Criterios de corrección de exámenes tradicionales de Física y Química: docentes en formación frente a la inteligencia artificial

Enric Ortega-Torres

Departament de Didàctica de les Ciències Experimentals i Socials. Facultat de Formació del Professorat. Universitat de València. España. enric.ortega@uv.es

[Recibido: 13 enero 2025, Revisado: 8 mayo 2025, Aceptado: 17 junio 2025]

Resumen: La evaluación de la Física y Química en las aulas de secundaria sigue usando como instrumentos habituales las pruebas de examen tradicionales. Pese al avance en la investigación sobre los métodos de evaluación y los beneficios de la evaluación formativa las resistencias al cambio en el tipo de instrumentos son férreas. El objetivo de este estudio fue comparar los criterios de corrección y calificación con los que parten los docentes de Física y Química en formación y los que integran las herramientas de Inteligencia Artificial (IA) ChatGPT y Gemini. Participan un total de 105 docentes de secundaria en formación. Los resultados muestran una falta de fiabilidad por la disparidad aplicada en los criterios de corrección de las pruebas sin diferencias importantes cuando se comparan con los usados por las herramientas de IA. Se aprecia mayor rigor en su aplicación por parte de la IA. Las conclusiones invitan a diversificar el tipo de instrumentos para evaluar el proceso de enseñanza y aprendizaje de la Física y Química de secundaria y a hacer uso de la IA para su corrección, en caso de seguir usando los exámenes tradicionales.

Palabras clave: Evaluación, Física y Química, Formación docente, Inteligencia Artificial.

Physics and chemistry assessment: A comparative study of traditional exam marking criteria applied by preservice teachers versus AI tools.

Abstract: The evaluation of Physics and Chemistry in secondary school classrooms continues to rely predominantly on traditional exam-based assessments. Despite advances in research on assessment methods and the benefits of formative evaluation, resistance to changing the types of assessment instruments remains strong. This study aimed to compare the correction and grading criteria used by preservice Physics and Chemistry teachers with those integrated into Artificial Intelligence (AI) tools, specifically ChatGPT and Gemini. A total of 105 secondary education preservice teachers participated in the study. The results reveal a lack of reliability due to inconsistencies in the application of correction criteria, with no significant differences compared to those employed by AI tools. However, the AI demonstrated greater rigor in applying these criteria. The findings encourage diversifying the types of instruments used to assess the teaching and learning process in secondary Physics and Chemistry and suggest leveraging AI for grading when traditional exams are maintained.

Keywords: Artificial Intelligence, Assessment, Physics and Chemistry, Teacher Training.

Para citar este artículo: Ortega-Torres, E. (2025). Criterios de corrección de exámenes tradicionales de Física y Química: docentes en formación frente a la inteligencia artificial. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias*. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 22*(2), 2302. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2025.v22.i2.2302

Introducción

La evaluación es un aspecto clave para la didáctica de las ciencias y especialmente importante en los procesos de enseñanza y aprendizaje que se llevan a cabo en las aulas de secundaria. Muchas veces la rigidez y la dificultad para aplicar cambios en los tipos de instrumentos y en los procesos de evaluación que existen en los claustros y en los departamentos de ciencias de los centros de secundaria suponen un obstáculo para poder

actualizar los métodos didácticos. Si no existe coherencia entre los cambios metodológicos y la evaluación que se aplique, no es posible conocer su efecto real sobre el aprendizaje del alumnado. La evaluación no se debe limitar a verificar los resultados del aprendizaje, sino que debe servir para identificar los objetivos alcanzados, las áreas de mejora y para evaluar la calidad del proceso de enseñanza.

A lo largo del siglo XX y XXI, las teorías pedagógicas han debatido ampliamente sobre qué evaluar, con qué propósito y cómo hacerlo. En las últimas décadas, la comunidad científica ha propuesto alternativas a los enfoques tradicionales, incorporando dimensiones estratégicas, técnicas, temporales e instrumentales (Mellado-Moreno et al., 2021) y con ello han ido apareciendo nuevos enfoques que orientan la evaluación hacia una mejora continua a través de la implicación de todo el proceso de enseñanza y aprendizaje de las ciencias (López-Lozano y Solís-Ramírez, 2016; Remesal, 2011). Pese a ello, la tendencia que sigue existiendo en las aulas de ciencias en cuanto a los tipos de pruebas de evaluación sigue estando representada por un predominio de las evaluaciones memorísticas con un enfoque en contenidos específicos (Mazzitelli et al., 2013). Esta realidad en las aulas supone que la transferencia real de los cambios que promuevan mejoras en la formación de los futuros docentes de ciencias en general y de Física y Química en particular se enfrente a una resistencia mayor.

El objetivo de este estudio es comparar los criterios de corrección y calificación de las pruebas de evaluación en forma de exámenes tradicionales con los que parten los docentes de Física y Química en formación y los que integran las herramientas de Inteligencia Artificial (IA) ChatGPT y Gemini de uso gratuito. Esta comparación pretende poner en consideración la problemática del uso habitual de este tipo de pruebas y su falta de fiabilidad, además de vislumbrar la posibilidad del uso de herramientas IA para la corrección de estas, sin diferencias en sus resultados.

A partir de la concepción que un instrumento con fiabilidad es aquel que proporciona una calificación replicable y consistente (Ruiz, 2020) se pretende evidenciar que la consistencia y replicabilidad de las pruebas de evaluación en forma de exámenes tradicionales de Física y Química es baja y no empeora si su corrección se lleva a cabo mediante el uso de IA. Esta constatación permitiría a los futuros docentes incorporar las herramientas IA, con las que ya tienen familiaridad, a su práctica docente como recurso facilitador para los procesos de corrección de exámenes tradicionales.

Pruebas de examen en la docencia de Física y Química en la etapa de secundaria

A día de hoy cuando hablamos de evaluación en Física y Química es habitual pensar en las pruebas tipo examen que realiza el alumnado al final de cada secuencia de aprendizaje para determinar si se ha conseguido alcanzar los objetivos didácticos previstos. Estas pruebas son habituales en todos los centros de secundaria y se realizan al final del proceso de aprendizaje y por tanto no tienen efecto directo sobre dicho proceso ya finalizado. Es decir, este tipo de instrumentos no contribuye a la mejora del aprendizaje del alumnado y solamente pueden ser útiles para certificar el nivel de aprendizaje alcanzado (Ruiz, 2020) en los aspectos particulares que dicha prueba pueda evaluar.

La investigación destaca la importancia de entender la evaluación como una representación de una interacción entre la enseñanza y el aprendizaje (López-Lozano y Solís-Ramírez, 2016) pero todavía hoy, existen resistencias importantes para el diseño de nuevos instrumentos de evaluación en las aulas de secundaria que sustituyan total o parcialmente a los exámenes tradicionales. El nuevo currículo LOMLOE (Ministerio de Educación y Formación Profesional, 2022) aporta cambios interesantes para la evaluación a través del Real Decreto 217/2022 en el que se definen por primera vez las Competencias Específicas de cada área y los Criterios de Evaluación asociados a dichas competencias específicas. Estos cambios introducidos por el Real Decreto se trasladan a la

Comunidad Valenciana, donde se realiza este estudio, a través de su Decreto Autonómico 107/2022 que mantiene y profundiza en ellos, tal y como destaca Quílez (2024). En su "Artículo 2. Definiciones" se definen los criterios de evaluación como los "referentes que indican los niveles de desempeño esperados en el alumnado en las situaciones o actividades a las que se refieren las competencias específicas de cada materia o ámbito en un momento determinado de su proceso de aprendizaje" (RD 217/2022; Decreto G.V. 107/2022) que se diferencian para las diferentes competencias específicas según los cursos 1º, 2º y 3º de ESO y otros para 4º de ESO. Este marco competencial debería implicar cambios en la evaluación, pero la adherencia a prácticas tradicionales (Mazzitelli et al., 2013) por parte de los docentes supone que se evalúe tal y como se evaluó al docente cuando fue alumno. La falta de formación en nuevas metodologías de evaluación puede ser una barrera (Pozuelo et al., 2023), así como la exigencia de cubrir los currículos y la preparación para exámenes externos siguen siendo justificaciones para la resistencia al cambio. Estas resistencias suponen que siga evidenciándose un enfoque tradicional en la evaluación que se lleva a la práctica con preguntas que priorizan la dimensión factual de la ciencia, de baja complejidad y en formatos de examen tradicional (Furman et al., 2013) que requieren respuestas directas y no esperan que los estudiantes analicen la información y establezcan conexiones entre conceptos.

En el área de ciencias hace ya tiempo que existe la concepción que es más sencillo evaluar con mayor objetividad y precisión los aprendizajes del alumnado por la propia naturaleza de los conocimientos a evaluar (Alonso et. al., 1996). Esta idea previa propicia que exista una cierta tendencia a evaluar aquello que sea más fácilmente medible y por ello los instrumentos en forma de exámenes tradicionales se centran en conocimientos fácticos que integran ejercicios de cálculo, aplicación de fórmulas o con respuestas unívocas para evitar posibles ambigüedades. De algún modo sigue perdurando la clásica visión elitista del aprendizaje de las ciencias que da por supuesto que un porcentaje de alumnos no podrá alcanzar los objetivos previstos y por ello el método o los instrumentos de evaluación no tienen influencia sobre estos resultados (Gil et al., 1991).

En general, los instrumentos de evaluación usados en una secuencia didáctica pretenden medir el nivel de consecución que el alumnado ha alcanzado para los objetivos de aprendizaje previstos. Las pruebas de evaluación no miden exactamente todo lo aprendido, sino que se basan en una estimación por muestreo de una parte de lo trabajado y con su resultado se infiere el aprendizaje global. Es decir, miden el nivel de consecución alcanzado para unos objetivos concretos previamente establecidos, pero no miden cualquier otra cosa que el alumnado pueda haber aprendido durante el proceso. Por ello, la coherencia entre el tipo de prueba de evaluación y lo que se pretende evaluar es un factor significativo en cualquier área (González et al., 2021) y también en Física y Química (Lupión y Caracuel, 2021) por lo que un examen no puede servir para evaluar cualquier tipo de aprendizaje en Física y Química, y en otras materias.

La evaluación final sumativa busca medir los aprendizajes al finalizar el proceso planificado y habitualmente se basan en pruebas de retención centradas en recordar partes de los saberes básicos trabajados (Mazzitelli et al., 2013). Este tipo de pruebas facilitan la calificación por parte de los docentes, pero tienen poco efecto sobre la mejora en el aprendizaje del alumnado. Las evaluaciones sumativas basadas en exámenes finales pretenden analizar el rendimiento del alumnado controlando los objetivos propuestos y el tiempo empleado como pretendida garantía de objetividad y obtención rigurosa de evidencias (Arancibia-Herrera et al., 2019) pero no suelen incluir pruebas de transferencia, con un enfoque más competencial y con mejores resultados para el aprendizaje tal y como constata la investigación (Brandsford et al., 2000). Cuando la prioridad es calificar, para cumplir con la norma establecida, se pierde eficiencia en el proceso de evaluación que sirva para el aprendizaje. No es lo mismo evaluar que calificar, la evaluación no requiere de calificación, pero no puede haber calificación sin una evaluación previa (López et al., 2004). Esto supone que el proceso

de enseñanza y aprendizaje de las ciencias quede condicionado debido a que el modo en que se aplican los procesos de evaluación tiene influencia sobre el proceso de aprendizaje del alumnado al condicionar su forma de estudio (Crooks, 1988) entre otros efectos. Cuando el alumnado conoce los instrumentos con los que va a ser evaluado por sus docentes, adapta sus formas de estudio a ellos; no es lo mismo prepararse para superar un examen de Física y Química que aprender Física y Química.

El sentido que debería tener la evaluación en los procesos de enseñanza y aprendizaje de las ciencias sería el de ser un elemento importante que incida en el propio proceso de aprendizaje, que sirva como indicador provisional destinado a favorecer la autorregulación del alumnado (Sanmartí, 2010) y por ello no puede basarse solamente en valoraciones terminales sino que debe incorporarse a todo el proceso con el fin de saber dónde y porqué se producen errores y el modo de corregirlos (Sanmartí, 2020). Existen evidencias interesantes sobre el impacto significativo que tiene la aplicación de métodos de evaluación formadora en estudiantes de Química (Babinčáková et al., 2019). Debemos tener en cuenta que una prueba de evaluación puede actuar de forma formativa en la medida en que sea utilizada por los docentes y estudiantes para obtener evidencias sobre el nivel de desarrollo alcanzado, y siempre que estas evidencias sean útiles para la toma decisiones sobre los siguientes pasos para continuar con el proceso de aprendizaje (William, 2011), pero esto no ocurre cuando solamente sirve para calificar.

Uso de IA para la evaluación

Los métodos de evaluación efectivos en la educación científica deben ser variados y alinearse con los objetivos didácticos definidos previamente, además deberían alentar la participación activa y aprovechar la tecnología para mejorar los resultados de aprendizaje. Las evaluaciones escritas, las herramientas de IA, los métodos colaborativos y los sistemas automatizados pueden ofrecer beneficios interesantes a incorporar. Algunas investigaciones actuales muestran la influencia positiva del uso de plataformas digitales en la evaluación (Morán et. al., 2024) por la posibilidad que ofrecen para promover una evaluación más dinámica y accesible (Tufiño y Cayambe, 2023). Todavía no existe amplia literatura científica sobre el uso de las herramientas IA para la evaluación, y específicamente para la evaluación en el aprendizaje de las ciencias. Pese a ello, si podemos transferir algunos resultados de investigación al área de didáctica de las ciencias que son de interés para este estudio. En este sentido Liao y colaboradores (2024) muestran el beneficio para la mejora de la autoeficacia que aportan las evaluaciones con IA en grupos numerosos gracias al tipo de retroalimentación que aportan en comparación con la que traslada el docente. También se muestran mejoras en el interés por la ciencia de los estudiantes cuando se usan métodos de evaluación interactivos (Abdulkadir et al., 2024).

Por otro lado, Alamr y colaboradores (2023) ponen de manifiesto la importancia de garantizar la integridad académica cuando se usan evaluaciones y calificaciones electrónicas automatizadas, prácticas que ya son habituales en los grados de informática, entre otros.

Existen ya algunos ejemplos de buenas prácticas en la incorporación de herramientas de inteligencia artificial en sistemas educativos, como el ejemplo de Australia que se describe por Furze y Roe (2024) donde se destaca el aumento de rendimiento de los estudiantes y su impacto positivo para la educación secundaria. También destaca este estudio una predisposición positiva para su uso por parte del profesorado que es coherente con la que menciona el trabajo de Yim y Wegerif (2024) para maestros en el sistema educativo chino. En otro trabajo de la misma investigadora se muestra la buena valoración descrita en las prácticas llevadas a cabo con el uso de inteligencia artificial para la evaluación de estudiantes de secundaria, mediante pruebas pre y post, cuestionarios y pruebas de conocimiento (Yim y Su, 2025). En el contexto español, en un estudio reciente (García-Perales et

al., 2025) se muestra una predisposición positiva por parte de los futuros maestros hacia el uso de herramientas IA, aunque con cierta desconfianza cuando se aplican a la evaluación. Todo ello pone en evidencia el papel cada vez más importante de la IA en la transformación de los métodos de evaluación sobre todo en la educación superior (Paiva et al., 2022). La revisión realizada por Paiva y colaboradores sugiere que la IA se está utilizando para mejorar la evaluación, la calificación y la retroalimentación en la educación superior por lo que sería de interés trasladar esta tendencia también a la educación secundaria; tanto en las aulas como en la formación de los y las futuras docentes de ciencias.

El estudio que se describe en este artículo pretende mostrar una vía de uso de la IA para la corrección de exámenes tradicionales en Física y Química tal y como ya se ha descrito previamente en sus objetivos.

Metodología

La investigación se enmarca en un paradigma postpositivista con un diseño exploratorio con muestreo representativo y con función de recolección e interpretación de datos por parte del investigador.

Muestra

Participan un total de 105 docentes de secundaria en formación (55 mujeres y 50 hombres) que cursan la especialidad de Física y Química del Máster de Secundaria en la Universidad de Valencia. Este grupo de futuros docentes ejercerá su profesión en un contexto de interacción habitual con herramientas de inteligencia artificial, motivo por el cual resulta de interés analizar sus concepciones sobre los criterios de corrección de pruebas tipo examen en comparación con dichas herramientas. Se empleó un muestreo censal, incluyendo a todos los estudiantes matriculados en el momento de la recogida de datos. No se aplicaron criterios de exclusión previos.

Del total de participantes el 51.43% provienen del grado o licenciatura de Química, el 13.33% de Física, el 17.14% de Ingenierías y el 18.10% restante proviene de Biotecnología, Bioquímica o Farmacia. La experiencia media en el aula no llega a un año.

La segunda parte de la investigación incluye una muestra de dos chatbots de uso gratuito: ChatGPT 3.0 y Gemini 1.5 Flash.

Instrumentos

Para el estudio se diseñaron dos instrumentos distintos: (1) Prueba resuelta de Física y Química nivel 3º ESO con errores habituales y (2) Formulario ad hoc para la recogida de información sobre la corrección de la prueba y percepciones sobre importancia de errores.

La prueba (1) se compone de 8 preguntas extraídas de exámenes de la asignatura de Física y Química de 3º de ESO diseñados por docentes en activo durante los últimos 4 cursos. Son preguntas que forman parte de pruebas tipo examen utilizadas para que alumnos previamente suspendidos puedan alcanzar el aprobado en la convocatoria extraordinaria. La selección de las preguntas se realiza en función de las más repetidas encontradas en los exámenes analizados provenientes de 3 centros educativos distintos para los cursos 20-21, 21-22, 22-23 y 23-24. La estructura del examen consta de dos preguntas teóricas y 6 cuestiones prácticas con cálculos requeridos. De éstas, 4 se asocian al temario de Química y 2 al de Física. La prueba está resuelta con *errores típicos* cometidos por alumnado de 3º de ESO en pruebas anteriores. Las preguntas se presentan sin cuantificar su valor para la calificación.

Los llamados errores típicos se categorizan en 6 tipologías diferentes a partir del proceso de validación de expertos realizada por 3 docentes en activo tras presentarles un listado previamente elaborado por el investigador. Este listado se fundamenta en diferentes estudios previos en los que se proponen diversas clasificaciones en la tipología de errores cometidos por estudiantes en Física (Slisko, 2013) o Matemáticas (Pochulu, 2009) descritos a partir de la categorización anterior de Rico (1995) que diferenciaba los 6 errores siguientes para la resolución de problemas matemáticos: (1) Datos mal utilizados; (2) Mala interpretación del lenguaje; (3) Inferencias sin lógica; (4) Definiciones deformadas; (5) Solución no verificada y (6) Errores técnicos.

En esta investigación, y tras la validación de los expertos consultados se determinó establecer la siguiente categorización para la tipología de errores del alumnado en la resolución de problemas de examen de Física y Química en 3º de ESO: Errores procedimentales o de cálculo (E1); Error en el cambio de unidades (E2); Errores conceptuales (E3); Error en la comprensión del enunciado (E4); Error en la aplicación de fórmulas (E5); Error en la explicación (E6)".

A pesar de tratarse de estudios con enfoques diferentes, en un caso para Matemáticas y en el otro para Física y Química puede apreciarse una relación clara entre los errores descritos por Rico (1995) y los establecidos por los docentes en activo en este estudio. Observamos una vinculación directa entre E2 y (1); E4 y (2); E3 y (3); E1 y (6), y una relación menos clara entre E5 y (4); E6 y (5). Además, incorporar este refinamiento de expertos en la categorización de los errores incide en la importancia didáctica que diversos autores (Buteler et al., 2008; Ardura y Zamora, 2014) otorgan a este proceso y por ello resulta efectivo en la formación del profesorado de Física y Química, participantes en este estudio.

La definición de los errores típicos descrita sirvió para seleccionar las respuestas a las preguntas de la prueba (instrumento 1) y asignar la tipología de error que éstas incluían tal y como se describe en la Tabla 1.

El formulario ad hoc (instrumento 2) diseñado se presenta en formato on-line (*google forms*) y se compone de 14 ítems diferenciados en 4 bloques. En el primer bloque se incluyen 3 cuestiones para conocer las características de los participantes: Género, Estudios Previos, Años de experiencia en el aula. El segundo bloque integra una pregunta con respuesta numérica para conocer la calificación que se otorga a la prueba corregida (números enteros del 0 al 10). En el tercer bloque se incluyen 4 cuestiones para conocer el nivel de importancia que el participante otorga a los diferentes errores que se cometen en un examen. El último bloque se compone de 5 cuestiones elaboradas para conocer el peso cuantitativo en la corrección que el participante otorga a los distintos errores habituales.

Tabla 1. Estructura de la prueba y errores tipo en las respuestas.

Tipología de pregunta	Competencia específica – Criterio evaluación*	Saberes básicos**	Error en la respuesta
P1- Pregunta teórica sobre las etapas del Método científico	CE2-2.1	BA-SBI	(E6) Falta información. La respuesta presenta información correcta pero incompleta.
P2- Cuestión de cálculo relacionada con leyes de gases	CE1-1.2	BB-SBI	Error en la realización del cambio de unidades (E2)
P3- Cuestión de cálculo sobre abundancia de isotopos	CE2-2.3	BB-SBIII	(E6) Error en la justificación de la respuesta. No se explica.
P4- Cuestión de cálculo relacionada con un movimiento MRUA	CE1-1.2	BD-SBI	(E1) Error de cálculo
P5- Cuestión de cálculo sobre carga eléctrica de átomos	CE3-3.2	BB-SBIII	(E4) Mala interpretación enunciado
P6- Completar una tabla sobre partículas que forman elementos químicos	CE3-3.1	BB-SB3	(E3) Confusión entre dos columnas de la tabla
P7- Cuestión de cálculo sobre MRU	CE1-1.2	BD-SBI	(E5) Respuesta incompleta
P8- Nomenclatura química	CE3-3.2	BB-SBV	(E2/E6) Respuestas correctas con mezcla de nomenclaturas

*Los criterios de evaluación y saberes básicos provienen de https://educagob.educacionyfp.gob.es/eu/curriculo/curriculo-lomloe/menu-curriculos-basicos/ed-secundaria-obligatoria/materias/fisica-quimica/criterios-eval-primer-tercer-curso.html

**Los Saberes básicos se categorizan con el código BA-SBI (En primer lugar, se denomina el bloque A, B, C, D, o E y a continuación muestra el saber básico en números romanos según el orden en el que aparece en el bloque. BA-SBI = Primer Saber básico del Bloque A). Conselleria de Educación, Cultura y Deporte (2022)

Procedimiento

Dentro del módulo sobre evaluación que se integra en la asignatura "Aprendizaje y Enseñanza en la Física y Química" del Máster de Secundaria de la UV se realizó la actividad en la tercera sesión. Previamente se había trabajado con los docentes en formación los tipos de evaluación, los cambios introducidos por la LOMLOE en la evaluación y los diferentes instrumentos de evaluación que pueden usarse en la asignatura de Física y Química. Tras ello se pide al profesorado en formación que corrija el examen resuelto (descrito en el apartado anterior) de forma individual con un tiempo aproximado de 45 min. Tras realizar la corrección se pide que respondan al cuestionario planteado (descrito también en el apartado anterior). El tiempo medio para dar respuesta al cuestionario es de 15 min.

Para la segunda parte del estudio se realizó una preparación paralela de los dos chatbots (ChatGPT 3.0 y Gemini 1.5) mediante la apertura de dos diálogos diferentes para cada uno de ellos, el primero para la corrección del examen y el segundo para la valoración de la importancia de los errores.

Para el desarrollo del primer dialogo se instruye a los dos chats del mismo modo mediante el *prompt* que les solicita que asuman el rol de docente de secundaria en la asignatura de Física y Química de 3º de ESO en la Comunidad Valenciana. Se les aporta el currículo específico de Física y Química extraído del Decreto Autonómico 107/2022 y a continuación se les pide que corrijan y califiquen las 8 preguntas del examen descrito. En un segundo diálogo se les asigna el mismo rol y se aporta el mismo documento y tras ello se les pide que valoren la importancia de los errores y sus

criterios de corrección con un diálogo basado en las mismas preguntas que las que se incluyen en el cuestionario usado con los docentes en formación.

El tiempo requerido en la primera parte (corrección examen) es inferior a 20 minutos en ambos chats. Para la segunda parte se requirió la espera de un día completo por haber superado el número de consultas máximo diario. Tras su restablecimiento el tiempo de respuesta de la segunda parte también fue inferior a los 10 minutos en cada chatbot.

Resultados y discusión

Criterios de calificación de la prueba

La primera decisión que debe tomar el participante en la corrección de la prueba es la de cuantificar el valor de cada una de las 8 preguntas que componen el examen. La forma de proceder del 70.47% de los participantes fue la de otorgar el mismo valor a las 8 preguntas. Solamente el 29.52% de los participantes asignaron valores diferentes a algunas preguntas.

En el caso de los chatbots, tanto ChatGPT 3.0 como Gemini 1.5 Flash asignan valores distintos a las preguntas según "*la importancia de estos bloques temáticos y la dificultad de las preguntas*" tal y como los mismos chats justifican. En este caso no existe una concordancia completa entre las valoraciones otorgadas. Se asigna el mismo valor a 5 de las 8 preguntas por parte de las dos IA, pero existe discrepancia en las otras 3. En dos de ellas Gemini asigna un valor mayor que ChatGPT y es a la inversa en la otra.

En esta primera decisión se puede apreciar un mayor rigor en el modo de proceder de los chatbots en comparación a los docentes en formación que es coherente con las investigaciones que muestran que el profesorado en formación parte de un conocimiento poco desarrollado sobre evaluación (Buck et al.,2010; Wang et al., 2010).

En relación a la calificación otorgada por los 105 docentes en formación a la misma prueba se obtuvo un valor promedio de las calificaciones de 5.02, con una mediana de 5, una desviación estándar de 1.19 y un rango de 5 (Máx.8 y Min.3).

Para la corrección del examen por parte de los chatbots se realizan dos correcciones diferentes: en el primer caso se solicita la corrección y calificación de las preguntas por separado (manteniendo la puntuación otorgada por el propio chat a cada una de ellas) y en segundo lugar se pide al mismo chatbot que realice la corrección y calificación global de todo el examen en su conjunto. Las calificaciones otorgadas se muestran en la tabla 2 a continuación:

Tabla 2. Calificaciones otorgadas por los chatbots al examen

Chatbox	Corrección por preguntas	Corrección de examen conjunto	Calificación media
ChatGPT 3.0	6.5	5.5	6
Gemini 1.5 Flash	5	7	6

Pese a las diferencias observadas en la asignación del peso a las preguntas del examen por parte de cada chatbot y las diferencias que se aprecian en cada corrección por separado, es interesante comprobar que la calificación media coincide en ambos chats.

Además, la media de las calificaciones otorgadas por los chatbots (6) es un punto superior a las otorgadas por los profesores de Física y Química en formación (5.02).

Si analizamos la distribución de las calificaciones asignadas por los docentes en formación mediante el histograma que se observa en la Figura 1 se puede apreciar una distribución con asimetría positiva, concentrada en el rango medio-bajo de calificaciones.

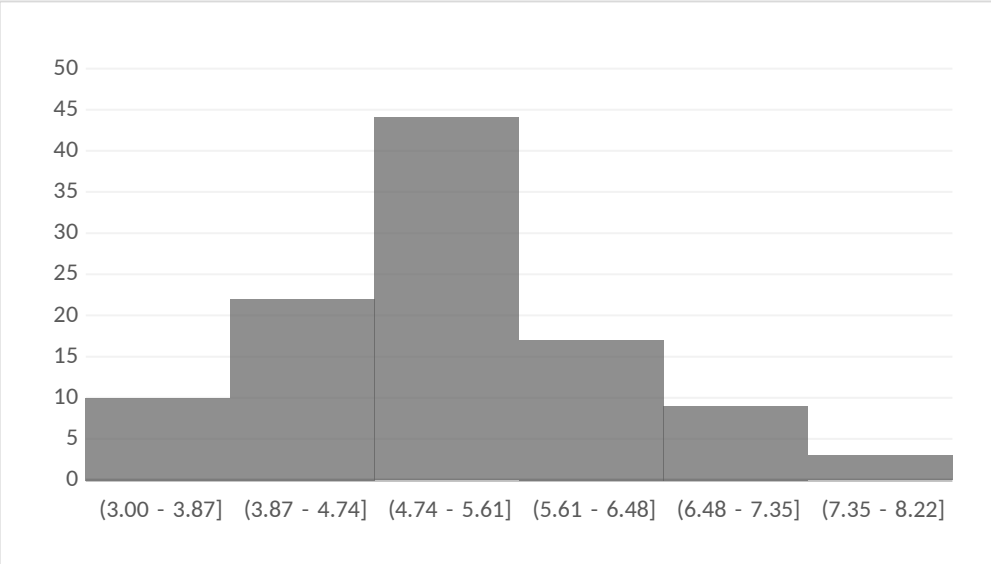

Figura 1. Histograma de la distribución de las calificaciones obtenidas en las 105 correcciones de la misma prueba realizadas por los docentes en formación.

Se realiza el cálculo de la prueba de Shapiro-Wilk y se obtiene un valor de $p = .001 < 0.05$ que indica que no existe una distribución normal de las calificaciones. Esto puede sugerir una tendencia a calificaciones más bajas y/o una posible dificultad del examen o la aplicación de criterios de evaluación estrictos por parte de los docentes en formación. La dispersión en las calificaciones que muestra el histograma revela una considerable falta de consenso entre los profesores en formación para la calificación del examen. Algunos profesores consideraron que el examen era de un nivel muy bajo (calificaciones cercanas a 3), mientras que otros lo calificaron con una nota mucho más alta (cercana a 8), lo cual sugiere poca objetividad y fiabilidad en cuanto al uso de este instrumento de evaluación sin haber revisado y clarificado previamente sus criterios de calificación con el fin de asegurar una mayor uniformidad en los resultados.

En el caso de incluir las calificaciones medias otorgadas por los chatbots en el histograma éstas ocuparían el intervalo de (5.61 - 6.48] que representa al 16.03% de los docentes en formación. Esta inclusión no mejoraría la dispersión de los resultados y seguiría evidenciando el bajo nivel de fiabilidad (Ruiz, 2020) del instrumento.

Se aprecia por tanto que la calificación otorgada por los chatbots es más alta (en cuanto al promedio) que la de los docentes en formación sin que exista una diferencia importante en cuanto a su asignación. La forma de calificar de las herramientas IA es asimilable a la de un grupo importante (16.03%) de los docentes en formación y mejora al 72.3% en la calificación otorgada al supuesto estudiante.

Con el fin de profundizar en las diferencias de los criterios de corrección aplicados por los docentes en formación y los chatbots se amplía el estudio para conocer la importancia que se atribuye a los errores en las respuestas a preguntas de examen de Física y Química que cometen los estudiantes.

A continuación, en la Tabla 3 podemos observar las consideraciones de importancia que asignan los y las futuras profesoras de Física y Química.

Tabla 3. Consideración del tipo de error más importante según los docentes en formación.

Tipología de error	% de profesores en formación
Error procedimental o de cálculo (E1)	0 %
Error en el cambio de unidades (E2)	8.57 %
Errores conceptuales (E3)	57.14 %
Error en la comprensión del enunciado (E4)	11.42 %
Error en la aplicación de fórmulas (E5)	5.71%
Error en la explicación (E6)	17.14 %

Este resultado es consistente cuando el formato de la pregunta es el de ordenar los 6 errores mostrados en la tabla 3 anterior por orden de importancia. En este caso el resultado según la importancia es el siguiente: E3 > E6 > E4 > E2 > E5 > E1

Podemos observar que coincide el orden de importancia del tipo de error asignado por los docentes en formación cuando se plantea la pregunta en ambas modalidades.

Para la ordenación de las 6 posibilidades se realiza el cálculo del coeficiente de Kendall con un resultado de W=0.6612 (para n=105 y k=6). Esto indica un nivel de concordancia moderado entre las asignaciones de importancia de los docentes en formación, con una tendencia a ordenar las opciones de forma similar.

Cuando se pregunta a los chatbots que ordenen los errores anteriores según su nivel de importancia los resultados son los siguientes:

Tabla 4. Importancia del tipo de error según la interpretación de los chatbots

Chatbot	Orden de importancia (de mayor a menor)
ChatGPT	E3 > E4 > E5 > E2 > E1 > E6
Gemini	E3 > E4 > E5 > E1 > E2 > E6

Se observa una interpretación similar entre ambas herramientas IA con la discrepancia en cuanto a la posición 4ª y 5ª donde aparecen en orden inverso los errores E1 y E2.

Cuando comparamos estos resultados con los ofrecidos por los docentes en formación que se muestran previamente (E3 > E6 > E4 > E2 > E5 > E1) se aprecia una diferencia importante en cuanto al valor de importancia que se otorga especialmente al error E6 “Error en la explicación”. Mientras los docentes en formación consideran este error como un error grave, las dos herramientas IA lo ubican en el último lugar de importancia.

Las justificaciones que aportan los chatbots son las siguientes:

> Aunque es relevante para el desarrollo de competencias comunicativas, este tipo de error es el menos grave desde el punto de vista conceptual, ya que suele ser una consecuencia de otros errores previos (conceptuales o procedimentales. (ChatGPT)

> Finalmente, el error en la explicación se refiere a la dificultad para justificar o razonar la respuesta. Si bien es importante que los alumnos puedan explicar sus razonamientos, considero que este error es menos grave que los anteriores, ya que no necesariamente indica una falta de comprensión conceptual. A menudo, se debe a la falta de práctica en la comunicación científica o a la dificultad para expresar ideas de forma clara y concisa. (Gemini)

Los resultados parecen indicar una mayor preocupación por parte de los futuros docentes en relación a que las explicaciones de los procedimientos y respuestas sean apropiadas. Los chatbots consideran estas explicaciones menos importantes y priorizan en la calidad de las respuestas en cuanto a su contenido científico y matemático.

Siguiendo el objetivo de entender mejor los criterios que se aplican en la corrección de las pruebas, se profundiza en el modo de penalizar en la corrección cuando se comete el error considerado más grave. Los docentes en formación consideraron que la penalización de dicho error supondría el 61% del valor de la pregunta (con una desviación estándar del 24%), mientras que para esta misma cuestión ChatGPT consideró que debía ser penalizado con el 100%, mientras que Gemini lo ha penalizado con el 75% de su valor.

Se observa que la penalización en ambos chatbots es mayor que la aplicada por los docentes en formación, pese a que sus criterios de corrección en la prueba fueron más laxos (véase Tabla 2).

Además, se preguntaba si dicha penalización se aplicaría del mismo modo en el caso que se hubiese realizado el mismo error de forma repetida en el mismo examen. Las respuestas de los docentes de Física y Química en formación mostraban que un 61.90% sí que aplicaría la misma penalización y un 15.24% no lo haría. El 22.86% restante respondió que dependería de diferentes situaciones, en la mayoría de casos *"depende del tipo de error"* (20%) o en menor medida *"del contexto del estudiante"* (2%).

En el caso de ChatGPT su respuesta fue que se penalizaría del mismo modo la repetición del mismo error "*Sí, penalizaría cada error de manera independiente, incluso si es del mismo tipo, porque cada pregunta del examen evalúa un contenido o competencia distinta*" de igual modo que Gemini "*…el objetivo de la evaluación no es solo calificar, sino también diagnosticar el aprendizaje del alumno. Si un alumno comete el mismo error en dos preguntas diferentes, esto indica que la comprensión del concepto o la habilidad involucrada es deficiente y requiere una atención específica".*

En este caso se sigue apreciando una mayor rigidez en los criterios de penalización en la corrección de la prueba por parte de los chatbots con relación a los docentes en formación pese a que en su aplicación previa esto no se reflejó en las calificaciones, tal y como ya se ha comentado.

Errores conceptuales (E3)

Viendo que los resultados de la valoración de la importancia de los errores consideran los Errores Conceptuales (E3) como los más graves en el aprendizaje de la Física y Química, tanto por parte de los docentes en formación (véase Tabla 3) como por los chatbots (véase Tabla.4) se consideró de interés profundizar en el significado que se asigna a este tipo de error. Para recabar esta respuesta se solicitó a 45 de los 105 docentes en formación que explicasen el significado de dicho error. La pregunta fue planteada de forma abierta y las respuestas se agrupan según su significado. Las respuestas con mayor representación se muestran en la Tabla 5.

Tabla 5. Significado de *error conceptual* según los docentes en formación

%	Significado E3 (error conceptual)
35.56%	Se muestra un entendimiento erróneo de ideas fundamentales
35.56%	Se muestra una aplicación errónea de un concepto a una situación problema
26.67%	Se muestra una confusión entre conceptos diferentes

Cuando se realiza la misma pregunta a los chatbots las respuestas expresadas son las que se recogen en la Tabla 6.

Tabla 6. Significado de *error conceptual* según los chatbots

Chatbot	Significado E3 (error conceptual)
ChatGPT	El error conceptual en el ámbito de la enseñanza de Física y Química se refiere a una falta de comprensión adecuada de los principios o conceptos fundamentales que sustentan un contenido científico. Este tipo de error ocurre cuando el alumno malinterpreta o no comprende correctamente las ideas clave, lo que lleva a aplicar incorrectamente esas ideas o a construir razonamientos que no son coherentes con los principios científicos.
Gemini	Un error conceptual se refiere a una falta de comprensión de los principios fundamentales, las leyes o los conceptos básicos que sustentan una determinada área de la Física o la Química. No se trata simplemente de equivocarse al aplicar una fórmula o al realizar un cálculo, sino de tener una idea errónea o incompleta sobre la naturaleza de un fenómeno, la relación entre diferentes conceptos o el significado de una ley física o química.

En el estudio de Slisko (2013) previamente citado se define el error conceptual como "El error conceptual ocurre cuando la formulación del problema o los pasos en su resolución revelan que el autor sostiene ideas erróneas sobre las cantidades, leyes o los procesos físicos relacionados con la situación" (pg 95). Tal y cómo muestran ambas tablas (5 y 6) de resultados, existe una coincidencia interesante en la interpretación del significado de error conceptual entre los chatbots y un porcentaje alto de los docentes en formación y con la definición de Slisko. También se observa como los docentes en formación asignan dicho error a la aplicación errónea de un concepto (entiéndase magnitud) en la resolución de un problema y en cambio los chatbots especifican que el error conceptual va más allá de esta mala aplicación exclusivamente. Partiendo de la semejanza en cuanto al significado del tipo de error y viendo la importancia que se le asigna, sería de interés evaluar el conocimiento conceptual vinculado al razonamiento científico mediante otros instrumentos más apropiados que las pruebas tradicionales, existen algunos ya testados con resultados interesantes (Riva-Riquelme, 2024; Ortega, 2019).

Conclusiones

A partir de los resultados presentados podemos comprobar que la comparación de los criterios de corrección de exámenes tradicionales entre los chatbots de uso gratuito y los docentes de Física y Química en formación muestra una semejanza alta con un mayor rigor en la aplicación de los criterios por parte de los chatbots, que además asigna calificaciones más altas. Esta similitud en los criterios y calificaciones no mejora la fiabilidad del instrumento que sigue mostrando dispersiones altas en su aplicación y falta de consenso en su modo de corrección. La persistencia de altas dispersiones en las calificaciones y la falta de consenso en la interpretación de los criterios ponen de manifiesto limitaciones en la capacidad de estos instrumentos para discriminar de forma precisa el nivel competencial del alumnado. Esta visualización de falta de fiabilidad es de interés para la formación de los futuros docentes de Física y Química porque aporta evidencias para promover la importancia que los futuros docentes deben otorgar a la necesidad de redefinir los instrumentos de evaluación que puedan usarse en sus futuras clases. Pese a ello, asumiendo el uso todavía generalizado de este tipo de pruebas examen, parece interesante trasladar su corrección a los chatbots con una instrucción previa adecuada. En este caso se ha comprobado que no se pierde fiabilidad en los criterios aplicados, incluso se aprecia un leve aumento del rigor aplicado al uso del instrumento por parte de las herramientas IA. Este hallazgo puede ser de interés para los docentes de Física y Química por el ahorro de tiempo que puede suponer en su función profesional, así como por la posibilidad que también ofrecen los chatbots de trasladar un feedback personalizado para cada estudiante. En este sentido las herramientas IA deben formar parte del conjunto de recursos

TIC de uso habitual para la docencia con fines de corrección y/o calificación y, por tanto, deberían incorporarse módulos específicos para conocer y aplicar su uso a las diferentes disciplinas didácticas, especialmente en ciencias experimentales.

La discrepancia de criterios encontrada en cuanto a la importancia que se otorga a la buena explicación de las respuestas por parte del alumnado, con los docentes en formación exigiendo mejores explicaciones que los chatbots, no hace más que constatar la necesidad de diseñar instrumentos más apropiados para evaluar la comunicación y el razonamiento científico. Es relevante destacar la necesidad por parte de la didáctica de la Física y Química de seguir mostrando la eficiencia del uso de otros instrumentos diferentes a los exámenes tradicionales para evaluar el desarrollo de las competencias específicas de nuestra disciplina. Este estudio ha querido poner en evidencia la baja fiabilidad de dichos exámenes para potenciar el uso de instrumentos de evaluación diversos. Pero sabiendo que siguen siendo los exámenes tradicionales los instrumentos más usados en las aulas de secundaria de Física y Química, el uso efectivo de la tecnología puede mejorar el rigor en su aplicación y de este modo se podría dedicar el tiempo ahorrado a la creación de nuevos instrumentos por parte del cuerpo docente o bien para mejorar el diseño de los propios exámenes con el fin de adecuarse a la certificación competencial que estos requieren.

Sería de interés ampliar esta investigación con un análisis de la retroalimentación que ofrecen estas herramientas tras la corrección de un examen tradicional, así como realizar un estudio en paralelo de los resultados que se obtienen tras la corrección de pruebas tipo examen por parte del profesorado en activo y por parte de los chatbots.

Declaración de autoría

Conceptualización: Enric Ortega; Metodología: Enric Ortega; Análisis formal: Enric Ortega; Investigación: Enric Ortega; Redacción, revisión y edición: Enric Ortega.

Declaración de uso responsable de herramientas de Inteligencia Artificial (IA)

Como forma de abordaje del objetivo de investigación se ha hecho uso de ChatGPT 3.0 y Gemini 1.5 para comparar los criterios de corrección de exámenes usados por la IA y profesores en formación.

Referencias

Abdulkadir, M., Osuwa, Y. y Ibrahim, A. (2024). Impact of Performance Assessment on Ssii Students' Interest and Academic Achievement in Physics, Chemistry and Biology in Katsina State, Nigeria. *International Journal of Research and Innovation in Applied Science, 9*(7), 647-667. https://doi.org/10.51584/ijrias.2024.907054

Alamr, S., León-Urrutia, M. y Carr, L. (2023). E-assessment in Computer Science Higher Education. *Proceedings of the 15th International Conference on Education Technology and Computers*. https://doi.org/10.1145/3629296.3629357

Alonso Sánchez, M., Daniel Gil Pérez, y Joaquín Martínez Torregrosa. (1996). Evaluar no es calificar. La evaluación y la calificación en una enseñanza constructiva de las ciencias. *Revista Investigación en la Escuela, 30*, 15-26. https://doi.org/10.12795/IE.1996.i30.02

Arancibia-Herrera, M., Novoa-Cáceres, V., y Casanova-Seguel, R. (2019). Concepciones sobre evaluación de docentes de Ciencias Naturales, Matemática, Lenguaje e Historia. *Revista Educación, 43*(1), 1-15. https://doi.org/10.15517/revedu.v43i1.30497

Ardura, D., Zamora, A. (2014) ¿En qué medida utilizan los estudiantes de Física de Bachillerato sus propios errores para aprender? Una experiencia de autorregulación en el aula de secundaria. *Enseñanza de las Ciencias, 32* (2), 253-268 https://doi.org/10.5565/rev/ensciencias.1067

Babinčáková, M., Ganajová, M., Sotáková, I. y Jurková, V. (2019). The implementation of formative assessment into chemistry education at secondary school. *Journal of Baltic Science Education, 19*(1), 36- 49. https://doi.org/10.33225/balticste/2019.09

Bransford, J. D., Brown, A. L. y Cocking, R. R. (2000). *How people learn: Brain, mind, experience, and school* (Vol. 11). National Academy Press.

Buck, G. A., Trauth-Nare, A., y Kaftan, J. (2010). Making formative assessment discernable topre-service teachers of science. *Journal of Research in Science Teaching, 47*(4), 402-421. https://doi.org/10.1002/tea.20344

Buteler, L., Coleoni, E., y Gangoso, Z. (2008). ¿Qué información útil arrojan los errores de los estudiantes cuando resuelven problemas de física?: Un aporte desde la perspectiva de recursos cognitivos. *Revista Electrónica de Enseñanza de las Ciencias, 7*(2), 349-365.

Conselleria de Educación, Cultura y Deporte (2022). *Decreto 107/2022, por el que se establece la ordenación y el currículo de Educación Secundaria Obligatoria*. DOGV (9403).

Crooks, T. J. (1988). The impact of classroom evaluation practices on students. *Review of educational research, 58*(4), 438-481.

Furman, M., Poenitz, M. V. y Podestá, M. E. (2013). ¿Qué saberes evalúan los formadores del profesorado de ciencias experimentales?: Una mirada sobre las preguntas de evaluación. En P. Membiela Iglesia, N. Casado Bailón y M. I. Cebreiros Iglesias (Coords.), *Experiencias de investigación e innovación en la enseñanza de las ciencias* (pp. 377–381). Educación Editora.

Furze, L. y Roe, J. (2024). The AI Assessment Scale (AIAS) in Australian K–12 Education. *Current office holders of the Teachers' Guild*, 17.

García-Perales, N., Hernández-Rincón, M.L. y Suárez-Lantarón, B. (2025). Docentes y tecnología: ¿cómo enfrenta el futuro profesorado el uso de la Inteligencia Artificial? *Revista Electrónica Interuniversitaria de Formación del Profesorado, 28*(1), 155-168. https://doi.org/10.6018/reifop.638431

Gil, D., Carrascosa, J., Furió, C., y Martinez-Torregrosa, J. (1991) *La enseñanza de las ciencias en secundaria (planteamientos didácticos generales y ejemplos de aplicación en las ciencias físico-químicas)*. Horsori.

González, A., Habechian, F., Bobadilla, E., Cancino, M., González, H., Crisóstomo, S., y Escobar, M. (2021). Rediseño del Currículo: ¿Garantizan la coherencia del método y la evaluación mejores oportunidades de aprendizaje en kinesiología? *REXE: Revista de Estudios y Experiencias en Educación, 20* (44), 428-444. https://doi.org/10.21703/0718-5162.v20.n43.2021.024

Liao, X., Zhang, X., Wang, Z. y Luo, H. (2024). Design and implementation of an AI-enabled visual report tool as formative assessment to promote learning achievement and self-regulated learning: An experimental study. *British Journal of Educational Technology, 55*, 1253-1276. https://doi.org/10.1111/bjet.13424

López-Lozano, L. y Solís-Ramírez, E. (2016). ¿Para qué, cómo y qué evalúa en ciencia elprofesorado de Primaria en formación? *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias, 13* (1), 102-120. http://hdl.handle.net/10498/18017

López Pastor, V. M., García-Peñuela de Miguel, A., Pérez Brunicardi, D., López Pastor, E. M. y Monjas Aguado, R. (2004). Las historias de vida en la formación inicial del profesorado de Educación Física. *Revista Internacional de Medicina y Ciencias de la Actividad Física y del Deporte, 4*(13), 45-57. https://uvadoc.uva.es/handle/10324/54003

Lupión Cobos, T. y Caracuel González, M. (2021). Competencias profesionales de futuros docentes de educación secundaria. Estudio de caso de la evaluación formativa promovida mediante e-rúbricas en la especialidad de física y química. *Profesorado. Revista de Currículum y Formación del Profesorado, 25*(1), 197–221. https://doi.org/10.30827/profesorado.v25i1.8374

Mazzitelli, C. A., Guirado, A. M. y Olivera, A. D. C. (2013). Las evaluaciones en física y en química: ¿Qué aprendizaje se favorece desde la enseñanza en la educación secundaria? *Investigações em Ensino de Ciências, 18*(1), 143–159. https://ienci.if.ufrgs.br/index.php/ienci/article/view/164

Mellado-Moreno, P. C., Sánchez-Antolín, P. y Blanco-García, M. (2021). Tendencias de la evaluación formativa y sumativa del alumnado en Web of Sciences. *Alteridad. Revista de educación, 16*(2), 170-183. https://doi.org/10.17163/alt.v16n2.2021.01

Ministerio de Educación y Formación Profesional (2022). Real Decreto 217/2022, por el que se establece la ordenación y las enseñanzas mínimas de la EducaciónSecundaria Obligatoria. *Boletín Oficial del Estado, 76*. https://www.boe.es/eli/es/rd/2022/03/29/217

Morán, E. G., Morán, P. C., Acosta, P. D., Morán, T. M. y Sánchez, C. E. (2024). El uso de plataformas digitales gamificadas para la evaluación formativa en educación. *Ciencia Latina Revista Científica Multidisciplinar, 8*(6), 3428-3438. https://doi.org/10.37811/cl_rcm.v8i6.15100

Ortega Torres, E., (2019). Un congreso científico para el alumnado de 3. º de ESO. *Alambique: Didáctica de las ciencias experimentales, 98*, 16-21.

Paiva, J. C., Leal, J. P. y Figueira, Á. (2022). Automated assessment in computer science education: A state-of-the-art review. *ACM Transactions on Computing Education (TOCE), 22*(3), 1-40. https://doi.org/10.1145/3513140

Pochulu, M. (2009). Análisis y categorización de errores en el aprendizaje de la matemática en alumnos que ingresan a la universidad. *Colección Digital Eudoxus, 8*, 1–15.

Pozuelo Muñoz, J. y Cascarosa Salillas, E. (2023). Diseño y uso de herramientas para el análisis del desarrollo de la Competencia Científica en el contexto de una Secuencia de Enseñanza Aprendizaje en Educación Secundaria. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias 21*(2), 2301. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2024.v21.i2.2301

Quílez, J. (2024). Análisis epistemológico del currículum LOMLOE de Química de la ESO de la Comunitat Valenciana. *Revista Eureka sobre Enseñanza y Divulgación de las Ciencias 21*(2), 3304. https://doi.org/10.25267/Rev_Eureka_ensen_divulg_cienc.2024.v21.i2.3304

Real Decreto 217/2022, de 29 de marzo, por el que se establece la ordenación y las enseñanzas mínimas de la Educación Secundaria Obligatoria. Boletín Oficial del Estado, 76, de 30 de marzo de 2022, páginas 41571 a 41789. https://www.boe.es/eli/es/rd/2022/03/29/217

Remesal, A. (2011). Primary and secondary teachers' conceptions of assessment: A qualitativestudy. *Teaching and Teacher Education, 27*(2), 472-482.

Rico, L (1995). Errores y dificultades en el aprendizaje de las matemáticas. En Kilpatrick, J., Rico, L., Gómez, P. (Eds.), *Educación Matemática. Errores y dificultades de los estudiantes. Resolución de problemas. Evaluación. Historia* (pp. 69-108). Bogotá: una empresa docente.

Riva-Riquelme, R., Nunez-Oviedo, M. C. y Flores-Morales, P. (2024). Identificación de ciclos de generación, evaluación y modificación en estequiometría. *Enseñanza de las Ciencias, 42*(3), 11-32 https://doi.org/10.5565/rev/ensciencias.5965

Ruiz Martín, H. (2020). *¿Cómo aprendemos? Una aproximación científica al aprendizaje y la enseñanza* (Vol. 1). Graó

Sanmartí, N. (2010). *Avaluar per aprendre: l'avaluació per millorar els aprenentatges de l'alumnat en el marc del currículum per competències*. Generalitat de Catalunya. Departament d'educació. Direcció General de l'educació Bàsica i Batxillerat

Sanmartí, N., y -Xarxa de Competències Bàsiques, (2020). *Avaluar és aprendre: l'avaluació per millorar els aprenentatges de l'alumnat en el marc del currículum per competències*.

Slisko, J. (2013). Errores comunes en problemas numéricos de la física escolar. *Didáctica de las Ciencias Experimentales y Sociales, 14*, 87-98.

Tufiño, M. y Cayambe, J. (2023). Evaluación de los aprendizajes mediante plataformas didácticas virtuales. *Ciencia Latina Revista Científica Multidisciplinar, 7*(3), 1709-1735. https://doi.org/10.37811/cl_rcm.v7i3.6306

Wang, J.-R., Kao, H.-L. y Lin, S.-W. (2010). Pre-service teachers' initial conceptions aboutassessment of science learning: The coherence with their views of learning science.*Teaching and Teacher Education, 26*(3), 522-529 https://doi.org/10.1016/j.tate.2009.06.014

William, D. (2011). Formative assessment: Definitions and relationships. *Studies in Educational Evaluation, 37*(1), 3-14 University of London.

Yim, I. H. Y. y Su, J. (2025). Artificial intelligence (AI) learning tools in K–12 education: A scoping review. *Journal of Computers in Education, 12*, 93–131. https://doi.org/10.1007/s40692-023-00304-9

Yim, I. H. Y. y Wegerif, R. (2024). Teachers' perceptions, attitudes, and acceptance of artificial intelligence (AI) educational learning tools: An exploratory study on AI literacy for young students. *Future in Educational Research, 2*(4), 318–345. https://doi.org/10.1002/fer3.65