Gamificación, Integridad y Validez en la Evaluación del Inglés: Un Análisis Psicométrico del Desempeño Universitario Mediante un *Serious Game* y TRI

Maricel Barrea Patiño – Universidad de Cundinamarca 0009-0002-4120-503X
Rafael Leonardo Cortes Lugo – Universidad de Cundinamarca 0000-0001-5938-9215

Recepción: 18.12.2025 | Aceptado: 06.02.2026

Correspondencia a través de **ORCID**: Maricel Barrera Patiño **0009-0002-4120-503X**

Citar: Barrea Patiño, M, & Cortes Lugo, RL (2026). Gamificación, Integridad y Validez en la Evaluación del Inglés: Un Análisis Psicométrico del Desempeño Universitario Mediante un *Serious Game* y TRI. *REIDOCREA, 15*(07), 77-83.
Financiación: proyecto financiado de la VIII convocatoria interna de proyectos de investigación de la Universidad de Cundinamarca, Colombia.
Área o categoría del conocimiento: Multidisciplinar psicología – Bilingüismo

Resumen: El presente estudio aborda el desafío de asegurar la validez de las mediciones en la enseñanza del inglés en entornos virtuales asincrónicos de educación superior. Se evaluó la eficacia de una intervención gamificada tipo *Serious Game* de cuatro semanas sobre la competencia léxica y gramatical mediante un diseño cuasi-experimental pretest-postest (n=124). Se aplicó un control de integridad y validación psicométrica con la Teoría de Respuesta al Ítem (TRI), confirmándose la estabilidad de la dificultad del instrumento. Contrario a lo esperado, se registró una disminución significativa en el puntaje global (pretest M=40.20; postest M=36.73; d=-0.42), con un índice de ganancia negativo (g=-0.06). El análisis por niveles (A1, A2, B1) mostró que la reducción se concentró en el nivel A2 (de 0.75 a 0.66), mientras que el tiempo de resolución se mantuvo constante (44 minutos). Estos hallazgos sugieren que la gamificación actuó como un filtro de integridad que corrigió una sobreestimación previa de la competencia intermedia, ofrece una métrica más realista del desempeño y evidencia al nivel A2 como la brecha crítica dentro del currículo.

Palabra clave: Gamificación

Gamification, Integrity, and Validity in English Language Assessment: A Psychometric Analysis of University Performance Using a Serious Game and IRT

Abstract: Teaching English in higher education faces the challenge of ensuring measurement validity in asynchronous virtual environments. This study analyzes the effectiveness of a four-week Serious Game gamified intervention on lexical and grammatical competence across levels A1, A2, and B1. Using a quasi-experimental design (pretest-posttest, n=124), strict integrity filtering and psychometric validation (IRT) were applied, confirming the instrument's difficulty invariance. Contrary to the expected gain, a significant decrease in the overall score was observed (pretest M=40.20 vs. posttest M=36.73; d=-0.42), resulting in a negative Hake gain index (g=-0.06). Disaggregated analysis by level revealed that this decline was driven by a sharp drop in level A2 (from 0.75 to 0.66). Since the test resolution time remained stable (44 min), the result is interpreted as a correction of measure that unmasked a *false intermediate competence*, severely questioning the reliability of unsupervised initial diagnoses. The study concludes that gamification functioned as an integrity filter, providing a real performance metric and identifying Level A2 as the critical curriculum gap.

Keyword: Gamification

Introducción

La incorporación de las tecnologías digitales en la enseñanza de lenguas extranjeras ha evolucionado progresivamente desde el uso de repositorios estáticos hacia entornos interactivos y de acceso ubicuo (Çakmak, 2019; Kohnke & Moorhouse, 2020). No obstante, en el ámbito universitario continúan siendo evidentes algunos retos de fondo, como las elevadas tasas de abandono en cursos en línea y las dificultades para demostrar avances significativos en competencias complejas que trasciendan el dominio léxico elemental. En este escenario, tanto la gamificación como el Aprendizaje

Basado en Juegos (GBL) se configuran como alternativas pedagógicas que combinan motivación, estructura cognitiva y posibilitan la práctica intencional, además de retroalimentación inmediata (Deterding et al., 2011; Sailer & Homner, 2020).

A diferencia de las plataformas comerciales de aplicación general, el diseño de *Serious Games* contextualizados facilita la alineación entre las dinámicas lúdicas y los resultados de aprendizaje prescritos en el currículo (Reinhardt, 2019). Sin embargo, los análisis críticos de la literatura advierten que gran parte de las investigaciones en este campo presentan limitaciones metodológicas ya que, habitualmente, se concentran en medir percepciones subjetivas de los participantes, sin comprobar la sensibilidad del instrumento para detectar cambios reales en el desarrollo de habilidades (Hamari et al., 2014; Dichev & Dicheva, 2017).

Por otro lado, la expansión de herramientas basadas en inteligencia artificial generativa ha introducido nuevos desafíos para la validez de las evaluaciones en entornos virtuales (UNESCO, 2023). En consecuencia, resulta imprescindible replantear los enfoques de medición del aprendizaje en escenarios educativos no supervisados. El estudio actual propone aportar evidencia empírica en esta línea, mediante el análisis del impacto de un videojuego educativo diseñado específicamente para estos fines y validar los resultados a partir de la Teoría de Respuesta al Ítem (TRI), lo que permite garantizar que las mejoras obtenidas reflejen aprendizajes genuinos y no se deban a variaciones en la dificultad de las pruebas o a conductas académicas deshonestas (Bond & Fox, 2015).

Objetivo

Evaluar la efectividad de una estrategia de intervención basada en un *Serious Game* para el desarrollo de la competencia lingüística de los niveles A1, A2 y B1 en estudiantes universitarios, empleando la Teoría de Respuesta al Ítem (TRI) junto con el análisis de patrones de respuesta, a fin de garantizar la validez y confiabilidad de los resultados obtenidos en las mediciones pretest y postest.

Método

Diseño de investigación: se adoptó un enfoque cuantitativo de tipo pre-experimental, con un esquema de medición pretest y postest en un único grupo. El propósito del estudio fue analizar la ganancia en el aprendizaje (Learning Gain) posterior a una intervención desarrollada durante cuatro semanas.

Participantes y procesamiento de datos: la muestra inicial estuvo integrada por 251 estudiantes pertenecientes al programa de Psicología. Con el fin de fortalecer la validez interna y ecológica se aplicó un proceso de depuración de datos sustentado en un análisis forense del tiempo empleado por los participantes. Se excluyeron los casos que no completaron ambas mediciones, así como aquellos con tiempos de respuesta por debajo del umbral estimado para una lectura comprensiva (<30 minutos). Tras este procedimiento, la muestra final quedó conformada por 107 participantes.

Instrumento y entorno tecnológico: la intervención se implementó mediante un entorno digital diseñado específicamente para el estudio, estructurado en niveles funcionales que conformaron un ecosistema adaptado a los objetivos de la investigación.***1) Experiencia de Usuario (Frontend):*** un videojuego 2D tipo Dungeon Crawler desarrollado en Unity. La narrativa sitúa al estudiante en un mapa de mazmorras donde el avance depende de la resolución de retos lingüísticos adaptativos, gestión (Middleware) de aplicación web en React para el seguimiento docente; ***2)***

Infraestructura (Backend): API REST en .NET Core y base de datos SQL Server en Azure, permitiendo la captura de telemetría de aprendizaje en tiempo real. Las pruebas diagnósticas (Pre y Post) constaron de 50 ítems alineados con el Marco Común Europeo de Referencia (MCER), cubriendo niveles A1, A2 y B1.

Procedimiento de Análisis: para garantizar la invarianza de la medición, la calibración de los ítems se realizó mediante el Modelo de Rasch (Embretson & Reise, 2000). Para el análisis de efectividad, se utilizó la prueba t de Student para muestras pareadas y se calculó el tamaño del efecto (d de Cohen) siguiendo los umbrales de interpretación para ciencias del comportamiento (Cohen, 1988; Plonsky & Oswald, 2014). Adicionalmente, se calculó el factor de ganancia normalizada de Hake (g) para determinar la proporción de aprendizaje alcanzado respecto al máximo posible (Hake, 1998).

Resultados

La presente sección expone los hallazgos tras la aplicación del filtro de integridad académica y el análisis estadístico de la ganancia de aprendizaje. Para garantizar la validez ecológica de los hallazgos, se aplicó un protocolo de depuración de datos excluyendo el 8.4% de la muestra inicial, correspondiente a patrones de respuesta anómalos (<30 minutos), identificados previamente como *falsos positivos* por uso de asistencia externa. El análisis se realizó sobre la muestra efectiva (n=107) que completó el ciclo de intervención gamificada (*Pretest → Gameplay en Unity → Postest).*

Calidad psicométrica del instrumento

Como paso esencial previo a la interpretación de los resultados, se validó la invarianza de la medición. El análisis mediante la Teoría de Respuesta al Ítem (TRI) confirmó la estabilidad de las propiedades métricas: los ítems mantuvieron parámetros de dificultad (b) y discriminación (a) estables entre el pretest y el postest. Las curvas características de los ítems (ICC) demostraron que la prueba discriminó adecuadamente en todo el espectro de habilidad (θ), descartando así que las variaciones en el puntaje se debieran a sesgos de dificultad entre las aplicaciones.

Desempeño global y ganancia de aprendizaje (Learning Gain)

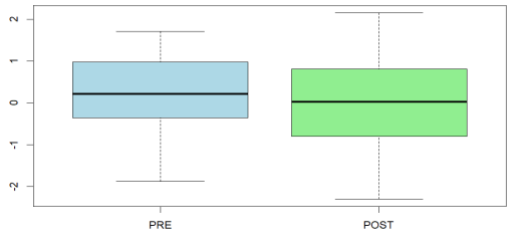
Para evaluar el impacto de la intervención gamificada en el desempeño académico, se analizaron los puntajes obtenidos por la muestra de estudiantes (n=124) que completaron satisfactoriamente ambas mediciones. Como se observa en la tabla 1, el puntaje promedio en el diagnóstico inicial sitúa al grupo en un rango de desempeño alto. Sin embargo, tras la intervención de cuatro semanas, la media en el postest descendió significativamente mostrando no solo una disminución en el rendimiento, sino un aumento en la dispersión de los datos. La prueba t de Student para muestras relacionadas confirmó que esta diferencia es estadísticamente significativa (t(123)=-4.73, p<.001), evidenciando un cambio negativo en el rendimiento global medido por el instrumento.

Tabla 1.
Estadísticos descriptivos y prueba de significancia (Pretest vs. Posttest)

Prueba pre (Calificación PRE)	Prueba post (Calificación POST)	PRE- POST- GAMIFICACION
n: 251 casos.	n: 124 casos.	n (con gamificación efectiva): 107 estudiantes.
Media: 39.23 puntos.	Media: 36.68 puntos.	Media: 36.51 puntos.
Mediana: 40.00 puntos.	Mediana: 38.64 puntos.	Mediana: 38.18 puntos.
Moda: 43.64 puntos.	Moda: 39.09 puntos.	Moda: 39.09 puntos.

En cuanto a la magnitud del cambio, el tamaño del efecto (d de Cohen) fue de d=-0.42, lo que indica un efecto moderado en sentido decreciente. De manera consistente, el índice de ganancia normalizada de Hake arrojó un valor negativo de g=-0.06. Este indicador señala que, en promedio, no hubo una ganancia neta de aprendizaje, sino una ligera regresión en el desempeño posterior a la intervención.

Figura 1.
Diagrama de cajas y bigotes (boxplot) pre vs post



Evolución diferencial por nivel de competencia (A1-B1)

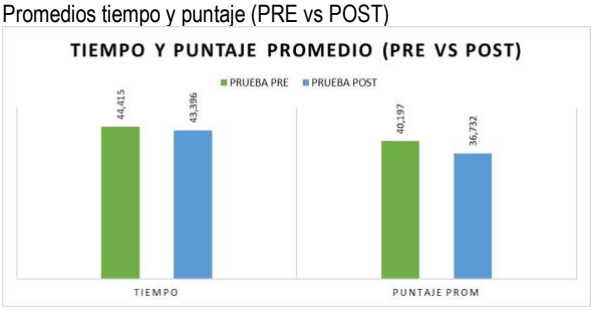
El análisis desagregado por subniveles de competencia (A1, A2 y B1) es crucial para matizar el descenso del puntaje global, dado que este puede no ser uniforme.

Para comprender las causas del descenso en el puntaje global, se desagregó el desempeño promedio por niveles de competencia del MCER, controlando la variable temporal. Es importante destacar que el tiempo promedio de dedicación a la prueba se mantuvo estable entre el Pretest (M=44.41min) y el Posttest (M=43.40 min), lo que descarta que la disminución en el rendimiento se deba a una resolución precipitada en la fase final.

La figura 2 presenta de forma conjunta la evolución del tiempo promedio de resolución y del puntaje medio obtenidos en las dos aplicaciones del instrumento: pretest y postest. En lo referente al tiempo, la diferencia entre ambas mediciones resulta mínima, lo que sugiere que los participantes invirtieron, prácticamente, el mismo periodo para completar las tareas, aun cuando el postest se llevó a cabo en un entorno controlado y con supervisión directa.

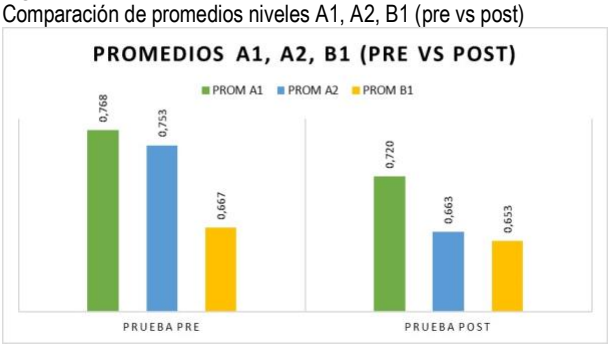
Por el contrario, el promedio de las calificaciones refleja una variación más evidente. El pretest fue administrado de manera virtual y autónoma, sin acompañamiento docente, mientras que el postest se realizó bajo condiciones de vigilancia y con limitación del uso de recursos externos; la disminución observada parece responder más a una corrección de posibles sobrevaloraciones iniciales que a un deterioro efectivo del aprendizaje alcanzado.

Figura 2.
Promedios tiempo y puntaje (PRE vs POST)



Al contrastar estos resultados con los promedios obtenidos en los niveles A1, A2 y B1, se aprecia que la disminución del puntaje es más evidente en el nivel intermedio (A2) y algo menos marcada en el nivel B1. Este comportamiento sugiere que la combinación de una evaluación posterior de mayor exigencia y un entorno de aplicación más controlado incide, especialmente, en las tareas de complejidad moderada. En conjunto, la figura 3 confirma que tanto el tipo de contexto evaluativo, sea autónomo o supervisado, como el ajuste en la dificultad del instrumento influyen, directamente, en la reducción de los promedios, mientras que el tiempo destinado a la prueba permanece constante. Esto refuerza la importancia de analizar las variaciones en el rendimiento considerando dichas condiciones metodológicas.

Figura 3.
Comparación de promedios niveles A1, A2, B1 (pre vs post)



Los resultados evidencian una respuesta desigual en función del grado de complejidad lingüística. En el nivel A1 (acceso), el desempeño se mantuvo, relativamente, estable; el promedio de aciertos disminuyó levemente de 0.77 en la evaluación inicial a 0.72 en la final. Este comportamiento sugiere que las destrezas básicas están sólidamente consolidadas y resisten cambios en las condiciones de aplicación. En el nivel B1 (umbral), se identificó un *efecto de suelo*; los participantes comenzaron con un puntaje bajo (0.67) y concluyeron en un valor semejante (0.65). Aunque la intervención no generó un aumento significativo en esta competencia superior, tampoco provocó una reducción considerable, lo que indica que dicho nivel podría representar el límite cognitivo efectivo del grupo.

El nivel A2 (plataforma) fue el principal responsable de la disminución general en los resultados. Se registró una bajada marcada desde 0.75 en el diagnóstico inicial hasta 0.66 en la prueba final. Mientras que en el pretest el rendimiento en A2 era similar al de A1, el postest reveló una distancia mayor entre ambos niveles, sugiriendo que la verdadera competencia en A2 se aproxima más al desempeño observado en B1.

La interpretación del patrón reflejado en la figura 3 debe considerar también las condiciones bajo las cuales se aplicaron las pruebas. El pretest se realizó de manera virtual y autónoma, sin supervisión directa, mientras que el postest se administró en un contexto presencial y controlado, con acompañamiento docente y control del tiempo de resolución. En la primera aplicación, la ausencia de vigilancia pudo haber propiciado el uso de apoyos externos, la colaboración entre estudiantes o una menor presión temporal, factores que probablemente aumentaron los puntajes, en especial en los niveles A1 y A2, cuyos ítems resultan más susceptibles a este tipo de ayudas.

En contraste, el postest se aplicó bajo condiciones estrictas de control (presencialidad, supervisión, acceso regulado a equipos y sin ayudas externas), por lo que los resultados reflejan con mayor fidelidad el desempeño real de los estudiantes frente a ítems de distinta complejidad. Desde esta perspectiva, la reducción de los promedios entre pre y post —más marcada en A2 y más moderada en B1— puede interpretarse menos como

una pérdida de aprendizaje y más como una corrección de posibles sobreestimaciones iniciales derivadas del contexto no supervisado de la prueba diagnóstica.

Este análisis por nivel es indispensable para mover la discusión más allá de la simple *caída de puntajes* y enfocarla en el desafío de consolidar el aprendizaje intermedio-avanzado en intervenciones cortas.

Discusión

Los hallazgos obtenidos en este estudio presentan un contraste notable respecto a la mayoría de las investigaciones sobre gamificación, que suelen señalar efectos predominantemente positivos (Zou et al., 2021; Parra-González et al., 2021). Luego de una intervención de cuatro semanas con un *Serious Game* de carácter inmersivo, se observó una disminución significativa en el rendimiento (d = -0.42). Este descenso, lejos de interpretarse como una deficiencia pedagógica, debe analizarse en el marco de la integridad académica (Sureda et al., 2009).

Uno de los argumentos más comunes frente a la reducción de calificaciones es la posible falta de compromiso durante el postest. No obstante, los registros de telemetría desmienten esa explicación, ya que el tiempo promedio de dedicación se mantuvo prácticamente igual (pretest: 44.4 min; postest: 43.4 min). Esto sugiere que la disminución no provino de un menor esfuerzo, sino de una medición más realista de las habilidades. La eliminación de variables externas, frecuentes en pruebas sin supervisión (OECD, 2021; Watson & Sottile, 2010), gracias a la inmersión del videojuego y al control del postest, permitió ajustar la sobrevaloración inicial a la verdadera capacidad cognitiva del estudiantado. Este resultado coincide con lo expuesto por Slade y Prinsloo (2013) sobre la relevancia ética en la analítica del aprendizaje.

El análisis detallado reveló el punto exacto donde se concentró la brecha de rendimiento. Mientras el nivel A1 se mantuvo estable (0.72) y el nivel B1 mostró un efecto suelo constante (0.65), el nivel A2 registró la caída más pronunciada (de 0.75 a 0.66). Esto indica que el A2 opera como una zona de confianza ilusoria: aunque los ambientes lúdicos fomentan la motivación, la apropiación de estructuras gramaticales complejas exige apoyos pedagógicos más estructurados. La intervención permitió evidenciar que las competencias intermedias eran menos sólidas de lo que reflejaban los diagnósticos previos, confirmando la efectividad de la Teoría de Respuesta al Ítem (TRI) en la identificación precisa de brechas cognitivas (De la Torre & Minchen, 2014).

Si bien la arquitectura tecnológica desarrollada (Unity/Azure) logró mantener una elevada tasa de retención, los resultados muestran que la inmersión lúdica por sí sola no basta para acelerar la adquisición de estructuras sintácticas avanzadas (nivel B1) en periodos breves. Esta aparente contradicción en la gamificación sugiere que, como plantean Acosta et al. (2024), los juegos resultan más eficaces como medio de refuerzo y motivación que como sustitutos directos de la instrucción explícita en contenidos complejos.

El aporte central de este trabajo es de tipo metodológico: la reducción significativa en los puntajes posteriores a la intervención evidencia la fragilidad de los diagnósticos iniciales realizados en entornos digitales. En tal sentido, la experiencia funcionó como un mecanismo de depuración que permitió obtener un indicador más fidedigno del desempeño, aspecto crucial para la planificación curricular.

De forma complementaria, se identificó el nivel A2 como el punto más sensible, donde se concentran las deficiencias persistentes. Se recomienda rediseñar las dinámicas del

videojuego para incorporar un andamiaje más sólido durante la transición entre A1 y A2, con retroalimentación inmediata y correctiva. Aunque el desarrollo en Unity evidenció ventajas frente a soluciones genéricas, se sugiere ampliar la duración de la intervención y combinar el *Serious Game* con sesiones sincrónicas de instrucción para atender, así, a los retos éticos que impone el aprendizaje mediado por inteligencia artificial.

Referencias

Acosta Santillán, JK, Romero Morales, JX, & Medina Gamboa, ME (2024). Impacto de la gamificación en el desarrollo del aprendizaje invisible: un enfoque lúdico para el fomento de habilidades y competencias en el aula. LATAM, 5(5), 4520–4530.

Çakmak, F (2019). Mobile learning and mobile assisted language learning in focus. Language and Technology, 2(1), 30-48.

Cohen, J (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.

Bond, TG, & Fox, CM (2015). Applying the Rasch Model: Fundamental Measurement in the Human Sciences (3rd ed.). Routledge.

De la Torre, J, & Minchen, N (2014). Cognitive diagnosis models and the assessment of learning gains. Journal of Educational Measurement, 51(3), 317-321.

Deterding, S, Dixon, D, Khaled, R, & Nacke, L (2011). From game design elements to gamefulness: Defining gamification. Proceedings of the 15th International Academic MindTrek Conference, 9-15.

Dichev, C, & Dicheva, D (2017). Gamifying education: what is known, what is believed and what remains uncertain: a critical review. International Journal of Educational Technology in Higher Education, 14(1), 9.

Embretson, SE, & Reise, SP (2000). Item Response Theory for Psychologists. Lawrence Erlbaum Associates.

Hake, RR (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. American Journal of Physics, 66(1), 64-74.

Hamari, J, Koivisto, J, & Sarsa, H (2014). Does gamification work? A literature review of empirical studies on gamification. 47th Hawaii International Conference on System Sciences, 3025-3034.

Kohnke, L, & Moorhouse, BL (2020). Facilitating synchronous online language learning through Zoom. RELC Journal, 53(1), 296-301.

OECD (2021). Digital Education Outlook 2021: Pushing the Frontiers with AI, Blockchain and Robots. OECD Publishing.

Parra-González, ME, López-Belmonte, J, Segura-Robles, A, & Moreno-Guerrero, AJ (2020). Gamification and flipped learning and their influence on aspects related to the teaching-learning process. Heliyon, 6(2), e03402.

Plonsky, L, & Oswald, FL (2014). How big is "big"? Interpreting effect sizes in L2 research. Language Learning, 64(4), 878-912.

Reinhardt, J (2019). Gameful Second and Foreign Language Teaching and Learning. Palgrave Macmillan.

Sailer, M, & Homner, L (2020). The gamification of learning: a meta-analysis. Educational Psychology Review, 32, 77-112.

Slade, S, & Prinsloo, P (2013). Learning analytics: Ethical issues and dilemmas. American Behavioral Scientist, 57(10), 1510-1529.

Sureda, J, Comas, R, & Morey, M (2009). Las causas del plagio académico entre el alumnado universitario según el profesorado. Revista Iberoamericana de Educación, 50, 197-220.

Watson, G & Sottile, J. (2010) Cheating in the Digital Age: Do Students Cheat More in Online Courses? Online Journal of Distance Learning Administration, Volume XIII, Number I.

Zou, D, Huang, Y, & Xie, H (2021). Digital game-based vocabulary learning: a systematic review and meta-analysis. Computer Assisted Language Learning, 34(8), 1-27.