



Ciencia Latina Revista Científica Multidisciplinar, Ciudad de México, México.
ISSN 2707-2207 / ISSN 2707-2215 (en línea), septiembre-octubre 2026,
Volumen 10, Número 5.

https://doi.org/10.37811/cl_rcm.v10i5

DE LA EVALUACIÓN DEL PRODUCTO A LA EVALUACIÓN DEL PROCESO: REDISEÑO DE LA EVALUACIÓN UNIVERSITARIA FRENTE A LA INTELIGENCIA ARTIFICIAL GENERATIVA

**FROM PRODUCT EVALUATION TO PROCESS
EVALUATION: REDESIGNING UNIVERSITY ASSESSMENT
IN THE FACE OF GENERATIVE ARTIFICIAL INTELLIGENCE**

Leandro Guerschberg

Universidad Nacional de José C. Paz, Argentina

DOI: https://doi.org/10.37811/cl_rcm.v10i5.26342

De la Evaluación del Producto a la Evaluación del Proceso: Rediseño de la Evaluación Universitaria frente a la Inteligencia Artificial Generativa

Leandro Guerschberg¹<https://orcid.org/0009-0005-9286-6358>leandro.guerschberg@docentes.unpaz.edu.arDCSyD, Universidad Nacional de José C. Paz
Argentina

RESUMEN

Este artículo postula que la respuesta sostenible de la universidad frente a la inteligencia artificial generativa no reside en la detección del fraude ni en la prohibición de las herramientas, sino en la revisión del objeto mismo que la evaluación toma como evidencia de aprendizaje. Se argumenta que buena parte de la tradición evaluativa de la educación superior descansa en el supuesto de que el producto entregado por el estudiante es un rastro fiable del trabajo cognitivo realizado, supuesto que la inteligencia artificial generativa debilita de manera especialmente severa en aquellas evaluaciones cuyo proceso de elaboración permanece opaco. Se distinguen analíticamente tres ejes presentes en la literatura sobre evaluación educativa (la finalidad, formativa o sumativa; el tipo de evidencia, centrada en el producto o en el proceso; y el grado de autenticidad de la tarea), y se muestra que la inteligencia artificial generativa obliga sobre todo a reconsiderar el segundo. A partir de una revisión integradora de un corpus acotado de literatura académica en español e inglés, publicada mayoritariamente entre 2023 y 2026, se examinan las respuestas institucionales documentadas y se discute con detalle el marco de cuatro pilares propuesto por Nikolić y Basta Nikolić (2026), del que se toma distancia crítica en tres aspectos. A partir de esa distancia se propone un marco alternativo de rediseño evaluativo organizado en dos niveles (el diseño de la tarea y la interacción evaluativa) y cuatro principios (trazabilidad del razonamiento, iteración documentada, comprensión verificable y metacognición explícita). El artículo cierra con una discusión de las tensiones que ese rediseño introduce, con un reconocimiento explícito de las limitaciones del corpus revisado, y con una agenda de investigación empírica pendiente.

Palabras clave: evaluación educativa; inteligencia artificial generativa; evaluación auténtica; evaluación formativa; integridad académica; educación superior.

¹ Autor principal.

Correspondencia: leandro.guerschberg@docentes.unpaz.edu.ar

From Product Evaluation to Process Evaluation: Redesigning University Assessment in the Face of Generative Artificial Intelligence

ABSTRACT

This article contends that the university's sustainable response to generative artificial intelligence does not lie in detecting fraud or banning these tools, but in revisiting the very object that assessment takes as evidence of learning. It is argued that much of the assessment tradition in higher education rests on the assumption that the product submitted by the student is a reliable trace of the cognitive work performed, an assumption that generative artificial intelligence weakens with particular severity in those assessments whose production process remains opaque. Three axes present in the literature on educational assessment are analytically distinguished (purpose, formative or summative; type of evidence, centered on the product or on the process; and the degree of authenticity of the task), and it is shown that generative artificial intelligence compels a reconsideration primarily of the second. Drawing on an integrative review of a bounded corpus of academic literature in Spanish and English, published mostly between 2023 and 2026, the article examines documented institutional responses and discusses in detail the four-pillar framework proposed by Nikolić and Basta Nikolić (2026), from which it takes critical distance on three points. Building on that distance, it proposes an alternative framework for assessment redesign organized into two levels (task design and evaluative interaction) and four principles (traceability of reasoning, documented iteration, verifiable understanding, and explicit metacognition). The article closes with a discussion of the tensions this redesign introduces, an explicit acknowledgment of the limitations of the corpus reviewed, and a pending empirical research agenda.

Keywords: educational assessment; generative artificial intelligence; authentic assessment; formative assessment; academic integrity; higher education.

*Artículo recibido 20 agosto 2026
Aceptado para publicación: 20 setiembre 2026*



INTRODUCCIÓN

Cuando un docente universitario recibe un trabajo escrito, da por sentado algo que rara vez formula de manera explícita: que el texto entregado es un rastro fiable del razonamiento de quien lo firma. Esa asunción no es trivial ni reciente. Sostiene, desde la consolidación del examen escrito y el trabajo entregable como dispositivos evaluativos estándar de la universidad moderna, buena parte de su edificio evaluativo: el examen escrito, el ensayo, el informe de laboratorio, la monografía de fin de curso. Todos estos instrumentos comparten una misma lógica inferencial, consistente en tomar un artefacto terminado como evidencia suficiente, y habitualmente única, del trabajo cognitivo que lo produjo. La calificación de ese artefacto se traduce, mediante un salto inferencial que la práctica institucional rara vez cuestiona, en una calificación del aprendizaje de quien lo entregó.

Ese salto inferencial nunca fue perfecto. La literatura sobre integridad académica documentó, mucho antes de la aparición de los sistemas de inteligencia artificial generativa, un repertorio amplio de prácticas capaces de desacoplar el producto entregado del trabajo efectivamente realizado por quien lo firma: la compra de trabajos a terceros, el uso de bancos de exámenes previos, la asistencia no declarada de familiares o tutores particulares, el plagio parcial o total de fuentes no citadas. La universidad convivió durante décadas con ese desacople porque su costo, económico, de tiempo y de riesgo de detección, mantenía la práctica dentro de márgenes tolerables, aunque nunca inexistentes. Quien deseaba desacoplar producto y proceso debía pagar por ello, buscar activamente a un tercero dispuesto a colaborar, o arriesgarse a una sanción cuyo umbral de detección, si bien imperfecto, imponía cierta fricción.

La disponibilidad masiva y gratuita de sistemas de inteligencia artificial generativa capaces de producir, en segundos y a partir de instrucciones mínimas, textos, códigos y análisis de calidad aceptable para buena parte de las tareas evaluativas universitarias no introduce una novedad cualitativa en el tipo de amenaza que representa para la validez de la evaluación. Introduce, en cambio, una transformación radical en su escala y en su costo. Lo que antes exigía una búsqueda activa, un desembolso económico o una red de contactos hoy está disponible en una pestaña del navegador, sin costo relevante y sin necesidad de involucrar a un tercero identificable. Esa caída abrupta del costo de desacoplar producto y proceso es, se argumentará en este artículo, la verdadera novedad que la inteligencia artificial



generativa introduce en la evaluación universitaria, y es también la razón por la cual las respuestas basadas en la detección o en la disuasión, adecuadas quizá para un régimen de costos altos, resultan insuficientes en un régimen de costo casi nulo.

Frente a esa transformación, la reacción institucional más extendida ha consistido en intentar restaurar, por vía tecnológica o normativa, el supuesto inferencial original: detectar cuándo un producto fue generado por inteligencia artificial, para poder seguir tratando el producto no asistido por estas herramientas como evidencia suficiente de aprendizaje. Este artículo sostiene que esa estrategia, además de enfrentar límites técnicos ya documentados por la literatura reciente, parte de un diagnóstico equivocado. El problema no es identificar qué productos fueron elaborados con asistencia de inteligencia artificial generativa y cuáles no, sino reconocer que el producto, como categoría evaluativa, dejó de ofrecer por sí solo una base suficiente para la inferencia sobre el aprendizaje en aquellas evaluaciones cuyas condiciones de producción no pueden reconstruirse, con independencia de si la inteligencia artificial intervino o no en su elaboración. Un examen presencial y supervisado, por ejemplo, sigue siendo un producto en sentido estricto, y sin embargo puede ofrecer evidencia razonablemente sólida del desempeño de quien lo resuelve, precisamente porque las condiciones de su producción están controladas y son, en ese sentido, trazables. Lo que la inteligencia artificial generativa debilita de manera radical no es cualquier producto, sino específicamente aquel elaborado fuera de condiciones controladas, cuya genealogía no puede reconstruirse una vez entregado. Un ensayo escrito sin ninguna asistencia tecnológica, pero copiado casi en su totalidad de un trabajo de un compañero de una cohorte anterior, plantea, en esas condiciones no controladas, exactamente el mismo problema de validez que un ensayo generado en segundos por un modelo de lenguaje: en ambos casos, el producto entregado no permite inferir de manera confiable el aprendizaje de quien lo firma.

La tesis que se defiende en este artículo es que la respuesta sostenible frente a esta situación no consiste en perfeccionar la detección del origen del producto, sino en desplazar el centro de gravedad de la evaluación, en aquellas instancias cuyas condiciones de producción no están controladas, del producto final hacia el proceso mediante el cual ese producto se construye. Ese desplazamiento no es una respuesta defensiva improvisada frente a una amenaza tecnológica puntual. Es, como se mostrará en la sección de resultados, una corrección de largo plazo hacia formas de evaluación que la tradición de la



evaluación formativa (Sadler, 1989; Black y Wiliam, 1998) venía reclamando desde antes de que la inteligencia artificial generativa volviera la cuestión urgente e ineludible. La inteligencia artificial generativa, en este sentido, no inaugura el problema, sino que lo agrava hasta un punto en el que la universidad ya no puede seguir postergando su resolución. Esta distinción entre la ocasión de un cambio y su fundamento último no es meramente académica: de ella depende, según se argumentará en la discusión, que el rediseño evaluativo propuesto pueda sostenerse en el tiempo con independencia de cómo evolucionen las capacidades técnicas de los sistemas de inteligencia artificial generativa en los próximos años.

El artículo se organiza de acuerdo con la siguiente secuencia. Tras precisar la aproximación metodológica empleada en la sección de método, la sección de resultados distingue tres ejes analíticos presentes en la literatura sobre evaluación educativa (finalidad, evidencia y autenticidad) y reconstruye, dentro de ellos, la distinción entre evaluación del producto y evaluación del proceso; examina el desafío específico que la inteligencia artificial generativa plantea a esa distinción con apoyo en la literatura empírica reciente sobre integridad académica; revisa las respuestas institucionales y pedagógicas documentadas hasta el momento; y presenta con detalle uno de los marcos conceptuales más completos identificados en la revisión sobre evaluación resiliente a la inteligencia artificial. La sección de discusión y conclusiones toma distancia crítica de ese marco en tres aspectos, apoyándose explícitamente en la distinción de ejes introducida en los resultados; propone, a partir de esa distancia, un marco alternativo de rediseño evaluativo organizado en dos niveles y cuatro principios; examina las tensiones que ese rediseño introduce; reconoce las limitaciones del corpus revisado; y cierra con una agenda de investigación empírica que la literatura disponible deja todavía abierta. El objetivo de este artículo es, en síntesis, analizar las limitaciones de las respuestas evaluativas centradas en la detección de la inteligencia artificial generativa y proponer, a partir de esa crítica, un marco conceptual de rediseño evaluativo fundamentado en evidencias del proceso de aprendizaje.

MÉTODO

Este trabajo adopta la forma de un ensayo teórico-argumentativo sustentado en una revisión integradora de un corpus acotado de literatura, siguiendo el procedimiento propuesto por Whittemore y Knafl (2005): identificación del problema, búsqueda de literatura, valoración cualitativa de la calidad y



pertinencia de las fuentes, análisis interpretativo y presentación de resultados. Se optó por este método, y no por una revisión sistemática de alcance PRISMA, porque el objetivo del artículo no es cartografiar de manera exhaustiva la producción empírica disponible, sino integrar un conjunto acotado de esa producción con literatura conceptual para sostener un argumento propositivo; la sección 4.4 explicita lo que esta elección implica en términos de alcance y las limitaciones que de ella se derivan.

La búsqueda se realizó entre julio y septiembre de 2026 en bases de acceso abierto (Redalyc, SciELO y Dialnet para la literatura en español) y en buscadores académicos generales de acceso abierto para la literatura en inglés. Las ecuaciones de búsqueda combinaron, en español, términos como "inteligencia artificial generativa", "evaluación universitaria", "evaluación auténtica", "integridad académica" y "educación superior", y en inglés términos como "generative AI", "assessment redesign", "academic integrity", "authentic assessment" y "higher education", ajustados de manera iterativa a medida que la revisión identificaba subtemas relevantes, entre ellos la evaluación oral como dispositivo de proceso y los marcos de evaluación resiliente a la inteligencia artificial. Se aplicaron los siguientes criterios de inclusión: documentos publicados entre 2023 y 2026, con dos excepciones fundacionales anteriores a ese período incorporadas por su valor teórico para la distinción entre evaluación del producto y evaluación del proceso (Sadler, 1989; Black y Wiliam, 1998), además de una referencia metodológica de 2005 (Whittemore y Knafl); documentos que abordaran de manera directa la relación entre evaluación educativa e inteligencia artificial generativa, o que aportaran un marco teórico directamente aplicable a esa relación; documentos con autoría, año y publicación verificables en su fuente primaria; y disponibilidad en español o en inglés. Se excluyeron los documentos exclusivamente técnicos sobre detección algorítmica de texto generado por inteligencia artificial sin relación con el diseño evaluativo, los materiales de divulgación sin proceso editorial identificable, y los documentos cuyos metadatos no pudieron verificarse de manera directa en la fuente primaria; este último criterio llevó a descartar dos fuentes inicialmente preseleccionadas cuyo acceso a la publicación original no pudo confirmarse, y a sustituir una tercera fuente, inicialmente incluida, tras detectarse una inconsistencia editorial en la revista que la publicó (un resumen y palabras clave en inglés correspondientes, según pudo verificarse directamente en la fuente, a un artículo no relacionado sobre gestión municipal) por un estudio equivalente publicado en una revista de mayor consolidación editorial.



De los registros identificados en las búsquedas realizadas (del orden de ochenta a noventa, considerando el conjunto de las distintas ecuaciones utilizadas), se examinó el texto completo de aproximadamente veinte, verificando en cada caso autoría, año, publicación y, cuando estuvo disponible, DOI, directamente en la fuente primaria antes de su incorporación a la síntesis. De ese subconjunto se retuvieron doce referencias: nueve documentos contemporáneos sobre evaluación universitaria e inteligencia artificial generativa, dos referencias teóricas fundacionales anteriores al período de búsqueda y una referencia metodológica. La valoración de la calidad de las fuentes fue de carácter cualitativo y atendió al tipo de diseño, el alcance del estudio, su carácter empírico, conceptual o de revisión, la pertinencia respecto del problema analizado y las limitaciones metodológicas declaradas o identificables en cada trabajo. No se aplicó una escala estandarizada de evaluación de calidad, dado el carácter heterogéneo del corpus; estos elementos se utilizaron para ponderar interpretativamente los aportes de cada fuente y se sintetizan en la Tabla 1. La síntesis se realizó de manera cualitativa y temática, agrupando los documentos retenidos alrededor de tres preguntas organizadoras: qué evidencia aporta cada fuente sobre la pérdida de validez de la evaluación centrada en el producto, qué respuestas institucionales o pedagógicas documenta, y qué principios de rediseño evaluativo propone o permite inferir, en un procedimiento consistente con la etapa de análisis e interpretación de Whitemore y Knafl (2005). La Tabla 1 sintetiza los nueve documentos contemporáneos que constituyen el núcleo empírico y conceptual del corpus.

Tabla 1. Síntesis del corpus contemporáneo revisado (2023-2026)

Autor/año	Tipo de estudio	Población/contexto	Aporte al problema evaluativo	Limitaciones
Gallent-Torres et al. (2023)	Análisis conceptual	Educación superior, perspectiva multiactor	Identifica desafíos éticos que involucran a estudiantes, docentes e instituciones	No aporta datos empíricos originales
Sánchez Mendiola (2023)	Editorial	Educación médica	Propuesta temprana de diversificación evaluativa	Género editorial, no es investigación empírica
Kofinas et al. (2025)	Estudio empírico	Educación superior, evaluaciones auténticas	Muestra la vulnerabilidad de las evaluaciones auténticas ante la IA generativa	Contexto institucional específico

Santiago Ruiz y Páramo Chávez (2025)	Estudio empírico cuantitativo	Muestra de estudiantes universitarios	Documenta prevalencia de uso deshonesto de IA y su correlación con habilidades de escritura	Diseño transversal, muestra acotada
Sánchez, et al. (2025)	Revisión sistemática (23 artículos)	Literatura hispanohablante sobre educación superior	Diagnóstico de disrupción evaluativa y llamado a reconceptualización	Reconoce escasez de evidencia empírica concluyente
Bañuelos Márquez y Romero Martínez (2024)	Estudio de caso exploratorio	Asignatura universitaria de psicología	Evidencia de discrepancias entre retroalimentación de IA y evaluación docente	Alcance reducido (cuatro trabajos evaluados)
Ncube et al. (2026)	Revisión sistemática	Educación superior, integridad académica	Identifica la evaluación centrada en el proceso como estrategia con mayor sustento	Depende de la calidad de los estudios primarios incluidos
Nikolić y Basta Nikolić (2026)	Análisis conceptual	Educación superior general	Marco de cuatro pilares para la evaluación resiliente a la IA	Propuesta teórica no sometida a prueba empírica
Yermaganbetova et al. (2026)	Estudio cuasiexperimental (N=120)	Curso universitario de modelado tridimensional	Evidencia de mejora de desempeño con plataforma adaptativa de IA	Mide desempeño en tareas prácticas, no trazabilidad del proceso cognitivo

Fuente: Elaboración propia

Este corpus es acotado y heterogéneo en su diseño metodológico, aspecto que se retoma en el apartado de limitaciones (4.4).

RESULTADOS

De la evaluación del producto a la evaluación del proceso: antecedentes conceptuales

La distinción entre evaluar el producto y evaluar el proceso no nace del debate sobre inteligencia artificial generativa; lo precede en varias décadas y pertenece al núcleo de la literatura sobre evaluación formativa. Conviene, sin embargo, evitar una identificación demasiado directa entre esa distinción y la que separa la evaluación formativa de la evaluación sumativa, porque ambas responden a preguntas distintas. La literatura sobre evaluación educativa permite distinguir, al menos, tres ejes analíticos independientes entre sí. El primero es el eje de la finalidad, que va de lo formativo a lo sumativo y describe la función que cumple la evaluación: intervenir sobre el aprendizaje mientras este todavía está en curso, o certificarlo una vez finalizado. El segundo es el eje de la evidencia, que va del producto al proceso y describe qué tipo de rastro se toma como base para la inferencia evaluativa: un artefacto



terminado, o las etapas intermedias que condujeron a él. El tercero es el eje de la autenticidad, que va de la tarea descontextualizada a la tarea auténtica y describe el grado de semejanza entre la actividad evaluativa y una situación del ejercicio profesional o disciplinar real.

Estos tres ejes son lógicamente independientes, aunque en la práctica evaluativa suelen aparecer combinados. Una evaluación sumativa puede incorporar evidencia de proceso sin dejar de ser sumativa: una defensa oral final, un portafolio con versiones sucesivas o una bitácora evaluada al término del curso son, en su función, sumativos, y sin embargo toman como evidencia el proceso y no solamente el producto. De manera simétrica, una evaluación formativa puede intervenir sobre un producto sin volverse, por eso, una evaluación de proceso en el sentido aquí definido: la retroalimentación que un docente ofrece sobre un borrador es formativa en su función, pero su objeto sigue siendo, en ese momento puntual, un producto terminado (el borrador) y no el proceso que condujo a él. Explicitar esta independencia importa porque permite formular con mayor precisión el argumento central de este artículo: lo que la inteligencia artificial generativa obliga a reconsiderar de manera prioritaria no es el eje de la finalidad, ya que no toda evaluación necesita volverse formativa, ni, necesariamente, el eje de la autenticidad, aunque una tarea auténtica pueda ayudar, como se discute más adelante, sin bastar por sí sola, sino específicamente el eje de la evidencia: la pregunta de si la evaluación, sumativa o formativa, auténtica o no, tiene o no tiene acceso al proceso que condujo al producto entregado.

Dentro de ese eje de la evidencia, Sadler (1989) formuló tempranamente la idea de que la función de la evaluación no se agota en la calificación de un resultado, sino que incluye la provisión de información capaz de modificar el proceso de aprendizaje mientras este todavía está en curso, y Black y Wiliam (1998), en la revisión que consolidó el campo de la evaluación formativa como área de investigación autónoma, mostraron que la evidencia disponible hasta entonces indicaba efectos consistentes de las prácticas de evaluación formativa sobre el aprendizaje, particularmente cuando la retroalimentación se ofrecía en momentos en que el estudiante todavía podía actuar sobre ella. Conviene precisar que ambos trabajos se sitúan, en rigor, sobre el eje de la finalidad, dado que su interés central es la función formativa de la evaluación, pero el cumplimiento de esa función presupone el acceso a evidencia del proceso: es difícil intervenir sobre un aprendizaje en curso si la única evidencia disponible es un producto ya



terminado. Esa relación indirecta entre ambos ejes, y no una identidad entre evaluación formativa y evaluación de proceso, es lo que este artículo retiene de la tradición formativa.

Conviene, por último, precisar los términos del eje de la evidencia antes de avanzar. Por evaluación del producto se entiende aquí la práctica evaluativa que toma como objeto de calificación un artefacto terminado (un examen resuelto, un ensayo entregado, un programa que compila), sin acceso directo a las decisiones intermedias que llevaron a su forma final. Por evaluación del proceso se entiende, en cambio, la práctica que incorpora como objeto de calificación, o al menos como evidencia relevante para la calificación, las etapas intermedias de producción: los borradores, las correcciones, las justificaciones de las decisiones tomadas, la interacción oral en torno al trabajo. La evaluación del proceso no excluye necesariamente al producto final; lo relativiza como única fuente de evidencia, y es compatible, como se mostró más arriba, tanto con una función sumativa como con una función formativa.

Esta distinción había permanecido, hasta hace pocos años, más discutida en la literatura especializada que aplicada de manera sistemática en la práctica evaluativa universitaria, en parte porque la evaluación del proceso exige un despliegue de tiempo docente y de instrumentos de seguimiento que la evaluación del producto no requiere. La inteligencia artificial generativa no introduce el eje de la evidencia ni la distinción que de él se deriva; lo que hace es volver insostenible, en un número creciente de disciplinas y de tipos de tarea, la opción de prescindir de esa distinción.

El desafío específico de la inteligencia artificial generativa a la validez evaluativa

Reducir el problema que plantea la inteligencia artificial generativa a una cuestión de integridad académica en sentido estricto (esto es, al riesgo de que un estudiante presente como propio un trabajo que no elaboró) capta solo una parte del fenómeno y deja fuera una cuestión central para la validez evaluativa. Gallent-Torres, et al. (2023) documentan que el desafío que estas herramientas plantean a la integridad académica en la educación superior involucra simultáneamente a estudiantes, docentes e instituciones, y que compromete cuestiones más amplias que la autoría del texto entregado: la fiabilidad de la información producida, la transparencia respecto de las fuentes utilizadas y la privacidad de los datos procesados. Santiago Ruiz y Páramo Chávez (2025), en un estudio realizado con una muestra de población universitaria, encontraron que aproximadamente el 54% de los estudiantes de esa muestra

reconoció haber utilizado inteligencia artificial de manera deshonesta en al menos una ocasión, y que, dentro del mismo grupo estudiado, quienes presentaban menores habilidades de escritura mostraban una correlación negativa moderada con conductas deshonestas relacionadas con el uso de estas herramientas. Este hallazgo, circunscripto a la población que los autores estudiaron, sugiere que la dependencia indebida de la inteligencia artificial generativa podría no distribuirse de manera uniforme entre el estudiantado, sino concentrarse en quienes llegan a la universidad con menor preparación escritural, lo cual introduce una dimensión de equidad que la discusión centrada exclusivamente en el fraude suele pasar por alto, aunque su generalización a poblaciones más amplias requeriría estudios adicionales.

Lo que estos datos ilustran, leídos en conjunto con la literatura sobre diseño evaluativo, es que el problema de fondo no es la deshonestidad individual, sino la pérdida de validez del producto como evidencia. Ncube, et al. (2026), en una revisión sistemática sobre los desafíos que la integración de inteligencia artificial plantea a la evaluación estudiantil, identifican, junto al riesgo de plagio asistido por inteligencia artificial, una preocupación menos discutida en el debate público: un posible debilitamiento del pensamiento crítico asociado al uso no reflexivo de estas herramientas, y las dificultades de equidad que introducen los sistemas automatizados de calificación cuando se los emplea sin criterio. Los autores concluyen que las evaluaciones resistentes a la inteligencia artificial, definidas como aquellas centradas en el proceso, entre las que mencionan explícitamente los exámenes orales y las evaluaciones de etapas múltiples, constituyen, junto con las políticas institucionales y los programas de alfabetización digital, la estrategia con mayor sustento en la literatura revisada por los autores.

Kofinas, et al. (2025) aportan una precisión importante al mostrar que el impacto de la inteligencia artificial generativa no se limita a las evaluaciones convencionales, sino que alcanza también a las evaluaciones diseñadas explícitamente como auténticas (aquellas que buscan replicar tareas del ejercicio profesional real), cuya legitimidad como respuesta al problema se había dado por sentada en buena parte de la literatura anterior a 2023. Si incluso las evaluaciones auténticas resultan vulnerables cuando su evidencia sigue siendo, en última instancia, un producto textual entregado sin trazabilidad del proceso que lo generó, entonces la autenticidad de la tarea no basta por sí sola para restituir la validez de la inferencia evaluativa; lo que se vuelve indispensable es la trazabilidad de cómo esa tarea auténtica



fue efectivamente resuelta. Este hallazgo, según se argumentará en la discusión, tiene consecuencias directas para la manera de organizar cualquier marco de rediseño evaluativo.

En el ámbito hispanohablante, la revisión sistemática de Sánchez, et al. (2025), realizada sobre 23 artículos publicados entre 2020 y 2025, es consistente con este diagnóstico y lo formula en términos cercanos a los de este artículo: los autores sostienen que los enfoques evaluativos tradicionales enfrentan una disrupción que obliga a las instituciones a reconceptualizar sus métodos pedagógicos, orientándolos hacia los procesos metacognitivos antes que hacia los productos finales, aunque advierten, con una prudencia metodológica que conviene retener, que la evidencia empírica concluyente sobre el impacto educativo real de estos rediseños todavía es escasa. Sánchez Mendiola (2023), desde una perspectiva editorial pero con apoyo en la literatura sobre evaluación educativa, llega a una conclusión convergente: ni la prohibición ni la evasión de estas herramientas constituyen respuestas sostenibles, y la alternativa consiste en diversificar los procedimientos evaluativos, incorporando exámenes orales, proyectos y observación directa, junto con una evaluación formativa más frecuente y un diálogo explícito entre docentes y estudiantes sobre los usos legítimos de la inteligencia artificial generativa.

Respuestas institucionales y pedagógicas documentadas

La respuesta institucional más inmediata frente a la inteligencia artificial generativa ha sido la detección y, en algunos casos, la prohibición explícita de su uso. Las fuentes del corpus que abordan esta cuestión señalan límites relevantes de esta estrategia: los detectores automáticos de texto generado por inteligencia artificial presentan tasas de error que los vuelven poco confiables como fundamento de decisiones disciplinarias, y la prohibición formal resulta, en la práctica, casi imposible de verificar cuando la herramienta es de acceso público y de uso doméstico. Ncube et al. (2026) señalan explícitamente que las herramientas de detección tienen limitaciones conocidas, y que apoyar en ellas el diseño evaluativo institucional traslada el problema de validez de la evaluación a la validez de la detección, sin resolverlo.

Frente a esos límites, la evaluación oral ha reaparecido en la literatura reciente como uno de los dispositivos más citados para restituir la trazabilidad del proceso, una tendencia que se examina en detalle en el apartado siguiente a propósito del marco propuesto por Nikolić y Basta Nikolić (2026). No toda incorporación de tecnología a la evaluación universitaria equivale, sin embargo, a una evaluación



centrada en el proceso, y conviene distinguir ambos fenómenos con cuidado, porque la literatura tiende a mezclarlos. Yermaganbetova, et al. (2026) reportan un diseño cuasiexperimental con grupos intactos (N=120) en un curso universitario de modelado tridimensional, en el que un grupo experimental que accedió durante ocho semanas a una plataforma que integraba distintos modelos de lenguaje para generar preguntas formativas y retroalimentación adaptativa mostró mejoras estadísticamente significativas en el desempeño posterior a la intervención, junto con una mayor autonomía percibida y el desarrollo de habilidades de aprendizaje autorregulado, en comparación con un grupo control que siguió instrucción tradicional. El hallazgo es valioso, y la referencia al aprendizaje autorregulado lo acerca, más que a otros estudios revisados, a la dimensión metacognitiva que este artículo defiende más adelante, pero conviene no confundirlo sin más con la evaluación de proceso propiamente dicha: lo que la plataforma optimiza es la calidad y la frecuencia de la retroalimentación sobre el desempeño en tareas prácticas sucesivas, no necesariamente la trazabilidad documentada del razonamiento que condujo a ese desempeño. Una plataforma adaptativa puede mejorar el desempeño medido sin que ese desempeño deje de evaluarse, en última instancia, como una serie de productos parciales.

Un matiz adicional, y de sentido inverso al que suele asumirse, proviene del estudio de Bañuelos Márquez y Romero Martínez (2024) sobre retroalimentación formativa generada por inteligencia artificial. Los autores evaluaron distintas configuraciones de un modelo conversacional aplicado a la corrección de trabajos estudiantiles y encontraron discrepancias entre las calificaciones sugeridas por el sistema y las otorgadas por la docente responsable de la asignatura, concluyendo que el sistema requiere un entrenamiento más específico y mayor precisión en las instrucciones antes de poder sustituir con fiabilidad el juicio experto. Este resultado introduce una cautela necesaria para cualquier propuesta de rediseño evaluativo que dependa, aunque sea parcialmente, de inteligencia artificial generativa para sostener el seguimiento del proceso a escala: la misma familia de herramientas que erosiona la validez de la evaluación del producto no ofrece todavía, según esta evidencia, garantías suficientes de fiabilidad cuando se la emplea del lado de la evaluación.

El marco de cuatro pilares de Nikolić y Basta Nikolić: una síntesis extendida

Entre los marcos conceptuales relevados para este artículo, el propuesto por Nikolić y Basta Nikolić (2026) es el que formula de manera más explícita y estructurada el desplazamiento del producto al



proceso que aquí se defiende, razón por la cual merece una presentación detallada antes de discutir sus alcances y sus límites en la sección de discusión.

Los autores organizan su propuesta alrededor de cuatro componentes, a los que denominan pilares y que presentan como interdependientes. El primero es la documentación del proceso, entendida como el requisito de que el estudiante conserve y entregue evidencia de las etapas intermedias de su trabajo, entre ellas borradores, notas y registros de búsqueda, de modo que la evaluación pueda acceder a algo más que el resultado final. El segundo es la defensa oral, concebida no como un examen adicional aislado, sino como una instancia en la que el estudiante debe explicar, justificar y eventualmente modificar en tiempo real las decisiones que su trabajo supone, lo cual permite verificar una comprensión que un producto escrito, por sí solo, no puede garantizar. El tercero es el diseño de tareas auténticas, es decir, actividades evaluativas que reproducen problemas o situaciones del ejercicio profesional real, bajo el supuesto de que ese tipo de tareas resulta más difícil de resolver mediante una consulta genérica a un sistema de inteligencia artificial generativa que una tarea abstracta o descontextualizada. El cuarto es la política transparente de uso de inteligencia artificial, que exige a las instituciones y a los docentes explicitar qué usos de estas herramientas están permitidos, en qué condiciones y con qué forma de declaración por parte del estudiante, en lugar de dejar la cuestión librada a una prohibición tácita e imposible de aplicar.

Nikolić y Basta Nikolić (2026) sostienen que la articulación de estos cuatro componentes permite desplazar la evaluación de la calidad del producto hacia el razonamiento demostrable, la toma de decisiones y la apropiación del aprendizaje por parte del estudiante. Su argumento central, y el más relevante para este artículo, es que la validez de la evidencia de aprendizaje no se ve sustancialmente disminuida por el acceso estudiantil a la inteligencia artificial generativa cuando la evaluación exige demostrar comprensión mediante modalidades múltiples: razonamiento en tiempo real, explicación del proceso seguido y aplicación a escenarios no anticipados por el estudiante al momento de preparar su trabajo. Se trata, hasta donde permite establecer la revisión realizada para este artículo, del marco más sistemático identificado en ella para convertir el desplazamiento del producto al proceso en un esquema operativo, susceptible de guiar decisiones concretas de diseño evaluativo, y no solamente en una



orientación general. Conviene notar, por su relevancia para la discusión que sigue, que los propios autores presentan su propuesta como un análisis conceptual, no sometido todavía a prueba empírica.

DISCUSIÓN Y CONCLUSIONES

El marco de cuatro pilares: una toma de distancia

El marco de Nikolić y Basta Nikolić (2026) ofrece, como se acaba de mostrar, la formulación más completa identificada en la revisión realizada del desplazamiento del producto al proceso, y buena parte de los dispositivos concretos que se proponen en el apartado siguiente son deudores de esa formulación. Conviene, sin embargo, señalar tres límites que justifican no adoptarlo sin más como marco de referencia de este artículo, y que motivan la propuesta alternativa que se desarrolla a continuación.

El primer límite es de orden conceptual y puede formularse con precisión a partir de la distinción de ejes introducida en los resultados. La documentación del proceso y la defensa oral operan sobre el eje de la evidencia, que es el que este artículo considera central. El diseño de tareas auténticas, en cambio, opera sobre el eje de la autenticidad, un eje distinto y, según se argumentó, independiente del anterior. Una tarea puede ser auténtica, en el sentido de reproducir con fidelidad un problema del ejercicio profesional, y seguir siendo evaluada exclusivamente a partir de un producto final entregado sin ningún registro de su elaboración, tal como lo muestra el propio hallazgo de Kofinas et al. (2025) respecto de la vulnerabilidad de las evaluaciones auténticas frente a la inteligencia artificial generativa. Si incluso las tareas auténticas resultan vulnerables cuando carecen de trazabilidad de proceso, la autenticidad de la tarea no puede tratarse como equivalente funcional de la documentación del proceso o de la defensa oral, y agruparla junto a esos dos componentes bajo la misma etiqueta de pilar mezcla dos ejes analíticos distintos, precisamente el tipo de asociación que este artículo procuró evitar al separar, en los resultados, la evidencia de la autenticidad.

El segundo límite, relacionado con el anterior, afecta a la política transparente de uso de inteligencia artificial. Ese componente no se deja ubicar con precisión en ninguno de los tres ejes propuestos, porque no describe una característica del diseño de la tarea evaluativa individual, sino una condición institucional que antecede a cualquier diseño particular y que un docente aislado difícilmente puede establecer por su cuenta. Tratar una política institucional y un dispositivo de diseño de tareas como componentes del mismo nivel analítico dificulta trasladar el marco a la práctica de un docente concreto,



que puede modificar de manera relativamente autónoma el diseño de sus evaluaciones, pero no la política institucional en la que esas evaluaciones se inscriben.

El tercer límite es de naturaleza genealógica y tiene consecuencias prácticas que no son menores. Nikolić y Basta Nikolić (2026) presentan su propuesta bajo la denominación de evaluación resiliente a la inteligencia artificial, lo cual sitúa el fundamento del marco en la amenaza tecnológica que se busca resistir, y no en una justificación pedagógica independiente de esa amenaza. Esa manera de fundamentar el rediseño evaluativo no es incorrecta, pero sí resulta frágil: si la justificación de documentar el proceso, exigir defensas orales o declarar el uso de herramientas de inteligencia artificial descansa en la necesidad de resistir a esa tecnología en particular, la mejora de los sistemas de detección, o la aparición de una tecnología distinta que vuelva obsoleta a la inteligencia artificial generativa tal como se la conoce en 2026, dejaría sin fundamento aparente al marco completo. Este artículo sostiene, en cambio, siguiendo la tradición inaugurada por Sadler (1989) y consolidada por Black y Wiliam (1998), que el valor de la evaluación centrada en el proceso es independiente de la amenaza que en este momento representa la inteligencia artificial generativa, y que esta última no es la razón última para adoptarla, sino la ocasión que ha vuelto políticamente posible e institucionalmente urgente una corrección que la literatura sobre evaluación formativa reclamaba desde antes de que esa tecnología existiera. Fundamentar el rediseño evaluativo en esa tradición, y no únicamente en la resiliencia frente a una tecnología particular, ofrece una base más estable para sostenerlo en el tiempo, incluso si las capacidades de los sistemas de inteligencia artificial generativa cambian de manera sustancial en los próximos años.

Hacia un marco de rediseño evaluativo centrado en el proceso

A partir de la distancia tomada en el apartado anterior, este artículo propone un marco de rediseño evaluativo organizado alrededor de un único eje, el de la evidencia, y de dos niveles en los que ese eje se despliega: el nivel del diseño de la tarea y el nivel de la interacción evaluativa entre docente y estudiante. La autenticidad de la tarea y la política institucional de uso de inteligencia artificial generativa, aunque relevantes y en buena medida deseables, se tratan aquí como condiciones que favorecen la eficacia de este marco, y no como principios de rediseño en sí mismos, en la medida en que pertenecen, según se argumentó, a ejes analíticos distintos.



En el nivel del diseño de la tarea se ubican dos principios. El primero es la trazabilidad del razonamiento: la evaluación debe poder acceder, en algún momento del ciclo evaluativo, a las decisiones intermedias que condujeron al resultado final, y no solamente al resultado. El segundo es la iteración documentada, es decir, la existencia de un registro, parcial y no necesariamente exhaustivo, de las versiones sucesivas de un trabajo, que permite reconstruir su desarrollo en el tiempo en lugar de recibirlo como un artefacto sin historia.

En el nivel de la interacción evaluativa se ubican los otros dos principios, y ambos requieren una precisión terminológica. El primero no consiste en verificar la autoría del trabajo entregado, algo que ya no puede garantizarse de manera general y que, en rigor, no es lo que interesa establecer. Lo que este principio busca es la comprensión verificable: la capacidad del estudiante de dar cuenta, en un intercambio directo con el docente, de las decisiones que el trabajo entregado supone, de justificarlas y de aplicarlas a variaciones no anticipadas del problema original. Un estudiante puede haber recurrido de manera intensiva a la inteligencia artificial generativa y, aun así, demostrar una apropiación plena de lo producido; ese es, precisamente, el caso que este principio busca poder distinguir de aquel en que la apropiación está ausente, con independencia del grado de asistencia tecnológica involucrado. El segundo principio del nivel de interacción es la metacognición explícita, que no se satisface con que el estudiante justifique de manera genérica el uso que hizo de una herramienta, sino que exige una reflexión más específica sobre el propio proceso: qué sabía antes de comenzar la tarea, qué estrategia eligió y por qué, en qué punto encontró dificultades, cómo verificó la validez de su resultado, qué modificaría si volviera a resolverla, y qué aprendió en el trayecto. Solo cuando incorpora estos elementos, la exigencia de justificar decisiones constituye, con propiedad, una exigencia metacognitiva, y no una simple declaración formal de uso de herramientas.

Estos cuatro principios (trazabilidad, iteración, comprensión verificable y metacognición explícita) se traducen en dispositivos concretos, varios de los cuales ya cuentan con respaldo en la literatura examinada. La Tabla 2 resume la correspondencia entre niveles, principios, tipo de evidencia y dispositivo asociado.



Tabla 2. Marco de rediseño evaluativo centrado en el proceso, organizado en dos niveles

Nivel	Principio	Evidencia posible	Dispositivo
Diseño de la tarea	Trazabilidad del razonamiento	Decisiones intermedias documentadas	Bitácora de proceso
Diseño de la tarea	Iteración documentada	Versiones sucesivas del trabajo	Historial de versiones
Interacción evaluativa	Comprensión verificable	Explicación y defensa de decisiones	Instancia oral estructurada
Interacción evaluativa	Metacognición explícita	Reflexión sobre estrategia, dificultades y aprendizaje	Informe reflexivo

Fuente: Elaboración propia

La evaluación oral estructurada, entendida no como examen final aislado sino como instancia recurrente de defensa de decisiones tomadas durante el proceso, es uno de los dispositivos más recurrentemente propuestos y con mayor respaldo conceptual entre las fuentes revisadas, y ocupa un lugar central tanto en el marco de Nikolić y Basta Nikolić (2026) como en la propuesta de diversificación de Sánchez Mendiola (2023). El control de versiones de la escritura, exigible en entornos digitales de manera relativamente sencilla, permite reconstruir el proceso de producción de un texto sin necesidad de vigilancia continua sobre el estudiante. Conviene, sin embargo, no sobrestimar su valor probatorio tomado de manera aislada: un historial de versiones es evidencia de proceso, pero no constituye por sí solo una prueba concluyente del proceso cognitivo, ya que las herramientas generativas también permiten producir borradores sucesivos y simulacros de iteración. La fuerza de este marco no reside, por lo tanto, en ninguno de sus dispositivos tomado de manera aislada, sino en la combinación de registros de proceso, iteraciones documentadas y capacidad de explicar decisiones ante otro: ninguna evidencia por sí sola garantiza la comprensión, pero la triangulación entre varias reduce de manera considerable el margen de simulación posible. Las bitácoras o diarios de proceso, en los que el estudiante documenta de manera periódica los obstáculos encontrados y las decisiones tomadas, cumplen una función semejante y tienen la ventaja adicional de fomentar la metacognición como práctica en sí misma, y no solo como requisito evaluativo. La evaluación entre pares centrada en el proceso, y no únicamente en el producto final como ocurre en buena parte de las prácticas de coevaluación actualmente extendidas, ofrece además una vía de escalamiento parcial para cursos



numerosos, aunque su fiabilidad depende de un entrenamiento previo del estudiantado que la literatura revisada no ha explorado en profundidad todavía.

Un elemento distintivo de este marco, y que lo separa de las respuestas meramente defensivas frente a la inteligencia artificial generativa, es el reconocimiento explícito de que estas herramientas pueden formar parte legítima del proceso evaluado, siempre que su uso quede documentado y sea objeto de reflexión por parte del estudiante. La política institucional de transparencia sobre el uso de inteligencia artificial, que Nikolić y Basta Nikolić (2026) incluyen entre sus cuatro pilares y que Sánchez Mendiola (2023) recupera bajo la forma de un diálogo explícito entre docentes y estudiantes, se entiende aquí como una condición institucional que hace posible este marco, y no como un principio equivalente a los cuatro anteriores. Esta posición se distingue tanto de la prohibición, cuya inviabilidad práctica ya se discutió, como de una aceptación acrítica que trate el producto generado por inteligencia artificial como equivalente al producto elaborado sin asistencia: lo que el marco aquí propuesto exige no es la ausencia de inteligencia artificial generativa en el proceso, sino la capacidad del estudiante de dar cuenta razonada de su intervención en ese proceso, sea cual sea el grado de asistencia tecnológica involucrado.

Tensiones, límites y objeciones

El rediseño evaluativo centrado en el proceso no está exento de costos, y sería intelectualmente deshonesto presentarlo como una solución sin fricciones. La objeción más inmediata, y probablemente la mejor fundada, es la de escalabilidad: la evaluación oral estructurada, el seguimiento de bitácoras y la revisión de versiones sucesivas exigen un tiempo docente considerablemente mayor que la corrección de un producto final, y ese costo se vuelve difícil de sostener en cursos masivos, particularmente en contextos institucionales donde la relación entre docentes y estudiantes ya se encuentra tensionada por restricciones presupuestarias. La evidencia revisada en este artículo no permite ofrecer una solución definitiva a este problema; sí permite señalar que las plataformas adaptativas de retroalimentación automatizada estudiadas por Yermaganbetova et al. (2026) constituyen una vía parcial de escalamiento, aunque, como se discutió antes, no sustituyen por sí solas la trazabilidad del proceso que este artículo defiende, y que el uso de inteligencia artificial generativa del lado docente para sostener ese seguimiento a escala enfrenta todavía, según Bañuelos Márquez y Romero Martínez (2024), problemas de fiabilidad no resueltos.



Una segunda tensión concierne a la equidad. Si la evaluación centrada en el proceso exige del estudiante una capacidad de articulación oral y una disposición a exponer su razonamiento de manera pública o semipública, existe el riesgo de que penalice de manera desproporcionada a quienes, por razones lingüísticas, culturales o de formación previa, se desenvuelven con mayor dificultad en esas instancias, incluso cuando su comprensión del contenido es sólida. Este riesgo no es hipotético: Santiago Ruiz y Páramo Chávez (2025) documentaron, en la muestra que estudiaron, que la dependencia indebida de la inteligencia artificial generativa se concentraba en los estudiantes con menores habilidades de escritura, un patrón que, de confirmarse en poblaciones más amplias, sugeriría que cualquier rediseño evaluativo que no atienda explícitamente a esas desigualdades de partida corre el riesgo de reproducirlas bajo una forma distinta, sustituyendo la desigualdad de acceso a la escritura por una desigualdad de acceso a la oralidad académica.

Una tercera tensión, de naturaleza más técnica, concierne a la validez y a la confiabilidad de la evaluación de proceso frente a los estándares con que tradicionalmente se ha evaluado la evaluación del producto. Los instrumentos de medición de la escritura y de la resolución de problemas cuentan con una tradición psicométrica considerablemente más desarrollada que los instrumentos disponibles para evaluar la calidad de un proceso de razonamiento documentado en una bitácora o expuesto en una defensa oral. La literatura revisada para este artículo no incluye estudios que hayan sometido estos instrumentos a procesos de validación con el rigor habitual en la evaluación educativa, lo cual constituye, como se retoma en las conclusiones, una limitación importante de la evidencia disponible antes que una objeción concluyente contra el rediseño propuesto.

Por último, cabe señalar el riesgo de sobrecarga evaluativa para el estudiante. Multiplicar las instancias de evaluación (bitácoras, versiones, defensas orales, evaluación entre pares) sin una reducción compensatoria en otras exigencias del curso puede traducirse en una intensificación del trabajo evaluativo que termine siendo contraproducente, tanto para el aprendizaje como para el bienestar del estudiantado. Este riesgo sugiere que el rediseño aquí propuesto no debe entenderse como una acumulación de instrumentos adicionales sobre la estructura evaluativa existente, sino como una sustitución: menos instancias de evaluación del producto, compensadas por instancias más selectivas, pero mejor diseñadas, de evaluación del proceso.



Limitaciones de la revisión

Las limitaciones discutidas en el apartado anterior conciernen al rediseño evaluativo propuesto. Corresponde, por separado, reconocer las limitaciones de la revisión de literatura sobre la que se apoya este artículo, que son de naturaleza distinta y no deben confundirse con las anteriores.

La primera es el tamaño y la composición del corpus. La revisión se apoyó en un número acotado de fuentes, identificadas mediante una búsqueda en bases de acceso abierto y buscadores académicos generales, y no en una base de datos indexada de cobertura exhaustiva ni en un protocolo de búsqueda replicable en el sentido de una revisión sistemática. Es razonable, en consecuencia, que otras búsquedas con términos, bases o períodos distintos pudieran identificar un conjunto parcialmente diferente de documentos y que esto pudiera matizar algunas de las afirmaciones aquí sostenidas.

La segunda es la heterogeneidad metodológica del corpus retenido, que combina, según muestra la Tabla 1, revisiones sistemáticas, estudios cuasiexperimentales, estudios de caso de alcance reducido, un editorial y un marco conceptual no sometido a prueba empírica. Esa heterogeneidad es coherente con la decisión metodológica de emprender una revisión integradora y no una revisión sistemática de un único tipo de diseño, pero limita la posibilidad de ponderar la evidencia de manera uniforme, y explica por qué las afirmaciones formuladas a lo largo del artículo se presentan como una tendencia observada en las fuentes examinadas, y no como un consenso demostrado de todo un campo de investigación.

La tercera es la rapidez con la que evoluciona la literatura sobre inteligencia artificial generativa y evaluación educativa, que vuelve probable que una parte del corpus revisado quede desactualizada, o sea complementada por nueva evidencia, en un plazo breve respecto de la fecha de esta revisión.

La cuarta es el predominio, dentro del corpus retenido, de literatura conceptual y propositiva por sobre la evidencia experimental controlada, un desequilibrio que este mismo artículo reproduce en su propia estructura argumentativa, y que, como se señaló en la sección de resultados, la literatura revisada reconoce como una carencia del campo en su conjunto y no como una limitación exclusiva de este trabajo.



Estas limitaciones no invalidan el argumento central del artículo, pero sí aconsejan leer sus afirmaciones sobre las fuentes examinadas o el corpus revisado como una descripción de tendencias observadas en un conjunto acotado de literatura, y no como el resultado de un mapeo exhaustivo del campo.

Conclusiones y agenda de investigación futura

El argumento desarrollado en este artículo puede resumirse de la siguiente manera: la inteligencia artificial generativa no crea el problema de la validez de la evaluación universitaria, pero vuelve insostenible una solución de compromiso que la práctica institucional había podido sostener mientras el desacople entre producto y proceso resultaba costoso y detectable, al menos en aquellas evaluaciones cuyas condiciones de producción no están controladas. Las fuentes examinadas, tanto en español como en inglés, muestran una convergencia notable en ese diagnóstico: en el corpus revisado, la evaluación centrada exclusivamente en el producto pierde validez como evidencia de aprendizaje cuando el proceso de elaboración permanece opaco. El marco de cuatro pilares de Nikolić y Basta Nikolić (2026), con las salvedades discutidas en el apartado 4.1, y las propuestas de diversificación evaluativa de Sánchez Mendiola (2023), ofrecen puntos de partida coherentes para un rediseño centrado en el proceso. La contribución específica de este artículo consiste en reorganizar esos puntos de partida separando explícitamente los ejes de finalidad, evidencia y autenticidad que buena parte de la literatura revisada tiende a mezclar, y en proponer, a partir de esa separación, un marco de dos niveles (diseño de la tarea e interacción evaluativa) y cuatro principios (trazabilidad del razonamiento, iteración documentada, comprensión verificable y metacognición explícita) que opera de manera consistente en un único eje analítico y que no requiere, para sostenerse, apelar a la amenaza tecnológica que lo ocasionó.

Esa convergencia diagnóstica no debe, sin embargo, leerse como evidencia de que el problema está resuelto, y las limitaciones señaladas en el apartado anterior piden prudencia adicional sobre este punto. Sánchez, et al. (2025) señalan que la evidencia empírica concluyente sobre el impacto educativo real de estos rediseños todavía es escasa, y la revisión realizada para este artículo, con las limitaciones ya reconocidas, es consistente con esa observación: la mayor parte de la literatura disponible es conceptual, propositiva o basada en estudios de caso de alcance reducido, y son todavía excepcionales los trabajos que hayan sometido las estrategias de evaluación centrada en el proceso a diseños de investigación capaces de establecer su efectividad comparativa frente a la evaluación tradicional, o de dar cuenta de



su sostenibilidad en el tiempo cuando se implementan a la escala de una institución completa. La agenda de investigación que se desprende de este diagnóstico incluye, como mínimo, tres líneas: estudios longitudinales sobre el efecto de la evaluación centrada en el proceso en el aprendizaje efectivo, más allá de la calificación obtenida; investigación sobre la equidad de estos dispositivos, particularmente en lo que respecta a la evaluación oral, y en poblaciones más amplias que las estudiadas hasta ahora; y el desarrollo y validación de instrumentos para evaluar procesos documentados, incluyendo estudios de validez, confiabilidad y, cuando corresponda, propiedades psicométricas, un terreno en el que la tradición de medición educativa tiene todavía un trabajo considerable por delante.

REFERENCIAS BIBLIOGRAFICAS

- Bañuelos Márquez, A. M., y Romero Martínez, E. (2024). Retroalimentación formativa con inteligencia artificial generativa: Un caso de estudio. *Wimb Lu*, 19(2), 103-122. <https://doi.org/10.15517/wl.v19i2.63262>
- Black, P., y Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7-74. <https://doi.org/10.1080/0969595980050102>
- Gallent-Torres, C., Zapata-González, A., y Ortego-Hernando, J. L. (2023). El impacto de la inteligencia artificial generativa en educación superior: una mirada desde la ética y la integridad académica. *RELIEVE*, 29(2). <https://www.redalyc.org/journal/916/91676028011/html/>
- Kofinas, A. K., Tsay, C.-H., y Pike, D. (2025). The impact of generative AI on academic integrity of authentic assessments within a higher education context. *British Journal of Educational Technology*, 56. <https://doi.org/10.1111/bjet.13585>
- Ncube, P. D. N., Dzvapatsva, G. P., Matobobo, C., y Ranga, M. M. (2026). Redefining student assessment in AI-infused learning environments: A systematic review of challenges and strategies for academic integrity. *AI and Ethics*, 6. <https://doi.org/10.1007/s43681-025-00871->
- Nikolić, D., y Basta Nikolić, M. (2026). Designing AI-resilient assessment in higher education: A four-pillar conceptual framework. *Frontiers in Artificial Intelligence*, 9, Artículo 1841682. <https://doi.org/10.3389/frai.2026.1841682>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18(2), 119-144. <https://doi.org/10.1007/BF00117714>



- Sánchez, N. E., Michay, G. C., y Calderón, J. V. (2025). Inteligencia artificial generativa en educación superior: Una revisión sistemática de literatura hispanohablante. *Revista Espacios*, 46(6), 14-25. <https://doi.org/10.48082/espacios-a25v46n06p02>
- Sánchez Mendiola, M. (2023). La inteligencia artificial generativa y la evaluación: ¿Qué pasará con los exámenes? *Investigación en Educación Médica*, 12(48), 5-8. <https://doi.org/10.22201/fm.20075057e.2023.48.23550>
- Santiago Ruiz, E., y Páramo Chávez, A. G. E. (2025). Inteligencia artificial en educación superior: Integridad académica y nuevas formas de escritura. *IE Revista de Investigación Educativa de la REDIECH*, 16, e2225. https://doi.org/10.33010/ie_rie_rediech.v16i0.2225
- Whittemore, R., y Knafl, K. (2005). The integrative review: Updated methodology. *Journal of Advanced Nursing*, 52(5), 546-553. <https://doi.org/10.1111/j.1365-2648.2005.03621.x>
- Yermaganbetova, M., Ashimbekova, A., Kaibassova, D., Akhitova, R., y Dyussembina, E. (2026). Evaluating the impact of an AI-integrated learning platform on student performance: A quasi-experimental study. *Frontiers in Education*, 11. <https://doi.org/10.3389/feduc.2026.1792353>

