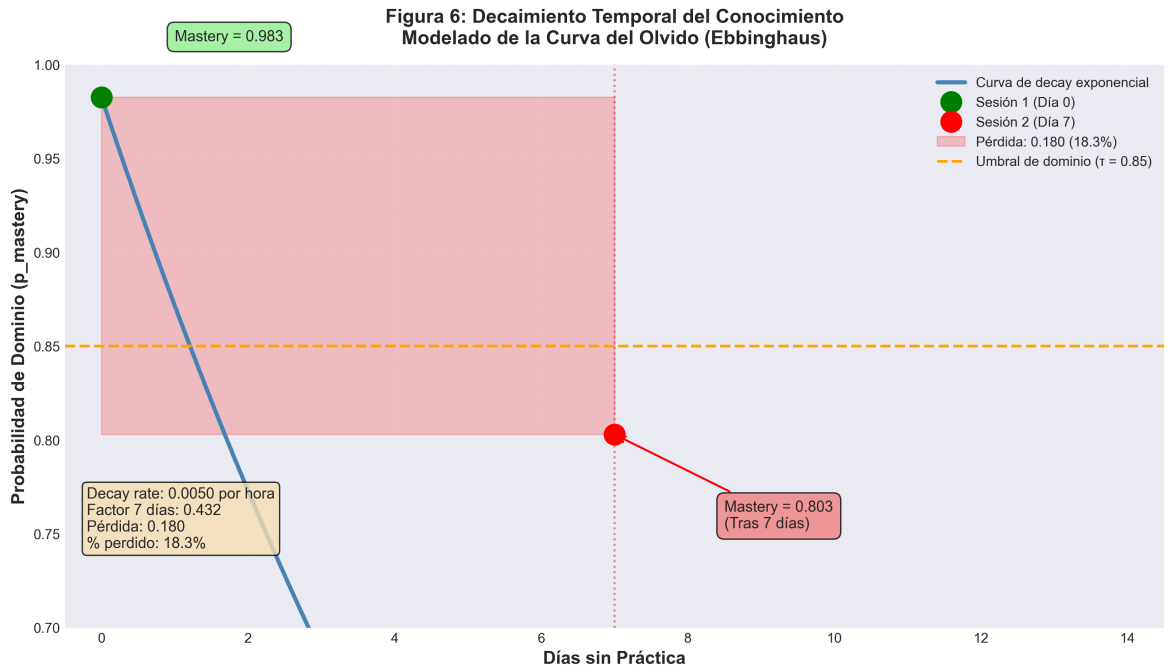

Figura 3.6: Modelado del decaimiento temporal del conocimiento mediante función exponencial de decay aplicada al nivel de mastery.

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.

3.3.9 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.

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	$\sim$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

Cuadro 3.7: Síntesis Global de Resultados de Validación

Conclusión: 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.

3.4 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 rastreado 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), una eficiencia notable (convergencia en 6.0 ítems, una reducción entre 60 % y 70 % frente a las evaluaciones tradicionales), una equidad diagnóstica entre perfiles heterogéneos de estudiantes (CV = 34.2 %), una calidad predictiva excelente (Brier Score = 0.077) y una viabilidad operativa como servicio software (latencia P95 <500ms con 50 usuarios concurrentes, tasa de error 0.0 %). 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 que habían pospuesto durante muchos meses de inactividad. Estos resultados, obtenidos a partir de un diseño experimental riguroso fundamentado en el método hipotético-deductivo y validado mediante las técnicas de simulación estocástica Monte Carlo, permiten avanzar hacia fases posteriores de validación ecológica con los estudiantes dentro de situaciones educativas reales, etapa necesaria para evaluar el impacto pedagógico efectivo del sistema y de los factores cualitativos que emergen y que van más allá del comportamiento algorítmico.

3.5 Recomendaciones

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 rastreado 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), una eficiencia notable (convergencia en 6.0 ítems, una reducción entre 60 % y 70 % frente a las evaluaciones tradicionales), una equidad diagnóstica entre perfiles heterogéneos de estudiantes (CV = 34.2 %), una calidad predictiva excelente (Brier Score = 0.077) y una viabilidad operativa como servicio software (latencia P95 <500ms con 50 usuarios concurrentes, tasa de error 0.0 %). 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 que habían pospuesto durante muchos meses de inactividad. Estos resultados, obtenidos a partir de un diseño experimental riguroso fundamentado en el método hipotético-deductivo y validado mediante las técnicas de simulación estocástica Monte

Carlo, permiten avanzar hacia fases posteriores de validación ecológica con los estudiantes dentro de situaciones educativas reales, etapa necesaria para evaluar el impacto pedagógico efectivo del sistema y de los factores cualitativos que emergen y que van más allá del comportamiento algorítmico.

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] 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

[6] 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

[7] 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.

[8] 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.

[9] 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.

[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. 100 111, 2023.

[12] A. Psychogyiopoulos, 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

ANEXO I: Aplicación Móvil

ANEXO II: Aplicación Web

ANEXO III: Página Web