

ENTRENAMIENTO METACOGNITIVO CON INTELIGENCIA ARTIFICIAL Y RENDIMIENTO EN BIOQUÍMICA: ESTUDIO CUASI-EXPERIMENTAL EN MEDICINA

AUTORES: MARINE PEDRAZA VILCHIS¹, JOSÉ ENRIQUE GARCÍA BOLL², MARÍA MONSERRAT PACHECO VÁZQUEZ³

¹ PROFESOR, ESCUELA DE MEDICINA, DIVISIÓN DE CIENCIAS DE LA SALUD, UNIVERSIDAD ANÁHUAC QUERÉTARO.

² COORDINADOR DE CIENCIAS BÁSICAS, ESCUELA DE MEDICINA, DIVISIÓN DE CIENCIAS DE LA SALUD, UNIVERSIDAD ANÁHUAC QUERÉTARO.

³ MÉDICO PASANTE DEL SERVICIO SOCIAL, ESCUELA DE MEDICINA, DIVISIÓN DE CIENCIAS DE LA SALUD, UNIVERSIDAD ANÁHUAC QUERÉTARO.

RESUMEN

Objetivo. Evaluar el efecto de una intervención semanal basada en resúmenes generados con inteligencia artificial (IA) y entrenamiento en estrategias metacognitivas sobre el rendimiento académico en bioquímica y la conciencia metacognitiva en estudiantes de primer semestre de la Licenciatura en Médico Cirujano de una universidad privada en Querétaro, México.

Metodología. Se realizó un estudio cuasi-experimental, longitudinal, con grupo de intervención ($n = 20$) y grupo de comparación no equivalente ($n = 87$). De una generación de 120 estudiantes, se invitó a participar a los inscritos en dos grupos de Bioquímica seleccionados por conveniencia; tras aplicar criterios de inclusión y exclusión, 20 completaron al menos el 50% + 1 de las sesiones. La intervención se desarrolló durante 13 semanas e incluyó cuatro componentes por sesión: lectura de un resumen generado con ChatGPT a partir de fuentes curadas por la docente, un mini-taller de estrategias metacognitivas, la aplicación práctica de la estrategia al tema semanal y una evaluación formativa (quiz) de 10 a 12 preguntas de opción múltiple generadas con IA y revisadas por la docente, sin valor en la calificación del curso. Se aplicaron 9 quizzes a partir de la semana 3. El rendimiento académico se midió con cuatro evaluaciones oficiales: pretest, parcial 1, intersemestral y examen colegiado final, el cual contenía 16 ítems ancla idénticos al pretest que no fueron utilizados en ninguna otra evaluación del curso.

La conciencia metacognitiva se evaluó con una versión adaptada al español del Inventario de Conciencia Metacognitiva (MAI) de 26 ítems (Likert 1-5) aplicada al inicio y al final del semestre. Se emplearon las pruebas U de Mann-Whitney para comparaciones entre grupos y Wilcoxon de rangos con signo para el análisis pareado del MAI, con un nivel de significancia de $\alpha = 0.05$.

Resultados. Ambos grupos fueron equivalentes en el pretest (intervención: 44.17 ± 7.40 ; control: 41.16 ± 10.06 ; $p = 0.176$). No se encontraron diferencias significativas en el parcial 1 ($p = 0.430$) ni en el intersemestral ($p = 0.343$). Sin embargo, el grupo de intervención obtuvo calificaciones significativamente superiores en el examen colegiado final (7.95 ± 1.16 vs. 7.36 ± 1.02 ; $p = 0.023$; d de Cohen = 0.56). Respecto al MAI ($n = 10$ pareados), el puntaje total aumentó significativamente tras la intervención (de 97.8 ± 9.1 a 105.3 ± 9.2 ; $p = 0.041$), con un incremento significativo en la subescala de Conocimiento de la cognición ($p = 0.012$); la subescala de Regulación de la cognición mostró tendencia positiva sin alcanzar significancia ($p = 0.125$).

Conclusiones. La intervención combinada de IA y entrenamiento metacognitivo produjo un efecto acumulativo positivo sobre la retención a largo plazo del conocimiento en Bioquímica y mejoró significativamente la conciencia metacognitiva de los participantes.

Estos hallazgos sugieren que integrar herramientas de IA como apoyo docente, junto con estrategias metacognitivas explícitas y práctica de recuperación, representa una estrategia viable y de bajo costo para fortalecer el aprendizaje en ciencias básicas médicas.

Palabras clave: educación médica, metacognición, inteligencia artificial, práctica de recuperación, Bioquímica, evaluación formativa.

PLANTEAMIENTO DEL PROBLEMA

La formación médica de pregrado se caracteriza por la necesidad de asimilar volúmenes elevados de información científica en periodos limitados, lo que exige a los estudiantes habilidades de aprendizaje autorregulado desde sus primeros semestres. La Bioquímica, por su naturaleza abstracta y su ubicación curricular al inicio de la carrera, representa uno de los principales retos académicos para los estudiantes de nuevo ingreso. De hecho, las tasas de reprobación y el bajo rendimiento en esta asignatura constituyen un problema recurrente en escuelas de Medicina a nivel nacional e internacional (Zheng y Sun, 2025). Muchos estudiantes ingresan sin estrategias de estudio efectivas, lo que genera frustración, bajo desempeño e incluso abandono temprano de la carrera.

En este contexto, la metacognición—entendida como la capacidad de planificar, monitorear y evaluar los propios procesos cognitivos—se ha consolidado como un predictor robusto del desempeño académico en ciencias de la salud. Chang et al. (2021) demostraron que la capacidad metacognitiva y la autorregulación explican el crecimiento longitudinal en exámenes de progreso en estudiantes de Medicina, mientras que Pucillo y Perez (2023) encontraron relaciones fuertes y significativas entre el puntaje del MAI y el promedio académico en estudiantes de profesiones sanitarias.

Más recientemente, una revisión sistemática de 59 estudios con más de 11,000 estudiantes universitarios reportó que el 91.5% de las intervenciones metacognitivas a nivel de curso produjeron al menos un hallazgo significativo de mejora en el rendimiento académico (Journal on Excellence in College Teaching, 2026). Estos hallazgos subrayan la importancia de incorporar la enseñanza explícita de estrategias metacognitivas en los primeros semestres de la formación médica.

El Inventario de Conciencia Metacognitiva (MAI), desarrollado por Schraw y Dennison (1994), es el instrumento más utilizado a nivel mundial para evaluar la metacognición en contextos educativos, con cerca de 5,000 citaciones en la literatura. Este inventario distingue dos dominios fundamentales: el Conocimiento de la cognición (lo que el estudiante sabe sobre sus propias estrategias y capacidades) y la Regulación de la cognición (la capacidad para planificar, monitorear y evaluar el propio aprendizaje). Estudios recientes han confirmado la validez y confiabilidad del MAI en poblaciones hispanohablantes de múltiples países latinoamericanos y España (Gutiérrez de Blume et al., 2024) y en estudiantes de ciencias de la salud (Omprakash et al., 2021), respaldando su aplicabilidad en el contexto del presente estudio.

Paralelamente, la práctica de recuperación—consistente en realizar evaluaciones frecuentes como estrategia de aprendizaje, también denominada "efecto de evaluación" o testing effect—ha demostrado efectos robustos sobre la retención a largo plazo. Meta-análisis recientes reportan tamaños de efecto medianos a grandes ($g = 0.51$; Adesope et al., 2017), y una revisión sistemática de estudios aplicados en contextos educativos reales encontró que el 57% de las investigaciones reportaron efectos medianos o grandes (Agarwal et al., 2021).

En educación médica, la práctica de recuperación resulta particularmente valiosa ya que proporciona retroalimentación objetiva sobre lo que el estudiante sabe y desconoce, facilitando así la autorregulación del aprendizaje (Nelson et al., 2024). Sin embargo, la elaboración de evaluaciones formativas de calidad es un proceso laborioso y costoso en términos de tiempo docente.

La irrupción de la inteligencia artificial generativa ha abierto nuevas posibilidades para superar esta barrera. Modelos de lenguaje como ChatGPT pueden generar preguntas de opción múltiple con propiedades psicométricas aceptables cuando se emplean prompts estructurados y revisión experta posterior. Stadler et al. (2024) presentaron una guía detallada para la ingeniería de prompts orientada a la generación de reactivos médicos, y una revisión narrativa reciente de 70 estudios concluyó que la IA generativa posee la capacidad de producir preguntas médicas de calidad cuando se utilizan prompts cuidadosamente diseñados, con aplicaciones que van desde la detección de errores en ítems hasta la facilitación de evaluaciones formativas y el aprendizaje personalizado (Qiu y Liu, 2025). Esta capacidad tecnológica permite ampliar la disponibilidad de material formativo sin multiplicar proporcionalmente la carga docente.

No obstante, la evidencia sobre intervenciones que combinen simultáneamente estrategias metacognitivas, práctica de recuperación e inteligencia artificial en ciencias básicas médicas es aún escasa. Si bien se reconoce que la metacognición y la práctica distribuida mejoran individualmente el rendimiento, no existen estudios que evalúen su efecto sinérgico en Bioquímica de primer semestre de Medicina en contextos latinoamericanos.

Este vacío resulta relevante porque las características demográficas, los estilos de aprendizaje y los contextos curriculares en universidades de México difieren sustancialmente de los reportados en la literatura anglosajona. Es importante abordar estos vacíos porque intervenciones exitosas en este nivel podrían tener implicaciones directas en la retención estudiantil, el desempeño en materias subsecuentes y la formación de hábitos de estudio que acompañen al futuro médico a lo largo de su carrera profesional.

PREGUNTA DE INVESTIGACIÓN.

¿Cuál es el efecto de una intervención semanal basada en resúmenes generados con inteligencia artificial y entrenamiento en estrategias metacognitivas sobre el rendimiento académico en Bioquímica y la conciencia metacognitiva en estudiantes de primer semestre de Medicina?

OBJETIVOS

Objetivo general

Evaluar el efecto de una intervención semanal basada en resúmenes generados con IA y entrenamiento metacognitivo sobre el rendimiento académico en Bioquímica y la conciencia metacognitiva en estudiantes de primer semestre de Medicina.

Objetivos específicos

1. Comparar el rendimiento académico (pretest, parcial 1, intersemestral y examen colegiado final con ítems ancla) entre el grupo de intervención y el grupo control.
2. Evaluar el cambio pre-post en la conciencia metacognitiva (MAI total y subescalas de Conocimiento y Regulación de la cognición) en los participantes del grupo de intervención.
3. Determinar si la intervención produce un efecto acumulativo a lo largo del semestre, evidenciado particularmente en la evaluación final que contenía ítems ancla del pretest.

SUJETOS Y METODOLOGÍAS

Se realizó un estudio cuasi-experimental, longitudinal, con grupo de intervención y grupo de comparación no equivalente, durante un semestre académico completo en una universidad privada de Querétaro, México. Este diseño es apropiado cuando la asignación aleatoria no es factible por razones logísticas y curriculares (Campbell y Stanley, 1963), situación habitual en educación médica donde los grupos ya están conformados institucionalmente.

PARTICIPANTES Y CRITERIOS DE SELECCIÓN

La población de estudio incluyó a los 120 estudiantes de primer semestre de la Licenciatura en Médico Cirujano inscritos en la materia de Bioquímica. Se invitó a participar a los estudiantes de dos grupos seleccionados por conveniencia (facilidad de horarios), correspondientes a aproximadamente la mitad de la generación. Los criterios de inclusión fueron: ser estudiante regular de primer semestre, estar inscrito en Bioquímica, presentar las cuatro evaluaciones del curso (pretest, parcial 1, intersemestral y colegiado final), participar en al menos el 50% + 1 de las sesiones de intervención y aceptar el aviso de privacidad. Los criterios de exclusión fueron: negarse a participar, no aceptar el aviso de privacidad o asistir a menos del 50% + 1 de las sesiones. Se eliminaron 12 estudiantes por no cumplir criterios generales (repetidores, bajas administrativas o evaluaciones incompletas). Inicialmente, aproximadamente 40 estudiantes aceptaron participar; sin embargo, dado que la participación era completamente voluntaria, la asistencia fue disminuyendo progresivamente a lo largo del semestre.

La muestra final quedó conformada por 20 estudiantes en el grupo de intervención que cumplieron el criterio mínimo de asistencia y 87 en el grupo control.

INTERVENCIÓN

La intervención se desarrolló en 13 sesiones semanales, cada una estructurada en cuatro componentes: (a) lectura y revisión de un resumen generado con IA (ChatGPT) a partir exclusivamente de capítulos y videos curados por la docente (20 minutos); (b) mini-taller metacognitivo donde se explicaba y practicaba una estrategia específica—como autoevaluación asistida por IA, mapas conceptuales elaborados a mano, flashcards generadas con IA o aprendizaje entre pares—(7 minutos); (c) aplicación de la estrategia metacognitiva al tema de la semana con reflexión breve (25 minutos); y (d) evaluación formativa semanal (quiz) de 10 a 12 reactivos de opción múltiple generados con IA y revisados por la docente, que no contaban para la calificación del curso (10 minutos). Las preguntas de los quizzes no fueron utilizadas en ninguna evaluación oficial del curso. Se aplicaron 9 quizzes en total a partir de la tercera semana. En la semana 6 se realizó una sesión especial de reflexión metacognitiva posterior a los exámenes intersemestrales. Para la generación de los quizzes se utilizó un prompt estructurado con las siguientes salvaguardas: 10 ítems de opción múltiple con 4 opciones y 1 correcta, cobertura de al menos 2 subtemas semanales, opciones de longitud similar, exclusión de ítems ancla y generación de metadatos por ítem (subtema, nivel Bloom, fuente y justificación).

El grupo control recibió docencia habitual sin sesiones adicionales de entrenamiento metacognitivo ni evaluaciones formativas semanales. Todos los estudiantes de ambos grupos presentaron las mismas cuatro evaluaciones oficiales del curso bajo condiciones idénticas.

INSTRUMENTOS Y VARIABLES

La variable principal de resultado fue el rendimiento académico, evaluado mediante cuatro exámenes: (1) pretest aplicado al inicio del semestre con contenido general de Bioquímica; (2) examen parcial 1; (3) examen intersemestral; y (4) examen colegiado final. Este último contenía 16 ítems ancla (20% del total) idénticos a los del pretest, seleccionados por dificultad media y buena discriminación, que no fueron utilizados en ninguna otra evaluación durante el curso. Esto permitió evaluar la ganancia en conocimiento de pretest a examen final sobre los mismos contenidos. La conciencia metacognitiva se evaluó mediante una versión adaptada al español del Inventario de Conciencia Metacognitiva (MAI) de Schraw y Dennison (1994), con 26 ítems en escala Likert de 5 puntos (1 = Nunca a 5 = Siempre), aplicada al inicio (semanas 0-1) y al final del semestre (semana 14). El MAI evalúa dos dimensiones: Conocimiento de la cognición (ítems 1-10, puntaje máximo 50) y Regulación de la cognición (ítems 11-26, puntaje máximo 80).

ANÁLISIS ESTADÍSTICO

Se realizó estadística descriptiva (medias y desviaciones estándar) para ambos grupos en cada evaluación.

Se verificó la normalidad de las distribuciones mediante la prueba de Shapiro-Wilk. Dado que algunas distribuciones del grupo control no cumplieron el supuesto de normalidad (parcial 1 e intersemestral: Shapiro-Wilk $p < 0.05$) y que los tamaños muestrales son asimétricos ($n = 20$ vs. $n = 87$), se utilizó la prueba U de Mann-Whitney para las comparaciones entre grupos independientes.

Para el análisis pre-post del MAI dentro del grupo de intervención ($n = 10$ participantes con datos pareados), se empleó la prueba de Wilcoxon de rangos con signo. Se calculó la d de Cohen como medida del tamaño del efecto para la comparación principal (examen colegiado final). Se estableció un nivel de significancia de $\alpha = 0.05$ para todas las pruebas. Los análisis se realizaron con GraphPad Prism.

CONSIDERACIONES ÉTICAS

La participación fue voluntaria y se obtuvo el consentimiento informado de todos los participantes. Los datos fueron anonimizados para el análisis. No hubo repercusiones académicas por participar o no en el protocolo, ya que las evaluaciones formativas (quizzes) no contaron para la calificación del curso. El banco de ítems ancla se mantuvo cerrado y no fue compartido con los estudiantes.

RESULTADOS Y ANÁLISIS

Equivalencia basal de los grupos

El análisis del pretest confirmó que ambos grupos partían de un nivel de conocimiento previo estadísticamente equivalente (grupo intervención: 44.17 ± 7.40 ; grupo control: 41.16 ± 10.06 ; $U = 1027.0$, $p = 0.176$), lo cual valida la comparabilidad de los grupos y justifica el uso de comparaciones directas entre ellos para las evaluaciones posteriores.

Rendimiento académico

La Tabla 1 presenta los resultados comparativos de las cuatro evaluaciones del curso entre el grupo de intervención y el grupo control.

Tabla 1.
Comparación del rendimiento académico entre grupo de intervención y grupo control

Evaluación	Intervención (n=20) Media ± DE	Control (n=87) Media ± DE	U de Mann-Whitney	p	d de Cohen	Sig.
Pretest (%)	44.17 ± 7.40	41.16 ± 10.06	1027.0	0.176	0.31	No
Parcial 1 (%)	78.57 ± 14.66	74.06 ± 18.55	968.5	0.430	0.25	No
Intersemestral	7.50 ± 1.38	7.20 ± 1.38	989.0	0.343	0.22	No
Colegiado final	7.95 ± 1.16	7.36 ± 1.02	1155.5	0.023*	0.56	Si

Nota: * p < 0.05, DE = Desviación estándar. Las calificaciones del pretest y parcial 1 se expresan en porcentaje; intersemestral y colegiado final en escala de 0 a 10.

Como se observa en la Tabla 1, no se encontraron diferencias estadísticamente significativas entre los grupos en el pretest (p = 0.176), lo cual confirma la equivalencia basal. Tampoco se hallaron diferencias en el parcial 1 (p = 0.430) ni en el intersemestral (p = 0.343). Sin embargo, en el examen colegiado final—que incluía las 16 preguntas ancla del pretest que no fueron utilizadas en ninguna otra evaluación durante el curso—el grupo de intervención obtuvo calificaciones significativamente superiores (7.95 ± 1.16 vs. 7.36 ± 1.02; p = 0.023) con un tamaño de efecto mediano (d de Cohen = 0.56). Este patrón de resultados sugiere un efecto acumulativo de la intervención que se manifestó particularmente en la evaluación que medía retención a largo plazo de los contenidos evaluados inicialmente en el pretest.

Conciencia metacognitiva (MAI)

De los 20 participantes del grupo de intervención, 10 completaron tanto el MAI al inicio como al final del protocolo, permitiendo un análisis pareado.

Los 10 restantes no completaron el cuestionario posterior por inasistencia el día de aplicación. La Tabla 2 presenta los resultados del cambio pre-post en conciencia metacognitiva.

Tabla 2.
Cambio pre-post en conciencia metacognitiva (MAI) del grupo de intervención (n = 10 pareados)

Dimensión MAI	PRE Media ± DE	POST Media ± DE	Diferencia Media ± DE	p (Wilcoxon)	Sig.
MAI Total (26-130)	97.80 ± 9.10	105.30 ± 9.24	+7.50 ± 9.14	0.041*	Si
Conocimiento de la cognición (10-50)	37.00 ± 2.94	40.70 ± 3.56	+3.70 ± 2.95	0.012*	Si
Regulación de la cognición (16-80)	60.80 ± 6.55	64.60 ± 6.92	+3.80 ± 7.18	0.125	No

Nota: * p < 0.05, DE = Desviación estándar. Los rangos teóricos de cada escala se indican entre paréntesis.

El puntaje total del MAI mostró un incremento estadísticamente significativo tras la intervención (p = 0.041). Al analizar las subescalas, el Conocimiento de la cognición presentó un aumento significativo (de 37.00 a 40.70; p = 0.012), lo que indica que los estudiantes desarrollaron mayor conciencia sobre sus propias estrategias, fortalezas y debilidades de aprendizaje. En contraste, la subescala de Regulación de la cognición mostró una tendencia positiva (de 60.80 a 64.60) que no alcanzó significancia estadística (p = 0.125), lo cual puede atribuirse tanto al menor tamaño de la muestra pareada como a que las habilidades regulatorias requieren más tiempo de práctica para consolidarse.

DISCUSION

El hallazgo principal de este estudio es que la combinación de resúmenes generados con IA, entrenamiento metacognitivo explícito y práctica de recuperación mediante evaluaciones formativas frecuentes produjo

una mejora significativa en el rendimiento académico a largo plazo y en la conciencia metacognitiva de estudiantes de primer semestre de Medicina cursando Bioquímica.

El patrón temporal de los resultados es particularmente revelador. Las diferencias entre grupos no fueron significativas en las evaluaciones intermedias (parcial 1 e intersemestral) pero sí en el examen colegiado final, que incluía ítems ancla del pretest. Este patrón es consistente con la literatura sobre el efecto de la práctica de recuperación (testing effect), que favorece la retención a largo plazo más que el desempeño inmediato. Adesope et al. (2017) reportaron en su meta-análisis que el efecto de la práctica de recuperación es especialmente pronunciado cuando la evaluación se realiza en intervalos largos. Asimismo, los efectos del entrenamiento metacognitivo tienden a manifestarse gradualmente conforme los estudiantes internalizan las estrategias y las aplican de manera autónoma. El tamaño del efecto encontrado ($d = 0.56$) se clasifica como mediano según la convención de Cohen y es congruente con los resultados reportados por Yang et al. (2021) para intervenciones basadas en práctica de recuperación en contextos educativos reales ($g = 0.33$) y con la revisión sistemática más reciente que encontró que más del 91% de las intervenciones metacognitivas universitarias producen mejoras significativas en el rendimiento (Journal on Excellence in College Teaching, 2026).

El incremento en la conciencia metacognitiva medida por el MAI es coherente con hallazgos previos en educación médica.

Shah y Modna (2022) reportaron correlaciones significativas entre el puntaje del MAI y el rendimiento académico en fisiología, particularmente en las subescalas de conocimiento declarativo y evaluación. Chang et al. (2021) demostraron que mientras el pensamiento crítico explica el desempeño inicial en exámenes de progreso, son la metacognición y la autorregulación las que explican el crecimiento longitudinal. En nuestro estudio, el mayor incremento se observó en Conocimiento de la cognición, reflejando que los estudiantes tomaron mayor conciencia de sus estrategias y capacidades de aprendizaje. Que la Regulación de la cognición no alcanzara significancia estadística es esperable, dado que las habilidades regulatorias—como la planificación de sesiones de estudio, el monitoreo en tiempo real de la comprensión y la evaluación posterior al aprendizaje—requieren mayor tiempo y práctica deliberada para consolidarse.

El uso de IA generativa como herramienta de apoyo docente para la elaboración de resúmenes y evaluaciones formativas demostró ser una estrategia viable y eficiente. Stadler et al. (2024) documentaron que las preguntas generadas por modelos de lenguaje pueden alcanzar propiedades psicométricas comparables a las elaboradas por expertos cuando se acompañan de revisión docente estructurada. Qiu y Liu (2025), en su revisión de 70 estudios, concluyeron que la IA generativa puede producir reactivos médicos de calidad con aplicaciones que van desde la evaluación formativa hasta el aprendizaje personalizado.

En el presente protocolo, todos los materiales generados por IA fueron curados y revisados por la docente antes de su aplicación, lo cual constituye una salvaguarda fundamental para garantizar la calidad y precisión del contenido, en concordancia con las recomendaciones de la literatura.

Estos resultados aportan evidencia desde un contexto latinoamericano, donde la investigación sobre metacognición en educación médica básica es aún incipiente. La validación reciente del MAI en poblaciones hispanohablantes de múltiples países (Gutiérrez de Blume et al., 2024) respalda la pertinencia del instrumento utilizado y la generalizabilidad potencial de los hallazgos a contextos similares en la región.

CONCLUSION

La intervención basada en resúmenes generados con IA y entrenamiento metacognitivo produjo una mejora estadísticamente significativa en el rendimiento académico en el examen colegiado final ($p = 0.023$, $d = 0.56$) y en la conciencia metacognitiva global medida por el MAI ($p = 0.041$), particularmente en la subescala de Conocimiento de la cognición ($p = 0.012$). El efecto acumulativo observado—ausencia de diferencias significativas en evaluaciones intermedias pero diferencia significativa en la evaluación final con ítems ancla—sugiere que la intervención favoreció la retención a largo plazo del conocimiento, un objetivo central en la formación médica.

El estudio presenta limitaciones que deben considerarse para la interpretación de los resultados. Primero, la muestra del grupo de intervención es pequeña ($n = 20$) y fue conformada por participación voluntaria, lo cual introduce un potencial sesgo de autoselección: es posible que los estudiantes más motivados o con mayor interés académico fueran quienes decidieron participar y permanecer en el protocolo. Segundo, la tasa de desgaste fue considerable—de aproximadamente 40 participantes iniciales a 20 que completaron el criterio mínimo de asistencia—, lo cual limita la generalización de los resultados y refleja los desafíos inherentes a las intervenciones voluntarias. Tercero, solo 10 de los 20 participantes completaron el MAI al final del protocolo, reduciendo el poder estadístico del análisis pareado. Cuarto, la ausencia de un diseño aleatorizado impide establecer relaciones causales definitivas, aunque la equivalencia confirmada en el pretest mitiga parcialmente esta limitación.

Futuras investigaciones deberían considerar muestras más grandes con diseños aleatorizados por conglomerados, que permitirían mayor control sobre las variables confusoras. Asimismo, sería valioso implementar mediciones longitudinales del impacto sobre materias subsiguientes para evaluar la transferencia de las habilidades metacognitivas adquiridas, así como incorporar componentes cualitativos que exploren las percepciones y experiencias de los estudiantes con las estrategias empleadas.

También se recomienda investigar si intervenciones similares implementadas de manera curricular—es decir, integradas en el curso regular y no como actividad voluntaria—reducen el sesgo de autoselección y mejoran la adherencia de los participantes.

En síntesis, la integración de inteligencia artificial como herramienta de apoyo docente, combinada con el entrenamiento explícito en estrategias metacognitivas y la práctica de recuperación, representa una estrategia prometedora, viable y de bajo costo para mejorar el aprendizaje en ciencias básicas médicas.

Formar estudiantes que aprendan a aprender constituye una inversión pedagógica cuyo impacto trasciende la asignatura y acompaña al futuro médico a lo largo de su formación y ejercicio profesional.

REFERENCIAS

1. Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research*, 87(3), 659–701. <https://doi.org/10.3102/0034654316689306>
2. Agarwal, P. K., Nunes, L. D., & Blunt, J. R. (2021). Retrieval practice consistently benefits student learning: A systematic review of applied research in school classrooms. *Educational Psychology Review*, 33, 1409–1453. <https://doi.org/10.1007/s10648-021-09595-9>
3. Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin. <https://archive.org/details/experimentalquas00camp>
4. Chang, C., Colón-Berlinger, M., Mavis, B., Laird-Fick, H. S., Parker, C., & Solomon, D. (2021). Medical student progress examination performance and its relationship with metacognition, critical thinking, and self-regulated learning strategies. *Academic Medicine*, 96(2), 278–284. <https://pubmed.ncbi.nlm.nih.gov/33003039/>
5. Gutiérrez de Blume, A. P., et al. (2024). Psychometric properties of the Metacognitive Awareness Inventory (MAI) in Spanish-speaking university students. *Metacognition and Learning*, 19, 367–395. <https://metacog-global.com/wp-content/uploads/2025/02/Gutierrez-de-Blume-et-al.-in-press-Standardization-of-MAI-in-Spanish-5.pdf>
6. Nelson, A., Mohammed, A., An, A., et al. (2024). Testing the effects of individual residents' retrieval practice on standardized examination scores. *Medical Science Educator*, 34, 647–652. <https://doi.org/10.1007/s40670-024-02031-x>
7. Omprakash, A., Kumar, R., Manickam, A., et al. (2021). Validation of Metacognitive Awareness Inventory from a private medical university in India. *Journal of Education and Health Promotion*, 10, 324. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8552251/>
8. Pucillo, E. M., & Perez, G. (2023). Metacognition and self-regulation influence academic performance in occupational and physical therapy students. *Journal of Occupational Therapy Education*, 7(1), Article 4. <https://doi.org/10.26681/jote.2023.070104>
9. Qiu, Y., & Liu, C. (2025). Capable exam-taker and question-generator: The dual role of generative AI in medical education assessment. *Global Medical Education*, 10, 1–8. <https://doi.org/10.1515/gme-2024-0021>
10. Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology*, 19(4), 460–475. <https://doi.org/10.1006/ceps.1994.1033>
11. Shah, D. K., & Modna, Y. (2022). The impact of medical students' metacognitive awareness level on their academic performance. *International Journal of Research in Medical Sciences*, 10(11), 2476–2483. <https://doi.org/10.18203/2320-6012.ijrms20222829>
12. Stadler, A., et al. (2024). Crafting medical MCQs with generative AI: A how-to guide on leveraging ChatGPT. *GMS Journal of Medical Education*, 41(2), Doc18. <https://doi.org/10.3205/zma001673>
13. Zheng, B., Sun, T. Self-regulated learning and learning outcomes in undergraduate and graduate medical education: a meta-analysis. *Teach Teach Educ*. 2025;157:104917.