ARTÍCULO ORIGINAL
doi: 10.35366/123064

Cirujano General
Enero-Marzo 2026
Vol. 48, núm. 1 / p. 7-17

Implementación y validación de la Evaluación Adaptativa Computarizada para la certificación en cirugía por el Consejo Mexicano de Cirugía General: de la prueba de concepto a la validación definitiva

Implementation and validation of a Computerized Adaptative Testing for general surgery certification by the Mexican Board of General Surgery: from the proof of concept to its definitive validation

Eduardo Prado,* Rafael Humberto Pérez-Soto,* Karina Sánchez-Reyes,* Elena López-Gavito,* Armando Hernández-Cendejas,* Jorge Kobeh,* David Velázquez-Fernández*

Palabras clave: prueba adaptativa computarizada, validación, certificación, cirugía, Consejo Mexicano de Cirugía General.

Keywords: *computerized adaptative testing, validation, certification, surgery, Mexican Board of General Surgery.*

RESUMEN

Introducción: la certificación en cirugía general exige el cumplimiento de tres fases (prueba piloto, implementación, validación) diseñadas para garantizar los estándares de competencia profesional en nuestro país. La EAC es una herramienta que permite la aplicación optimizada de reactivos en forma dinámica con base en el desempeño de los sustentantes en tiempo real, optimizando el tiempo, número de reactivos, equipo de cómputo y el recurso humano, además de reducir la fatiga asociada con evaluaciones prolongadas. **Objetivo:** implementar y analizar la eficiencia de la EAC para determinar la suficiencia académica de los cirujanos generales que aplican para la certificación del CMCG. **Material y métodos:** utilizamos una metodología de tres fases: prueba piloto, de implementación y de validación en cohortes independientes. La primera incluyó a 322 sustentantes, la segunda se aplicó en 569 sustentantes independientes, en los que se comparó la prueba convencional de entrenamiento con la EAC, mientras que, para la fase de validación, se utilizaron 1,194 sustentantes con dos tipos de EAC diferentes que fueron sorteadas aleatoriamente de un *pool* de 1,200 reactivos con diferentes niveles de dificultad, los cuales fueron contrastados con el examen de entrenamiento de cada sustentante de manera pareada y global. Para el análisis estadístico utilizamos el software IBM® SPSS® Statistics v26, considerando significativo cualquier valor de $p < 0.05$ para una prueba de hipótesis de

ABSTRACT

Introduction: *the board certification in general surgery requires the full compliance of 3 phases (pilot test, implementation, validation), designed to guarantee the quality standards of professional competence in our country. CAT is a tool that allows the optimized application of items in a dynamic way based on their performance in real time, optimizing the time, number of items, computer equipment and human resources plus reducing the potential fatigue associated with longer periods of time.* ***Objective:*** *to implement and analyze the efficiency of CAT to determine the academic proficiency of general surgeons who apply for the CMCG certification.* ***Material and methods:*** *we used a three-phase methodology: a pilot test, with the test implementation and a final validation in independent cohorts. The first included 322 supporters, the second was applied to 569 independent applicants in which the results from our conventional test were contrasted with the CAT, while for the validation phase 1,194 applicants with two different CATs were analyzed with randomly drawn items obtained from a pool of 1,200 items with different levels of difficulty that were contrasted with the conventional test of each sustainer in a paired and global analysis. For statistical analysis we used IBM SPSS® Statistics® v26 software considering any p value < 0.05 as statistically significant for a two-tie hypothesis test.* ***Results:*** *RACs resulted similar between*

* Consejo Mexicano de Cirugía General A.C.

Recibido: 15/12/2025
Aceptado: 22/01/2026

Citar como: Prado E, Pérez-Soto RH, Sánchez-Reyes K, López-Gavito E, Hernández-Cendejas A, Kobeh J et al. Implementación y validación de la Evaluación Adaptativa Computarizada para la certificación en cirugía por el Consejo Mexicano de Cirugía General: de la prueba de concepto a la validación definitiva. Cir Gen. 2026; 48 (1): 7-17. https://dx.doi.org/10.35366/123064

dos colas. **Resultados:** las RAC resultaron muy cercanas entre la prueba convencional y la EAC (30.32 ± 3.89 vs 30.67 ± 6.83 respectivamente), y por sustentante, con una delta de 0.35 ± 7.26 (p = 0.38) y un EEM = 0.40. Para la fase de implementación, se demostró un tiempo menor entre los grupos (117.2 ± 12.6 vs 72.05 ± 18.2 minutos, respectivamente), con un promedio menor de reactivos contestados (100.56 ± 15.64 vs 193.69 ± 15.78) y con una calificación promedio similar (50.84 ± 7.9 vs 54.53 ± 7.2), pero capaz de discernir entre AAA y BAA (51.34 ± 7.5 vs 36.42 ± 5.5). En la fase de validación, no existió diferencia en tiempo, número de reactivos contestados ni en calificación, pero se mantuvo la capacidad de discriminación para discernir entre AAA y BAA (p < 0.0001). **Conclusiones:** la EAC demostró ser una herramienta no sólo factible, sino válida y con una mayor eficiencia que el método convencional, al lograr resultados equiparables, con una reducción significativa en el tiempo y número de reactivos utilizados, manteniendo su capacidad para discriminar rápidamente a los sustentantes con mayor asertividad.

the conventional test and CAT (30.32 ± 3.89 vs 30.67 ± 6.83 respectively) with a delta of 0.35 ± 7.26; p = 0.38 and SEM = 0.40. For the implementation phase, a shorter time was documented between groups (117.2 ± 12.6 vs 72.05 ± 18.2 minutes, respectively), with a lower average of answered items (100.56 ± 15.64 vs 193.69 ± 15.78), but with a similar average score (50.84 ± 7.9 vs 54.53 ± 7.2), but able to discern between AAA vs BAA applicants (51.34 ± 7.5 vs 36.42 ± 5.5). In the validation phase, there was no difference in time, number of items answered, or difficulty grade, but the ability to discriminate between AAA and BAA applicants was maintained (p < 0.0001). ***Conclusions:*** *setting CAT proved to be a useful tool not only feasible, but also valid with greater efficiency than the conventional tests, achieving comparable results, but with a significant reduction in the time and number of items utilized, maintaining its ability to quickly discriminate the applicants with greater academic assertiveness.*

Abreviaturas:
AAA = alta asertividad académica
BAA = baja asertividad académica
CMCG = Consejo Mexicano de Cirugía General
EAC = Evaluación Adaptativa Computarizada
EEM = error estándar de la media
PUEM = Plan Único de Especialidades Médica
RAC = Razón de la Calificación Aritmética
TRI = Teoría de Respuesta al Ítem

INTRODUCCIÓN

El Consejo Mexicano de Cirugía General (CMCG), fundado en 1977, es el único organismo reconocido para la certificación de los médicos especialistas en cirugía general formados en México, así como de aquellos cirujanos con entrenamiento en el extranjero que buscan ejercer legalmente en nuestro país.

La certificación de primera vez en México exige la aprobación de tres fases de evaluación complementarias, diseñadas para garantizar que el cirujano general cumpla con los estándares nacionales de formación y competencia profesional. Estas incluyen: 1) evaluación curricular de los documentos oficiales que acreditan la trayectoria universitaria, hospitalaria, académica y formativa del candidato; 2) valoración de la suficiencia académica en el dominio cognitivo de las diversas áreas del conocimiento quirúrgico mediante un examen computarizado, y 3) evaluación de competencias clínicas y transversales por parte de pares profesionales, orientada a determinar la capacidad del sustentante para aplicar conocimientos y habilidades en contextos propios comunes de la práctica quirúrgica.

Adicionalmente, el CMCG realiza anualmente exámenes de recertificación para cirujanos con certificación previa, así como un examen de entrenamiento dirigido a los médicos residentes de tercer año de la especialidad, este último con el propósito de identificar áreas de oportunidad y preparar al futuro sustentante para el examen oficial de certificación.[1]

La Evaluación Adaptativa Computarizada (EAC) es una herramienta que permite realizar evaluaciones de manera más eficiente mediante la aplicación optimizada de reactivos o ítems a cada sustentante, seleccionados de forma dinámica con base en su desempeño en tiempo real. Este enfoque evita utilizar la misma cantidad y tipo de ítems para todos los sustentantes, lo que se traduce en un uso más racional del tiempo, del equipo de cómputo y del recurso humano durante el proceso de evaluación. Además, reduce la probabilidad de fatiga en los sustentantes asociada con evaluaciones prolongadas.[2]

La EAC se sustenta en la Teoría de Respuesta al Ítem (TRI), la cual establece un modelo matemático que permite estimar la probabilidad de que un individuo obtenga un resultado satisfactorio —en este caso, demostrar la suficiencia

académica— a partir de sus respuestas correctas o incorrectas.[3] Esta estimación deriva en un inicio del desempeño del sustentante ante un ítem asignado aleatoriamente, cuya dificultad después aumenta o disminuye en función del acierto o error en el reactivo previo. De esta manera, es posible determinar con alto grado de precisión la capacidad cognitiva del sustentante, al identificar el nivel de dificultad de los ítems que puede resolver adecuadamente.[4]

La primera descripción de la EAC tiene su origen en 1905 con Binet y Simon, quienes desarrollaron el Binet IQ Test para evaluar el coeficiente intelectual (IQ) en la población pediátrica.[5] En la actualidad su aplicación se ha extendido, especialmente en el campo de las ciencias de la salud. Sólo por mencionar algunos ejemplos: el examen de certificación del *American Society of Clinical Pathology* (ASCP), el examen para obtener la licencia del *National Council of State Boards of Nursing* (NCSBN) y el *American Pharmacist Licensure Examination* (NAPLEX®), entre otros, utilizan esta metodología en sus instrumentos de evaluación.

En México, hasta la fecha no existe evidencia científica publicada sobre la utilización de la EAC como herramienta tecnológica de evaluación en los procesos de certificación de médicos especialistas.

Objetivos del estudio

Implementar una EAC en los exámenes de certificación en cirugía general del CMCG para su validación y estandarización; demostrar la precisión, exactitud y validez que tiene esta prueba respecto al estándar de oro que es la prueba convencional de 200 reactivos de opción múltiple cerrada, que ya está validada, y aplicarla en una cohorte independiente para verificar si vuelve a tener la misma precisión, exactitud y confiabilidad.

MATERIAL Y MÉTODOS

Para cumplir con los objetivos de este estudio y responder a nuestra pregunta de investigación, se realizó una metodología de investigación de tres fases: 1) prueba piloto preliminar, para ajustes y prueba de concepto, 2) prueba de implementación y 3) prueba de validación en una cohorte independiente.

Prueba piloto

Para esta prueba, se invitaron a 322 sustentantes, a los que se les solicitó que aplicaran para la prueba convencional y para la prueba adaptativa por computadora. En relación con la primera prueba, se utilizaron 200 reactivos de opción múltiple cerrada de las diferentes áreas de las ciencias quirúrgicas, que incluye el Plan Único de Especialidades Médicas (PUEM) y otros exámenes estándares de la especialidad de cirugía general. En relación con la prueba adaptativa por computadora, se utilizó un banco de reactivos ya validados en las pruebas convencionales, los cuales fueron sorteados de manera aleatoria por sustentante (se sortearon tanto los reactivos como las 4 opciones de respuesta cerradas múltiples). Todos los reactivos fueron clasificados en los seis diferentes niveles incluidos en la taxonomía de Bloom modificada,[6] que son: 1) recordar, que consiste en recuperar información relevante de la memoria; 2) comprender, que incluye la construcción de significados a partir de información nueva, interpretándola o clasificándola; 3) aplicar, que consiste en la utilización de una habilidad o conocimiento en una nueva situación, usándola o implementándola; 4) analizar, que se trata de descomponer la información en partes más pequeñas para identificar patrones o relaciones, por ejemplo, en un escenario o caso clínico, o comparando conceptos; 5) evaluar, que consiste en realizar juicios de valor basados en criterios, estándares, pruebas, escalas, etcétera, formulando una hipótesis o diagnóstico tentativo respecto al diferencial, y 6) crear, que consiste en producir algo nuevo combinando los elementos de las taxonomías previas para formar un diagnóstico o abordaje coherente y estructurado, por ejemplo, elaborando un plan de abordaje, diagnóstico o tratamiento para un paciente con una patología particular.

Prueba de implementación

Para esta fase, se utilizaron las pruebas de 566 sustentantes, comparando la prueba convencional de entrenamiento de cada uno de ellos y contrastando estos resultados con las mismas EAC para la certificación de estos alumnos.

El contraste estadístico fue ligado de manera individual y global. Para los EAC, se utilizaron 200 reactivos de opción múltiple cerrada de las diferentes áreas de las ciencias quirúrgicas, que incluye el PUEM y otros exámenes estándares de la especialidad de cirugía general. En ambos exámenes se clasificaron todos los reactivos por dificultad, con base en los niveles de la taxonomía de Bloom modificada que comentamos anteriormente. Tanto los reactivos como las respuestas fueron sorteados de manera aleatoria entre los sustentante para ambos exámenes.

Prueba de validación

Para esta fase, se utilizaron los exámenes de 1,164 sustentantes, comparando dos diferentes tipos de exámenes adaptativos computarizados (A y B), y contrastando ambos con el examen de entrenamiento previamente realizado por esos mismos sustentantes. En este examen se sortearon 1,200 reactivos estandarizados y validados en dos diferentes exámenes de máximo 200 reactivos de opción múltiple cerrada en las diferentes áreas del conocimiento de las ciencias quirúrgicas que incluyen el PUEM y otros exámenes estándares recomendados para la certificación en la especialidad de cirugía general. Se clasificaron los reactivos por su dificultad y con base en los niveles de la taxonomía de Bloom modificada. Tanto los reactivos como las respuestas fueron sorteadas aleatoriamente entre los sustentantes para todos los exámenes mencionados. En esta ocasión, se requería de al menos el 30% de aciertos por nivel taxonómico para poder pasar al siguiente nivel de dificultad, de tal manera que, nuevamente, el número de preguntas y reactivos contestados fue variable entre exámenes y sustentantes. Ambos exámenes (A y B) se aplicaron en 8 diferentes turnos en días distintos.

Para las tres fases se utilizó estadística descriptiva para medidas de tendencia central, dispersión y posición, así como contrastes estadísticos tanto paramétricos como no paramétricos, dependiendo del escalamiento natural de las variables incluidas en el análisis. Utilizamos contrastes pareados (entre cada sustentante y entre las pruebas duplicadas: en la primera fase, para la prueba adaptativa por computadora vs la prueba estándar, y en las últimas dos fases, para los mismos sustentantes, entre la prueba adaptativa computarizada vs la prueba de certificación estándar). Para los contrastes, se utilizaron pruebas como t pareada, pruebas de correlación, concordancia, gráfico de Bland-Altman y χ^2. Para este análisis, utilizamos el paquete de software estadísticos IBM® SPSS® Statistics versión 26. Todo contraste inferencial que obtuvo un valor de p para un error alfa inferior al 5% o 0.05 se consideró como estadísticamente significativo para una prueba de hipótesis de dos colas.

RESULTADOS

Como se comentó en el apartado anterior, realizamos tres fases para este proyecto, que consistió en una prueba piloto preliminar, para ajustes y prueba del concepto, una prueba de implementación y una de validación en cohortes de sustentantes independientes.

Prueba piloto

En esta fase se obtuvieron marcadores de tendencia central muy similares; por ejemplo, la razón de la calificación aritmética (RAC)/número de reactivos contestados como subrogada de la eficiencia de un sustentante (n = 322). En este marcador se obtuvo un promedio ± DE de 30.32 ± 3.86 para la prueba convencional, frente a 30.67 ± 6.83 global para la prueba adaptativa computarizada. Cuando se compararon los promedios y la RAC por cada sustentante de manera pareada mediante una prueba de t pareada, se obtuvo una diferencia de promedios ± DE de −6.80 ± 8.73 (error estándar de la media [EEM] = 0.46) para los promedios de cada examen, con una $p < 0.0001$, y una diferencia de promedios ± DE de 0.35 ± 7.26 (EEM = 0.40) para las RAC por sustentante de manera pareada, con una p = 0.38.

Los gráficos de Bland-Altman para estos contrastes se muestran en la *Figura 1*. En este gráfico se puede apreciar la mayor dispersión de la calificación promedio de los sustentantes, en contraste con la RAC. Sin embargo, el acuerdo por valor de kappa fue igual a 0.15 ($p < 0.0001$), pero con una correlación baja ($r_{Pearson} = 0.22$; $p < 0.0001$). Esto muy probablemente se debió a la utilización de

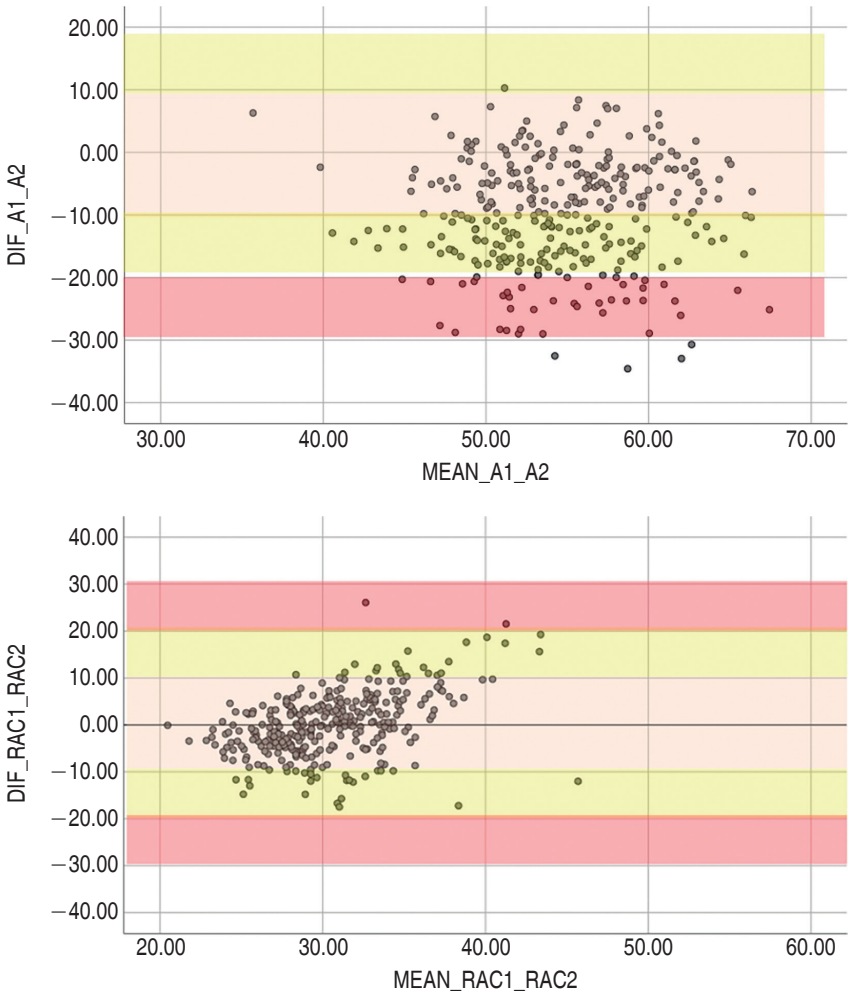

Figura 1:

Gráficos de Bland-Altman comparando la dispersión de los promedios aritméticos (arriba) y las RAC (abajo) obtenidos por cada sustentante (de manera pareada) entre la prueba convencional vs la EAC.
EAC = Evaluación Adaptativa Computarizada.
RAC = Razón de la Calificación Aritmética.

diferentes puntos de corte debido a la diferencia en el número de reactivos que se usaron como denominador. Sin embargo, cuando se utilizó una prueba de W de Kendall de rangos para contrastar la concordancia de los grupos ordinales, se obtuvo un valor W Kendall de 0.67, con una $p < 0.0001$.

Con base en estos resultados pudimos concluir que el uso de la prueba adaptativa computarizada no sólo era factible sino más eficiente, puesto que requirió un menor tiempo de aplicación (−20 minutos), con resultados de concordancia/correlación limitadas pero comparables. En esta fase, el uso de la RAC al parecer tuvo mayor utilidad que la simple comparación de los promedios aritméticos por sustentante.

Prueba de implementación

En esta fase, el tiempo fue significativamente menor entre la prueba de entrenamiento convencional versus la adaptativa computarizada de certificación (117.2 ± 12.6 minutos vs 72.05 ± 18.2 minutos, respectivamente) sin correlacionar con el número de aciertos ni la calificación aritmética ni la RAC ($r_{Pearson}$ = −0.08, −0.26 y −0.26, con valores de p = 0.71, 0.21 y 0.21, respectivamente). Sin embargo, llama la atención la tendencia a la correlación negativa entre estos marcadores respecto al tiempo, como hemos observado previamente en los exámenes convencionales.

La *Figura 2* muestra la tendencia a la correlación lineal negativa en esta prueba entre

el promedio aritmético y la RAC, basada en la clasificación baja asertividad académica (BAA) versus alta asertividad académica (AAA). Aunque en estos casos no se alcanzó la significancia matemática, llama la atención la clara tendencia a la correlación negativa en todas las comparaciones. El promedio ± DE de reactivos contestados fue también significativamente menor para la segunda fase adaptativa en contraste con la primera (100.56 ± 15.64 vs 163.66 ± 15.78, respectivamente). El promedio ± DE de reactivos contestados fue también significativamente menor para la segunda fase adaptativa en contraste con la primera (100.56 ± 15.64 vs 163.66 ± 15.78, respectivamente). Sin embargo, el promedio aritmético fue muy similar entre ambas (50.84 ± 7.6 vs 54.53 ± 7.2, respectivamente), a pesar de la diferencia entre el grupo de cirujanos con AAA vs BAA (51.34 ± 7.5 vs 36.42 ± 5.5 para el promedio aritmético y 26.64 ± 8.2 vs 13.81 ± 4.1 para el RAC, respectivamente); estos resultados demuestran que la prueba fue tan eficiente como la convencional para identificar dichos contrastes.

De manera interesante, 24 minutos fueron suficientes para demostrar suficiencia académica en los sustentantes que aplicaron la prueba adaptativa computarizada, mientras

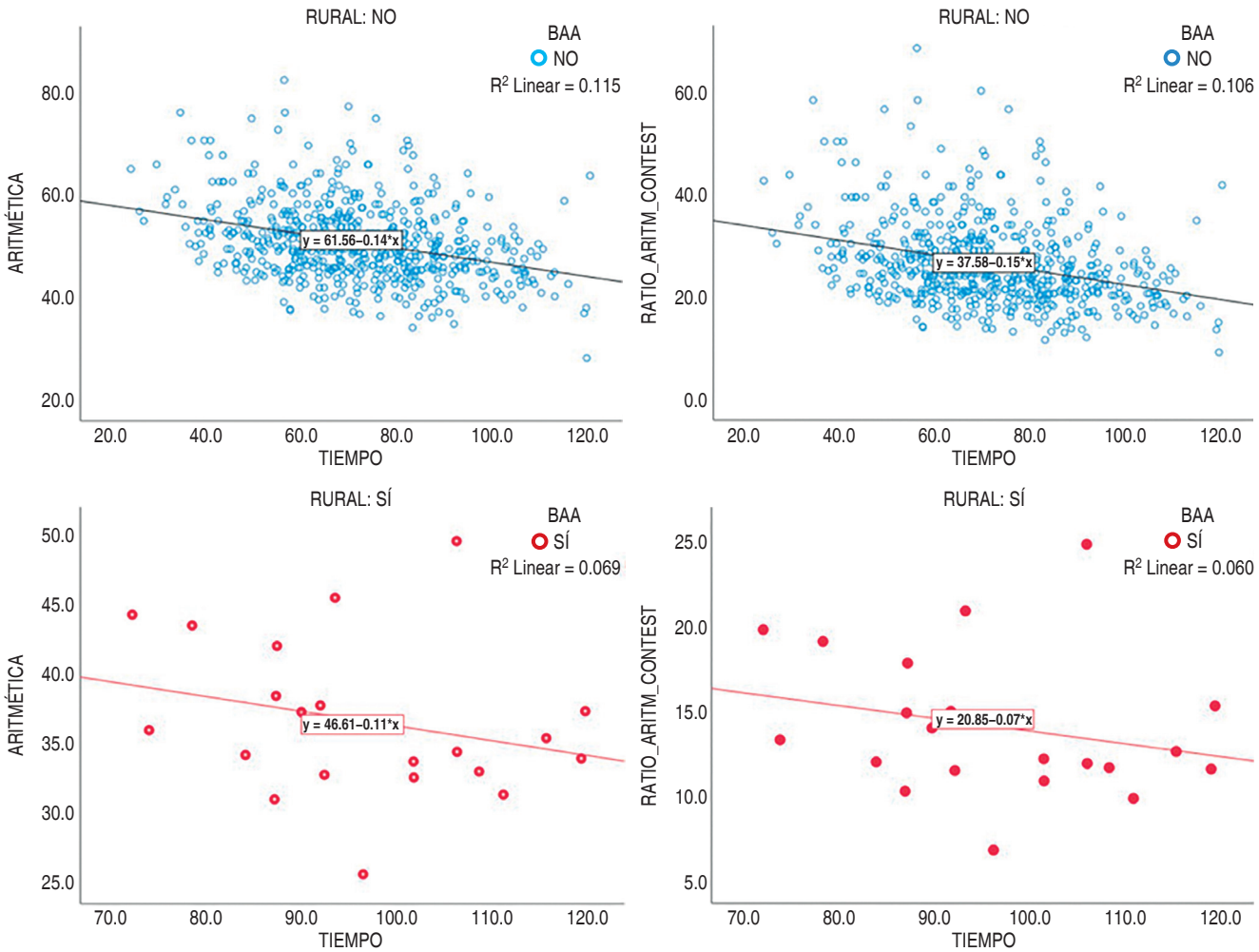

Figura 2: Gráficas de correlación con líneas de tendencia para las correlaciones obtenidas en la segunda fase entre sustentantes de cirugía con AAA en contraste con los cirujanos con BAA para el promedio aritmético y la RAC.
AAA = Alta Asertividad Académica. BAA = Baja Asertividad Académica. RAC = Razón de la Calificación Aritmética.

que los sustentantes que no alcanzaron la suficiencia académica pudieron tomar hasta 118 minutos, lo que confirma la hipótesis de que, con base en la RAC y en el promedio aritmético, el mayor tiempo es un factor en detrimento del resultado obtenido y, por tanto, respecto a la suficiencia académica obtenida por los sustentantes en ambas pruebas. En esta fase se volvió a demostrar la alta eficiencia tiempo-asertividad, siendo la prueba adaptativa computarizada la de mayor eficiencia en menor tiempo y mayor aleatorización del banco de reactivos validados, lo que permite una mayor eficiencia general respecto a la prueba convencional.

Prueba de validación

Para esta fase, se utilizaron las pruebas de 1,164 sustentantes, de los cuales 818 (68.5%) fueron varones; sólo 45 (3.8%) pertenecían al curso de cirugía de los servicios rurales; 24 (2%) eran extranjeros. La edad promedio ± DE del grupo fue de 32.3 ± 4.2 años.

A 630 (52.8%) sustentantes se les aplicó el examen tipo A y a los 564 (47.2%) restantes el tipo B. No hubo diferencias estadísticas para la edad (32.6 ± 16.1 vs 32.3 ± 4.41; p = 0.30), ni para el número de reactivos contestados (66.6 ± 16.2 vs 102.7 ± 16.1; p = 0.55), ni para el tiempo de aplicación (66.4 ± 16.2 vs 72.0 ± 16; p = 0.61), ni para el tiempo por pregunta (0.7 ± 0.16 vs 0.7 ± 0.17; p = 0.23), aunque sí para el número de aciertos (46.6 ± 0.8 vs 46.8 ± 1.1; p = 0.03), para la calificación aritmética (51.4 ± 8.5 vs 46.7 ± 7.8; p = 0.04) y para la RAC (27.2 ± 6.1 vs 25.3 ± 7.6; p = 0.004). Estas diferencias son en realidad un error alfa por un alto poder estadístico pero no significativas de manera práctica, ya que no tuvieron ninguna injerencia en la suficiencia académica o no, y en realidad son más similares de lo que uno esperaría.

La *Figura 3A* muestra una gráfica de cajas y bigotes con estos contrastes entre los resultados del examen A contra el B. Estos dos tipos de exámenes fueron aplicados en ocho diferentes turnos: 156 (13.3%) para el primer turno, 154 (12.6%) para el segundo, 154 (12.6%) para el tercero, 156 (13.1%) para el cuarto, 155 (13%) para el quinto, 165 (13.8%) para el sexto, 68 (8.2%) para el séptimo y 153 (12.8%) para el octavo turno. De similar manera, la edad, número de aciertos y aciertos por taxonomía no difirieron estadísticamente entre los ocho turnos, sin embargo, si difirieron estadísticamente el número de reactivos contestados (p = 0.002), la calificación aritmética (p = 0.002), la RAC (p = 0.003), el tiempo (p = 0.001), el tiempo por pregunta (p = 0.0001) y la razón de momios entre la aritmética y el tiempo (p = 0.004) sin embargo ninguno de los grupos difirió en más de seis unidades para cualquiera de estos rubros.

En la *Figura 3B* se muestran gráficos de cajas y bigotes con las diferencias entre los ocho turnos, que fueron más sutiles que significativas, en sentido matemático. Cuando se contrastaron los cirujanos AAA vs BAA, las pruebas fueron capaces de discriminar categóricamente entre estos grupos, con una significancia estadística menor a 0.0001 en todos los rubros, excepto para el tiempo invertido por pregunta, con resultados similares en los cirujanos con AAA vs BAA (0.70 ± 0.18 vs 0.72 ± 0.16; p = 0.41).

En la *Figura 4* se muestra cómo el grupo de cirujanos de BAA tuvo un mayor número de reactivos contestados con un mayor tiempo, pero con una menor calificación aritmética y una RAC mucho menor, lo que le confiere nuevamente una mayor precisión para discernir en un grupo con BAA respecto al grupo de AAA, como ya se había demostrado con exámenes convencionales de certificación en cirugía general. Un análisis de correlación lineal con r de Pearson demostró correlaciones fuertemente negativas entre el tiempo total, el tiempo por pregunta en relación con el número de reactivos contestados ($r_{Pearson} = -0.34$; p = 0.0001), la calificación aritmética ($r_{Pearson} = -0.34$; p = 0.0001) y la RAC ($r_{Pearson} = -0.34$; p = 0.0001) para este grupo, confirmando la capacidad de esta prueba adaptativa para diferenciar entre los grupos conocidos por su asertividad (AAA vs BAA), de manera similar al examen convencional de 200 preguntas totales.

Aunque, en términos generales, esta tendencia se observó tanto para los cirujanos AAA como para los BAA, en ambos grupos un mayor tiempo se relaciona con una menor calificación aritmética ($r_{Pearson} = -0.36$ vs -0.40, respectivamente; $p < 0.006$), menor

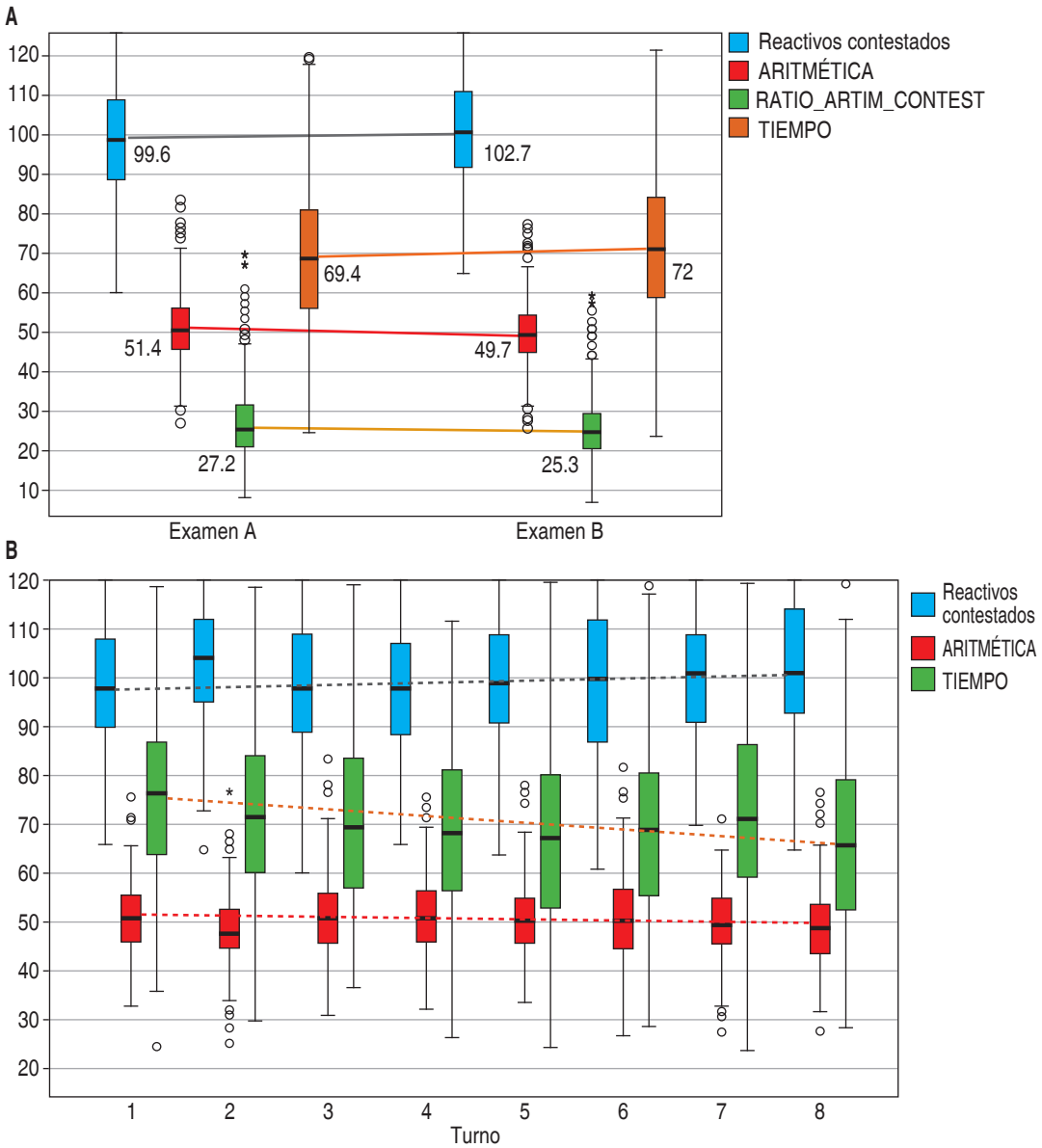

Figura 3: Gráfica de cajas y bigotes, representando: **A)** La mediana e intervalo intercuartilar para el número de reactivos contestados, calificación aritmética, razón de la calificación aritmética (RAC) y tiempo entre los sustentantes que aplicaron para el examen A respecto al B y que difirieron en el número de reactivos, estructura y composición; **B)** Las diferencias entre los ocho turnos en los que se aplicaron estos dos tipos de exámenes en relación con el número de reactivos contestados, la calificación aritmética y el tiempo invertido por el total del examen en cada turno.

RAC ($r_{Pearson} = -0.35$ vs -0.31, respectivamente; $p < 0.04$) y menor razón de momios de la calificación aritmética en relación con el tiempo total ($r_{Pearson} = -0.86$ vs -0.84 respectivamente; $p < 0.0001$), como se puede observar en la *Figura 5*. Esta observación fue documentada también en las pruebas convencionales, por lo que, de nueva cuenta, es un indicador de validez externa.

DISCUSIÓN

En este estudio de factibilidad, eficiencia y validez de la EAC, comprobamos que esta

herramienta evalúa las mismas competencias cognitivas que el examen en formato convencional, pero lo hace en mucho menor tiempo, pudiendo discriminar apropiadamente aquellos sustentantes con AAA en comparación con aquellos con BAA, mostrando ser equiparables el EAC respecto al convencional. Hasta donde sabemos, este es el primer estudio realizado en México para la validación de la EAC en un examen de alto impacto, como lo es el examen de certificación de una especialidad médica.

Si bien el CMCG es el primer consejo de especialidad en validar y aplicar esta herramienta de evaluación en México, la EAC ha sido respaldada por la literatura científica y académica internacional desde hace aproximadamente tres décadas. En 1995, Lunz y colaboradores demostraron que los exámenes basados en EAC son equivalentes a los exámenes tradicionales, en este caso particular los exámenes realizados a papel y lápiz, generando estimaciones de habilidades o competencias, tasas de aprobación y decisiones de suficiencia y no-suficiencia comparables; estos hallazgos concuerdan con nuestros resultados.[7] Además, en este mismo trabajo, documentan cómo un banco de reactivos desarrollado y calibrado para evaluaciones no basadas en EAC muestra estabilidad en criterios psicométricos durante su validación en la EAC, siempre que se hayan diseñado adecuadamente. Esta validación derivó en el reemplazo de las evaluaciones basadas en papel y lápiz por la herramienta EAC en exámenes de alto impacto.[8]

Un ejemplo de esto es el examen de certificación y licencia profesional de enfermería de Estados Unidos (NCLEX). En estos exámenes de alta exigencia, la EAC ha demostrado, al igual que en nuestro estudio, alta precisión para estimar la competencia del sustentante, utilizando un número sustancialmente reducido de ítems, disminuyendo la fatiga del sustentante y aportando un mecanismo adicional de seguridad al banco de ítems; acompañado de un alto coeficiente de correlación entre puntajes obtenidos por el examen EAC y el convencional.[9] Algunos otros ejemplos en el área de la salud incluyen el examen del *National Registry of Emergency Medical Technicians* y el del *National Health Personnel Licensing Examination Board* en Estados Unidos y la República de Corea, respectivamente.[10,11]

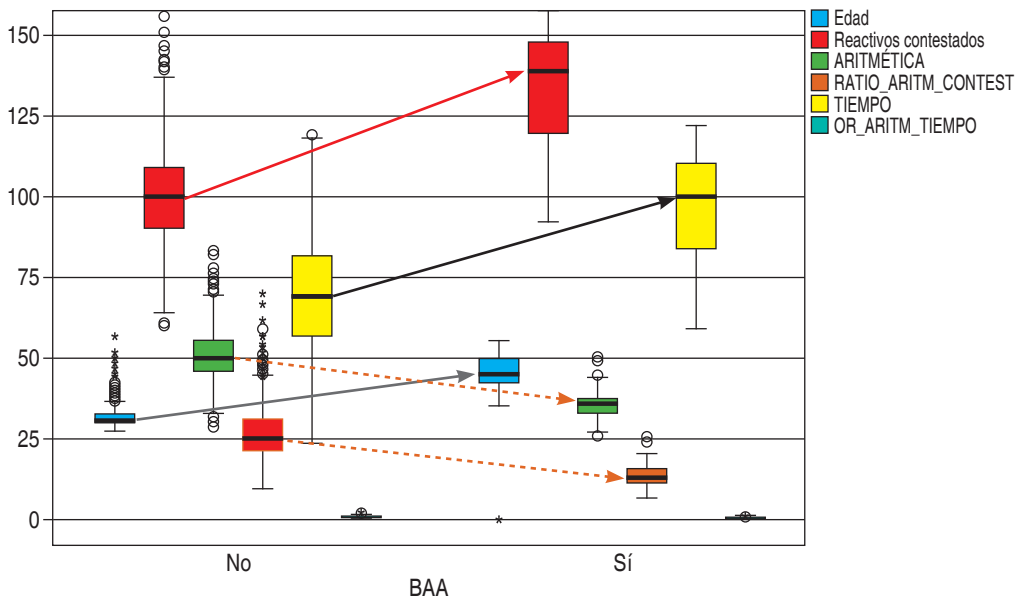

Figura 4: Gráfica de cajas y bigotes que muestran las diferencias entre las medianas e intervalos intercuartilares entre el grupo de cirujanos con AAA vs BAA para la edad, número de reactivos contestados, calificación aritmética, RAC, tiempo total y razón de momios de la calificación aritmética respecto al tiempo total invertido en la prueba.
AAA = Alta Asertividad Académica. BAA = Baja Asertividad Académica. RAC = Razón de la Calificación Aritmética.

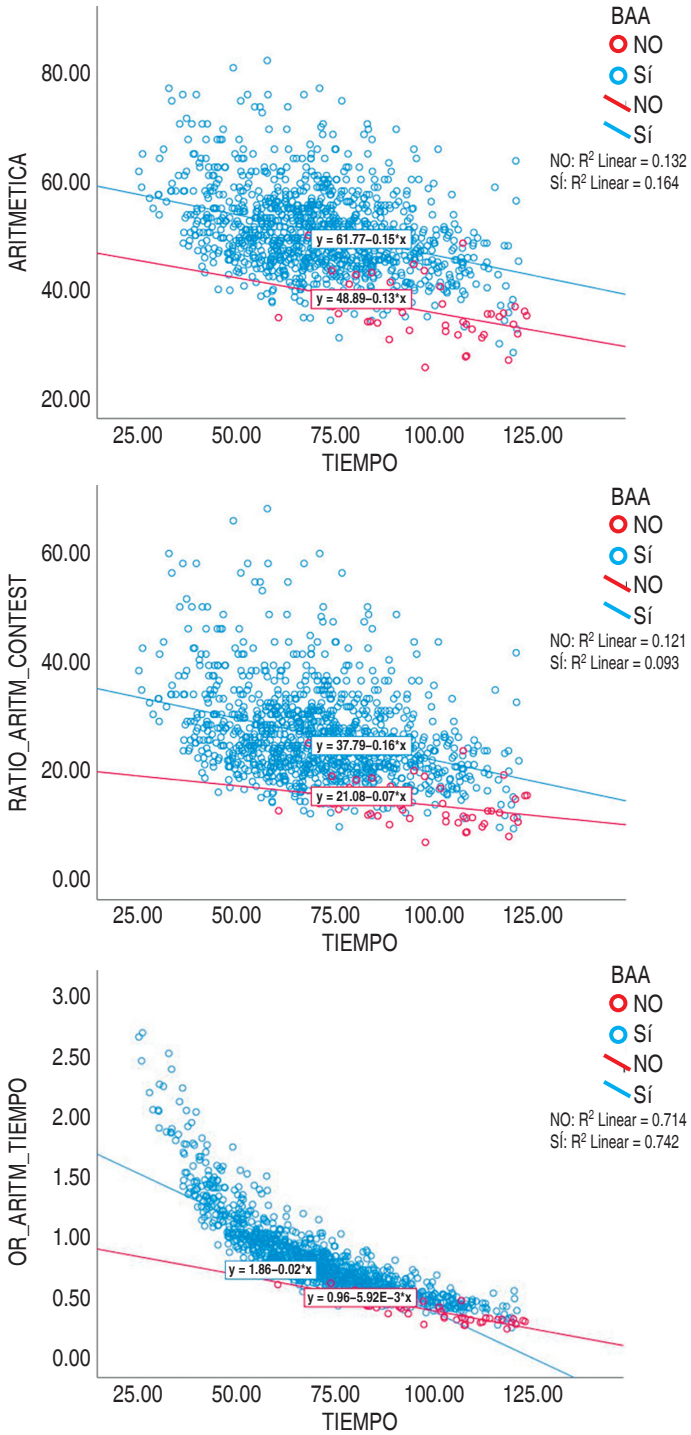

Figura 5: Gráfica de dispersión en el que se puede apreciar las correlaciones lineales por la prueba r de Pearson entre los resultados de los cirujanos con AAA vs BAA para la calificación aritmética (arriba), RAC (en medio) y la razón de momios entre la calificación aritmética respecto al tiempo total (abajo). Todos los coeficientes de Pearson así como los valores de p fueron significativos.
AAA = Alta Asertividad Académica. BAA = Baja Asertividad Académica. RAC = Razón de la Calificación Aritmética.

Es importante mencionar que, adicionalmente a la validación realizada del instrumento, nuestro análisis mostró un nuevo indicador para dictaminar suficiencia académica, la RAC. Este indicador demostró ser estable, con una menor variabilidad, compensando el efecto de la longitud (duración) que un examen puede tener respecto al resultado aritmética simple.

Por otro lado, la EAC no es una herramienta infalible y su aplicación requiere de algunas consideraciones especiales. Debe ser considerada como un instrumento de evaluación diagnóstica orientado exclusivamente a la suficiencia teórico-cognitiva del sustentante, y no sustituye la necesidad de evaluar de manera sumativa el resto de las competencias de los cirujanos generales.[12]

Dado esto, el CMCG considera el uso de esta herramienta validada como parte del proceso en tres fases antes mencionado para la certificación de primera vez. Además, agrega un grado de complejidad al mantenimiento del banco de reactivos debido al riesgo de sobreexposición de los reactivos informativos y óptimos (con mejores parámetros psicométricos) ante los sustentantes, lo que requiere de estrategias de seguridad adicionales para evitar la fuga de información entre aplicaciones.[13]

Desde la perspectiva del sustentante, existe el riesgo de tener una percepción errónea sobre un mal desempeño en el examen, incluso cuando no es así, debido al incremento progresivo en la dificultad de los reactivos que se le van presentando, lo que puede derivar en ansiedad y bajas tasas de satisfacción durante la evaluación.[14,15] Mecanismos de comunicación y divulgación de información como redes sociales, vídeos con explicaciones claras de la estrategia de la EAC y explicaciones a través de los instructivos del examen son algunas de las estrategias utilizadas por el CMCG para mitigar este potencial efecto perjudicial.

Nuestro estudio muestra fortalezas: 1) su diseño metodológico en tres fases para la validación sistemática de la herramienta, 2) el número (tamaño muestral) considerable de sustentantes en certificación para cirugía general, 3) la validación de la herramienta en un escenario real e independiente, lo que favorece la validación tanto interna como externa de las inferencias, 4) el análisis pareado por sustentan-

te, lo que permite evaluar el efecto directo de la evaluación de manera individual, eliminando la variabilidad de los sustentantes muestreados, y, finalmente, 5) el uso de un banco amplio de reactivos previamente estandarizados y validados en cohortes independientes. Sin embargo, el estudio también posee ciertas limitaciones: por un lado, el banco de reactivos utilizados no fue inicialmente validado o calibrado bajo el esquema de EAC y la TRI. Si bien la taxonomía de Bloom de los reactivos es una buena aproximación, no es el estándar para esta herramienta. Además, existe cierto grado de baja representatividad de algunos grupos dentro de la muestra, especialmente el de cirujanos extranjeros. Por último, la aleatorización de los reactivos podría entregar exámenes sutilmente distintos en cuanto a estructura en las diferentes áreas de conocimiento relacionadas con la cirugía general, pero consideramos que esto no es un factor propio de la EAC *per se*. Sin embargo, la aleatorización de un banco grande de reactivos con una estructura similar al PUEM permite que los exámenes extraídos de este gran banco tengan una estructura semejante al banco universal de reactivos ya validados previamente en exámenes convencionales, lo que de alguna manera reduce significativamente la posibilidad de que dos individuos presenten el mismo examen. Esto sin duda garantiza una mayor seguridad del examen, con una mínima probabilidad de copiar durante su aplicación, lo que incrementa la eficiencia de la EAC respecto a exámenes convencionales similares entre sustentantes.

CONCLUSIONES

En nuestro estudio, la EAC demostró ser una herramienta no sólo factible, sino válida y con una mayor eficiencia que el método de evaluación convencional, al lograr resultados equiparables, con una reducción significativa en el tiempo y número de reactivos utilizados, manteniendo su capacidad para discriminar rápidamente a los sustentantes con AAA. Además, la RAC mostró ser un indicador superior a la simple calificación aritmética para categorizar eficientemente a los sustentantes en evaluaciones de longitudes variables en cuanto al tiempo y número de reactivos, mostrando correlaciones negativas estadísticamente significativas entre el tiempo de evaluación, la eficiencia del sustentante y su grado de suficiencia académica.

REFERENCIAS

1. Zermeño-Gómez MG, Kobeh-Jirash JA, Moreno-Guzmán A, Jiménez-Chavarría E, Pantoja-Millán JP, Noyola-Villalobos H, et al. La certificación en Cirugía General a 42 años de la fundación del Consejo Mexicano de Cirugía General. Cir Gen. 2016; 41: 314-321.
2. Morris S, Bass M, Lee M, Neapolitan RE. Advancing the efficiency and efficacy of patient reported outcomes with multivariate computer adaptative testing. J Am Med Inform Assoc. 2017; 24: 867-902.
3. Bamikole OI. Item Response Theory (IRT): A Modern Statistical Theory for Solving Measurement Problem in 21st Century. International Journal of Scientific Research in Education, 2018; 11: 627-635.
4. Linacre JM. Computer-adaptive testing: a methodology whose time has come. South Korea: Komesa Press; 2000. Disponible en: https://www.rasch.org/memo69.pdf
5. Cicciola E, Foschi R, Lombardo GP. Making up intelligence scales: De Sanctis's and Binet's tests, 1905 and after. Hist Psychol. 2014; 17: 223-236.
6. Adams NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc. 2015; 103: 152-153.
7. Lunz ME, Bergstrom BA. Equating computerized adaptive certification examinations: the Board of Registry series of studies. Eric.ed.gov. 1995. Disponible en: http://files.eric.ed.gov/fulltext/ED388696.pdf
8. Zaglaniczny KL. The transition of the national certification examination from paper and pencil to computer adaptive testing. AANA J. 1996; 64: 9-14.
9. Wise SL, Kingsbury GG. Practical Issues in Developing and Maintaining a Computerized Adaptative Testing Program. Psicológica. 2000; 21: 135-155.
10. Seo DG. Overview and current management of computerized adaptive testing in licensing/certification examinations. J Educ Eval Health Prof. 2017; 14: 17.
11. Huh S. Preparing the implementation of computerized adaptive testing for high-stakes examinations. J Educ Eval Health Prof. 2008; 5: 1.
12. Van Der Vleuten CPM, Schuwirth LWT, Driessen EW, Dijkstra J, Tigelaar D, Baartman LKJ, et al. A model for programmatic assessment fit for purpose. Med Teach. 2012; 34: 205-214.
13. Stocking ML, Lewis C. Controlling item exposure conditional on ability in computerized adaptive testing. J Educ Behav Stat. 1998; 23: 57-75.
14. Tonidandel S, Quiñones MA, Adams AA. Computer-adaptive testing: the impact of test characteristics on perceived performance and test takers' reactions. J Appl Psychol. 2002; 87: 320-332.
15. Wise SL, Plake BS. Research on the effects of administering tests via computers. Educ Meas Issu Pr. 1989; 8: 5-10.

Correspondencia:
Dr. David Velázquez Fernández
E-mail: cmcg.academico@gmail.com