

Validez de contenido, usabilidad y accesibilidad de un recurso HTML interactivo para la enseñanza de la entropía.

Content Validity, Usability, and Accessibility of an Interactive HTML Resource for Teaching Entropy.

Valarezo-Chamba, Bryan Stalin¹; Guerrero-Zambrano, Marcos².

¹ Universidad Estatal de Milagro; Ecuador, Milagro; <https://orcid.org/0009-0001-3907-247X>; bvalarezoc@unemi.edu.ec

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

¹ Autor Correspondencia

 <https://doi.org/10.63618/omd/isj/v4/n3/394>

Cita: Valarezo-Chamba, B. S., & Guerrero-Zambrano, M. (2026). Validez de contenido, usabilidad y accesibilidad de un recurso HTML interactivo para la enseñanza de la entropía. *Innova Science Journal*, 4(3), 975-989. <https://doi.org/10.63618/omd/isj/v4/n3/394>

Recibido: 27/02/2026

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: Los recursos digitales pueden representar ideas termodinámicas complejas; aun así, su incorporación educativa exige valorar de antemano el rigor científico, la coherencia pedagógica, la facilidad de uso y las condiciones de acceso. La investigación examinó la validez del contenido, la usabilidad y la accesibilidad de Entrop-IA v5, una aplicación HTML creada para enseñar entropía en bachillerato y universidad. Se adoptó un enfoque mixto convergente, mediante un diseño no experimental, transversal y descriptivo-evaluativo. Participaron cinco jueces, quienes remitieron cinco instrumentos en torno a su uso; después de eliminar dos registros totalmente duplicados, la muestra analítica quedó integrada por tres evaluadores independientes. La validez de contenido se examinó mediante 25 ítems valorados en seis criterios, lo que produjo 18 valoraciones por ítem. En las cinco dimensiones se obtuvo evidencia preliminar favorable, con valores V de Aiken entre 0.801 y 0.986. Las revisiones prioritarias correspondieron a la incorporación de preguntas guía, la explicación de microestados y la diferenciación entre la rapidez del modelo y la temperatura termodinámica. El grado de acuerdo entre los evaluadores fue moderado y alcanzó significación estadística ($W = 0.513$; $p = 0.045$). La evaluación heurística detectó solo dificultades cosméticas o menores, sin registrar problemas mayores ni críticos. El checklist de accesibilidad obtuvo una puntuación descriptiva ponderada de 81.1%, y el principio perceptible presentó el rendimiento más reducido. La integración de los hallazgos justifica avanzar hacia una etapa de optimización técnica y didáctica, seguida de pruebas con usuarios. Sin embargo, los resultados son preliminares por el reducido número de evaluadores únicos y por los ajustes conceptuales, pedagógicos, operativos, visuales, adaptativos y textuales aún requeridos.

Palabras clave: entropía; tecnología educativa; simulación por computadora; usabilidad; accesibilidad web.

Abstract: Digital resources can facilitate the representation of complex thermodynamic concepts. However, their integration into educational settings requires the prior evaluation of scientific accuracy, pedagogical coherence, usability, and accessibility. This study assessed the content validity, usability, and accessibility of Entrop-IA v5, an HTML-based application designed to support the teaching of entropy at upper-secondary and university levels. A convergent mixed-methods approach was employed within a non-experimental, cross-sectional, descriptive-evaluative design. Five expert judges participated and submitted five evaluation instruments related to the application. After two fully duplicated records were excluded, the analytical sample comprised three independent evaluators. Content validity was assessed through 25 items evaluated according to six criteria, resulting in 18 ratings per item. Preliminary favorable evidence was obtained across the five dimensions, with Aiken's V coefficients ranging from 0.801 to 0.986. The main areas requiring revision included the incorporation of guiding questions, a clearer explanation of microstates, and a more explicit distinction between simulation speed and thermodynamic temperature. Inter-rater agreement was moderate and statistically significant ($W = 0.513$, $p = 0.045$). The heuristic evaluation identified only cosmetic or minor usability issues, with no major or critical problems reported. The accessibility checklist yielded a weighted descriptive score of 81.1%, with the Perceivable principle showing the lowest level of compliance. The integrated findings support progression to a phase of technical and pedagogical optimization, followed by user testing. Nevertheless, the results should be considered preliminary because of the limited number of independent evaluators and the conceptual, pedagogical, operational, visual, adaptive, and textual refinements that remain necessary.

Keywords: entropy; educational technology; computer simulation; usability; web accessibility.

1. Introducción

La integración de tecnologías digitales en la educación científica ha favorecido la creación de recursos capaces de combinar textos, ecuaciones, gráficas, animaciones y simulaciones. Estas herramientas pueden ampliar las posibilidades de aprendizaje cuando permiten observar relaciones entre variables, formular predicciones y contrastar de inmediato las consecuencias de una decisión. Sin embargo, la interactividad no garantiza por sí sola calidad educativa: el diseño tecnológico debe articularse con propósitos científicos, didácticos y comunicativos claramente definidos (Battaglia et al., 2023).

En física y química, las simulaciones adquieren especial relevancia porque numerosos fenómenos no pueden observarse de manera directa o exigen relacionar niveles macroscópicos, microscópicos, simbólicos y matemáticos. La evidencia reciente indica que estos recursos pueden fortalecer la comprensión cuando se integran en secuencias que orientan la exploración y la reflexión del estudiantado (Diab et al., 2024). En termodinámica, además, los apoyos visuales deben diseñarse considerando atención, memoria, organización de la información y carga cognitiva (Delgado-Sanchez et al., 2023).

La comprensión de la entropía demanda articular distintos niveles de representación y reconocer las condiciones bajo las cuales ocurren los procesos termodinámicos. Entre las dificultades documentadas se encuentran la interpretación de los cambios entrópicos en el sistema y sus alrededores, la diferenciación entre procesos reversibles e irreversibles y la conexión entre las propiedades macroscópicas y las configuraciones microscópicas. En este contexto, la noción cotidiana de desorden puede emplearse como una analogía introductoria, siempre que se expliciten sus limitaciones y no se presente como una definición suficiente del concepto. Una revisión sistemática con metaanálisis señaló que la abstracción conceptual, la coexistencia de interpretaciones, las concepciones alternativas y las exigencias matemáticas continúan siendo obstáculos centrales para enseñar entropía y segunda ley de la termodinámica (Natalis & Leyh, 2025).

Las dificultades tampoco se limitan al uso de ecuaciones. Estudiantes universitarios pueden resolver ejercicios algorítmicos sin interpretar adecuadamente procesos reversibles, irreversibles, de equilibrio y de no equilibrio (Weber et al., 2024). Por tanto, un recurso HTML sobre entropía debe distinguir claramente las variables del modelo de las magnitudes físicas reales, precisar el uso de sus analogías e incorporar apoyos que favorezcan la predicción, la comparación, justificación e interpretación.

Antes de emplear un recurso de este tipo con estudiantes, resulta muy útil analizar su validez de contenido. Esta describe el grado en que los componentes del material representan de forma pertinente, relevante, clara, coherente, suficiente y rigurosa el campo conceptual que se busca desarrollar. El juicio de expertos es un procedimiento habitual para obtener esta evidencia, aunque su calidad depende de la pertinencia de los evaluadores, la claridad de las tareas y la transparencia del análisis son fundamentales (Beck, 2020). Estudios recientes también han utilizado el juicio experto y la V de Aiken para documentar validez de instrumentos educativos y tecnológicos (Gutiérrez-Castillo et al., 2023; Kania et al., 2024; Merino-Soto, 2023).

La usabilidad constituye otra dimensión decisiva. Un recurso científicamente correcto puede resultar poco funcional cuando su navegación es confusa, los controles no son reconocibles o la retroalimentación no permite comprender lo ocurrido. Las diez heurísticas de Nielsen ofrecen un marco para identificar problemas de interfaz y priorizarlos según severidad (Nielsen, 1994). Su aplicación en sistemas de información y plataformas educativas ha permitido localizar problemas desde detalles cosméticos hasta barreras críticas (Jeddi et al., 2020; Kusuma et al., 2024).

La accesibilidad complementa esta valoración tecnológico-pedagógica. Las Web Content Accessibility Guidelines agrupan los lineamientos de inclusión en cuatro principios: perceptible, operable, comprensible y robusto (World Wide Web Consortium, 2018, 2023). En un recurso orientado a la entropía, las ecuaciones, gráficas, representaciones de partículas y animaciones requieren alternativas textuales; los controles han de funcionar mediante teclado; y la información no puede depender únicamente del color. Las revisiones automatizadas deben combinarse con inspección minuciosa, pues numerosas barreras solo se interpretan atendiendo al contexto de empleo y a la intencionalidad pedagógica (Campoverde-Molina et al., 2023; Timbi-Sisalima et al., 2024).

La literatura muestra avances en simulaciones científicas, enseñanza de la entropía, usabilidad y accesibilidad, pero estas dimensiones suelen examinarse por separado. El vacío de investigación se encuentra en la limitada articulación de validez de contenido, usabilidad y accesibilidad dentro de un mismo proceso de evaluación aplicado a un recurso HTML para la enseñanza de la entropía. Esta articulación es importante porque impide asumir que un recurso posee valor educativo apenas por presentar contenidos correctos o funcionar sin dificultades técnicas.

El objetivo del estudio fue valorar la validez de contenido, la usabilidad y la accesibilidad de Entrop-IA v5, un recurso HTML interactivo para la enseñanza de la entropía en bachillerato y educación superior, para identificar fortalezas, limitaciones y prioridades de mejora antes de usarlo con estudiantes. La pregunta central fue: ¿qué grado de validez de contenido, usabilidad y accesibilidad presenta el recurso? A partir de este marco se formularon tres preguntas concretas: 1) ¿qué nivel de validez científica, didáctica y comunicativa se estima según el criterio de expertos?; 2) ¿qué dificultades de usabilidad muestran las heurísticas de Nielsen y cuál es su severidad?; y 3) ¿en qué grado el recurso satisface los criterios del checklist agrupados según los principios perceptible, operable, comprensible y robusto?

2. Materiales y Métodos

2.1. Enfoque y diseño

Se efectuó un estudio evaluativo de un recurso educativo digital con enfoque mixto convergente, diseño no experimental y corte transversal. El componente cuantitativo estuvo constituido por las puntuaciones de validez de contenido, los niveles de severidad heurística y las categorías de cumplimiento del checklist de accesibilidad. El componente cualitativo correspondió a las observaciones y recomendaciones consignadas por los evaluadores. Ambos componentes se analizaron inicialmente por separado y luego se integraron mediante una matriz conjunta, de acuerdo con la lógica de los diseños convergentes y de los joint displays (Fetters & Tajima, 2022; Peters et

al., 2024). El estudio se orientó por un paradigma pragmático, dado que se combinaron procedimientos destinados a producir evidencia útil para mejorar un producto educativo.

2.2. Objeto y contexto de evaluación

El recurso evaluado fue Entrop-IA v5, diseñado con HTML, CSS y JavaScript. Su estructura abarca una pantalla de inicio, un módulo orientador y tres escenas interactivas: difusión de perfume, transferencia térmica y fusión del hielo. La unidad analítica integró los componentes científicos, didácticos, visuales, interactivos y accesibles del recurso, entre ellos textos, analogías, ecuaciones, gráficas, controles, estados del sistema, ayudas y elementos de programación vinculados con la accesibilidad. La evaluación se desarrolló virtualmente, mediante la ejecución local del archivo en navegadores web actualizados en un entorno controlado.

2.3. Participantes y depuración del panel

Se aplicó un muestreo no probabilístico intencional. Se invitó a cinco evaluadores con experiencia en física, enseñanza de las ciencias, tecnología educativa, diseño de recursos digitales, usabilidad o accesibilidad. Todos remitieron los instrumentos requeridos. Durante la depuración se observó que dos casos reproducían exactamente los metadatos, puntuaciones y observaciones de registros recibidos previamente. Para asegurar la independencia de las respuestas, ambos duplicados fueron excluidos antes del análisis; en consecuencia, el panel analítico quedó integrado por tres evaluadores únicos. Los registros incluidos cumplieron los criterios de participación y aportaron evidencia de revisión de los tres módulos del recurso.

Los criterios de inclusión comprendieron: formación de posgrado o experiencia profesional pertinente; experiencia docente o investigativa en física, ciencias, tecnología educativa o evaluación digital; experiencia en diseño, validación o revisión de materiales educativos; y disponibilidad para examinar el recurso en su totalidad. Se excluyeron instrumentos con más del 10% de respuestas omitidas, sin identificación suficiente del perfil profesional, sin evidencia de exploración completa, duplicados exactos o enviados fuera del periodo de recolección sin justificación documentada.

2.4. Variables y dimensiones

Se analizaron tres variables: validez de contenido, usabilidad y accesibilidad. La validez se organizó en cinco dimensiones: rigor y precisión científica; coherencia didáctica y objetivos; diseño de interacción y control; representación de las escenas; y comunicación, legibilidad y carga cognitiva. La usabilidad se examinó mediante las diez heurísticas de Nielsen. La accesibilidad se estructuró según los principios perceptible, operable, comprensible y robusto de WCAG.

2.5. Instrumentos

2.5.1. Matriz de validez de contenido

El instrumento de validez estuvo compuesto por 25 ítems distribuidos en cinco dimensiones. Cada ítem se examinó según seis criterios: relevancia, claridad, coherencia, suficiencia, precisión científica y pertinencia didáctica. Se utilizó una escala ordinal de cuatro niveles: 1 = no cumple, 2 = cumple poco, 3 = cumple y 4 = cumple totalmente. Así, cada evaluador asignó seis puntuaciones por ítem y, con tres

evaluadores, cada ítem acumuló 18 valoraciones. Esta organización explica que los coeficientes por ítem no se restrinjan a incrementos de 1/9, puesto que la unidad de cálculo correspondió a la valoración criterio-ítem y no a una puntuación general de cada juez. El instrumento incorporó espacios para observaciones y recomendaciones de mejora.

2.5.2. Matriz de evaluación heurística

El segundo instrumento se apoyó en las diez heurísticas de Nielsen. Cada evaluador examinó la pantalla inicial y las escenas de perfume, transferencia térmica y fusión del hielo. Para cada heurística se registraron la ubicación del hallazgo, su descripción, la severidad y una recomendación muy concreta. La severidad se valoró de 0 a 4: 0 = sin problema; 1 = cosmético; 2 = menor; 3 = mayor; y 4 = crítico. El instrumento exigió evidencia detallada en los casos con severidad 3 o 4.

2.5.3. Checklist de accesibilidad

El tercer instrumento integró 15 criterios agrupados según los cuatro principios WCAG. Se examinaron contraste, legibilidad, jerarquía visual, empleo del color, navegación mediante teclado, foco visible, manejo de controles, reinicio, claridad de etiquetas, instrucciones, tecnicismos, retroalimentación, compatibilidad con navegadores, diseño adaptable y alternativas de interacción. Cada criterio se calificó como Sí, Parcial, No o No aplica. La revisión contempló al menos dos navegadores actualizados, tamaño de pantalla de escritorio y ventana estrecha simulada, así como pruebas con Tab, Enter y barra espaciadora. Las pautas WCAG 2.1 y 2.2 se utilizaron como marco de referencia, sin atribuir a la puntuación del checklist el carácter de certificación formal de conformidad.

2.6. Procedimiento

Se comprobó el funcionamiento técnico de Entrop-IA v5 y la operación de sus tres escenas. Se enviaron a los cinco evaluadores el archivo HTML, instrucciones, consentimiento y los tres instrumentos digitales. Cada evaluador examinó el recurso de forma independiente y entregó sus registros sin intercambiar valoraciones con los demás. Se revisaron los datos mediante comparación de metadatos, valoraciones y observaciones; se excluyeron dos duplicados exactos. Se examinaron por separado los resultados de validez, usabilidad, accesibilidad y las observaciones cualitativas. Se reunieron los resultados en una matriz conjunta para determinar coincidencias, complementariedades, prioridades de rediseño y meta-inferencias.

2.7. Análisis de datos

Para la validez de contenido se calcularon medias y medianas descriptivas por dimensión, así como V de Aiken por ítem y dimensión. La V se estimó sobre las valoraciones criterio-ítem mediante $V = \sum s / [n(c-1)]$, donde $s = r - l_0$, r es la puntuación asignada, l_0 la puntuación mínima, n el número de valoraciones incluidas en la unidad analítica y c el número de categorías de respuesta. Por ítem, n fue igual a 18 valoraciones (tres evaluadores por seis criterios). Debido al reducido número de evaluadores únicos, los coeficientes se interpretaron como evidencia preliminar y diagnóstica. Como criterio operativo, se conservaron ítems con $V \geq 0.80$, se revisaron valores entre 0.70 y 0.79 y se priorizaron para reformulación los valores inferiores a 0.70.

La concordancia entre evaluadores se examinó mediante W de Kendall sobre los 25 ítems.

Los hallazgos de usabilidad se clasificaron por heurística, módulo y severidad. Se resumieron mediante media, mediana, severidad máxima y número de evaluadores que reportaron problema. Para accesibilidad se asignaron valores descriptivos de 1 a Sí, 0.5 a Parcial y 0 a No; la puntuación ponderada se calculó como $PC = (\sum p_i / N) \times 100$, donde N fue el número de criterios aplicables. Esta puntuación se utilizó exclusivamente como indicador descriptivo del checklist y no como certificación WCAG.

Las observaciones se examinaron mediante análisis de contenido dirigido. Las categorías iniciales abarcaron precisión científica, coherencia didáctica, interacción y control, usabilidad y accesibilidad. Se agruparon recomendaciones frecuentes y se relacionaron con los resultados cuantitativos en una matriz conjunta. La integración se consideró convergente cuando ambos componentes indicaron el mismo problema, y complementaria cuando las observaciones aclararon o ampliaron una puntuación.

2.8. Consideraciones éticas

La participación fue voluntaria y estuvo precedida por consentimiento informado. Se comunicaron el propósito del estudio, las tareas solicitadas, el uso académico de las respuestas y el derecho a retirarse antes del cierre de la base. Las identidades se protegieron mediante códigos y en el artículo se reportan únicamente datos agregados. Esta fase no involucró estudiantes, menores ni expedientes académicos. Los archivos originales se conservaron para auditoría y los duplicados se mantuvieron como respaldo, sin incluirlos en los análisis.

3. Resultados

3.1. Depuración de los registros

Se recopilaron cinco conjuntos de instrumentos aportados por los cinco evaluadores seleccionados. La comparación de metadatos, fechas, puntuaciones y observaciones mostró que dos envíos duplicaban exactamente registros ya recibidos. Ambos se excluyeron antes del análisis, de modo que la muestra analítica quedó constituida por tres evaluadores únicos.

Tabla 1.

Depuración de los instrumentos recibidos

Registro	Condición	Decisión analítica	Justificación
E1	Original	Incluido	Evaluador único
E2	Original	Incluido	Evaluador único
E3	Original	Incluido	Evaluador único
E4	Duplicado exacto	Excluido	Coincidencia de metadatos, puntuaciones y observaciones
E5	Duplicado exacto	Excluido	Coincidencia de metadatos, puntuaciones y observaciones

Nota. Elaborado por los autores

3.2. Validez científica, didáctica y de contenido

La evaluación global aportó evidencia preliminar favorable en las cinco áreas. Interacción y control alcanzó el coeficiente superior ($V = 0.986$), seguida por representación de las escenas ($V = 0.903$). La coherencia didáctica y los objetivos

registraron el valor inferior ($V = 0.801$), apenas superior al umbral operativo. Estos resultados deben leerse como indicadores diagnósticos del recurso y no como una validación definitiva, debido al reducido tamaño del panel. El análisis por dimensión se expone en la Tabla 2 y se resume visualmente en la Figura 1.

Tabla 2.

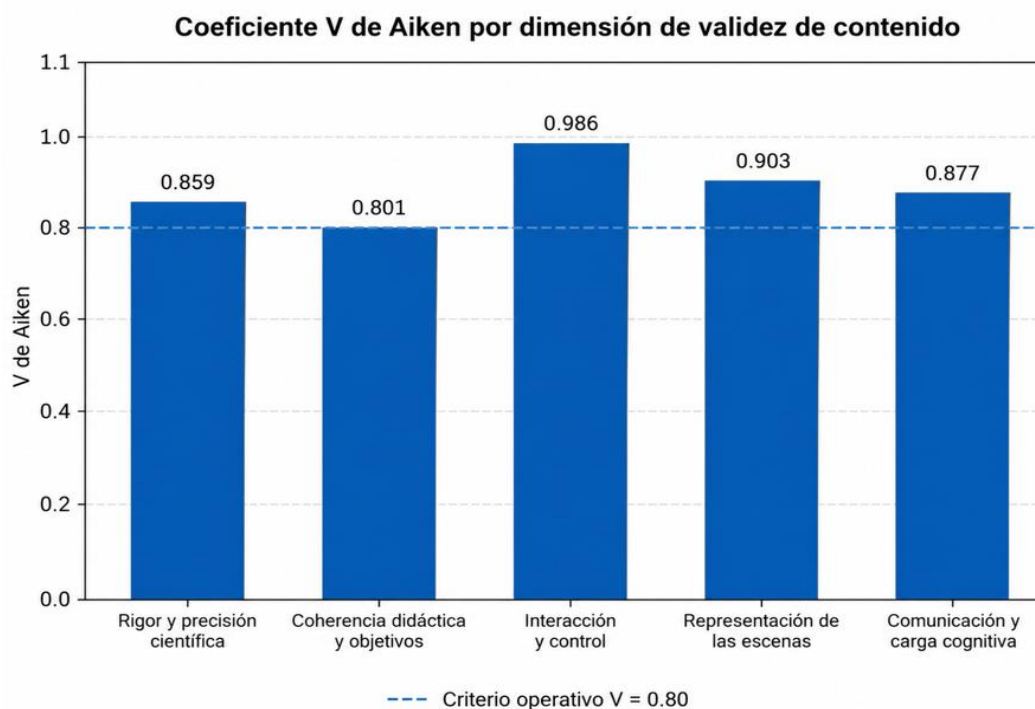
Validez de contenido por dimensión

Dimensión	Media	Mediana	V de Aiken
Rigor y precisión científica	3.58	4.00	0.859
Coherencia didáctica y objetivos	3.40	4.00	0.801
Interacción y control	3.96	4.00	0.986
Representación de cada escena	3.71	4.00	0.903
Comunicación y carga cognitiva	3.63	4.00	0.877

Nota. Escala de 1 a 4. Panel analítico: tres evaluadores únicos. Cada ítem reunió 18 valoraciones criterio-ítem.

Figura 1.

Coeficiente V de Aiken por dimensión de validez de contenido



Nota. Elaborado por los autores

Como se muestra en la Figura 1, las cinco dimensiones superaron el criterio operativo de $V = 0.80$, lo que respalda una valoración global favorable del recurso. No obstante, la proximidad de coherencia didáctica y objetivos al umbral confirma la necesidad de fortalecer el andamiaje pedagógico, mientras que la superioridad de interacción y control sugiere que la arquitectura funcional del recurso se encuentra más consolidada que su mediación didáctica explícita.

El análisis por ítem estableció dos prioridades críticas: B3, asociada con la incorporación de preguntas guía, y D6, vinculada con la distinción entre rapidez del modelo y temperatura termodinámica. Asimismo, debían revisarse la explicación de

microestados, la suficiencia de las instrucciones y la relación entre parámetros del modelo y magnitudes físicas.

Tabla 3.

Ítems con menor coeficiente de validez

Ítem	Contenido resumido	Media	V de Aiken	Decisión
B3	Incorpora preguntas guía que orientan la indagación sin revelar respuestas	3.00	0.667	Reformular
D6	Diferencia la rapidez del modelo de la temperatura termodinámica real	3.06	0.685	Reformular
A4	Evita equiparar entropía con energía o con desorden de manera simplista	3.28	0.759	Revisar
E2	Ofrece instrucciones suficientes para uso autónomo	3.28	0.759	Revisar
A5	Las variables y etiquetas no inducen interpretaciones científicamente incorrectas	3.39	0.796	Revisar
B4	Favorece discusión, comparación y justificación de resultados	3.39	0.796	Revisar
B2	La secuencia explorar-observar-interpretar favorece aprendizaje conceptual	3.44	0.815	Conservar con ajustes

Nota. Elaborado por los autores

La concordancia entre los tres evaluadores fue moderada y estadísticamente significativa, W de Kendall = 0.513, $\chi^2(24) = 36.90$, $p = 0.045$. El resultado muestra un patrón común de valoración, aunque con diferencias razonables en precisión conceptual, orientación didáctica y comunicación visual.

3.3. Usabilidad y severidad

De las 30 valoraciones heurísticas posibles diez heurísticas por tres evaluadores, ocho indicaron ausencia de problemas, 11 se clasificaron como cosméticas y 11 como menores. No se identificaron severidades mayores ni críticas. Las severidades medias más altas del estudio se observaron en correspondencia con el mundo real, control y libertad del usuario, reconocimiento antes que recuerde, y ayuda o documentación, como se expone en la Tabla 4 y se sintetiza en la Figura 2.

Tabla 4.

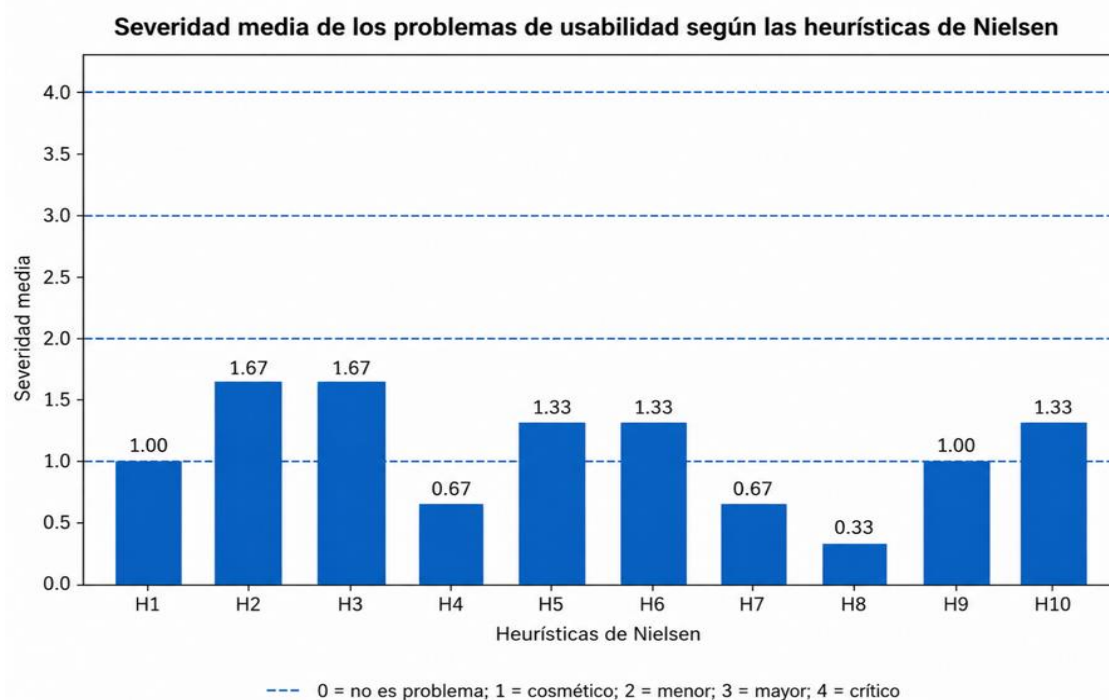
Resultados de la evaluación heurística

H	Heurística	Media	Mediana	Máxima	Evaluadores con problema
H1	Visibilidad del estado del sistema	1.00	1.0	2	2/3
H2	Correspondencia con el mundo real	1.67	2.0	2	3/3
H3	Control y libertad del usuario	1.67	2.0	2	3/3
H4	Consistencia y estándares	0.67	1.0	1	2/3
H5	Prevención de errores	1.33	2.0	2	2/3
H6	Reconocer mejor que recordar	1.33	1.0	2	3/3
H7	Flexibilidad y eficiencia	0.67	1.0	1	2/3
H8	Estética y diseño minimalista	0.33	0.0	1	1/3
H9	Ayuda a reconocer y recuperar errores	1.00	1.0	2	2/3
H10	Ayuda y documentación	1.33	2.0	2	2/3

Nota. Escala de severidad: 0 = no es problema; 1 = cosmético; 2 = menor; 3 = mayor; 4 = crítico.

Figura 2.

Severidad media de los problemas de usabilidad según las heurísticas de Nielsen



Nota. Elaborado por los autores

La Figura 2 indica que las heurísticas H2 y H3 presentaron la mayor severidad media (1.67), seguidas de H5, H6 y H10 (1.33). Este patrón indica que los cambios clave no derivan de fallos críticos, sino de ajustes de diseño orientados a fortalecer la coherencia conceptual del recurso, ampliar el control del usuario sobre la interacción e incorporar apoyos de contexto que permitan una exploración más autónoma y comprensible.

Las observaciones cualitativas se organizaron en cuatro líneas de mejora: 1) precisar la relación entre las metáforas, la rapidez del modelo y la temperatura; 2) ampliar el control mediante opciones de pausa, continuación o avance paso a paso; 3) estandarizar etiquetas, unidades y títulos de las gráficas; y 4) incorporar ayudas contextuales y preguntas orientadoras. La ausencia de puntuaciones 3 y 4 indica que las limitaciones no impidieron completar las tareas esenciales, aunque podrían incrementar la carga cognitiva o disminuir la autonomía del usuario.

3.4. Accesibilidad

La puntuación descriptiva ponderada del checklist logró 81.1%. El principio perceptible obtuvo 75.0%, mientras operable, comprensible y robusto obtuvieron 83.3%. Ningún criterio se valoró como incumplido; las limitaciones fueron consignadas como cumplimiento parcial. Estos porcentajes corresponden únicamente al checklist aplicado y no constituyen una certificación formal de conformidad WCAG.

Tabla 5.

Puntuación del checklist de accesibilidad por principio

Principio	Sí	Parcial	No	Puntuación ponderada
Perceptible	6	6	0	75.0%
Operable	8	4	0	83.3%
Comprensible	8	4	0	83.3%
Robusto	6	3	0	83.3%

Nota. Ponderación descriptiva: Sí = 1; Parcial = 0.5; No = 0. Los conteos corresponden a tres evaluadores.

Los criterios con mayor frecuencia de cumplimiento parcial fueron contraste texto-fondo, legibilidad en pantalla estrecha, dependencia parcial del color, foco visible, operación de controles deslizantes mediante teclado y diseño adaptable. Las recomendaciones priorizaron teclado completo, foco visible, reorganización vertical de animación, controles y gráfica en pantallas pequeñas, y retroalimentación textual que complemente la información gráfica.

3.5. Integración de resultados

Los criterios reportados con mayor frecuencia como cumplimiento parcial fueron contraste texto-fondo, legibilidad en ventana estrecha, dependencia parcial del color, foco visible, uso de mandos deslizantes por teclado y diseño adaptable. Las recomendaciones priorizaron operación completa por teclado, foco perceptible, disposición vertical de animaciones, controles y gráficas en pantallas pequeñas, además de retroalimentación textual complementaria.

Tabla 6.

Matriz conjunta de integración

Dimensión	Resultado cuantitativo	Evidencia cualitativa	Relación	Meta-inferencia y acción
Coherencia didáctica	V = 0.801; B3 = 0.667	Ausencia de preguntas antes, durante y después de las escenas	Convergencia	La interacción requiere andamiaje; incorporar preguntas de predicción, comparación y explicación.
Precisión conceptual	D6 = 0.685; A4 = 0.759	Confusión potencial entre velocidad del modelo, temperatura y metáfora del desorden	Convergencia	Distinguir parámetros del modelo y magnitudes físicas; explicar microestados y alcance de analogías.
Control y ayuda	H2 y H3 = 1.67; H10 = 1.33	Necesidad de pausa, avance paso a paso, tooltips y ayuda contextual	Complementariedad	La interfaz funciona, pero debe aumentar autonomía y control.
Accesibilidad perceptible	75.0%	Textos secundarios pequeños, contraste y dependencia parcial del color	Convergencia	Aumentar contraste, tamaño adaptable y equivalentes textuales.
Operabilidad	83.3%	Foco poco visible y sliders dependientes del mouse	Complementariedad	Implementar operación completa por teclado y foco persistente.

Nota. Elaborado por los autores

Globalmente, Entrop-IA v5 mostró una base preliminar que permite avanzar hacia una fase de rediseño y una nueva evaluación. Las prioridades comprenden ajustes conceptuales, incorporación de orientación didáctica, mejora del control y la navegación, y fortalecimiento de la accesibilidad perceptible y operable.

4. Discusión

Los hallazgos aportan evidencia de que Entrop-IA v5 presenta fortalezas tecnológicas y pedagógicas; sin embargo, el tamaño del panel evaluador y las mejoras aún impiden considerar finalizado su proceso de validación. La convergencia de resultados favorables en contenido, la ausencia de problemas heurísticos mayores o críticos y una puntuación de accesibilidad de 81.1% justifican avanzar hacia una etapa de

perfeccionamiento, aunque no permiten asumir el recurso como una versión definitiva ni atribuirle efectos sobre el aprendizaje.

La principal fortaleza de la validez se observó en interacción y control ($V = 0.986$) y en representación de las escenas ($V = 0.903$), mientras que coherencia didáctica y objetivos registró el valor más bajo ($V = 0.801$). Este comportamiento muestra que la arquitectura interactiva y la representación visual recibieron mejores valoraciones que la mediación pedagógica explícita. El resultado concuerda con Delgado-Sanchez et al. (2023), quienes señalan que una visualización atractiva no asegura comprensión cuando no organiza la atención, la memoria y la selección de información. Asimismo, coincide con Natalis y Leyh (2025), quienes documentaron dificultades persistentes en la interpretación de la entropía y la segunda ley.

La baja valoración de las preguntas guía (B3, $V = 0.667$) muestra que la exploración necesita consignas de predicción, observación, comparación, justificación y explicación. Las simulaciones ofrecen mejores oportunidades de razonamiento cuando la manipulación se integra en secuencias de aprendizaje orientadas (Battaglia et al., 2023; Diab et al., 2024). Por ello, las preguntas deberían distribuirse antes, durante y después de cada escena y conectarse con los cambios observables en el modelo.

La diferenciación entre rapidez del modelo y temperatura termodinámica (D6, $V = 0.685$), junto con las observaciones sobre microestados y la metáfora del desorden, constituye el principal riesgo conceptual. Las analogías pueden ser útiles, pero deben explicitar sus límites. Weber et al. (2024) evidenciaron dificultades al relacionar descripciones macroscópicas y microscópicas y comprender la reversibilidad y el equilibrio. Por ello, Entrop-IA debe presentar cada control como parámetro del modelo, precisar cuándo una ecuación constituye una aproximación y vincular animaciones con explicaciones conceptuales y matemáticas adecuadas al nivel educativo.

La concordancia moderada ($W = 0.513$) mostró un patrón compartido, aunque reflejó diferencias de criterio entre los evaluadores. Esta variación no invalida el diagnóstico; más bien evidencia que la revisión experta requiere transparencia en la selección del panel, la estructura del instrumento y la interpretación de los juicios (Beck, 2020). La exclusión de dos duplicados evitó sobreestimar el acuerdo, aunque el panel final de tres personas obliga a considerar los coeficientes como evidencia preliminar.

La evaluación heurística detectó únicamente problemas cosméticos y menores. La ausencia de severidades 3 y 4 indica que las tareas esenciales pueden ejecutarse, aunque las limitaciones encontradas podrían elevar la carga cognitiva o favorecer interpretaciones poco precisas. Jeddi et al. (2020) y Kusuma et al. (2024) señalan que este procedimiento permite reconocer y jerarquizar los problemas antes de una implementación amplia. Las mejoras propuestas pausa, avance gradual, etiquetas consistentes, opciones visibles y apoyo contextual corresponden a ajustes moderados capaces de fortalecer la autonomía y la eficiencia de uso.

La valoración global de accesibilidad mostró una base funcional, aunque no implica conformidad formal con WCAG. El principio perceptible fue el más débil por contraste, legibilidad y dependencia parcial del color. Además, persistieron problemas de foco visible, control por teclado y diseño adaptable. Campoverde-Molina et al. (2023) señalan que las herramientas automáticas requieren complementarse con revisión manual,

mientras Timbi-Sisalima et al. (2024) sostienen que la accesibilidad debe integrarse desde el diseño. En el ámbito latinoamericano, Celi-Meza et al. (2026) e Ibarra-Rivas et al. (2025) destacan que la inclusión tecnológica depende de recursos adaptables, alternativas de acceso y apoyo institucional.

El aporte metodológico central consiste en integrar la evaluación de contenido, usabilidad y accesibilidad. La matriz conjunta permitió traducir puntajes y observaciones en prioridades de mejora, conforme con el papel analítico de los joint displays (Fetters & Tajima, 2022; Peters et al., 2024). La secuencia propone un orden prudente: primero se diagnostica y mejora el recurso; luego se desarrollan pruebas con usuarios; finalmente, se examina su contribución al aprendizaje mediante un diseño de intervención.

Los resultados requieren una interpretación atenta a varias limitaciones. El panel analítico quedó formado por tres evaluadores únicos, porque los coeficientes dependen del tamaño muestral. La inspección heurística no reemplaza las pruebas con estudiantes. El checklist tampoco constituye una certificación WCAG. El estudio consideró una única versión del recurso y no evaluó aprendizaje, motivación ni transferencia conceptual. Estudios posteriores deberán efectuar otra ronda de validación, incluir estudiantes y usuarios de tecnologías de asistencia, registrar tareas y tiempos de uso, y valorar posteriormente el aprendizaje mediante diseños cuasiexperimentales o experimentales.

5. Conclusiones

Entrop-IA v5 mostró una base preliminar favorable en validez de contenido, usabilidad y accesibilidad que justifica continuar con su rediseño y someter la versión corregida a una nueva ronda de evaluación. Esta conclusión se sustenta en valores de Aiken superiores o próximos al criterio operativo en las cinco dimensiones, la ausencia de problemas heurísticos mayores o críticos y una puntuación descriptiva ponderada de accesibilidad de 81.1%.

Las prioridades conceptuales y didácticas incluyen añadir preguntas guía, explicar microestados, acotar la idea de desorden y distinguir los controles del modelo frente a las magnitudes termodinámicas reales. En usabilidad, las mejoras deben ampliar el control del usuario, unificar etiquetas y brindar apoyo contextual. En accesibilidad, se requiere mejorar contraste, legibilidad, equivalencia textual, foco visible, uso por teclado y diseño adaptable.

La integración del juicio experto, la inspección heurística y el checklist de accesibilidad transformó valoraciones cuantitativas y observaciones cualitativas en decisiones específicas de rediseño. Sin embargo, las conclusiones se circunscriben a Entrop-IA v5, los instrumentos empleados y los tres evaluadores únicos considerados en el análisis.

Se recomienda aplicar las mejoras priorizadas y realizar una segunda evaluación. Después, el recurso deberá someterse a pruebas de usabilidad y accesibilidad con estudiantes y personas usuarias de tecnologías de asistencia. Únicamente entonces será metodológicamente apropiado examinar su efecto sobre la comprensión de la entropía mediante un diseño de intervención con muestra suficiente, grupo comparativo y mediciones del aprendizaje.

Referencias Bibliográficas

- Adams, A., Miller-Lewis, L., & Tieman, J. (2023). Learning designers as expert evaluators of usability: Understanding their potential contribution to improving the universality of interface design for health resources. *International Journal of Environmental Research and Public Health*, 20(5), 4608. <https://doi.org/10.3390/ijerph20054608>
- Battaglia, O. R., Fazio, C., Guastella, A. A., & Tarantino, G. (2023). Outcomes of a teaching learning sequence on modelling surface phenomena in liquids. *Education Sciences*, 13(4), 425. <https://doi.org/10.3390/educsci13040425>
- Beck, K. (2020). Ensuring content validity of psychological and educational tests: The role of experts. *Frontline Learning Research*, 8(6), 1–37. <https://doi.org/10.14786/flr.v8i6.517>
- Campoverde-Molina, M., Luján-Mora, S., & Valverde, L. (2023). Accessibility of university websites worldwide: A systematic literature review. *Universal Access in the Information Society*, 22(1), 133–168. <https://doi.org/10.1007/s10209-021-00825-z>
- Celi-Meza, V. A., Garcés-Yanzapanta, S. D., Bombón-Chico, H. S., & Marín-Delgado, L. D. (2026). Aplicaciones móviles en la enseñanza de las ciencias para estudiantes con discapacidad intelectual en educación secundaria. *Innova Science Journal*, 4(2), 494–505. <https://doi.org/10.63618/omd/isj/v4/n2/290>
- Delgado-Sanchez, J.-M., Benítez-Temiño, B., Moreno-Tejera, S., Larrañeta, M., Silva-Pérez, M., & Lillo-Bravo, I. (2023). Visual resources for learning thermodynamics: A neuroeducation perspective. *Education Sciences*, 13(10), 1030. <https://doi.org/10.3390/educsci13101030>
- Diab, H., Daher, W., Rayan, B., Issa, N., & Rayan, A. (2024). Transforming science education in elementary schools: The power of PhET simulations in enhancing student learning. *Multimodal Technologies and Interaction*, 8(11), 105. <https://doi.org/10.3390/mti8110105>
- Eser, M. T., & Aksu, G. (2022). Comparison of the results of the generalizability theory and inter-rater reliability methods. *International Journal of Curriculum and Instruction*, 14(2), 1629–1643. <https://files.eric.ed.gov/fulltext/EJ1340783.pdf>
- Fetters, M. D., & Tajima, C. (2022). Joint displays of integrated data collection in mixed methods research. *International Journal of Qualitative Methods*, 21, 1–13. <https://doi.org/10.1177/16094069221104564>
- Guo, M., Wang, Y., Yang, Q., Li, R., Zhao, Y., Li, C., Zhu, M., Cui, Y., Jiang, X., Sheng, S., Li, Q., & Gao, R. (2023). Normal workflow and key strategies for data cleaning toward real-world data: Viewpoint. *Interactive Journal of Medical Research*, 12, e44310. <https://doi.org/10.2196/44310>
- Gutiérrez-Castillo, J. J., Palacios-Rodríguez, A., Martín-Párraga, L., & Serrano-Hidalgo, M. (2023). Development of digital teaching competence: Pilot experience and validation through expert judgment. *Education Sciences*, 13(1), 52. <https://doi.org/10.3390/educsci13010052>

- Ibarra-Rivas, J. G., Inga-Aguagallo, C. F., Castillo-Ordóñez, M. A., Lema-Muyulema, S. J., & Gusqui-Mata, N. J. (2025). Entornos virtuales y educación inclusiva: Experiencias desde la ruralidad. *Innova Science Journal*, 3(4), 708–719. <https://doi.org/10.63618/omd/isj/v3/n4/168>
- Jeddi, F. R., Nabovati, E., Bigham, R., & Farrahi, R. (2020). Usability evaluation of a comprehensive national health information system: A heuristic evaluation. *Informatics in Medicine Unlocked*, 19, 100332. <https://doi.org/10.1016/j.imu.2020.100332>
- Kania, N., Kusumah, Y. S., Dahlan, J. A., Nurlaelah, E., Gürbüz, F., & Bonyah, E. (2024). Constructing and providing content validity evidence through the Aiken's V index based on the experts' judgments of the instrument to measure mathematical problem-solving skills. *Research and Evaluation in Education*, 10(1). <https://doi.org/10.21831/reid.v10i1.71032>
- Kusuma, H., Rue, F. S., Rumagit, R. Y., & Pratama, G. D. (2024). Usability evaluation of Ruangguru online learning mobile application using heuristic method. *Procedia Computer Science*, 245, 1777–1786. <https://doi.org/10.1016/j.procs.2024.10.241>
- Merino-Soto, C. (2023). Aiken's V coefficient: Differences in content validity judgments. *MHSalud: Revista en Ciencias del Movimiento Humano y Salud*, 20(1), 1–10. <https://doi.org/10.15359/mhs.20-1.3>
- Moran, K. (2023). *How to conduct a heuristic evaluation*. Nielsen Norman Group. <https://www.nngroup.com/articles/how-to-conduct-a-heuristic-evaluation/>
- Natalis, V., & Leyh, B. (2025). Improving the teaching of entropy and the second law of thermodynamics: A systematic review with meta-analysis. *Chemistry Education Research and Practice*, 26, 9–33. <https://doi.org/10.1039/D4RP00158C>
- Nielsen, J. (1994). Enhancing the explanatory power of usability heuristics. En *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 152–158). <https://doi.org/10.1145/191666.191729>
- Peters, M., Baroud, J., Crossland, A., Buli-Holmberg, J., & Schmid, R. F. (2024). Missed opportunities in mixed methods EdTech research? Visual joint display development as an analytical strategy for achieving integration in mixed methods studies. *Educational Technology Research and Development*, 72, 1213–1238. <https://doi.org/10.1007/s11423-023-10234-z>
- Timbi-Sisalima, C., Sánchez-Gordón, M., Otón-Tortosa, S., & Mendoza-González, R. (2024). Self-assessment guide to quality in accessible virtual education: An expert validation. *Sustainability*, 16(22), 10011. <https://doi.org/10.3390/su162210011>
- Weber, I., Borić, T., Mardešić, J., Bilušić, A., & Zoranić, L. (2024). Enhancing thermodynamics education: Insights from student knowledge assessments on (ir)reversible processes and (non)equilibrium phenomena. *Education Sciences*, 14(12), 1395. <https://doi.org/10.3390/educsci14121395>
- World Wide Web Consortium. (2018). *Introduction to understanding WCAG 2.1*. <https://www.w3.org/WAI/WCAG21/Understanding/intro>

World Wide Web Consortium. (2023). *Web Content Accessibility Guidelines (WCAG)*
2.2. <https://www.w3.org/TR/WCAG22/>

CONFLICTO DE INTERESES

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