

# Enterolandia y la resolución de problemas con números enteros: estudio preexperimental en octavo año de Educación General Básica.

*Enterolandia and integer problem solving: A pre-experimental study in eighth-grade basic education.*

Avecillas-Mendieta, Sonia Maritsa<sup>1</sup>; Guerrero-Zambrano, Marcos<sup>2</sup>.

<sup>1</sup> Universidad Nacional de Educación; Ecuador, Azogues; <https://orcid.org/0009-0002-0155-7655>; [smavecillas@unae.edu.ec](mailto:smavecillas@unae.edu.ec)

<sup>2</sup> Universidad Estatal de Milagro; Ecuador, Milagro; <https://orcid.org/0000-0002-5617-6836>; [mguerreroz@unemi.edu.ec](mailto:mguerreroz@unemi.edu.ec)

<sup>1</sup> Autor Correspondencia

 <https://doi.org/10.63618/omd/isj/v4/n3/391>

**Cita:** Avecillas-Mendieta, S., & Guerrero-Zambrano, M. (2026). Enterolandia y la resolución de problemas con números enteros: estudio preexperimental en octavo año de Educación General Básica. *Innova Science Journal*, 4(3), 933-947. <https://doi.org/10.63618/omd/isj/v4/n3/391>

**Recibido:** 24/02/2026

**Aceptado:** 14/07/2026

**Publicado:** 31/07/2026



**Copyright:** © 2026 por los autores. Este artículo es un artículo de acceso abierto distribuido bajo los términos y condiciones de la **Licencia Creative Commons, Atribución-NoComercial 4.0 Internacional. (CC BY-NC)**.

(<https://creativecommons.org/licenses/by-nc/4.0/>)

**Resumen:** Estimar la diferencia en el desempeño para resolver problemas con números enteros antes y después de implementar Enterolandia, una intervención gamificada mediada por Genially. Se realizó un estudio cuantitativo aplicado, con diseño preexperimental de un grupo y mediciones pretest-posttest. Participaron 18 estudiantes de octavo año, organizados en nueve parejas. La intervención comprendió dos sesiones de 45 minutos y 16 retos con retroalimentación diferenciada, pistas graduadas, reintentos y roles rotativos. El desempeño se evaluó mediante dos pruebas de 16 ítems. Se aplicó una prueba t para muestras relacionadas, un análisis de sensibilidad de Wilcoxon y una estimación del tamaño del efecto. La puntuación media aumentó de 10,11 a 11,72 puntos sobre 16. La diferencia media fue de 1,61 puntos, intervalo de confianza IC del 95 % es decir varía entre 0,59 y 2,64, prueba estadística  $t(17) = 3,315$ , nivel de significancia  $p = 0,004$ , con coeficiente de cohen  $d_z = 0,781$ . Doce estudiantes mejoraron, cuatro mantuvieron su puntuación y dos disminuyeron. Ninguna dimensión conservó significación después del ajuste de Holm. Se observó un cambio global favorable en el grupo estudiado. Sin embargo, la ausencia de un grupo comparativo, el tamaño muestral reducido y la equivalencia no demostrada de las formas impiden atribuir causalidad o generalizar los resultados.

**Palabras clave:** gamificación; enseñanza de la matemática; resolución de problemas; números enteros; educación secundaria.

**Abstract:** To estimate the difference in integer problem-solving performance before and after the implementation of Enterolandia, a gamified intervention delivered through Genially. A quantitative applied study with a single-group pre-experimental pretest-posttest design was conducted. Eighteen eighth-grade students participated in nine pairs. The intervention comprised two 45-minute sessions and 16 challenges with differentiated feedback, progressively scaffolded hints, repeated attempts, and rotating roles. Performance was assessed using two 16-item tests. A paired-samples t test, a Wilcoxon signed-rank sensitivity analysis, and a standardized within-participant effect size were calculated. The mean score increased from 10.11 to 11.72 out of 16 points. The mean difference was 1.61 points, with a 95% confidence interval ranging from 0.59 to 2.64, statistical test  $t(17) = 3.315$ , significance level of  $p = .004$ , with Cohen's coefficient  $d_z = 0.781$ . Twelve students improved, four maintained their score, and two obtained a lower score. None of the domain-level comparisons remained statistically significant after the Holm adjustment. A favorable overall change was observed in the study group. However, the lack of a comparison group, the small sample, and the unconfirmed equivalence of the two test forms preclude causal attribution and broad generalization.

**Keywords:** Gamification; Mathematics Instruction; Problem Solving; Integers; Secondary School Mathematics.

## 1. Introducción

Resolver problemas con números enteros requiere interpretar signos, reconocer puntos de referencia y representar desplazamientos que pueden atravesar el cero. Estas exigencias superan la aplicación mecánica de reglas operatorias, porque obligan a relacionar los datos con una situación y a comprobar si el resultado conserva sentido. Haji Ismail et al. (2023) documentaron que las dificultades con los enteros pueden persistir después de la instrucción y señalaron la utilidad de combinar representaciones visuales con acompañamiento pedagógico.

En octavo año de Educación General Básica, los números enteros sostienen aprendizajes posteriores de álgebra, funciones y ciencias. También permiten modelar temperaturas, altitudes, deudas, saldos, ganancias y pérdidas. El problema educativo no se limita, por tanto, a recordar una regla; se manifiesta en la distancia entre ejecutar una operación y emplearla para representar, resolver y justificar una situación contextualizada.

La gamificación incorpora elementos del diseño de juegos en actividades que mantienen objetivos curriculares explícitos. Sailer y Homner (2020) advirtieron que su aporte educativo depende de la relación entre esos elementos y los mecanismos de aprendizaje. En consecuencia, una propuesta gamificada no se define únicamente por puntos, insignias o recompensas, sino por la forma en que organiza metas, retos, retroalimentación y progresión.

Zeng et al. (2024) identificaron un efecto positivo moderado de la gamificación sobre el desempeño académico en estudios publicados entre 2008 y 2023. No obstante, la magnitud varió según el nivel educativo, el entorno y la configuración de los elementos lúdicos. Li et al. (2023) también encontraron resultados favorables, aunque subrayaron la heterogeneidad entre intervenciones y la necesidad de interpretar los efectos dentro de cada contexto.

La motivación puede favorecer la permanencia en tareas matemáticas exigentes cuando los objetivos y las ayudas se encuentran alineados con el contenido. Fadda et al. (2022) informaron un efecto positivo de los juegos digitales sobre la motivación matemática en educación escolar. Este antecedente no implica que toda actividad digital produzca aprendizaje por sí misma; indica que el interés y la expectativa de logro pueden sostener el esfuerzo.

La resolución de problemas exige análisis, toma de decisiones y transferencia. Cai et al. (2025) mostraron que el aprendizaje basado en juegos digitales puede favorecer esta capacidad cuando las actividades no se limitan a repetir respuestas. Aunque el aprendizaje basado en juegos y la gamificación no son equivalentes, ambos enfoques coinciden en la necesidad de diseñar tareas cognitivamente significativas.

El desarrollo de recursos gamificados requiere comprobar su pertinencia didáctica y su viabilidad. Lukman et al. (2023) señalaron que los materiales de matemática deben vincular los componentes lúdicos con habilidades específicas de pensamiento crítico y resolución de problemas. Esta perspectiva desplaza el interés desde la herramienta tecnológica hacia la secuencia, la retroalimentación y las oportunidades de explicación.

En el contexto latinoamericano, Rizo-Padilla et al. (2025) describieron que los juegos didácticos pueden favorecer el interés y el pensamiento lógico-matemático cuando se articulan con propósitos formativos. Chamorro-Andrade et al. (2025) destacaron el valor de las explicaciones paso a paso y los recursos visuales para comprender operaciones matemáticas. Zambrano-Alay (2025) resaltó que la retroalimentación continua puede orientar la revisión de la tarea y fortalecer la participación.

La retroalimentación es especialmente relevante cuando permite identificar el origen del error y ofrece información para corregirlo. Wisniewski et al. (2020) concluyeron que su efecto educativo varía según el tipo de información proporcionada y la forma en que el estudiante la utiliza. Por ello, el reintento adquiere valor pedagógico cuando está acompañado por pistas que orientan la revisión del procedimiento.

Las plataformas de autoría interactiva permiten integrar narrativa, navegación, imágenes y respuesta inmediata. Pastás Hernández et al. (2024) documentaron el uso de Genially en actividades gamificadas de matemática. Putra y Afrina (2023) mostraron que la legibilidad y la organización de las rutas de navegación son condiciones importantes del diseño multimedia. En educación secundaria, Uchima-Marin et al. (2025) implementaron un juego matemático con misiones y colaboración, y destacaron que el acompañamiento pedagógico debe considerarse al interpretar los resultados.

A partir de estos antecedentes, el vacío de interés no consiste en una ausencia general de investigaciones sobre gamificación, sino en la disponibilidad todavía limitada de estudios contextualizados que combinen, en intervenciones breves de educación secundaria, problemas con números enteros, retroalimentación específica por error, pistas graduadas, reintentos y roles colaborativos. Además, la valoración de una innovación debe distinguir el cambio en el desempeño de la aceptación subjetiva de la experiencia.

Enterolandia se diseñó como una intervención gamificada mediada por Genially. La propuesta integró cuatro misiones sobre comparación y orden, desplazamientos, ganancias y pérdidas y problemas integradores. Cada misión incluyó retos contextualizados, retroalimentación diferenciada, dos niveles de pistas, oportunidad de reintento, explicación del procedimiento y una pregunta de transferencia. El trabajo en parejas se organizó mediante los roles rotativos de Analista y Navegante.

La pregunta de investigación fue: ¿qué diferencia se observa en el desempeño para resolver problemas con números enteros entre el pretest y el posttest después de implementar Enterolandia en un grupo de estudiantes de octavo año de Educación General Básica? El objetivo fue estimar esa diferencia intrapersonal y describir, de manera exploratoria, los cambios en las cuatro dimensiones del desempeño. La formulación se concentró en cambios observados y no en efectos causales, porque el diseño de un solo grupo no permite descartar explicaciones alternativas.

## 2. Materiales y Métodos

### 2.1. Contexto y participantes

La investigación se realizó durante el periodo lectivo 2026 en un curso intacto de octavo año de una unidad educativa. La intervención fue presencial con apoyo digital y utilizó

la presentación interactiva Enterolandia, equipos informáticos o proyección colectiva, una recta numérica visible y hojas para operaciones breves.

Participaron 18 estudiantes, organizados en nueve parejas. Se incluyó de manera exhaustiva al alumnado elegible del aula accesible. El aula fue seleccionada mediante muestreo no probabilístico por conveniencia, debido a la disponibilidad del grupo y a la integración de la experiencia en la planificación escolar.

Bhattacharjee (2012) indicó que el muestreo por conveniencia puede utilizarse en estudios aplicados cuando el acceso está delimitado por el contexto y sus restricciones se explicitan. López-Roldán y Fachelli (2015) advirtieron que este procedimiento no permite extender automáticamente los resultados a poblaciones más amplias. Por ello, las inferencias se circunscribieron al grupo estudiado.

Los criterios de inclusión fueron estar matriculado en el curso, disponer del consentimiento del representante, otorgar asentimiento informado y completar las mediciones inicial y final. La ausencia en uno de los dos momentos de evaluación excluyó al participante del contraste pareado, pero no restringió su acceso a la actividad pedagógica.

## 2.2. Diseño y alcance

Se realizó un estudio cuantitativo aplicado, de campo, con diseño preexperimental de un grupo y dos mediciones repetidas. La finalidad fue evaluar una propuesta construida para atender una necesidad didáctica concreta y estimar la diferencia entre dos mediciones del mismo grupo.

La estructura del diseño fue  $O_1$ -X- $O_2$ , donde  $O_1$  representa la medición inicial, X la intervención Enterolandia y  $O_2$  la medición posterior. Jhangiani et al. (2019) señalaron que los diseños de un solo grupo permiten describir cambios intraparticipante, pero mantienen amenazas de historia, maduración, efecto de prueba, instrumentación y regresión estadística. En consecuencia, el estudio no se clasificó como explicativo y no estableció causalidad.

## 2.3. Intervención Enterolandia

Enterolandia se implementó como una intervención didáctica gamificada mediada por Genially. La plataforma se utilizó como soporte de navegación e interactividad; el componente pedagógico se estructuró mediante problemas contextualizados, progresión de dificultad, retroalimentación específica, pistas, reintentos, transferencia y trabajo colaborativo.

La intervención comprendió dos sesiones de 45 minutos. La primera misión abordó comparación y orden; la segunda, desplazamientos; la tercera, ganancias, pérdidas, deudas y saldos; y la cuarta, problemas de varios pasos. Cada misión incluyó cuatro retos.

Las alternativas incorrectas activaron mensajes asociados con errores previsibles, como inversión del sentido en la recta numérica, omisión del signo inicial, suma de valores absolutos o interpretación incorrecta de una penalización. Después del primer error se ofreció una pista conceptual. Si la dificultad persistía, se presentaba una segunda orientación más explícita o visual.

Los estudiantes trabajaron en parejas con los roles de Analista y Navegante. El Analista leía el problema, identificaba los datos y proponía una estrategia. El Navegante controlaba la presentación, verificaba el acuerdo y registraba el uso de pistas. Los roles se intercambiaron después de dos retos. Las insignias reconocieron comprensión, persistencia y colaboración, sin repercusión en la calificación ni premios por rapidez.

#### **2.4. Fases de aplicación**

La aplicación se desarrolló en ocho fases. Primero, se asignaron códigos y se administró individualmente el pretest Forma A. Después, se activaron conocimientos previos sobre la recta numérica, el cero como referencia y las situaciones de ganancia o pérdida. En la tercera fase se presentó la narrativa, las reglas, el sistema de pistas y los roles. En la cuarta, las parejas resolvieron los 16 retos. En la quinta fase se aplicó la retroalimentación diferenciada y se habilitaron los reintentos. La sexta fase consistió en una reflexión metacognitiva sobre estrategias, errores y distribución de decisiones. En la séptima se administró individualmente el postest Forma B. Finalmente, se recogieron medidas de percepción, motivación y observación de la implementación.

#### **2.5. Instrumentos**

El desempeño se evaluó mediante un pretest Forma A y un postest Forma B, integrados por 16 ítems de elección múltiple con tres alternativas. Cada prueba distribuyó cuatro ítems en las dimensiones de comparación y orden, desplazamientos, ganancias y pérdidas y problemas integradores. Las respuestas correctas recibieron un punto; las incorrectas o no contestadas, cero. La puntuación total osciló entre 0 y 16 y cada dimensión entre 0 y 4. Las pruebas se aplicaron individualmente, sin calculadora y durante 20 minutos.

Las formas mantuvieron la misma distribución dimensional, el mismo número de ítems y el mismo formato de respuesta. Esta correspondencia estructural no se interpretó como equivalencia psicométrica, porque no se dispuso de un estudio independiente de equiparación de dificultad.

También se administraron dos cuestionarios de 20 ítems y una ficha de observación por parejas. Sin embargo, la base conservada contenía únicamente puntuaciones agregadas que no coincidían con la estructura final de los cuestionarios, y el rango de la observación no correspondía con la versión definitiva. Por razones de integridad analítica, esas medidas no se utilizaron para sustentar los resultados ni las conclusiones del presente manuscrito.

Taber (2018) explicó que la consistencia interna debe interpretarse a partir de la matriz de respuestas y de la estructura del instrumento, no como un umbral aislado. Viladrich et al. (2017) señalaron que el alfa y el omega requieren supuestos y datos que permitan relacionar la puntuación con el constructo. Al no disponerse de respuestas por ítem, no se estimó la confiabilidad de los cuestionarios.

#### **2.6. Validez de contenido y consistencia de las pruebas**

Los instrumentos se elaboraron a partir de una matriz de correspondencia entre objetivo, variables, dimensiones, indicadores e ítems. La valoración de contenido se organizó mediante matrices dirigidas a cinco jueces y consideró claridad, pertinencia, coherencia, relevancia y adecuación.

Pedrosa et al. (2013) indicaron que la evidencia de validez de contenido debe examinar la representatividad y suficiencia de los elementos respecto del dominio. En este estudio, las valoraciones disponibles fueron globales por criterio y no por ítem; por tanto, se interpretaron como evidencia preliminar de valoración general y no como demostración de la validez de cada reactivo.

El acuerdo se resumió mediante la V de Aiken. Merino Soto y Livia Segovia (2009) recomendaron interpretar este coeficiente junto con intervalos de confianza, especialmente cuando participa un número reducido de jueces.

La consistencia interna de las pruebas dicotómicas se estimó mediante KR-20. Anselmi et al. (2019) advirtieron que los índices de consistencia responden a supuestos diferentes y deben interpretarse junto con la estructura de la prueba. En consecuencia, los coeficientes se consideraron preliminares por el tamaño muestral y por la presencia de cuatro dominios.

## 2.7. Procedimiento

La investigación comenzó con la autorización institucional y la información a representantes y estudiantes. Después del consentimiento y el asentimiento, se asignaron códigos alfanuméricos que sustituyeron los nombres. La lista de correspondencia entre códigos e identidades se mantuvo separada de la base de análisis.

El pretest se administró antes de la intervención. Posteriormente se organizaron las parejas, se eligieron los avatares y se distribuyeron los roles iniciales. Durante las dos sesiones se completaron las cuatro misiones, se revisó la retroalimentación, se utilizaron pistas cuando fue necesario y se rotaron los roles.

El posttest se administró en un momento independiente y sin mostrar las respuestas de la actividad. Antes del análisis se revisaron códigos duplicados, valores fuera de rango, datos faltantes e inconsistencias entre respuestas y puntuaciones totales. El análisis pareado incluyó a quienes completaron ambas mediciones.

## 2.8. Procesamiento y análisis de datos

Los datos se procesaron mediante Python 3 en Google Colab. Google describe Colab como un servicio alojado de Jupyter que permite ejecutar código desde el navegador. La documentación de Python Software Foundation se utilizó para registrar la sintaxis del entorno.

El cuaderno integró importación, depuración, transformación, análisis y generación de gráficos. Downey (2014) destacó que el análisis con Python permite vincular exploración, visualización y comprobación computacional en un flujo reproducible.

El análisis descriptivo incluyó frecuencias, porcentajes, media, desviación estándar, mediana, mínimo, máximo y rango intercuartílico. La selección de medidas y pruebas consideró la distribución y el carácter pareado de los datos, conforme a los principios descritos por Çetinkaya-Rundel et al. (2019).

El cambio individual se definió como  $D_i = \text{[postest]}_i - \text{[pretest]}_i$ . La distribución de las diferencias se examinó mediante histograma, gráfico cuantil-cuantil, detección de

valores atípicos y prueba de Shapiro-Wilk. Cuando la distribución fue compatible con normalidad, se aplicó una prueba t para muestras relacionadas. La prueba de rangos con signo de Wilcoxon se utilizó como análisis de sensibilidad.

Se estableció un nivel bilateral de significación de 0,05. Se informó la diferencia media, su intervalo de confianza del 95 % y el tamaño del efecto  $d_z = \sqrt{D/s_D}$ . Lakens (2013) recomendó reportar la magnitud del efecto junto con la significación para facilitar la interpretación y la acumulación de evidencia.

Las cuatro dimensiones se analizaron como resultados secundarios y exploratorios. Los valores p se ajustaron mediante el procedimiento de Holm. La interpretación principal se concentró en la puntuación total.

## 2.9. Consideraciones éticas

La participación fue voluntaria e independiente de la asistencia, las calificaciones y el trato institucional. Los representantes recibieron información sobre el propósito, los procedimientos, los riesgos mínimos, los beneficios posibles, la confidencialidad y el derecho a retirar la autorización. El estudiantado recibió un documento de asentimiento en lenguaje accesible.

Los registros utilizaron códigos en lugar de nombres y los resultados se presentaron de forma agregada. No se utilizaron fotografías, audios o videos con fines investigativos sin consentimiento adicional. Los datos se mantuvieron bajo acceso restringido.

## 2.10. Control de sesgos y limitaciones del diseño

La aplicación individual de las pruebas redujo la influencia directa de la pareja sobre las puntuaciones. El uso de formas diferentes buscó limitar el recuerdo literal, aunque introdujo la necesidad de comprobar su equivalencia. La rotación de roles procuró disminuir la dominancia de un integrante.

El diseño careció de grupo de comparación; la selección fue por conveniencia; el tamaño muestral fue reducido y el posttest fue inmediato. Además, la posible desigualdad de dificultad entre las formas limita la atribución del cambio a la intervención. Estas condiciones sitúan los resultados como evidencia contextual y exploratoria.

## 3. Resultados

### 3.1. Control de calidad y valoración global de contenido

La base incluyó 18 registros completos. No se identificaron códigos duplicados ni datos ausentes en las puntuaciones globales. La diferencia registrada coincidió en todos los casos con la resta entre posttest y pretest.

Las matrices de juicio experto incluyeron cinco valoraciones por criterio en una escala de 1 a 4. La V de Aiken promedio osciló entre 0,856 para la propuesta Enterolandia y 0,907 para el cuestionario de percepción. Los valores más bajos correspondieron a factibilidad temporal y duración de la implementación, ambos con  $V = 0,733$ . Debido a que las valoraciones fueron globales por criterio, no permiten concluir que cada ítem individual haya sido validado.

Tabla 1.

**Valoración global por criterios de los instrumentos y de la propuesta**

| Componente                 | Criterios | V promedio | Rango de V  | Observación principal                 |
|----------------------------|-----------|------------|-------------|---------------------------------------|
| Pretest                    | 5         | 0,880      | 0,800–0,933 | Revisar distractores                  |
| Posttest                   | 5         | 0,880      | 0,800–0,933 | Equilibrar dificultad                 |
| Cuestionario de percepción | 5         | 0,907      | 0,800–1,000 | Simplificar una pregunta abierta      |
| Cuestionario de motivación | 5         | 0,880      | 0,800–0,933 | Evitar solapamiento                   |
| Ficha de observación       | 5         | 0,867      | 0,800–0,933 | Incluir ejemplos y segundo observador |
| Propuesta Enterolandia     | 6         | 0,856      | 0,733–0,933 | Ajustar tiempo y retroalimentación    |
| Validación pedagógica      | 10        | 0,887      | 0,733–1,000 | Ajustar duración e intentos           |

**Nota.** Las valoraciones fueron globales por criterio y no por ítem; por tanto, constituyen evidencia preliminar.

**3.2. Desempeño global antes y después de la intervención**

En el pretest, la puntuación media fue de 10,11 puntos sobre 16 (DE = 2,93; IC del 95 % [8,65, 11,57]), con mediana de 10,00 y rango de 4 a 15. En el posttest, la media fue de 11,72 puntos (DE = 2,49; IC del 95 % [10,48, 12,96]), con mediana de 11,50 y rango de 7 a 15. La diferencia media fue de 1,61 puntos (DE = 2,06; IC del 95 % [0,59, 2,64]). Doce estudiantes mejoraron, cuatro mantuvieron su puntuación y dos disminuyeron.

La distribución de las diferencias fue compatible con normalidad,  $W = 0,966$ ,  $p = 0,717$ . La prueba  $t$  para muestras relacionadas indicó una diferencia estadísticamente significativa,  $t(17) = 3,315$ ,  $p = 0,004$ . El análisis de sensibilidad mediante Wilcoxon produjo un resultado concordante,  $W = 9,5$ ,  $p = 0,007$ . El tamaño del efecto fue  $d_z = 0,781$ , con IC bootstrap del 95 % [0,388, 1,354], compatible con un cambio de magnitud moderada a grande, aunque estimado con escasa precisión.

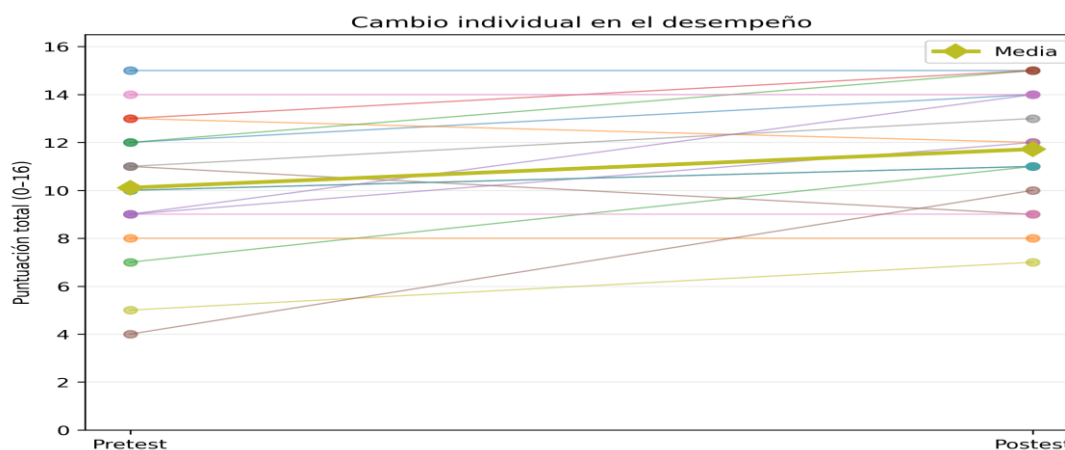
Tabla 2.

**Estadísticos descriptivos de las puntuaciones globales**

| Medición   | n  | Media | DE   | Mediana | Q1–Q3       | Rango | IC 95 % de la media |
|------------|----|-------|------|---------|-------------|-------|---------------------|
| Pretest    | 18 | 10,11 | 2,93 | 10,00   | 9,00–12,00  | 4–15  | 8,65–11,57          |
| Posttest   | 18 | 11,72 | 2,49 | 11,50   | 10,25–14,00 | 7–15  | 10,48–12,96         |
| Diferencia | 18 | 1,61  | 2,06 | 1,50    | 0,00–2,75   | –2–6  | 0,59–2,64           |

**Nota.** El intervalo de la fila «Diferencia» corresponde a la diferencia media pareada.

Figura 1.

**Cambio individual entre el pretest y el posttest**

**Nota.** La línea de mayor grosor representa la media del grupo.

**3.3. Resultados por dimensión**

El mejor desempeño inicial correspondió a comparación y orden, con una media de 3,11 sobre 4, mientras que los problemas integradores presentaron el promedio más bajo, 1,94. En el posttest, comparación y orden alcanzó una media de 3,61 y los problemas integradores permanecieron como la dimensión de menor puntuación, con 2,17. La mayor diferencia descriptiva se observó en desplazamientos, con 0,78 puntos.

Las comparaciones dimensionales se consideraron secundarias por el reducido rango de las subescalas. Desplazamientos presentó un valor *p* sin ajustar de 0,020; sin embargo, ninguna dimensión conservó significación después de aplicar la corrección de Holm. Por tanto, las diferencias por dominio se interpretaron de manera descriptiva y exploratoria.

Tabla 3.

**Comparación pretest-posttest por dimensión**

| Dimensión              | Pre M (DE)  | Pre Md [Q1–Q3]   | Post M (DE) | Post Md [Q1–Q3]  | Dif. media | W    | p exacta | p Holm | r biserial |
|------------------------|-------------|------------------|-------------|------------------|------------|------|----------|--------|------------|
| Comparación y orden    | 3,11 (1,08) | 3,00 [3,00–4,00] | 3,61 (0,50) | 4,00 [3,00–4,00] | 0,50       | 6,0  | 0,133    | 0,398  | 0,667      |
| Desplazamientos        | 2,11 (1,28) | 2,00 [1,00–3,00] | 2,89 (0,83) | 3,00 [2,25–3,00] | 0,78       | 20,0 | 0,020    | 0,081  | 0,667      |
| Ganancias y pérdidas   | 2,94 (1,06) | 3,00 [2,25–4,00] | 3,06 (1,00) | 3,00 [3,00–4,00] | 0,11       | 18,0 | 0,781    | 0,810  | 0,200      |
| Problemas integradores | 1,94 (1,16) | 2,00 [1,00–2,75] | 2,17 (1,15) | 2,00 [2,00–3,00] | 0,22       | 38,0 | 0,405    | 0,810  | 0,276      |

**Nota.** M = media; DE = desviación estándar; Md = mediana; Q1–Q3 = rango intercuartílico; W = estadístico de Wilcoxon. Los valores *p* de Holm corrigen cuatro comparaciones. El coeficiente biserial por rangos se presenta como medida exploratoria de magnitud.

3.4. Consistencia interna y comportamiento de los ítems

La consistencia interna del pretest fue KR-20 = 0,666, con IC bootstrap del 95 % [0,206, 0,813]. En el postest se obtuvo KR-20 = 0,596, con IC bootstrap del 95 % [0,237, 0,746]. La amplitud de los intervalos refleja una estimación imprecisa, esperable con 18 participantes y una prueba que integra cuatro dominios.

La proporción de aciertos del pretest varió entre 0,222 y 0,889. En el postest osciló entre 0,444 y 0,944. Las correlaciones punto-biserales corregidas fueron negativas en los ítems 6 y 15 del pretest y en los ítems 2, 5 y 7 del postest. Otros ítems presentaron valores inferiores a 0,20. Estos resultados indican la necesidad de revisar claves, distractores y equivalencia entre formas antes de una aplicación confirmatoria.

Tabla 4.  
*Dificultad y discriminación corregida de los ítems*

| Ítem | Pre p | Pre r_pb | Post p | Post r_pb | Revisión por r_pb < 0,20 |
|------|-------|----------|--------|-----------|--------------------------|
| 1    | 0,778 | 0,263    | 0,944  | 0,287     | —                        |
| 2    | 0,889 | 0,354    | 0,944  | −0,022    | Post                     |
| 3    | 0,778 | 0,587    | 0,944  | 0,079     | Post                     |
| 4    | 0,667 | 0,294    | 0,778  | 0,160     | Post                     |
| 5    | 0,722 | 0,578    | 0,778  | −0,012    | Post                     |
| 6    | 0,611 | −0,020   | 0,889  | 0,280     | Pre                      |
| 7    | 0,444 | 0,324    | 0,667  | −0,129    | Post                     |
| 8    | 0,333 | 0,143    | 0,556  | 0,487     | Pre                      |
| 9    | 0,833 | 0,498    | 0,889  | 0,280     | —                        |
| 10   | 0,722 | 0,525    | 0,778  | 0,160     | Post                     |
| 11   | 0,778 | 0,212    | 0,667  | 0,393     | —                        |
| 12   | 0,611 | 0,146    | 0,722  | 0,333     | Pre                      |
| 13   | 0,611 | 0,062    | 0,611  | 0,464     | Pre                      |
| 14   | 0,500 | 0,142    | 0,556  | 0,016     | Pre, post                |
| 15   | 0,611 | −0,020   | 0,556  | 0,064     | Pre, post                |
| 16   | 0,222 | 0,218    | 0,444  | 0,513     | —                        |

**Nota.** p = proporción de aciertos; r\_pb = correlación punto-biserial corregida. Los valores negativos o inferiores a 0,20 requieren revisión del ítem.

3.5. Medidas secundarias no incluidas en la interpretación

Los cuestionarios de percepción y motivación y la ficha de observación no se incorporaron a los resultados finales. La base disponible conservó puntuaciones agregadas que no coincidían con las dimensiones de los instrumentos descritos, y la escala observacional registrada no correspondía con el rango de la versión definitiva. Esta decisión evitó sustentar conclusiones en puntuaciones cuya trazabilidad no podía garantizarse.

4. Discusión

4.1. Interpretación del hallazgo principal

El estudio identificó un incremento medio de 1,61 puntos entre el pretest y el postest. La mayoría de los participantes mejoró y el tamaño del efecto intrapersonal fue  $d_z = 0,781$ . Este patrón es compatible con la hipótesis pedagógica de la intervención, pero no permite estimar un efecto causal, debido a la ausencia de grupo comparativo y a la equivalencia no demostrada entre las formas.

Zeng et al. (2024) observaron resultados favorables de la gamificación sobre el desempeño académico, aunque con variación entre contextos. Li et al. (2023) también

informaron heterogeneidad en los efectos y subrayaron que la configuración de las actividades condiciona los resultados. El cambio observado en este estudio debe situarse dentro de esa variabilidad y no interpretarse como evidencia definitiva.

#### **4.2. Desempeño matemático y dimensiones**

El diagnóstico inicial mostró mayor desempeño en comparación y orden y mayores dificultades en problemas integradores y desplazamientos. Esta configuración sugiere que reconocer relaciones básicas entre enteros puede resultar menos exigente que coordinar signo, referencia y secuencia de operaciones en una situación contextualizada.

Haji Ismail et al. (2023) señalaron que las representaciones visuales pueden apoyar la comprensión de los enteros cuando se integran con orientación docente. En Enterolandia, la recta numérica, los recorridos y las pistas visuales se diseñaron para apoyar ese proceso.

La mayor diferencia descriptiva se presentó en desplazamientos, pero ninguna dimensión mantuvo significación después del ajuste de Holm. Por ello, no puede afirmarse que la intervención haya producido un cambio específico en ese dominio. Las subescalas de cuatro ítems y el tamaño muestral pudieron limitar la sensibilidad de los contrastes.

#### **4.3. Mecanismos pedagógicos plausibles**

El interés de Enterolandia no reside únicamente en la narrativa o las insignias. La propuesta articuló problemas contextualizados, progresión de dificultad, mensajes vinculados con errores previsibles, dos niveles de pistas, reintentos, transferencia y roles rotativos.

Sailer y Homner (2020) sostuvieron que los efectos de la gamificación dependen de la relación entre elementos lúdicos y mecanismos de aprendizaje. En este estudio, la narrativa organizó la progresión, mientras que la retroalimentación y las pistas orientaron la revisión de procedimientos.

Lukman et al. (2023) destacaron que los materiales gamificados de matemática deben vincularse con habilidades específicas de resolución. Esa condición se buscó mediante retos que exigían interpretar la situación antes de seleccionar una operación.

Wisniewski et al. (2020) mostraron que la retroalimentación es más útil cuando aporta información que puede emplearse para ajustar la tarea. En Enterolandia, el reintento acompañado por una pista pudo funcionar como mecanismo de regulación. Esta interpretación es plausible, pero no fue probada mediante un análisis de mediación ni mediante registros completos de uso de pistas.

Genially aportó navegación, integración visual y respuesta inmediata. Pastás Hernández et al. (2024) describieron posibilidades semejantes para organizar actividades matemáticas. Sin embargo, el soporte tecnológico no debe confundirse con el tratamiento pedagógico, porque la secuencia y la mediación continúan siendo determinantes.

#### 4.4. Calidad de la medición

Las valoraciones globales de contenido fueron favorables, pero no se realizaron por ítem. Pedrosa et al. (2013) advirtieron que la validez de contenido requiere examinar la representatividad de cada elemento respecto del dominio. Por consiguiente, los valores de V de Aiken deben considerarse preliminares.

Los coeficientes KR-20 presentaron intervalos amplios y varios ítems mostraron discriminación negativa o baja. Anselmi et al. (2019) señalaron que la consistencia interna no debe interpretarse sin considerar la estructura de la prueba. La revisión de distractores y claves es prioritaria antes de atribuir el cambio a diferencias de desempeño.

La comparabilidad entre las formas A y B constituye una limitación central. Mantener la misma estructura no garantiza dificultad equivalente. Estudios posteriores deberían incorporar ítems ancla, contrabalanceo o un pilotaje independiente de las formas.

#### 4.5. Implicaciones, fortalezas y limitaciones

Para la práctica docente, la propuesta sugiere cuatro principios transferibles: situar los enteros en contextos con referentes explícitos, asociar la retroalimentación con errores conceptuales, graduar las ayudas antes de mostrar la solución y organizar la interacción mediante roles que exijan explicar y verificar.

Entre las fortalezas se encuentran la descripción de la lógica de intervención, la comparación pareada, el análisis de sensibilidad, el reporte de intervalos y tamaño del efecto y la decisión de no interpretar medidas secundarias cuya trazabilidad era insuficiente.

Las limitaciones restringen el alcance del hallazgo. El diseño careció de grupo de comparación, la selección fue por conveniencia, el tamaño muestral fue reducido y el posttest fue inmediato. Además, las formas no fueron equiparadas psicométricamente y algunos ítems mostraron funcionamiento deficiente.

Futuras investigaciones deberían incorporar varias aulas y un grupo comparativo, aplicar una medición diferida, revisar las formas de prueba y conservar respuestas por ítem. También conviene registrar la fidelidad por sesión y el uso de pistas para examinar los mecanismos de aprendizaje.

### 5. Conclusiones

En el grupo analizado se observó un incremento en la puntuación total de resolución de problemas con números enteros entre el pretest y el posttest. La diferencia fue estadísticamente significativa y el tamaño del efecto intrapersonal fue compatible con una magnitud moderada a grande.

El resultado constituye evidencia preliminar de un cambio temporalmente asociado con la aplicación de Enterolandia. No permite atribuirlo exclusivamente a la intervención ni generalizarlo a otras poblaciones, debido a la ausencia de grupo comparativo, el tamaño muestral, el posttest inmediato y la equivalencia no demostrada de las formas.

Las diferencias por dimensión fueron descriptivas. Aunque desplazamientos presentó el mayor incremento medio, ninguna comparación conservó significación después del ajuste de Holm.

Enterolandia ofrece un diseño pedagógico que integra narrativa, problemas contextualizados, retroalimentación específica, pistas graduadas, reintentos, transferencia y roles colaborativos. Su potencial deberá evaluarse mediante estudios comparativos con instrumentos revisados y mediciones de seguimiento

### Referencias Bibliográficas

- Anselmi, P., Colledani, D., & Robusto, E. (2019). A comparison of classical and modern measures of internal consistency. *Frontiers in Psychology*, 10, Article 2714. <https://doi.org/10.3389/fpsyg.2019.02714>
- Bhattacharjee, A. (2012). *Social science research: Principles, methods, and practices*. University of South Florida. [https://digitalcommons.usf.edu/oa\\_textbooks/3/](https://digitalcommons.usf.edu/oa_textbooks/3/)
- Cai, Z., Zhang, X., Liu, C., & Zhan, J. (2025). Effects of digital game-based learning on students' problem-solving ability: A three-level meta-analysis. *Journal of Computer Assisted Learning*, 41(2), Article e70002. <https://doi.org/10.1111/jcal.70002>
- Chamorro-Andrade, R. D., Mendoza-Toala, E. F., & Jama-Zambrano, V. R. (2025). Estrategias neurodidácticas para la enseñanza de operaciones básicas de cálculo en Educación Básica Media en la Unidad Educativa Cinco de Mayo del cantón Chone. *Innova Science Journal*, 3(4), 126–134. <https://doi.org/10.63618/omd/isj/v3/n4/126>
- Çetinkaya-Rundel, M., Diez, D. M., & Barr, C. D. (2019). *OpenIntro statistics* (4th ed.). OpenIntro. <https://www.openintro.org/book/os/>
- Downey, A. B. (2014). *Think stats: Exploratory data analysis* (2nd ed.). Green Tea Press. <https://greenteapress.com/wp/think-stats-2e/>
- Fadda, D., Pellegrini, M., Vivanet, G., & Zandonella Callegher, C. (2022). Effects of digital games on student motivation in mathematics: A meta-analysis in K–12. *Journal of Computer Assisted Learning*, 38(1), 304–325. <https://doi.org/10.1111/jcal.12618>
- Google. (s. f.). *Google Colaboratory: Frequently asked questions*. <https://research.google.com/colaboratory/faq.html>
- Haji Ismail, N. F., Shahrill, M., & Asamoah, D. (2023). Learning through virtual manipulatives: Investigating the impact of Gizmos-based lessons on students' performance in integers. *Contemporary Mathematics and Science Education*, 4(1), Article ep23009. <https://doi.org/10.30935/conmaths/12857>
- Jhangiani, R. S., Chiang, I.-C. A., Cuttler, C., & Leighton, D. C. (2019). *Research methods in psychology* (4th ed.). Kwantlen Polytechnic University. <https://kpu.pressbooks.pub/psychmethods4e/>

- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *Frontiers in Psychology*, 4, Article 863. <https://doi.org/10.3389/fpsyg.2013.00863>
- Li, M., Ma, S., & Shi, Y. (2023). Examining the effectiveness of gamification as a tool promoting teaching and learning in educational settings: A meta-analysis. *Frontiers in Psychology*, 14, Article 1253549. <https://doi.org/10.3389/fpsyg.2023.1253549>
- López-Roldán, P., & Fachelli, S. (2015). *Metodología de la investigación social cuantitativa*. Universitat Autònoma de Barcelona. <https://dialnet.unirioja.es/servlet/libro?codigo=857471>
- Lukman, H. S., Agustiani, N., & Setiani, A. (2023). Gamification of mathematics teaching materials: Its validity, practicality and effectiveness. *International Journal of Emerging Technologies in Learning*, 18(20), 4–22. <https://doi.org/10.3991/ijet.v18i20.36189>
- Merino Soto, C., & Livia Segovia, J. (2009). Intervalos de confianza asimétricos para el índice de validez de contenido: Un programa Visual Basic para la V de Aiken. *Anales de Psicología*, 25(1), 169–171. <https://revistas.um.es/analesps/article/view/71631>
- Pastás Hernández, X. del R., Zumba Novay, E. G., Imbacuán Gordón, D. F., & Peña Robles, C. J. (2024). Genially gamification tool for teaching and learning mathematics. *Revista Imaginario Social*, 7(1). <https://doi.org/10.59155/is.v7i1.151>
- Pedrosa, I., Suárez-Álvarez, J., & García-Cueto, E. (2013). Evidencias sobre la validez de contenido: Avances teóricos y métodos para su estimación. *Acción Psicológica*, 10(2), 3–18. <https://doi.org/10.5944/ap.10.2.11820>
- Putra, L. D., & Afrina, N. (2023). The development of Genially-based interactive learning multimedia for elementary school students. *Jurnal Fundadikdas (Fundamental Pendidikan Dasar)*, 6(2), 138–151. <https://doi.org/10.12928/fundadikdas.v6i2.8413>
- Rizo-Padilla, Z. P., Zambrano-Valdez, G. C., & Jama-Zambrano, V. R. (2025). Los juegos didácticos y su influencia en el pensamiento lógico matemático en niños de básica media. *Innova Science Journal*, 3(4), 254–268. <https://doi.org/10.63618/omd/isj/v3/n4/134>
- Sailer, M., & Homner, L. (2020). The gamification of learning: A meta-analysis. *Educational Psychology Review*, 32, 77–112. <https://doi.org/10.1007/s10648-019-09498-w>
- Taber, K. S. (2018). The use of Cronbach's alpha when developing and reporting research instruments in science education. *Research in Science Education*, 48, 1273–1296. <https://doi.org/10.1007/s11165-016-9602-2>
- Uchima-Marin, C., Ospina, J., Ospina, V., Salvador-Acosta, L., & Acosta-Vargas, P. (2025). Design and implementation of a gamified math game for learning whole numbers in secondary education using Genially. *Sustainability*, 17(21), Article 9759. <https://doi.org/10.3390/su17219759>

- Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Anales de Psicología*, 33(3), 755–782. <https://doi.org/10.6018/analesps.33.3.268401>
- Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, 10, Article 3087. <https://doi.org/10.3389/fpsyg.2019.03087>
- Zambrano-Alay, A. G. (2025). La evaluación formativa y su impacto en la motivación por el aprendizaje en estudiantes de séptimo grado: Revisión bibliográfica. *Innova Science Journal*, 3(4), 15–26. <https://doi.org/10.63618/omd/isj/v3/n4/119>
- Zeng, J., Sun, D., Looi, C.-K., & Fan, A. C. W. (2024). Exploring the impact of gamification on students' academic performance: A comprehensive meta-analysis of studies from the year 2008 to 2023. *British Journal of Educational Technology*, 55(6), 2478–2502. <https://doi.org/10.1111/bjet.13471>

### CONFLICTO DE INTERESES

“Los autores declaran no tener ningún conflicto de intereses”.