



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_rem.v10i5

EVALUACIÓN TRADICIONAL Y PENSAMIENTO CRÍTICO EN ESTUDIANTES DE UNA UNIVERSIDAD PERUANA

**TRADITIONAL ASSESSMENT AND CRITICAL THINKING IN
STUDENTS AT A PERUVIAN UNIVERSITY**

Isabel Deycy Capillo Lucar

Universidad Nacional del Santa, México

Gladis Consuelo Vásquez Pereyra

Universidad Nacional del Santa, México

Pedro Gustavo Reyes Carrera

Universidad Nacional del Santa, México

DOI: https://doi.org/10.37811/cl_rcm.v10i5.25978

Evaluación Tradicional y Pensamiento Crítico en Estudiantes de una Universidad Peruana

Isabel Deycy Capillo Lucar¹icapillo@uns.edu.pe<https://orcid.org/0000-0002-9197-426X>Universidad Nacional del Santa, Facultad de
Educación y Humanidades. Nuevo Chimbote,
Perú**Gladis Consuelo Vásquez Pereyra**gvasquez@uns.edu.pe<https://orcid.org/0000-0002-8338-9781>Universidad Nacional del Santa, Facultad de
Educación y Humanidades. Nuevo Chimbote,
Perú**Pedro Gustavo Reyes Carrera**preyesca@uns.edu.pe<https://orcid.org/0000-0003-4854-2952>Universidad Nacional del Santa, Facultad de
Educación y Humanidades. Nuevo Chimbote,
Perú

RESUMEN

Las prácticas de evaluación no solo determinan cómo se califica el aprendizaje; también orientan qué información priorizan los estudiantes y qué formas de razonamiento ponen en juego. El objetivo de este estudio fue comparar dos perfiles evaluativos —tradicional y orientado al pensamiento crítico— según el desempeño en interpretación, análisis y argumentación. Se desarrolló una investigación cuantitativa, no experimental, transversal y descriptivo-comparativa con 168 estudiantes de cuatro aulas de Didáctica General I de la Universidad Nacional del Santa, Perú. Se utilizaron un cuestionario de 22 ítems, una ficha de revisión documental, una prueba de cuatro casos con doce preguntas abiertas y una rúbrica analítica. En las aulas de perfil tradicional predominaron los exámenes escritos, las preguntas cerradas, la reproducción de contenidos y el énfasis en la calificación final; en las aulas orientadas al pensamiento crítico fueron más frecuentes los casos contextualizados, las preguntas abiertas, los criterios explícitos y la retroalimentación orientada a la mejora. El puntaje total fue mayor en este segundo grupo ($M = 26,54$; $DE = 3,87$) que en el tradicional ($M = 19,05$; $DE = 4,67$), $t(166) = -11,31$; $p < ,001$; $d = 1,75$. La misma tendencia se observó en las tres dimensiones evaluadas. Sin embargo, debido al diseño del estudio y a que cada perfil estuvo asociado con un docente, los resultados deben interpretarse como una asociación y no como evidencia causal.

Palabras clave: evaluación educativa, pensamiento crítico, educación superior, evaluación formativa, formación docente

¹ Autor principal

Correspondencia: icapillo@uns.edu.pe

Traditional Assessment and Critical Thinking in Students at a Peruvian University

ABSTRACT

Assessment practices do more than determine how learning is graded; they also influence what students attend to and the kinds of reasoning they use. This study compared two assessment profiles—traditional and critical thinking-oriented—according to students' performance in interpretation, analysis, and argumentation. A quantitative, non-experimental, cross-sectional, descriptive-comparative design was used with 168 students from four General Didactics I classes at the National University of Santa, Peru. Data were collected through a 22-item questionnaire, a documentary review checklist, a four-case test containing twelve open-ended questions, and an analytical rubric. Traditional-profile classes relied more heavily on written examinations, closed questions, content reproduction, and final grading, whereas critical thinking-oriented classes more frequently used contextualized cases, open-ended questions, explicit criteria, and improvement-oriented feedback. Total performance was higher in the latter group ($M = 26.54$, $SD = 3.87$) than in the traditional group ($M = 19.05$, $SD = 4.67$), $t(166) = -11.31$, $p < .001$, $d = 1.75$. The same pattern was observed across the three dimensions. Nevertheless, because of the study design and the fact that each profile was linked to a different instructor, the findings should be interpreted as an association rather than as evidence of a causal effect.

Keywords: educational assessment, critical thinking, higher education, formative assessment, teacher education

Artículo recibido 15 julio 2026
Aceptado para publicación: 23 agosto 2026



INTRODUCCIÓN

En educación superior, evaluar no es un acto neutro. Cada instrumento transmite al estudiante qué vale la pena aprender, qué tipo de respuesta será reconocida y cuánto margen existe para justificar, revisar o mejorar una idea. Cuando la evaluación se reduce a una nota final, es comprensible que parte del esfuerzo se concentre en responder al formato y cumplir con el requisito inmediato. En cambio, cuando se explicitan los criterios y la retroalimentación puede utilizarse en una tarea posterior, la evaluación se integra con mayor claridad al aprendizaje. Esta diferencia está ampliamente respaldada por la literatura sobre evaluación formativa y retroalimentación (Black & Wiliam, 1998; Hattie & Timperley, 2007; Sadler, 1989). Además, desde la teoría de la autodeterminación, los entornos educativos que favorecen la autonomía y la percepción de competencia tienden a sostener formas de motivación más autónomas, mientras que los contextos excesivamente controladores pueden desplazar el interés hacia el cumplimiento externo (Ryan & Deci, 2020).

Los exámenes escritos y las preguntas cerradas cumplen una función legítima: permiten comprobar conocimientos declarativos, asegurar condiciones relativamente homogéneas y gestionar grupos numerosos. La dificultad aparece cuando se convierten en la principal —o casi única— fuente de evidencia en asignaturas que declaran resultados vinculados con el análisis, el juicio profesional o la resolución de problemas. En esos casos, un instrumento puede medir con precisión cuánto recuerda el estudiante, pero decir poco sobre cómo interpreta una situación, relaciona evidencias o sostiene una decisión. El metaanálisis de Abrami et al. (2015) mostró que el pensamiento crítico se desarrolla con mayor consistencia cuando las tareas incorporan problemas, diálogo y acompañamiento explícito. En la misma línea, Paul et al. (2023) advierten que los formatos exclusivamente cerrados pueden ocultar aspectos relevantes de la calidad del razonamiento y proponen recurrir a respuestas abiertas cuando se pretende valorar procesos críticos complejos.

Cuando el aprendizaje esperado exige interpretar, analizar o argumentar, las evaluaciones basadas en desempeño ofrecen una ventaja importante: hacen visible el razonamiento. Un caso, una pregunta abierta o una decisión argumentada obligan a seleccionar información, reconocer relaciones y explicar por qué una respuesta resulta defendible. No obstante, el formato abierto, por sí solo, no garantiza una demanda cognitiva alta; la tarea necesita criterios comprensibles y condiciones que orienten la elaboración de la



respuesta. Kleemola et al. (2022) encontraron que las evaluaciones de desempeño permiten observar componentes del pensamiento crítico que permanecen menos visibles en formatos de reconocimiento. De manera complementaria, Van Damme et al. (2023), a partir de una evaluación internacional de habilidades genéricas, sostienen que el pensamiento crítico puede medirse de forma válida mediante tareas de desempeño, mientras que Cheng et al. (2023) muestran que la evaluación dialógica entre pares puede favorecer el pensamiento crítico cuando el intercambio se encuentra adecuadamente estructurado. La retroalimentación resulta decisiva porque conecta lo que el estudiante hizo con lo que puede hacer mejor después. Una devolución que se limita a señalar errores o comunicar una calificación informa sobre el resultado, pero no siempre ofrece una ruta de mejora. La retroalimentación formativa, en cambio, ayuda a comparar el desempeño con un criterio, reconocer la distancia respecto del nivel esperado y tomar decisiones para una tarea posterior. Esta lógica está presente tanto en los planteamientos clásicos de la evaluación formativa como en los modelos de autorregulación del aprendizaje (Nicol & Macfarlane-Dick, 2006; Sadler, 1989). En formación inicial docente, Baroudi et al. (2023) observaron que el uso de feedforward —retroalimentación orientada explícitamente a la siguiente tarea— se relacionó con mejoras en pensamiento crítico y escritura académica, lo que refuerza la idea de que una devolución es más valiosa cuando puede transformarse en una acción concreta.

En América Latina y, de manera particular, en la educación superior peruana, esta discusión adquiere especial importancia por el tránsito hacia currículos orientados a competencias. Si se espera que el estudiante movilice conocimientos para interpretar situaciones y tomar decisiones, la evaluación necesita recoger evidencias de ese uso y no limitarse al dominio declarativo. Arenales Solis et al. (2024) vinculan de forma explícita el pensamiento crítico con la coherencia entre currículo y evaluación en el contexto peruano. A la vez, experiencias recientes muestran resultados favorables cuando las metodologías y las tareas exigen analizar, contrastar y justificar: Chero-Santisteban et al. (2025) reportaron mejoras del pensamiento crítico en una experiencia de aula invertida, mientras que Ossa Cornejo et al. (2022) encontraron avances en habilidades científicas de estudiantes de pedagogía luego de un programa explícito de pensamiento crítico.

A pesar de estos antecedentes, todavía son menos frecuentes los estudios que comparan configuraciones evaluativas completas dentro de una misma asignatura. Esta distinción es relevante porque, en la



experiencia real del estudiante, la evaluación no se reduce a un instrumento aislado: se vive como una combinación de exámenes, tareas, rúbricas, preguntas, criterios, devoluciones y oportunidades de revisión. El interés del presente trabajo fue observar esa configuración en cuatro aulas de Didáctica General I de la Universidad Nacional del Santa, donde coexistían dos perfiles predominantes de evaluación.

Con ese propósito, el estudio comparó la evaluación tradicional y la evaluación orientada al pensamiento crítico según el desempeño de los estudiantes en interpretación, análisis, argumentación y puntaje total. Se planteó como hipótesis que quienes pertenecían a las aulas con prácticas evaluativas orientadas al pensamiento crítico obtendrían puntuaciones superiores. Sin embargo, las aulas ya estaban conformadas y cada perfil correspondía a un docente distinto; por ello, la comparación se concibió desde el inicio como una asociación contextual y no como una demostración causal del efecto del tipo de evaluación.

METODOLOGÍA

Diseño y contexto

Se desarrolló una investigación aplicada, de enfoque cuantitativo, con diseño no experimental, transversal y descriptivo-comparativo. El equipo investigador no manipuló las prácticas evaluativas ni asignó aleatoriamente a los estudiantes a una condición. En lugar de ello, observó dos perfiles que ya estaban presentes durante el semestre académico 2025-I y comparó a los grupos mediante una medición común del desempeño.

El estudio se realizó en la Facultad de Educación y Humanidades de la Universidad Nacional del Santa, en Chimbote, Perú. Participaron cuatro aulas intactas de Didáctica General I, correspondientes al III ciclo de Educación Secundaria, en las especialidades de Lengua y Literatura; Historia, Geografía y Ciencias Sociales; Idiomas: Inglés-Francés; y Matemática, Computación y Física. Dos aulas estuvieron a cargo del docente que utilizaba predominantemente una evaluación tradicional y las otras dos de la docente que aplicaba con mayor frecuencia una evaluación orientada al pensamiento crítico. Esta organización constituye una limitación importante, porque no permite separar por completo las características del profesor, del aula y del perfil evaluativo.



Población, muestra y criterios de participación

La población accesible estuvo formada por 180 estudiantes matriculados en las cuatro aulas. Debido a que las secciones tenían tamaños diferentes y la menor contaba con 42 estudiantes, se estableció una cuota de 42 participantes por aula. De este modo, la muestra final quedó conformada por 168 estudiantes: 84 de aulas con evaluación tradicional y 84 de aulas orientadas al pensamiento crítico.

El muestreo fue no probabilístico e intencional. En el aula de menor tamaño participaron todos los estudiantes y, en las otras tres, se seleccionó el mismo número con el fin de mantener grupos equivalentes. Los 12 estudiantes no incluidos respondieron a este procedimiento de igualación y no a criterios de rendimiento académico. Se consideraron como criterios de inclusión la matrícula vigente, la participación regular en la asignatura, el consentimiento informado y la realización de los instrumentos programados.

Definición de los perfiles evaluativos

El perfil tradicional se definió por el predominio de exámenes escritos, preguntas cerradas, reproducción de conceptos, énfasis en la calificación final y retroalimentación centrada en la corrección. El perfil orientado al pensamiento crítico incorporó con mayor frecuencia casos y problemas contextualizados, preguntas abiertas, justificación de respuestas, criterios explícitos, rúbricas analíticas y retroalimentación orientada a la mejora.

La pertenencia a uno u otro perfil se estableció en el nivel del aula y del docente; no se clasificó individualmente a cada estudiante. Para comprobar que la caracterización atribuida a cada aula fuera coherente con la práctica observada, se utilizaron dos fuentes complementarias: un cuestionario de percepción estudiantil y una revisión documental de los productos de evaluación desarrollados durante el semestre.

Técnicas e instrumentos

El cuestionario de prácticas evaluativas estuvo compuesto por 22 ítems con cinco opciones de respuesta, desde 1 (nunca) hasta 5 (siempre). Diez ítems describieron prácticas tradicionales y doce prácticas orientadas al pensamiento crítico. Como las subescalas tenían diferente número de ítems, la interpretación se realizó a partir de la media de cada conjunto y no mediante una suma directa.



La ficha de revisión documental incluyó ocho indicadores valorados de 0 (ausente) a 3 (presencia predominante). Cuatro indicadores examinaron preguntas cerradas, reproducción de contenidos, calificación del producto final y retroalimentación limitada; los otros cuatro consideraron casos contextualizados, análisis y argumentación, criterios explícitos y retroalimentación orientada a la mejora. En total se revisaron 40 evidencias académicas: 20 de las dos aulas de perfil tradicional y 20 de las dos aulas orientadas al pensamiento crítico. Se incluyeron exámenes, tareas, rúbricas, productos académicos y registros de retroalimentación.

La prueba de desempeño se contextualizó en contenidos de Didáctica General I y estuvo compuesta por cuatro casos educativos con doce preguntas abiertas: cuatro de interpretación, cuatro de análisis y cuatro de argumentación. Cada respuesta se valoró de 0 a 3 mediante una rúbrica analítica. El puntaje máximo por dimensión fue de 12 puntos y el puntaje total alcanzó 36. Para describir los niveles se utilizaron los rangos 0-4, 5-8 y 9-12 en cada dimensión, y 0-12, 13-24 y 25-36 para el total.

La validez de contenido se examinó mediante el juicio de tres expertos, quienes valoraron la claridad, pertinencia y coherencia de los ítems en una escala ordinal de cuatro categorías. La V de Aiken por ítem osciló entre ,89 y 1,00 y el coeficiente global fue de ,96; todos los ítems superaron el criterio mínimo de ,80 establecido en los borradores del estudio.

Procedimiento y análisis de datos

Una vez coordinada la aplicación con los docentes, se explicó a los estudiantes el propósito del estudio y se recogieron los consentimientos informados. El cuestionario y la revisión documental se emplearon para caracterizar el perfil predominante de cada aula. Posteriormente, la misma prueba de pensamiento crítico se administró a los cuatro grupos bajo instrucciones y condiciones equivalentes. Las respuestas se calificaron con una única rúbrica y los registros fueron codificados sin identificadores personales.

Se calcularon frecuencias, porcentajes, medias y desviaciones estándar. La normalidad de los puntajes se examinó con la prueba de Shapiro-Wilk y la homogeneidad de varianzas con la prueba de Levene. Como los valores p fueron superiores a ,05 en las medidas evaluadas, se utilizó la prueba t de Student para muestras independientes. La magnitud de las diferencias se estimó mediante la d de Cohen y se adoptó un nivel de significación de $\alpha = ,05$. Los valores reportados como ,000 por el programa estadístico se expresaron como $p < ,001$.



La unidad estadística analizada fue el estudiante. No obstante, los 168 participantes se encontraban agrupados en solo cuatro aulas y cada perfil estuvo vinculado a un docente. Con un número tan reducido de conglomerados no resultaba estable estimar un modelo multinivel que separara los efectos del aula, del profesor y del perfil evaluativo. Por esta razón, las diferencias se interpretaron con cautela y únicamente como asociaciones entre condiciones observadas.

Consideraciones éticas

La participación fue voluntaria. Antes de responder los instrumentos, los estudiantes recibieron información sobre el propósito del estudio, la confidencialidad de sus respuestas y su derecho a retirarse sin consecuencias académicas. La base utilizada para el análisis fue codificada y los resultados se presentan de forma agregada, sin información que permita identificar a los participantes.

RESULTADOS Y DISCUSIÓN

Perfiles evaluativos observados

El cuestionario y la revisión documental permitieron reconocer dos configuraciones evaluativas diferenciadas. En las aulas de perfil tradicional se concentraron los instrumentos de respuesta cerrada, la reproducción de contenidos y la calificación del producto final. En las aulas orientadas al pensamiento crítico aparecieron con mayor frecuencia los casos contextualizados, la justificación de respuestas, las rúbricas y la retroalimentación. La Tabla 1 sintetiza ambos perfiles. Los porcentajes de la revisión documental no deben interpretarse como si fueran directamente comparables, porque cada conjunto corresponde a indicadores distintos.

Tabla 1. Perfiles evaluativos identificados mediante cuestionario y revisión documental

Perfil	Prácticas predominantes	Cuestionario	Revisión documental
Tradicional	Exámenes escritos, preguntas cerradas, memorización, reproducción y calificación final	M = 3,91 DE = 0,54	Presencia global: 76,3 %
Orientado al pensamiento crítico	Casos contextualizados, preguntas abiertas, justificación, rúbricas y retroalimentación	M = 4,00 DE = 0,51	Presencia global: 77,5 %

Nota. Las medias del cuestionario se expresan en una escala de 1 a 5. Los porcentajes documentales corresponden a indicadores diferentes en cada perfil y no representan una medida directa de superioridad.

El contraste observado coincide con la literatura que diferencia una evaluación centrada principalmente en certificar resultados de otra que utiliza tareas, criterios y retroalimentación para apoyar el aprendizaje. Black y Wiliam (1998) y Sadler (1989) ya señalaban que la evaluación cambia de función cuando la



información obtenida se utiliza para ajustar el desempeño. En este estudio, por tanto, la diferencia más relevante no fue simplemente la presencia o ausencia de exámenes, sino el conjunto de prácticas que acompañó a cada perfil y el tipo de respuesta que esas prácticas pedían al estudiante.

La mayor presencia de casos y preguntas abiertas en las aulas orientadas al pensamiento crítico también resulta coherente con estudios sobre evaluación de desempeño. Kleemola et al. (2022) y Paul et al. (2023) muestran, desde aproximaciones diferentes, que las tareas construidas permiten observar la calidad del razonamiento con mayor detalle que los formatos centrados únicamente en seleccionar una respuesta. A ello se suma la retroalimentación: cuando ayuda al estudiante a contrastar su producción con un criterio y a decidir qué modificar, deja de ser un comentario final y se convierte en parte del proceso de aprendizaje (Nicol & Macfarlane-Dick, 2006).

Niveles de desempeño en pensamiento crítico

La distribución de los niveles mostró tendencias opuestas entre los dos perfiles. En las aulas tradicionales predominó el nivel medio en las tres dimensiones y en el puntaje total; además, la argumentación concentró la mayor proporción relativa de estudiantes en el nivel bajo. En las aulas orientadas al pensamiento crítico, en cambio, predominó el nivel alto. La Tabla 2 presenta la distribución completa junto con las medias y desviaciones estándar.

Tabla 2. Niveles de desempeño según perfil evaluativo

Perfil	Dimensión	Bajo n (%)	Medio n (%)	Alto n (%)	M	DE
Tradicional	Interpretación	13 (15,5)	54 (64,3)	17 (20,2)	6,72	1,82
Tradicional	Análisis	16 (19,0)	55 (65,5)	13 (15,5)	6,38	1,77
Tradicional	Argumentación	21 (25,0)	51 (60,7)	12 (14,3)	5,95	1,88
Tradicional	Total	10 (11,9)	62 (73,8)	12 (14,3)	19,05	4,67
Pensamiento crítico	Interpretación	3 (3,6)	29 (34,5)	52 (61,9)	9,05	1,58
Pensamiento crítico	Análisis	4 (4,8)	31 (36,9)	49 (58,3)	8,89	1,64
Pensamiento crítico	Argumentación	6 (7,1)	34 (40,5)	44 (52,4)	8,60	1,70
Pensamiento crítico	Total	2 (2,4)	25 (29,8)	57 (67,9)	26,54	3,87

Nota. Cada dimensión tuvo un rango de 0 a 12 y el puntaje total un rango de 0 a 36.

La argumentación obtuvo la media más baja en ambos perfiles. Este resultado es relevante porque argumentar exige algo más que reconocer información pertinente: supone vincular razones con evidencias, sostener una conclusión y mantener coherencia entre los distintos componentes de la respuesta. La dificultad compartida sugiere que incluso en aulas donde se emplean preguntas abiertas

conviene enseñar de manera explícita cómo se construye, revisa y mejora un argumento. La evaluación dialógica puede contribuir a ese aprendizaje cuando los estudiantes cuentan con orientaciones claras para examinar el razonamiento de sus pares (Cheng et al., 2023).

El patrón general es compatible con investigaciones que han encontrado mejores resultados cuando las actividades académicas exigen una participación cognitiva más activa. Abrami et al. (2015) identificaron efectos más consistentes en intervenciones que combinaban problemas auténticos, diálogo y acompañamiento. En contextos de educación superior, Van Damme et al. (2023) sostienen que las habilidades de pensamiento crítico pueden evaluarse de manera válida mediante tareas de desempeño, aunque sus resultados también muestran que el desarrollo de estas habilidades no ocurre de forma automática durante la trayectoria universitaria. En el ámbito latinoamericano, Chero-Santisteban et al. (2025) reportaron mejoras asociadas al aula invertida y Ossa Cornejo et al. (2022) observaron avances en habilidades científicas de estudiantes de pedagogía después de un programa explícito de pensamiento crítico. El presente estudio no evaluó una estrategia aislada, pero sus resultados convergen con una idea común: las oportunidades reiteradas para analizar, justificar y revisar se asocian con un mejor desempeño cuando esas mismas operaciones son requeridas en la evaluación.

Comparación entre grupos

Las comparaciones inferenciales mostraron diferencias estadísticamente significativas en interpretación, análisis, argumentación y puntaje total. En las cuatro medidas, los tamaños del efecto fueron grandes. Para facilitar la lectura, la Tabla 3 presenta las medias de cada perfil, la diferencia de medias con su intervalo de confianza del 95 %, el estadístico t y la d de Cohen.

Tabla 3. Comparación de puntuaciones según perfil evaluativo

Dimensión	Tradicional (DE)	M Pensamiento crítico M (DE)	Diferencia (IC 95 %)	t(166)	p	d
Interpretación	6,72 (1,82)	9,05 (1,58)	2,33 [1,81; 2,85]	-8,86	< ,001	1,37
Análisis	6,38 (1,77)	8,89 (1,64)	2,51 [1,99; 3,03]	-9,54	< ,001	1,47
Argumentación	5,95 (1,88)	8,60 (1,70)	2,65 [2,10; 3,20]	-9,59	< ,001	1,48
Total	19,05 (4,67)	26,54 (3,87)	7,49 [6,18; 8,80]	-11,31	< ,001	1,75

Nota. Los intervalos de confianza se calcularon a partir de las medias, desviaciones estándar y tamaños de grupo informados. Los valores negativos de t se deben al orden de sustracción utilizado en la comparación.



La diferencia total de 7,49 puntos sobre un máximo de 36 es estadísticamente clara y, por su magnitud estandarizada, también tiene relevancia práctica. Sin embargo, no debe interpretarse como una estimación causal del efecto de la evaluación. Los estudiantes estaban distribuidos en cuatro aulas y cada perfil se encontraba asociado con un docente. La experiencia del profesor, la forma de acompañar las tareas, las expectativas del grupo y otras características no medidas pudieron contribuir a la diferencia observada.

También debe considerarse un posible efecto de familiaridad con el formato. Los estudiantes acostumbrados a resolver casos y justificar sus respuestas encontraron una prueba más cercana a las actividades que habían practicado durante el semestre. Esto no invalida la comparación —de hecho, muestra la coherencia entre lo que se practica y lo que luego se solicita—, pero impide atribuir la diferencia a un solo componente. Por ello, la conclusión más defendible es que ambos perfiles evaluativos estuvieron asociados con desempeños distintos, no que uno de ellos haya sido demostrado como la única causa.

Desde una perspectiva pedagógica, los hallazgos apoyan una combinación más equilibrada de instrumentos. Los exámenes pueden conservarse cuando se busca comprobar conocimientos básicos, pero deberían convivir con casos, proyectos, ensayos, debates u otros productos revisables cuando el resultado de aprendizaje exige analizar o argumentar. Del mismo modo, una rúbrica resulta más útil cuando describe la calidad del razonamiento y no solo el orden o el cumplimiento formal. La retroalimentación cobra mayor valor cuando ofrece información que el estudiante puede usar en la siguiente tarea; esta idea aparece tanto en la literatura clásica sobre feedback (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006) como en experiencias recientes de feedforward con futuros docentes (Baroudi et al., 2023).

En el contexto universitario peruano, esta lectura es especialmente pertinente para el enfoque por competencias. Arenales Solís et al. (2024) plantean que el desarrollo del pensamiento crítico requiere coherencia entre currículo y evaluación, mientras que Vendrell Morancho (2024) destaca la necesidad de ofrecer oportunidades sostenidas para analizar, contrastar y decidir. Solís Trujillo et al. (2025), por su parte, subrayan la relevancia de la evaluación formativa y la retroalimentación en educación superior. A esta discusión puede añadirse una dimensión motivacional: cuando los criterios son comprensibles, el



estudiante dispone de oportunidades para mejorar y percibe que puede progresar, la evaluación puede apoyar su sensación de competencia y una participación más autónoma; cuando todo se concentra en la sanción o la nota, aumenta el riesgo de que el esfuerzo se oriente principalmente al cumplimiento externo (Ryan & Deci, 2020). Aunque este estudio no midió motivación, esta perspectiva ayuda a comprender por qué el diseño de la evaluación puede influir en la manera en que el estudiante se implica con las tareas.

El estudio aporta evidencia situada en una universidad pública peruana y en una asignatura de formación docente, pero su alcance es necesariamente acotado. La muestra fue intencional, no se controlaron por completo las diferencias entre especialidades y aulas, y el reducido número de conglomerados impidió separar estadísticamente los efectos del profesor y del perfil evaluativo. Estas limitaciones restringen la generalización y obligan a interpretar los resultados dentro del contexto en que fueron obtenidos.

CONCLUSIONES

Las cuatro aulas examinadas mostraron perfiles evaluativos claramente diferenciados. En dos predominaron los exámenes escritos, las preguntas cerradas, la reproducción de contenidos y la calificación final; en las otras dos fueron más frecuentes los casos contextualizados, las preguntas abiertas, los criterios explícitos y la retroalimentación orientada a la mejora. La diferencia, por tanto, no estuvo en un instrumento aislado, sino en la combinación de prácticas que organizó la experiencia de evaluación.

Los estudiantes de las aulas orientadas al pensamiento crítico obtuvieron puntuaciones superiores en interpretación, análisis, argumentación y puntaje total. La diferencia global fue de 7,49 puntos sobre 36 y estuvo acompañada por un tamaño del efecto elevado. Aun así, la argumentación presentó la media más baja en ambos perfiles, lo que revela una dificultad compartida y refuerza la necesidad de enseñar de manera más explícita cómo se relacionan razones, evidencias y conclusiones.

En conjunto, los resultados respaldan la conveniencia de alinear las tareas de evaluación con las habilidades que se pretende desarrollar. No demuestran, sin embargo, que el tipo de evaluación sea la única causa de las diferencias, porque las aulas no fueron asignadas al azar y cada perfil estuvo asociado con un docente. Futuras investigaciones deberían incorporar mediciones iniciales, un mayor número de aulas y análisis capaces de controlar el efecto de agrupamiento. También sería pertinente replicar el



estudio en otras asignaturas y carreras, e incluir variables como motivación y autorregulación, para examinar con mayor precisión cómo las prácticas evaluativas se relacionan con el desarrollo del pensamiento crítico.

Declaración de conflicto de intereses

Los autores declaran que no existen conflictos de intereses de carácter personal, académico, institucional, laboral o financiero que hayan condicionado el desarrollo, el análisis o la presentación de los resultados.

REFERENCIAS BIBLIOGRÁFICAS

- Abrami, P. C., Bernard, R. M., Borokhovski, E., Waddington, D. I., Wade, C. A., & Persson, T. (2015). Strategies for teaching students to think critically: A meta-analysis. *Review of Educational Research*, 85(2), 275–314. <https://doi.org/10.3102/0034654314551063>
- Arenales Solis, J. C. E., Villegas Paz, M. C., Torero Soralez, M. J., & Rimarachin Montesinos, L. A. (2024). Pensamiento crítico y currículo por competencias: Reflexiones desde la educación superior peruana. *Revista Venezolana de Gerencia*, 29(12), 1667–1679. <https://doi.org/10.52080/rvgluz.29.e12.49>
- Baroudi, S., Aoun, S., & Hamam, D. (2023). Using feedforward to improve pre-service teachers' academic writing and critical thinking skills. *Frontiers in Education*, 8, Article 1126594. <https://doi.org/10.3389/feduc.2023.1126594>
- Black, P., & Wiliam, D. (1998). Assessment and classroom learning. *Assessment in Education: Principles, Policy & Practice*, 5(1), 7–74. <https://doi.org/10.1080/0969595980050102>
- Cheng, L., Li, Y., Su, Y., & Gao, L. (2023). Effect of regulation scripts for dialogic peer assessment on feedback quality, critical thinking and climate of trust. *Assessment & Evaluation in Higher Education*, 48(4), 451–463. <https://doi.org/10.1080/02602938.2022.2092068>
- Chero-Santisteban, Y. A., Moreno-Núñez, P. J., Saldaña-Taboada, H. J., & Nina-Cuchillo, E. E. (2025). Impacto del aula invertida en el pensamiento crítico de estudiantes en una universidad privada de Lima (Perú). *Formación Universitaria*, 18(2), 11–24. <https://doi.org/10.4067/S0718-50062025000200011>



- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Kleemola, K., Hyytinen, H., & Toom, A. (2022). Exploring internal structure of a performance-based critical thinking assessment for new students in higher education. *Assessment & Evaluation in Higher Education*, 47(4), 556–569. <https://doi.org/10.1080/02602938.2021.1946482>
- Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, 31(2), 199–218. <https://doi.org/10.1080/03075070600572090>
- Ossa Cornejo, C., Palma Luengo, M., Díaz Mujica, A., & Páez Rovira, D. (2022). Promoting pedagogy students' scientific skills through a critical thinking program. *Electronic Journal of Research in Educational Psychology*, 20(57), 263–288. <https://doi.org/10.25115/ejrep.v20i57.5272>
- Paul, J. A., Sinha, M., & Cochran, J. D. (2023). Instruments to assess students' critical thinking—A qualitative approach. *Decision Sciences Journal of Innovative Education*, 21(3), 123–143. <https://doi.org/10.1111/dsji.12295>
- Ryan, R. M., & Deci, E. L. (2020). Intrinsic and extrinsic motivation from a self-determination theory perspective: Definitions, theory, practices, and future directions. *Contemporary Educational Psychology*, 61, Article 101860. <https://doi.org/10.1016/j.cedpsych.2020.101860>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. *Instructional Science*, 18, 119–144. <https://doi.org/10.1007/BF00117714>
- Solis Trujillo, B. P., Velarde-Camaqui, D., Gonzales Nuñez, C. A., Castillo Silva, E. V., & Gonzalez Said de la Oliva, M. del P. (2025). The current landscape of formative assessment and feedback in graduate studies: A systematic literature review. *Frontiers in Education*, 10, Article 1509983. <https://doi.org/10.3389/feduc.2025.1509983>
- Van Damme, D., Zahner, D., Cortellini, O., Dawber, T., & Rotholz, K. (2023). Assessing and developing critical-thinking skills in higher education. *European Journal of Education*, 58(3), 369–386. <https://doi.org/10.1111/ejed.12563>



Vendrell Morancho, M. (2024). Desarrollo del pensamiento crítico en la universidad: Estrategias para superar desafíos y fomentar su crecimiento. *Revista de la Educación Superior*, 53(210), 1–20.

<https://doi.org/10.36857/resu.2024.210.2848>

