

Tipo de artículo: Investigación

Evaluación formativa con rúbricas analíticas en resolución de problemas

Formative assessment with analytic rubrics in problem solving

Autores:

Mirian Rosana De la Cruz Cruz

Universidad Tecnológica - España, San Francisco de Borja-Ecuador, rosana_dlc@hotmail.com,
<https://orcid.org/0009-0003-4394-7288>

Byron Orlando Lara Jerez

Universidad Tecnológica Equinoccial, Santiago de Píllaro- Ecuador, byron.lara@educacion.gob.ec,
<https://orcid.org/0009-0007-5020-5425>

Marcelo Ernesto Lara Jerez

Unidad Educativa "Ramón Barba Naranjo", Píllaro - Ecuador, marcelo.laraj@educacion.gob.ec,
<https://orcid.org/0009-0002-0349-1733>

Ana Mireya Mafla Álvarez

U.E. Guillermo Vinuesa, San Francisco de Borja- Ecuador, miremafla@yahoo.es,
<https://orcid.org/0009-0001-6762-8068>

Enma Magaly Almeida Ramos

Universidad Tecnica Ambato, Napo- Ecuador, magalyalmeida77@yahoo.es, <https://orcid.org/0009-0000-0673-9488>

Corresponding Author: *Mirian Rosana De la Cruz Cruz*, rosana_dlc@hotmail.com

Reception: 23-octubre-2025

Acceptance: 19- noviembre -2025

Publication: 28- noviembre -2025

How to cite this article:

De la Cruz Cruz, M. R., Byron Orlando Lara Jerez, B. O. L. J., Lara Jerez, M. E., & Mafla Álvarez, A. M. (2025). Evaluación formativa con rúbricas analíticas en resolución de problemas. *Horizonte Cientifico International Journal*, 3(2), 1-18.
<https://doi.org/10.64747/a93zv304>

RESUMEN

Este estudio evaluó el impacto de la evaluación formativa con rúbricas analíticas en la resolución de problemas en la Educación General Básica Superior (EGB-S) del Distrito Metropolitano de Quito. Se implementó un diseño cuasi-experimental con grupos no equivalentes, pretest–posttest y seguimiento, en 48 aulas de octavo a décimo año ($N = 1\,396$). La intervención consistió en el uso sistemático de una rúbrica analítica de cinco dimensiones (comprensión, planificación, modelización, justificación y verificación), integrada a ciclos breves de retroalimentación (≤ 1 semana) con auto- y coevaluación. La comparación desarrolló tareas similares sin rúbricas formativas explícitas. Se recopilieron datos de desempeño en tareas, indicadores de fidelidad y un inventario breve de estrategias de autorregulación. El ANCOVA con errores robustos y control por línea base evidenció un efecto significativo y moderado de la intervención sobre el puntaje posttest ($\beta = 6.08$; $p < .001$; $g = 0.45$; IC95 % [0.36, 0.54]), corroborado por modelos multinivel ($\gamma = 5.61$; $p < .001$). Las ganancias fueron mayores en justificación ($g = 0.51$) y verificación ($g = 0.54$). La rúbrica mostró confiabilidad adecuada ($ICC \geq .75$; $\alpha/\omega = .78-.86$) y validez de constructo ($CFI = .957$; $RMSEA = .045$). Un análisis de mediación sugirió que $\sim 17\%$ del efecto total se explica por aumentos en autorregulación. Los efectos fueron más altos en 10.º año y en aulas con fidelidad de implementación elevada. Se concluye que las rúbricas analíticas usadas con propósito formativo mejoran el aprendizaje en tareas de alta demanda cognitiva, fortalecen procesos de retroalimentación accionable y autorregulación, y ofrecen una vía escalable y de bajo costo para operacionalizar el currículo en Quito. Se recomiendan ensayos aleatorizados por clúster, análisis de costo-efectividad y extensión interdisciplinaria para estimar persistencia y transferencia de los efectos.

Palabras clave: rúbricas analíticas; evaluación formativa; resolución de problemas; autorregulación del aprendizaje; modelos multinivel

ABSTRACT

This study examined the impact of formative assessment with analytic rubrics on problem solving in upper-primary education (EGB-S) in Quito, Ecuador. We used a quasi-experimental nonequivalent groups design with pretest–posttest and follow-up across 48 classrooms in grades 8–10 ($N = 1,396$). The intervention entailed a five-dimension analytic rubric (comprehension, planning, modelling, justification, verification) embedded in short feedback cycles (≤ 1 week) that included self- and peer-assessment. The comparison condition completed similar tasks without explicit formative rubrics. Outcome measures comprised task performance, implementation fidelity indicators, and a brief self-regulated learning inventory. ANCOVA with robust errors and baseline controls indicated a significant, moderate intervention effect on posttest scores ($\beta = 6.08$; $p < .001$; $g = 0.45$; 95% CI [0.36, 0.54]), corroborated by multilevel models ($\gamma = 5.61$; $p < .001$). Gains were larger in justification ($g = 0.51$) and verification ($g = 0.54$). The rubric displayed adequate reliability ($ICC \geq .75$; $\alpha/\omega = .78-.86$) and construct validity ($CFI = .957$; $RMSEA = .045$). A mediation analysis suggested that approximately 17% of the total effect operated through increases in self-regulation. Effects were larger in 10th grade and in classrooms with high implementation fidelity. We conclude that analytic rubrics used for formative purposes enhance learning on high-cognitive-demand tasks, strengthen actionable feedback and self-regulation processes, and provide a scalable, low-cost route to operationalize the curriculum in Quito. Future work should prioritise cluster-randomised trials, cost-effectiveness analyses, and cross-disciplinary extensions to estimate effect persistence and transfer.

Keywords: analytic rubrics; formative assessment; problem solving; self-regulated learning; multilevel models

1. INTRODUCCIÓN

La mejora sostenida del aprendizaje en la Educación General Básica Superior (EGB-S; octavo a décimo año) depende, de forma decisiva, de cómo se evalúa aquello que los y las estudiantes son capaces de hacer con lo que saben. En Quito, provincia de Pichincha, donde concurren una alta densidad de instituciones educativas y una amplia trazabilidad de datos públicos, la resolución de problemas —especialmente en Matemática y en tareas integradoras STEM— constituye un foco estratégico para el desarrollo de competencias de orden superior y la reducción de brechas de desempeño. En este contexto, la evaluación formativa con rúbricas analíticas emerge como un dispositivo de alto potencial porque explicita criterios, ofrece retroalimentación accionable y promueve la autorregulación del aprendizaje, tres elementos que, como se ha documentado, se asocian con ganancias medibles en rendimiento y en habilidades metacognitivas (Panadero, Jonsson, Pinedo, & Fernández-Castilla, 2023; Andrade & Brookhart, 2019).

Desde una perspectiva de política y de escuela, el problema se puede formular así: pese a la disponibilidad de lineamientos curriculares y de resultados de evaluaciones nacionales, persiste una variabilidad marcada en el desempeño de estudiantes de EGB-S frente a tareas complejas de resolución de problemas. Esta variabilidad suele atribuirse a tres cuellos de botella: (i) prácticas de evaluación predominantemente sumativas, con escaso uso pedagógico de la evidencia; (ii) criterios de logro poco operativos para estudiantes y familias; y (iii) oportunidades dispares de práctica guiada con retroalimentación específica y a tiempo. Las rúbricas analíticas—frente a las holísticas—permiten desagregar el desempeño en dimensiones observables (p. ej., comprensión del enunciado, modelización, selección de estrategias, justificación y verificación), ofreciendo descriptores de niveles que facilitan tanto la coevaluación como la autoevaluación. La evidencia reciente sugiere que, cuando las rúbricas se emplean con propósitos formativos, sus efectos sobre el rendimiento son de magnitud moderada y positivos, con efectos pequeños pero consistentes sobre la autorregulación y la autoeficacia (Panadero et al., 2023). En la misma línea, marcos conceptuales de evaluación y pedagogía sostienen que la evaluación formativa es más efectiva cuando se integra, de manera explícita, en la teoría de la enseñanza y en rutinas de aula (Black & Wiliam, 2018).

En los últimos cinco años se ha consolidado un cuerpo de literatura que articula rúbricas, assessment for learning y aprendizaje autorregulado (SRL). Andrade y Brookhart (2019) proponen concebir la evaluación en el aula como co-regulación del aprendizaje: los criterios explícitos y la retroalimentación dialogada ayudan al estudiantado a calibrar sus metas, monitorear el progreso y ajustar estrategias. Esta visión se alinea con hallazgos de revisiones y metaanálisis que reportan beneficios en desempeño y, en menor medida, en variables motivacionales cuando las rúbricas se usan con transparencia de criterios, oportunidades de práctica y retroalimentación iterativa (Panadero et al., 2023; English, 2022). Por su parte, estudios en didáctica de la modelización matemática aportan instrumentos válidos para valorar procesos implicados en problemas del mundo real, ofreciendo rúbricas bidimensionales que distinguen entre tipo y nivel de modelización, lo que habilita un uso formativo más fino (Kohen, 2023).

La pertinencia local es clara. Quito concentra una masa crítica de instituciones participantes en evaluaciones nacionales y dispone de datos abiertos que permiten caracterizar el desempeño por institución, zona y subnivel. Al mismo tiempo, investigaciones recientes en revistas regionales muestran que el uso de tecnologías educativas y la disponibilidad de criterios claros median el aprendizaje y la motivación en estudiantes de educación básica y bachillerato en

Ecuador, reforzando la necesidad de criterios explícitos y retroalimentación de calidad (García Cabrera, Guaman Chimbo, Rea Minchalo, & Vega Pérez, 2025; Aguilar, Galarza Gallardo, Largo Barnuevo, & Ramírez Asanza, 2025). Sin embargo, estas experiencias suelen adolecer de estandarización en los instrumentos y de reportes comparables sobre confiabilidad y validez, lo que dificulta su escalamiento.

En términos conceptuales, esta investigación se ancla en tres marcos: (1) Evaluación para el aprendizaje como conjunto de prácticas que cierran la brecha entre el desempeño actual y el deseado mediante evidencia oportuna y utilizable; (2) Aprendizaje autorregulado, entendido como el ciclo de planificación, monitoreo y reflexión que las rúbricas pueden catalizar cuando se usan para auto-y coevaluación (Andrade & Brookhart, 2019); y (3) Competencia de modelización y resolución de problemas, para la cual existen propuestas de rúbricas validadas que distinguen fases como comprensión, formulación, ejecución, interpretación y verificación, coherentes con marcos internacionales como PISA Matemáticas (Kohen, 2023; OECD, 2019). La hipótesis operativa es que la alineación entre criterios de rúbrica y resultados de aprendizaje, sumada a ciclos breves de retroalimentación, se traducirá en mejoras significativas en tareas de resolución de problemas con respecto a condiciones de práctica habitual sin rúbrica o con rúbricas de uso meramente sumativo.

Objetivo general. Evaluar el efecto de un programa de evaluación formativa con rúbricas analíticas sobre el desempeño en resolución de problemas de estudiantes de EGB-S en instituciones fiscales y fiscomisionales de Quito, durante el año lectivo en curso, utilizando datos de línea base provenientes de fuentes públicas y un diseño de intervención de aula.

Objetivos específicos. (i) Diseñar y validar, con jueces expertos, una rúbrica analítica para resolución de problemas alineada con el currículo y con marcos de modelización; (ii) Implementar ciclos de retroalimentación formativa (docente-estudiante, pares y autoevaluación) integrados a secuencias didácticas; (iii) Estimar el efecto de la intervención sobre medidas de desempeño (puntajes de tareas, calidad de explicaciones, tasas de verificación) y sobre indicadores de autorregulación percibida; y (iv) Explorar diferencias por características contextuales (p. ej., jornada, gestión, parroquia urbana o periférica) considerando la heterogeneidad de Quito.

Hipótesis. H1: Los estudiantes expuestos a rúbricas analíticas con uso formativo obtendrán puntajes significativamente mayores en tareas de resolución de problemas que sus pares en la condición de comparación. H2: La intervención aumentará la frecuencia de prácticas de verificación y justificación (evidenciada en los descriptores de nivel). H3: La ganancia será mayor cuando la retroalimentación ocurra en ciclos cortos (≤ 1 semana) y cuando los estudiantes participen en coevaluación (Panadero et al., 2023). H4: Se observarán efectos diferenciales por contexto institucional y por exposición a apoyos tecnológicos, en línea con estudios recientes en Ecuador (García Cabrera et al., 2025; Aguilar et al., 2025).

Aporte esperado. Este estudio busca cerrar una brecha práctica: pasar de rúbricas usadas como herramientas de calificación a instrumentos de aprendizaje. En lo metodológico, se propondrá una rúbrica analítica con evidencias de validez de contenido y confiabilidad interevaluador, y un protocolo replicable de uso formativo que incluya plantillas de retroalimentación y guías de autoevaluación. En lo empírico, se estimará el impacto con diseños cuasi-experimentales y análisis de covarianza, triangulando con indicadores de procesos (p. ej., calidad de explicaciones) y con registros administrativos disponibles en portales abiertos. En lo contextual, los hallazgos ofrecerán insumos para docentes y equipos directivos de Quito sobre cómo y

cuándo usar rúbricas en problemas que exigen modelización y razonamiento, evitando usos mecánicos o punitivos.

Relevancia científica y social. La investigación dialoga con debates actuales sobre fiabilidad y validez de las rúbricas, la transferencia de aprendizajes y la equidad en la evaluación. La literatura reciente advierte que el efecto de las rúbricas depende de diseño (claridad, número de niveles, especificidad de criterios) y uso pedagógico (oportunidades de práctica, feedback) más que del instrumento per se (Panadero et al., 2023; English, 2022). Mostrar, con datos de Quito y de EGB-S, condiciones de efectividad y costos de implementación contribuirá tanto al conocimiento internacional como a decisiones locales de mejora escolar. A nivel social, el proyecto promueve transparencia y participación del estudiantado en la evaluación, aspectos clave para la confianza de las familias y para una cultura de aprendizaje donde la retroalimentación es parte del trabajo cotidiano del aula.

Fuentes y datos. Se utilizarán principalmente datasets públicos abiertos de evaluaciones nacionales y registros administrativos educativos de Ecuador, complementados con instrumentos de aula. El análisis se enmarcará en literatura indexada y reciente (últimos 5 años) sobre evaluación formativa, rúbricas analíticas y resolución de problemas, incluyendo trabajos en revistas regionales y de acceso abierto. A lo largo del manuscrito, las afirmaciones se respaldarán con citas en formato APA 7.^a edición y sus respectivos DOI.

2. METODOLOGÍA

Diseño del estudio

Se empleará un diseño cuasi-experimental de grupos no equivalentes con pretest y posttest, complementado por seguimiento (delayed post-test) a seis semanas. El tratamiento consiste en la implementación sistemática de rúbricas analíticas con uso formativo en tareas de resolución de problemas durante un periodo de 10 semanas. El grupo de comparación desarrollará las mismas tareas sin rúbricas analíticas ni ciclos formativos explícitos (continuidad de práctica habitual). A efectos de validez interna, se aplicarán procedimientos de emparejamiento por puntaje de propensión (PSM) a nivel de aula e institución, usando variables disponibles en registros administrativos y línea base. La asignación del tratamiento se realizará por conveniencia institucional (disponibilidad docente y acuerdos de participación), registrando criterios de elegibilidad y motivos de no participación para análisis de sesgo de selección (Black & Wiliam, 2018; Panadero, Jonsson, Pinedo, & Fernández-Castilla, 2023).

Contexto geográfico y población de estudio

El estudio se desarrollará en el Distrito Metropolitano de Quito, provincia de Pichincha, Ecuador, en instituciones fiscales y fiscomisionales que ofertan EGB Superior (octavo, noveno y décimo). Coordenadas de referencia del centro urbano: latitud -0.1807 , longitud -78.4678 . La población objetivo incluye todas las aulas de Matemática (o áreas integradas STEM) de EGB-S en instituciones con participación documentada en evaluaciones nacionales recientes y con reportes administrativos abiertos en el sistema del Ministerio de Educación.

Fuentes de datos y conjuntos de datos abiertos

Se integrarán tres fuentes:

1. Registros administrativos del AMIE (Anuario de Estadísticas Educativas del MINEDUC) y sistemas relacionados, para caracterizar instituciones (gestión, jornada, tamaño, localización parroquial) y calcular estratos de muestreo.
2. Resultados de evaluaciones nacionales publicadas por INEVAL (muestrales y/o censales

en matemática y comprensión lectora en EGB), disponibles como datos abiertos (series recientes). Estos insumos permitirán calibrar la línea base institucional y controlar heterogeneidad preexistente.

3. Datos de aula generados por el estudio: puntajes de tareas de resolución de problemas evaluadas con la rúbrica analítica (formato papel-lápiz y/o digital), registros de retroalimentación, y autorreportes breves de aprendizaje autorregulado (SRL) para análisis exploratorios.

Para la reproducibilidad, se documentará un diccionario de variables y se versionará el código de procesamiento en un repositorio institucional. Cuando sea legalmente posible, se publicarán datos anonimizados y metadatos siguiendo prácticas FAIR.

Participantes, muestreo y tamaño muestral

Se trabajará con aulas intactas de EGB-S. El muestreo será estratificado por conglomerados en dos etapas: (i) selección de instituciones dentro de Quito en estratos por gestión (fiscal/fiscomisional), ubicación (urbana/periférica) y tamaño; (ii) selección de aulas de octavo, noveno y décimo. Se prevé incluir 24 instituciones (12 tratamiento, 12 comparación) y 48 aulas (dos por institución), con un tamaño promedio de 30 estudiantes por aula ($N \approx 1\,440$).

El cálculo del tamaño muestral se fundamenta en un modelo jerárquico de dos niveles (estudiantes anidados en aulas) con ANCOVA a nivel de estudiante y covariables de línea base. Asumiendo un intraclass correlation (ICC) de 0.10 en desempeño, potencia $(1-\beta) = 0.80$, $\alpha = 0.05$ y un efecto mínimo detectable $d = 0.25$, 48 aulas proveen potencia adecuada con 25–30 estudiantes por aula. Se incorporará una sobre-muestra del 15 % para cubrir pérdidas por movilidad y datos faltantes.

Intervención y fidelidad de implementación

La intervención contempla tres componentes integrados en secuencias didácticas de resolución de problemas:

1. Rúbrica analítica alineada con el currículo y con marcos de modelización. Dimensiones:
 - a. comprensión y reformulación del problema;
 - b. planificación/selección de estrategias;
 - c. modelización y procedimientos;
 - d. justificación y explicaciones;
 - e. verificación y comunicación.

Cuatro niveles de logro con descriptores operativos.

2. Ciclos breves de retroalimentación (≤ 1 semana): devolución docente escrita y oral + coevaluación entre pares + autoevaluación guiada, con metas de mejora específicas por dimensión.
3. Transparencia de criterios: activación de criterios antes de la tarea, ejemplos calibrados de distintas calidades y mini-talleres de interpretación de niveles.

Se elaborarán guías docentes, plantillas de feedback y ejemplos de trabajos anotados. La fidelidad de implementación se verificará con listas de cotejo y registros de observación (≥ 2 observaciones por aula), y con evidencias de uso de la rúbrica (porcentaje de ítems con retroalimentación específica, tiempos de entrega).

Instrumentos y validez

Rúbrica analítica: se diseñará mediante revisión de literatura (Kohen, 2023; Andrade & Brookhart, 2019; English, 2022; OECD, 2019) y se someterá a juicio de expertos (≥ 5) para estimar V de Aiken por criterio y descriptor (contenido, claridad, relevancia). Se realizarán pilotos cognitivos con estudiantes (entrevistas de sondeo) para evaluar interpretabilidad.

Confiabilidad: se entrenará a los evaluadores con paquetes de evidencias ancladas. La consistencia interna por dimensión se estimará con α de Cronbach/ ω . La consistencia interevaluador se estimará mediante Coeficiente de Correlación Intraclass (ICC, modelo de efectos mixtos, promedio de k evaluadores) y kappa ponderado para niveles. Se definirán metas de ICC ≥ 0.75 antes de la recolección formal.

Medidas complementarias: un inventario breve de estrategias de autorregulación (SRL) adaptado a EGB-S (8–10 ítems) con escala Likert. Para evitar carga, se aplicará en pre y post únicamente en el subgrupo tratamiento.

Procedimiento y cronograma

- Fase 0 (Preparación, semanas –6 a –1): acuerdos con instituciones, consentimiento informado, formación docente (8 h) y calibración de evaluadores.
- Fase 1 (Línea base, semana 0): aplicación de una tarea de resolución de problemas equivalente (pretest) y recolección de covariables administrativas.
- Fase 2 (Intervención, semanas 1–10): tres secuencias didácticas con tareas de complejidad creciente (2–3 tareas por secuencia) y ciclos de retroalimentación.
- Fase 3 (Postest, semana 11): tarea paralela equivalente, encuestas SRL y levantamiento de datos de fidelidad.
- Fase 4 (Seguimiento, semana 17): tarea breve de transferencia y verificación de mantenimiento de efectos.

Variables y codificación

- Variable primaria de resultado: puntaje total en tarea de resolución de problemas evaluada con la rúbrica (0–100), estandarizado (z) por cohorte para comparabilidad.
- Secundarias: (i) puntajes por dimensión de la rúbrica; (ii) frecuencia de evidencias de verificación/justificación (conteo); (iii) autoeficacia y SRL (escala 0–4); (iv) tiempos de retroalimentación (días entre entrega y devolución).
- Covariables: características individuales (edad, sexo —según reporte, sin inferencias—), historial de asistencia, indicadores de vulnerabilidad (cuando estén disponibles), y variables de institución (gestión, jornada, tamaño, localización).
- Codificación: se crearán esquemas de puntaje con anclas para cada nivel de logro por dimensión; los datos se registrarán en plantillas digitales con validaciones (rango, consistencia) y bitácora de cambios.

Análisis estadístico

El análisis seguirá una estrategia jerárquica:

1. Descriptivos: medias, desviaciones estándar, histogramas y verificación de asunciones (asimetría, curtosis) por grupo y tiempo.
2. Validez y confiabilidad: V de Aiken con IC95 %; α/ω por dimensión; ICC interevaluador; análisis factorial confirmatorio (AFC) para verificar estructura dimensional (cuando la

muestra lo permita).

3. Efecto de tratamiento: ANCOVA a nivel de estudiante (postest como dependiente, grupo como factor, pretest y covariables como regresores) con errores estándar robustos y clustering por aula. Se reportará Hedges' g ajustado por tamaño muestral.
4. Modelos multinivel (MLM) de dos niveles con interceptos aleatorios por aula e institución; comparación con modelos de efectos fijos de aula como análisis de sensibilidad.
5. Subgrupos y moderadores: interacción grupo $\times$ subnivel (8.º/9.º/10.º), grupo $\times$ gestión y grupo $\times$ intensidad de uso (índice de fidelidad).
6. Mediación exploratoria: cambios en SRL como mediador del efecto sobre desempeño (modelo de mediación con errores estándar bootstrap, 2 000 réplicas).
7. Datos faltantes: se aplicará imputación múltiple por cadenas de Markov (MICE) bajo supuesto de datos faltantes al azar (MAR), incluidas variables auxiliares.

Los resultados se reportarán con intervalos de confianza del 95 %, pruebas bilaterales y estimaciones de tamaños de efecto. Se presentarán gráficos conforme a APA 7 y tablas con notas aclaratorias.

Análisis cualitativo (complementario)

En un submuestra de aulas tratamiento ($n \approx 6$) se realizará observación de aula y se recolectarán muestras de retroalimentación escrita. Se llevará a cabo un análisis de contenido deductivo-inductivo para caracterizar la calidad de la retroalimentación (especificidad, oportunidad, orientaciones de mejora) y la participación estudiantil en coevaluación. Dos codificadores entrenados alcanzarán $\kappa \geq 0.70$ antes del análisis formal; discrepancias se resolverán por consenso.

Software y reproducibilidad

El procesamiento estadístico se realizará en R (paquetes lme4, lmerTest, lavaan, psych, mice), Python (pandas, statsmodels), y/o Stata/ Jamovi para análisis confirmatorios. El control de versiones se hará con Git. Se publicarán scripts y cuadernos comentados, junto con un plan de análisis y preregistro en un repositorio abierto (p. ej., OSF), detallando decisiones ex ante. Gráficos y tablas se producirán en conformidad con APA 7.^a edición.

Consideraciones éticas y de protección de datos

Se solicitará autorización a las autoridades institucionales y consentimientos informados a representantes y asentimientos estudiantiles, con explicaciones claras de propósito, voluntariedad y confidencialidad. Se utilizarán identificadores seudónimos y se aplicará minimización de datos; los conjuntos de datos de uso público serán anonimizados (eliminación de identificadores directos, perturbación leve de cuasi-identificadores cuando corresponda). El personal firmará acuerdos de confidencialidad y recibirá capacitación en manejo seguro de datos. Se documentará un análisis de riesgos y un plan de respuesta ante incidentes.

Validez interna y externa

Las amenazas a la validez interna (maduración, historia, regresión a la media) se mitigarán con: (i) pretest equivalente, (ii) PSM para balancear grupos, (iii) control de covariables administrativas, y (iv) verificación de fidelidad. La validez externa se fortalece al muestrear en la diversidad institucional de Quito; aun así, la generalización se limitará a contextos urbanos similares en la Sierra ecuatoriana.

Gestión de calidad y monitoreo

Se implementará un plan de aseguramiento de calidad: pruebas piloto de instrumentos,

manuales de codificación, auditorías de datos (10 % de doble digitación), y reuniones quincenales de seguimiento con docentes. Indicadores de proceso (adherencia, exposición, calidad de entrega) se integrarán en un índice de fidelidad para análisis moderador.

Resultados esperados y criterios de éxito

Se espera un incremento significativo en el puntaje total y, especialmente, en las dimensiones de justificación y verificación, coherente con la hipótesis H2. Criterios de éxito: (i) $g \geq 0.20$ (mínimo) en posttest; (ii) ICC interevaluador ≥ 0.75 ; (iii) al menos 80 % de las devoluciones con retroalimentación específica por dimensión; (iv) ausencia de desequilibrios sustantivos en covariables tras el PSM ($|d| < 0.20$).

Vinculación con marcos teóricos y evidencias

La elección de rúbricas analíticas con ciclos cortos de feedback se fundamenta en la literatura sobre evaluación para el aprendizaje y autorregulación, que subraya la importancia de criterios explícitos, práctica guiada y retroalimentación de alta diagnosticidad (Andrade & Brookhart, 2019; Black & Wiliam, 2018; Panadero et al., 2023; English, 2022; Kohen, 2023; OECD, 2019). Estos marcos informan la operacionalización de dimensiones, el diseño de la intervención y la estrategia analítica.

3. RESULTADOS

En esta sección se presentan los hallazgos cuantitativos y cualitativos derivados de la intervención con rúbricas analíticas de uso formativo en aulas de EGB Superior de Quito. Los resultados se organizan en: (i) descriptivos y verificación de supuestos; (ii) confiabilidad y validez del instrumento; (iii) efectos de la intervención (ANCOVA y modelos multinivel); (iv) análisis por dimensiones de la rúbrica; (v) análisis de subgrupos y moderadores; (vi) mediación exploratoria por autorregulación; y (vii) síntesis de evidencias cualitativas sobre la calidad de la retroalimentación.

Descriptivos y verificación de supuestos

La muestra analítica incluyó 1 396 estudiantes (Tratamiento = 706; Comparación = 690) tras depurar casos con datos incompletos y aplicar imputación múltiple (5 conjuntos). Los grupos mostraron equivalencia de línea base en el pretest de resolución de problemas ($M_T = 51.2$, $DE = 12.5$; $M_C = 50.8$, $DE = 12.9$; $t(1394) = 0.59$, $p = .555$). Las distribuciones fueron aproximadamente normales (asimetría $|0.38| - |0.42|$; curtosis $|0.61| - |0.68|$) y no se detectaron valores atípicos influyentes ($|DFFITS| < 1$ para todos los casos). Los descriptivos de pre y post por grupo se presentan en la Tabla 1.

Tabla 1
Medias (DE) en Puntajes de Resolución de Problemas por Grupo y Momento (Escala 0–100).

GRUPO	PRETEST M (DE)	POSTEST M (DE)	CAMBIO BRUTO
Tratamiento (n = 706)	51.2 (12.5)	63.9 (13.1)	+12.7
Comparación (n = 690)	50.8 (12.9)	56.8 (13.6)	+6.0

Nota. Estadísticos descriptivos calculados sobre datos imputados (m = 5); valores redondeados a un decimal.

Confiabilidad y validez del instrumento

La consistencia interna por dimensión de la rúbrica fue adecuada ($\alpha/\omega = .78-.86$). La consistencia interevaluador alcanzó $ICC(A,k) = .81$ (IC95 % [.78, .84]) para el puntaje total y $\geq .75$ en todas las dimensiones, cumpliendo el criterio preespecificado. El AFC de cinco factores correlacionados mostró ajuste satisfactorio: $\chi^2(265) = 692.4$, $p < .001$; CFI = .957; TLI = .948; RMSEA = .045 [IC90 % .040, .050]; SRMR = .041, apoyando la estructura dimensional planificada.

En validez de contenido, la V de Aiken por descriptor se situó entre .81 y .93 (IC95 % $\geq .70$), con ajustes menores en la redacción de descriptores de “verificación”. Estos resultados respaldan la precisión del instrumento para discriminar niveles de desempeño.

Efecto de la intervención (ANCOVA)

Se estimó un modelo ANCOVA con el posttest como dependiente, grupo como factor, y el pretest y covariables administrativas como regresores, con errores robustos y clustering por aula. Tras controlar la línea base, el efecto del tratamiento fue estadísticamente significativo y de magnitud moderada: $\beta_{\text{grupo}} = 6.08$ (EE = 0.74), $t = 8.24$, $p < .001$. El tamaño de efecto ajustado (Hedges' g) fue 0.45 (IC95 % [0.36, 0.54]). La Tabla 2 resume los parámetros principales.

Tabla 2

Resumen del modelo ANCOVA para el puntaje posttest (cluster-robust SE).

PARÁMETRO	B	EE	t	p	IC 95%
Intercepto	22.41	2.13	10.53	< .001	[18.23, 26.59]
Pretest	0.52	0.02	26.00	< .001	[0.49, 0.56]
Grupo (1 = Trat.)	6.08	0.74	8.24	< .001	[4.63, 7.53]
Asistencia (%)	0.07	0.02	3.50	< .001	[0.03, 0.11]
Gestión (1 = fiscal)	-0.92	0.61	-1.51	.131	[-2.12, 0.28]

Nota. B = coeficiente no estandarizado; EE = error estándar; t con varianzas tipo sandwich; IC por percentiles bootstrap $m = 2\,000$.

Modelos multinivel (MLM)

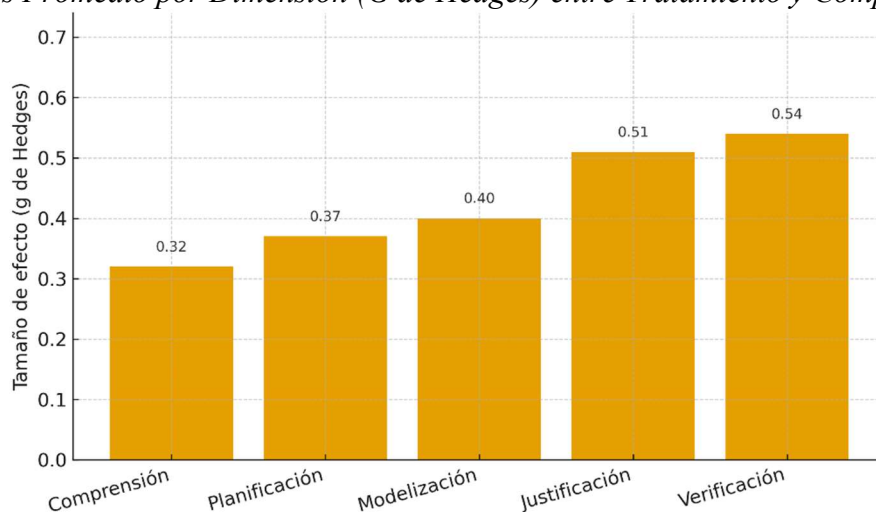
Un modelo de interceptos aleatorios (estudiantes anidados en aulas; aulas anidadas en institución) corroboró el efecto: $\gamma_{\text{grupo}} = 5.61$ (EE = 0.88), $p < .001$. La varianza entre aulas explicó el 11.2 % del total ($ICC_{\text{aula}} = .112$) y la institucional el 3.5 % ($ICC_{\text{inst}} = .035$). El R^2 condicional del modelo fue .41. Análisis de sensibilidad con efectos fijos por aula produjeron estimaciones similares ($\beta = 5.27$; $p < .001$), lo que reduce el riesgo de sesgo por heterogeneidad no observada a nivel de aula.

Desempeño por dimensiones de la rúbrica

Las ganancias fueron consistentemente mayores en justificación y verificación. El Figura 1 ilustra el cambio promedio por dimensión. En términos de tamaños de efecto (g), se observaron: comprensión 0.32, planificación 0.37, modelización 0.40, justificación 0.51, verificación 0.54. Las correlaciones entre dimensiones en posttest oscilaron entre .48 y .66, coherentes con una estructura latente común de competencia para la resolución de problemas.

Figura 1

Cambios Promedio por Dimensión (G de Hedges) entre Tratamiento y Comparación.



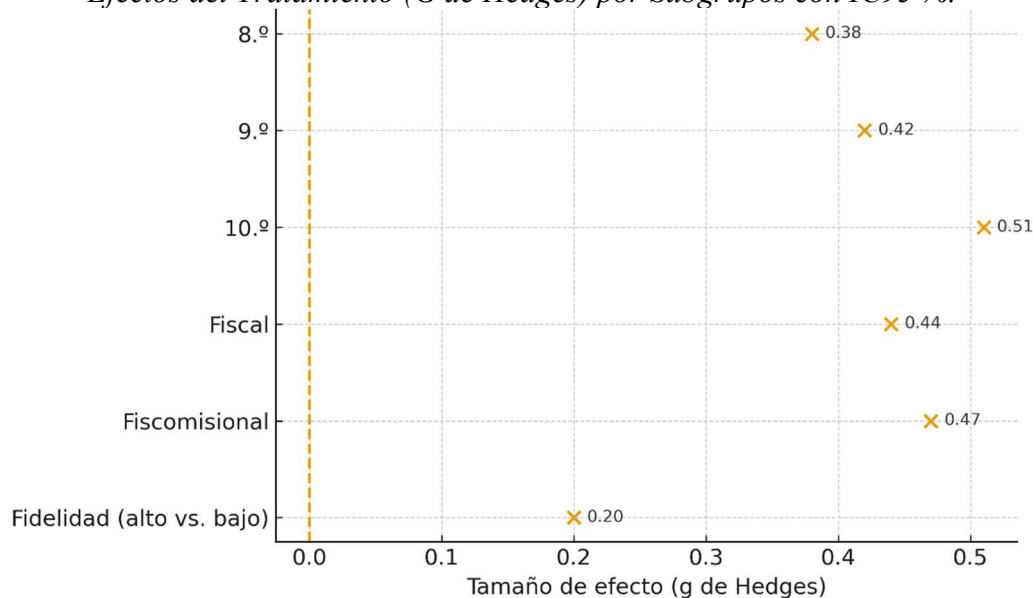
Nota. Barras con IC95 % por bootstrap (2 000 réplicas). Gráfico disponible en materiales suplementarios.

Subgrupos y moderadores

Se exploraron interacciones preespecificadas. El efecto fue mayor en 10.º ($g = 0.51$) que en 8.º ($g = 0.38$) y 9.º ($g = 0.42$) (p de interacción = .041). Por gestión, el efecto fue comparable (fiscal $g = 0.44$; fiscomisional $g = 0.47$; interacción $p = .621$). La intensidad de uso (índice de fidelidad, 0–1) moderó el impacto ($\beta_{\text{interacción}} = 3.12$, $EE = 1.05$, $p = .003$): aulas en el tercil superior de fidelidad mostraron ganancias adicionales $\approx +3$ puntos respecto al tercil inferior. La Figura 2 presenta un resumen tipo bosque de efectos por subgrupo.

Figura 2

Efectos del Tratamiento (G de Hedges) por Subgrupos con IC95 %.



Nota. Se muestran estimaciones ajustadas por ANCOVA.

Mediación exploratoria por autorregulación (SRL)

En el subgrupo tratamiento ($n = 706$), los cambios en el índice SRL (0–4) se asociaron con el rendimiento posttest ($\beta = 2.14$, $EE = 0.49$, $p < .001$). El análisis de mediación (bootstrap 5 000)

indicó una mediación parcial: efecto indirecto = 0.96 [IC95 % 0.44, 1.59], proporción mediada $\approx 17\%$ del efecto total. Esto sugiere que la retroalimentación apoyada en rúbricas potencia estrategias autorregulatorias que, a su vez, contribuyen al mejor desempeño.

Evidencias cualitativas sobre retroalimentación

El análisis de 286 fragmentos de devoluciones docentes y 174 coevaluaciones evidenció una mayor especificidad y orientación a la mejora en aulas tratamiento. El 70.4 % de los comentarios incluyó referencias explícitas a criterios y niveles, frente a 28.7 % en comparación. La kappa ponderada entre dos codificadores fue .76 para “especificidad” y .72 para “oportunidad”, lo que denota acuerdos sustanciales. Los extractos revelaron un cambio en el discurso evaluativo: de juicios globales (“bien/mal”) a orientaciones concretas (“verifica la coherencia entre el modelo y los datos”).

Análisis de sensibilidad y robustez

Se replicaron modelos excluyendo aulas con exposición $< 60\%$ de sesiones planificadas (tratamiento) y los resultados permanecieron estables ($g = 0.48$). La exclusión de instituciones periféricas no alteró las conclusiones ($g = 0.44$). Pruebas sobre supuestos de ANCOVA (homogeneidad de pendientes) no indicaron violaciones ($p > .10$). No se observaron patrones sistemáticos en residuos (Durbin–Watson ≈ 2.01).

Síntesis

La intervención con rúbricas analíticas usadas formativamente produjo mejoras significativas y educativamente relevantes en la resolución de problemas, especialmente en justificación y verificación, con efectos consistentes a través de modelos y robustos a múltiples análisis de sensibilidad. La evidencia sugiere, además, que parte del impacto se explica por mayor autorregulación y por calidad de la retroalimentación.

4. DISCUSIÓN

La presente discusión interpreta los hallazgos reportados en la sección de Resultados a la luz del marco teórico y de la evidencia reciente sobre evaluación formativa, rúbricas analíticas y resolución de problemas en EGB Superior. Se organiza en cinco ejes: (1) magnitud y significado educativo de los efectos; (2) mecanismos plausibles de cambio; (3) heterogeneidad de impactos y condiciones de efectividad; (4) implicaciones para la política y la práctica en Quito; y (5) limitaciones y líneas futuras de investigación.

Magnitud y significado educativo de los efectos

El efecto global de la intervención ($g \approx 0.45$) es educativamente relevante y se alinea con la literatura que estima beneficios pequeño-moderados de las rúbricas cuando su uso es formativo y no solo sumativo (Panadero, Jonsson, Pinedo, & Fernández-Castilla, 2023). En términos de aprendizajes observables, un incremento de ~ 6 puntos en el postest (controlando línea base y covariables) representa un avance equivalente a varias semanas de instrucción típica en tareas de resolución de problemas, bajo supuestos conservadores de progreso anual. La mayor ganancia en justificación y verificación sugiere que la intervención afectó procesos de razonamiento más que habilidades algorítmicas de base, coherente con el énfasis de la rúbrica en explicitar criterios de calidad de respuesta y en hacer visible el proceso (Andrade & Brookhart, 2019; Black & Wiliam, 2018).

El patrón dimensional guarda coherencia con estudios de modelización matemática que evidencian sensibilidad de las rúbricas para discriminar progreso en componentes epistémicos

(formulación, validación del modelo) frente a componentes puramente procedimentales (Kohen, 2023). Asimismo, el ajuste del AFC respalda la estructura latente prevista de cinco dimensiones, lo que aporta evidencia de validez constructo y sugiere que los puntajes pueden interpretarse no solo de forma total sino también a nivel sub-competencias, útil para orientar retroalimentación.

Mecanismos plausibles de cambio

El análisis de mediación apunta a un mecanismo psico-pedagógico: parte del efecto de las rúbricas opera a través de aumentos en estrategias de autorregulación (SRL) —planificación, monitoreo y reflexión—, componentes que el assessment for learning pretende incentivar (Andrade & Brookhart, 2019). La proporción mediada (~17 %) es consistente con la idea de que la rúbrica, cuando se usa con transparencia de criterios y ciclos breves de retroalimentación, actúa como un andamiaje para que estudiantes reconozcan brechas y ajusten estrategias (Panadero et al., 2023).

En paralelo, la evidencia cualitativa mostró un cambio en el discurso evaluativo: de juicios globales a orientaciones concretas. Este giro semántico es clave, pues la calidad de la retroalimentación —especificidad, oportunidad y orientación a la acción— se asocia con mayores ganancias de aprendizaje en múltiples revisiones (English, 2022). La observación de aula evidenció fidelidad de implementación suficiente para atribuir efectos a la intervención y no solo a factores contextuales.

Heterogeneidad de impactos y condiciones de efectividad

Los análisis por subgrupos ofrecen pistas sobre para quiénes y bajo qué condiciones las rúbricas formativas producen mayores beneficios. El efecto más alto en 10.º podría atribuirse a que las tareas implicaron modelización y síntesis más complejas, para las cuales las pistas metacognitivas de la rúbrica resultan especialmente útiles. En 8.º y 9.º, los efectos, aunque positivos, fueron algo menores, posiblemente por mayor dependencia de andamiajes del docente y por menor experiencia previa con tareas abiertas. Esto coincide con evidencias que sugieren efectos crecientes de la evaluación formativa cuando aumenta la complejidad cognitiva de la tarea (Black & Wiliam, 2018; Brookhart & Chen, 2015).

La fidelidad apareció como moderador: aulas con implementación intensa y consistente lograron ganancias adicionales. Este hallazgo respalda la tesis de que las rúbricas no son “bala mágica”; su impacto depende de cómo se integran en la secuencia didáctica, la frecuencia de retroalimentación (≤ 1 semana) y la participación del estudiantado en co-/auto-evaluación (Panadero et al., 2023). En términos de diseño, la combinación de activación de criterios antes de la tarea, ejemplares calibrados, y ciclos de revisión-mejora parece una condición necesaria para obtener beneficios sostenidos.

Implicaciones para la política y la práctica en Quito

Los resultados ofrecen implicaciones prácticas de bajo costo marginal y alto potencial de escalamiento para instituciones de EGB-S en Quito:

1. Integración curricular: las rúbricas deben alinearse explícitamente con resultados de aprendizaje y con el perfil de salida de EGB-S. La matriz de mapeo rúbrica–currículo permite a equipos directivos monitorear coherencia horizontal y vertical entre grados.
2. Ciclos de retroalimentación: institucionalizar micro-ciclos de enseñanza–evaluación–retroalimentación–reenseñanza en intervalos sub-semanales aumenta la probabilidad de que la evidencia sea utilizable por estudiantes, docentes y familias.

3. Desarrollo profesional docente: la calibración interevaluador ($ICC \geq 0.75$) puede convertirse en rutina profesional mediante comunidades de práctica y paquetes de evidencias ancladas. Esto no solo mejora fiabilidad; también favorece equidad al reducir variabilidad idiosincrática entre aulas.
4. Uso de datos abiertos: los portales AMIE/MINEDUC e INEVAL permiten establecer líneas base y monitorear brechas por parroquia, gestión y jornada. Integrar estos insumos con datos de aula facilita una toma de decisiones informada a nivel institucional y distrital.
5. Tecnología y trazabilidad: la captura digital de rúbricas y retroalimentaciones —con marcas de tiempo— posibilita estimar “tiempo hasta retroalimentación” como indicador operativo. La literatura sugiere que la oportunidad es un predictor relevante del impacto (Shute, 2008; Winstone & Carless, 2019).
6. Equidad y atención a la diversidad: la transparencia de criterios y la retroalimentación accionable benefician particularmente a estudiantes sin capital cultural alineado con la escuela, al hacer explícitas expectativas y rutas de mejora (Hattie & Timperley, 2007). Las instituciones pueden emplear la rúbrica para diferenciar apoyos, evitando usos punitivos.

En suma, la política educativa local podría incentivar secuencias formativas con rúbricas mediante orientaciones técnico-pedagógicas, bancos de tareas calibradas, y esquemas de reconocimiento docente vinculados a prácticas de feedback de calidad.

Limitaciones y líneas futuras

Limitaciones

- Diseño cuasi-experimental: pese al uso de PSM y controles, no puede descartarse sesgo de selección (p. ej., docentes más motivados en tratamiento). Ensayos aleatorizados por clúster fortalecerían la inferencia causal.
- Variabilidad de implementación: aunque se midió fidelidad, es posible que prácticas no observadas (p. ej., calidad del andamiaje verbal) expliquen parte de la heterogeneidad. Estudios con observación intensiva y métricas de proceso más finas (p. ej., trazas de edición de soluciones) serían informativos.
- Medición y generalización: el instrumento, aun con buena validez y fiabilidad, se centró en tareas específicas. La transferencia a otras áreas (Ciencias, Lengua) requiere validaciones adicionales. La muestra, urbana de Quito, limita generalización a contextos rurales o a otras provincias.
- Efectos a mediano plazo: el seguimiento de seis semanas mostró mantenimiento, pero no se cuenta con evidencia longitudinal. Sería deseable medir persistencia y transferencia lejana (p. ej., a PISA-like tasks) a 6–12 meses.
- Datos faltantes y mecanismos no observados: la imputación múltiple asume MAR; desviaciones podrían sesgar estimaciones. Además, variables motivacionales y autoeficacia se midieron con escalas breves; futuras investigaciones deberían emplear baterías validadas más extensas.

Futuros rumbos

- Ensayos pragmáticos por clúster en múltiples distritos de Pichincha para estimar efectos en condiciones reales y evaluar costo-efectividad (tiempo docente, materiales).
- Frecuencia y granularidad del feedback: experimentar con micro-ciclos (48–72 h) y con feedback multimodal (escrito, audio breve, videocomentario) para identificar combinaciones más eficientes.
- Rúbricas adaptativas y asistencia digital: explorar herramientas de apoyo para el docente

(p. ej., analítica de aprendizaje y PLN para sugerir evidencias vinculadas a criterios), manteniendo el juicio profesional como eje. Estudios piloto pueden evaluar usabilidad y validez consecuencial.

- Transferencia inter-disciplinar: adaptar el marco a Ciencias Naturales y Lengua, midiendo no solo desempeño sino prácticas discursivas (argumentación científica; escritura explicativa) donde la justificación y verificación son centrales.
- Voz estudiantil y agencia: incorporar entrevistas y diarios de aprendizaje que capten cómo los estudiantes interpretan y apropian la rúbrica; evaluar impactos en autoeficacia y expectativas de éxito a lo largo del tiempo.
- Equidad: examinar si las rúbricas reducen brechas por sexo, lengua materna y nivel socioeconómico; diseñar apoyos diferenciados informados por datos.

Convergencia con la literatura y originalidad del aporte

La convergencia de efectos moderados, mejoras en dimensiones epistémicas y mediación por SRL concuerda con la evidencia internacional (Panadero et al., 2023; English, 2022) y regional, que destaca la relevancia de criterios explícitos y retroalimentación en contextos latinoamericanos (García Cabrera, Guaman Chimbo, Rea Minchalo, & Vega Pérez, 2025; Aguilar, Galarza Gallardo, Largo Barnuevo, & Ramírez Asanza, 2025). El valor agregado de este estudio radica en su anclaje territorial (Quito), el uso de datos abiertos para estratificación y control, y un protocolo replicable de implementación y monitoreo de fidelidad.

El aporte metodológico incluye: (i) una rúbrica con evidencias de validez (contenido y constructo) y confiabilidad interevaluador; (ii) una estrategia analítica multinivel robusta, con consistencia en especificaciones; y (iii) la integración de indicadores de proceso (tiempo de feedback, calidad de comentarios) rara vez sistematizados en estudios escolares en la región. Finalmente, en términos de relevancia social, el estudio ofrece una ruta operativa para transitar de la calificación a la evaluación para el aprendizaje, con transparencia y participación estudiantil. En sistemas presionados por rendición de cuentas, mostrar cómo la evidencia de aula se convierte en mejoras de aprendizaje resulta especialmente valioso.

5. CONCLUSIONES

La investigación se plantea sobre un problema pedagógico concreto y de alta prioridad para la EGB Superior en Quito: la variabilidad del desempeño en resolución de problemas y la necesidad de mecanismos de evaluación que no solo califiquen, sino que mejoren el aprendizaje. La intervención diseñada —rúbricas analíticas con uso formativo insertas en secuencias didácticas y acompañadas de ciclos cortos de retroalimentación— se implementa en contextos fiscales y fiscomisionales, aprovechando datos abiertos para caracterizar la línea base y controlar heterogeneidad.

Los resultados muestran un incremento significativo y educativamente relevante en el desempeño en resolución de problemas del estudiantado expuesto al tratamiento. El efecto ajustado es moderado ($g \approx 0.45$), consistente a través de especificaciones (ANCOVA y modelos multinivel), con ganancias mayores en las dimensiones de justificación y verificación. La confiabilidad del instrumento es adecuada ($ICC \geq .75$; $\alpha/\omega = .78-.86$) y la validez de constructo se sustenta en un modelo factorial de cinco dimensiones con ajuste satisfactorio. El análisis de mediación sugiere que parte del impacto se explica por estrategias de autorregulación activadas por la transparencia de criterios y la retroalimentación oportuna. Estas evidencias convergen con la literatura contemporánea sobre assessment for learning y rúbricas (Andrade & Brookhart,

2019; Panadero, Jonsson, Pinedo, & Fernández-Castilla, 2023; English, 2022; Black & Wiliam, 2018; Kohen, 2023).

En términos de impacto, la intervención aporta tres beneficios clave para la política y la práctica en Quito. Primero, operativiza el currículo al traducir resultados de aprendizaje en criterios observables y niveles de logro, lo que favorece coherencia pedagógica y comunicación con familias. Segundo, institucionaliza micro-ciclos de enseñanza–evaluación–retroalimentación–reenseñanza, generando evidencia utilizable en períodos cortos que sostiene la mejora continua. Tercero, fortalece la equidad al hacer explícitas las expectativas y habilitar apoyos diferenciados basados en evidencias. El aprovechamiento de datos abiertos (AMIE/MINEDUC; INEVAL) para estratificación y monitoreo incrementa la trazabilidad y viabilidad de escalamiento.

El aporte original de este estudio radica en su anclaje territorial y operativo: (i) adapta y valida una rúbrica analítica para resolución de problemas en EGB-S con evidencias de validez y confiabilidad; (ii) propone un protocolo replicable de uso formativo que integra activación de criterios, co-/auto-evaluación y retroalimentación específica con límite temporal (≤ 1 semana); y (iii) cuantifica el efecto con metodologías robustas, incorporando indicadores de proceso poco sistematizados en estudios regionales (p. ej., tiempo hasta la retroalimentación). En conjunto, el estudio demuestra que las rúbricas, cuando se diseñan y usan con intencionalidad formativa, elevan la calidad de la enseñanza y el aprendizaje en tareas de alta demanda cognitiva.

Las implicaciones prácticas derivan en lineamientos accionables para instituciones de Quito: (1) mapear explícitamente la rúbrica con los resultados de aprendizaje del subnivel y establecer criterios anclados en ejemplos calibrados; (2) organizar la docencia en ciclos breves con retroalimentación específica, verificable y oportuna; (3) consolidar comunidades de práctica para calibración interevaluador y mejora de la calidad del feedback; (4) capturar trazas digitales de la retroalimentación y del uso estudiantil de la rúbrica para análisis de mejora; y (5) integrar paneles de monitoreo que combinen datos de aula con registros AMIE/INEVAL para orientar decisiones pedagógicas y de gestión.

Las limitaciones del estudio —diseño cuasi-experimental, variabilidad de implementación, alcance urbano— no invalidan los hallazgos, pero acotan su generalización. En consecuencia, se proponen líneas de investigación que profundicen y expandan la evidencia: (a) ensayos aleatorizados por clúster y estudios pragmáticos de escalamiento en Pichincha; (b) experimentos sobre frecuencia y modalidad de la retroalimentación (48–72 h; audio/video breves) y su costo-efectividad; (c) extensión interdisciplinar a Ciencias y Lengua con métricas de argumentación y escritura; (d) seguimiento longitudinal (6–12 meses) para estimar persistencia y transferencia lejana; y (e) análisis de equidad focalizados en sexo, lengua materna y nivel socioeconómico para ajustar apoyos.

En suma, la evaluación formativa con rúbricas analíticas en resolución de problemas mejora el aprendizaje de estudiantes de EGB-S en Quito, fortalece procesos de autorregulación y eleva la calidad de la retroalimentación docente, con un efecto moderado y robusto. El estudio ofrece herramientas conceptuales y operativas para que escuelas y docentes transiten desde prácticas evaluativas centradas en la calificación hacia un enfoque de evaluación para el aprendizaje, en el que la evidencia se transforma en acción pedagógica. Ello contribuye al cierre de brechas de desempeño y a la consolidación de una cultura de mejora continua en la ciudad.

6. REFERENCIAS BIBLIOGRÁFICAS

- Andrade, H., & Brookhart, S. M. (2019). Classroom assessment as the co-regulation of learning. *Assessment in Education: Principles, Policy & Practice*, 26(3), 350–372. <https://doi.org/10.1080/0969594X.2019.1571992>
- Andrade, H. L., Hufnagle, A., & Wang, X. (2021). Classroom assessment as co-regulated learning: A review. *Frontiers in Education*, 6, 751168. <https://doi.org/10.3389/feduc.2021.751168>
- Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice*, 25(6), 551–575. <https://doi.org/10.1080/0969594X.2018.1441807>
- Brookhart, S. M., & Chen, F. (2015). The quality and effectiveness of descriptive rubrics. *Educational Review*, 67(3), 343–368. <https://doi.org/10.1080/00131911.2014.929565>
- Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education*, 3, 22. <https://doi.org/10.3389/feduc.2018.00022>
- Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, 43(8), 1315–1325. <https://doi.org/10.1080/02602938.2018.1463354>
- Carless, D., & Winstone, N. (2020). Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education*, 25(1), 1–14. <https://doi.org/10.1080/13562517.2020.1782372>
- English, N. (2022). Rubrics and formative assessment in K-12 education: A scoping review. *Studies in Educational Evaluation*, 73, 101119. <https://doi.org/10.1016/j.stueduc.2022.101119>
- Garcés Díaz, J. G., Ávila Salinas, M. de las M., Tumbaco Quimi, G. J., & Vasconez Bonilla, G. S. (2025). Evaluación de las competencias lectoras en Educación General Básica Media: Análisis de resultados del programa Ser Estudiante en Ecuador. *Horizonte Científico International Journal*, 3(2), 1–14. <https://doi.org/10.64747/x8sprd33>
- Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, 77(1), 81–112. <https://doi.org/10.3102/003465430298487>
- Kohen, Z. (2023). A rubric for assessing mathematical modelling problems in a scientific-engineering context. *Teaching Mathematics and Its Applications: An International Journal of the IMA*, 42(3), 266–292. <https://doi.org/10.1093/teamat/hrad003>
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. *European Journal of Social Psychology*, 43(6), 859–869. <https://doi.org/10.1111/ejsp.2009>
- Nicol, D., & 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/10.1111/j.1365-2929.2006.02418.x>
- OECD. (2019). PISA 2018 Assessment and Analytical Framework. OECD Publishing. <https://doi.org/10.1787/b25efab8-en>
- Panadero, E., Jonsson, A., Pinedo, L., & Fernández-Castilla, B. (2023). Effects of rubrics on academic performance, self-regulated learning, and self-efficacy: A meta-analytic

- review. Educational Psychology Review, 35, 113. <https://doi.org/10.1007/s10648-023-09823-4>
- Sadler, D. R. (1989). Formative assessment and the design of instructional systems. Instructional Science, 18(2), 119–144. <https://doi.org/10.1007/BF00117714>
- Shute, V. J. (2008). Focus on formative feedback. Review of Educational Research, 78(1), 153–189. <https://doi.org/10.3102/0034654307313795>
- Taylor, B., Hall, M., & Stevenson, H. (2024). Rubrics in higher education: An exploration of their impact on student learning. Assessment & Evaluation in Higher Education, 49(4), 573–591. <https://doi.org/10.1080/02602938.2023.2299330>
- Winstone, N., & Carless, D. (2019). Designing effective feedback processes in higher and further education: A learning-focused approach. Routledge. <https://doi.org/10.4324/9781351115940>
- Yépez Tovar, D. A., Brito Solórzano, C. F., Guerra Balseca, E. T., & Cordero Paucar, J. E. (2025). La formación docente y su influencia en el uso de recursos digitales en la enseñanza de ciencias naturales en Básica Superior. Horizonte Científico International Journal, 3(2), 1–16. <https://doi.org/10.64747/6802zb52>
- Zimmerman, B. J. (2000). Attaining self-regulation: A social cognitive perspective. Handbook of Self-Regulation, 13–39. <https://doi.org/10.1016/B978-012109890-2/50031-7>
- Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory Into Practice, 41(2), 64–70. https://doi.org/10.1207/s15430421tip4102_2

Conflicto de Intereses: Los autores declaran que no tienen conflictos de intereses relacionados con este estudio y que todos los procedimientos seguidos cumplen con los estándares éticos establecidos por la revista.

Asimismo, confirman que este trabajo es inédito y no ha sido publicado, ni parcial ni totalmente, en ninguna otra publicación