Estudio comparativo de la aplicación de un examen diagnóstico en posgrado, utilizando dos formatos: por computadora e impreso

María Eugenia Ponce de León-C.* Armando Ortiz-Montalvo** María del Carmen Ruíz-Alcocer ***

Recepción versión modificada 12 de junio de 2002; aceptación 14 de enero de 2003

Resumen

En la Facultad de Medicina de la Universidad Nacional Autónoma de México se aplicó un examen diagnóstico a todos los residentes de tercer año de la especialidad de Medicina Interna, el examen se estructuró con 42 casos clínicos y 210 reactivos de opción múltiple, divididos en dos cuadernillos, cada uno con 21 casos clínicos y 105 reactivos. Los residentes fueron asignados a dos grupos en forma aleatoria, grupo A (examen impreso) y grupo B, (examen en computadora). Se buscaron diferencias en el rendimiento académico de los residentes y en el comportamiento del examen. Como indicadores del rendimiento académico se usaron: media, desviación estándar, mínimo y máximo de aciertos; del comportamiento del examen: el grado de dificultad, poder de discriminación y confiabilidad. Se aplicó un cuestionario de opinión a los residentes y se llevó un control del tiempo de respuesta. Los resultados mostraron que con el examen impreso se obtuvo un mayor rendimiento académico de los residentes en tanto que el comportamiento del examen no mostró diferencias significativas; la confiabilidad del examen fue mayor en la modalidad de computadora, la opinión de los residentes fue favorable hacia el uso de esta tecnología.
Posteriormente, se analizaron y compararon por separado cada uno de los dos cuadernillos que integraban el examen y se observó que las diferencias mencionadas estaban dadas por el resultado obtenido en el primer cuadernillo, en el segundo los resultados eran muy semejantes. Se concluyó que probablemente la inexperiencia en el manejo de la computadora fue un factor determinante de los resultados.

Palabras clave: *Examen computarizado, examen impreso, rendimiento académico.*

Summary

A diagnostic test was applied at the National Autonomous University of Mexico School of Medicine using a two-part format with 42 clinical cases and 210 multiple-choice items. Residents were randomized into two groups: group A (printed test) and group B (computerized test). Academic performance was measured by determination of media, standard deviation, and correct-answer score; the exam was measured by reliability, difficulty index, and discrimination index and question/item assessment were calculated for item analysis.
Residents answered a survey questionnaire and time taken to answer was controlled. Results showed that the printed test had higher achievement; there was better reliability for computerized format, and resident opinion was more favorable toward computer use. The two parts of the test were analyzed and results were produced for the first part of the test; in part two, results were very similar. We conclude that lack of experience in computer use could be a determining factor in our results.

Key words: *Computerized test, printed test, academic performance.*

**Secretaria de Educación Médica de la Facultad de Medicina de la UNAM.*
***Jefe del Departamento de Evaluación Educativa de la Secretaría de Educación Médica de la Facultad de Medicina de la UNAM.*
****Departamento de Evaluación Educativa de la Secretaría de Educación Médica de la Facultad de Medicina, de la UNAM*
Correspondencia y solicitud de sobretiros: Dra. María Eugenia Ponce de León C. Secretaría de Educación Médica. Facultad de Medicina, UNAM. Tercer piso, edificio B. 5623 2448, 5623 2449, Fax 5616 2346. e-mail: mepdl@servidor.unam.mx

Introducción

En la actualidad las computadoras y la tecnología informática son parte esencial de nuestra vida diaria. En el campo de la educación, el impacto de ambas se inició aproximadamente en la década de los setenta, como consecuencia de la disminución en el costo y tamaño del equipo, así como por la facilidad en su manejo. Los avances en la tecnología computacional y el diseño de programas para la educación médica, han favorecido en primer lugar el desarrollo de programas educativos autodirigidos, algunos de ellos a distancia; en segundo término, los programas de autoevaluación o evaluaciones formales, para las cuales se hace necesario el manejo de bancos de preguntas muy amplios, en donde es factible incorporar la retroinformación inmediata.

Actualmente se cuenta con suficientes estudios internacionales que analizan las actitudes de los alumnos hacia esta modalidad,[1] lo que ha favorecido que las escuelas de medicina consideren dentro de su currículo la inclusión de programas de cómputo en los primeros semestres.[2,3]

En un estudio dirigido a residentes de pediatría en la Universidad de Carolina del Norte, se encontró que la actitud hacia este tipo de exámenes era fuertemente positiva y que la mayoría de los residentes preferían los exámenes en computadora a los aplicados por escrito. En este estudio se buscó analizar la efectividad de los exámenes computarizados desde cuatro puntos de vista: el técnico, que demostró que sí es posible construir un examen computarizado de opción múltiple; el de actitud, que mostró que es factible que los alumnos se adapten al uso de la computadora y la encuentren más atractiva que el examen escrito; el de conducta, en el cual se corroboró que los alumnos frecuentemente recurren a la retroinformación ofrecida por el programa y el de conocimientos, en el se verificó que los alumnos aprenden de la retroinformación proporcionada.[4,5]

En su artículo sobre la actitud de los alumnos hacia la evaluación computarizada en un curso de ciencias básicas, Ogilvie[6] menciona que autores como Rattan, Miller y Legler coinciden en haber encontrado una actitud positiva similar en los alumnos de pregrado que fueron sometidos a la experiencia de exámenes computarizados. En estos estudios no sólo se realizó la aplicación del examen, también se utilizó la computadora para dar retroinformación inmediata a los alumnos y para introducir imágenes al texto del examen, específicamente en el examen de biología celular e histología en la Universidad de Carolina del Sur. Ahí se demostró también una modificación en los hábitos de estudio de los alumnos, quienes consideraron que la prueba les había resultado accesible y que contestarla les había consumido menos tiempo.

Para este modelo de examen comúnmente se utilizan reactivos de opción múltiple en tres de sus versiones: de cinco opciones de respuesta con una verdadera, de correlación de columnas y de falso-verdadero. Así se puede evaluar a un mayor número de alumnos y de temas en cortos periodos, con una gran confiabilidad y objetividad. Cuando son elaborados con acuciosidad pueden medir niveles complejos de conocimiento.[7,8] Entre los inconvenientes de estos reactivos está que someten al alumno a un proceso en el cual debe buscar la respuesta a un problema entre cinco opciones previamente planteadas, en contraposición con el proceso que sigue el médico, quien utiliza sus conocimientos para proponer una respuesta al problema.[9,10]

Norcini en 1985, mencionó que los reactivos de opción múltiple eran de los más confiables y eficientes, específicamente cuando se utilizan haciendo referencia a una situación clínica.[11] La autenticidad del caso o situación clínica es de suma importancia, ya que la información clara y real orienta al alumno a que tome una decisión manteniendo su interés y motivación en el caso.[12,13]

Un estudio realizado por Miller con alumnos de Anatomía Patológica, analizó la opinión de los profesores respecto al uso de la computadora para la aplicación de exámenes, observó que la administración del examen fue fácil y de gran beneficio para los alumnos, ya que el proporcionarles retroinformación les permite corregir sus deficiencias. Las desventajas encontradas fueron: 1) la dificultad del estudiante para enfrentar de primera intención a la máquina, lo que le genera un impacto brusco, que vence al adaptarse al manejo de la misma; 2) el costo del equipo y 3) el hecho de que no quedaba una memoria de los resultados.[14]

Comparar el uso de dos metodologías en la aplicación de un examen, requiere sustentar el análisis en la confiabilidad y factibilidad de los métodos,[1,15,16] así como en la dificultad y discriminación obtenida por los reactivos,[17] ambos índices relacionados con la estructura de la pregunta y con los procesos cognitivos que se demandan del alumno.

Si se considera lo hasta aquí mencionado, se puede asegurar que la automatización de los procesos de evaluación, ya sea diagnóstica, formativa o sumativa permite ampliar la cobertura y los hace accesibles a cualquier sitio donde se encuentren un procesador y una línea telefónica, lo cual con el tiempo lógicamente va a repercutir en los costos y en la efectividad de la evaluación.

En la Facultad de Medicina de la Universidad Nacional Autónoma de México, la División de Estudios de Posgrado en conjunto con la Secretaría de Educación Médica desde hace tres años realiza exámenes diagnósticos a todos los residentes de las cuatro especialidades médicas troncales (Medicina Interna, Pediatría, Gineco-Obstetricia y Cirugía). En la aplicación de febrero de 2001 se realizó la investigación motivo del presente artículo.

El objetivo del estudio fue investigar si existían o no diferencias en el comportamiento de un examen y en el

rendimiento de los residentes de tercer año de la especialidad de Medicina Interna, cuando un mismo examen se aplicó en dos formatos diferentes, uno en computadora y otro impreso en papel.

Material y métodos

Se realizó un estudio comparativo de asignación aleatoria a dos formatos de aplicación de un examen. Se evaluaron 130 residentes de tercer año de Medicina Interna, provenientes de 22 hospitales del Distrito Federal. El examen fue elaborado por los profesores, con casos clínicos reales, cada uno se estructuró con cinco reactivos de opción múltiple, los cuales se relacionan directamente con el caso clínico. Los contenidos de los casos correspondían a los contenidos del Programa Único de Especialidades Médicas de Medicina Interna del tercer año. El examen se integró con 210 reactivos que correspondían a 42 casos.

Los residentes participantes estaban inscritos en la Facultad de Medicina y se asignaron en forma aleatoria a cada uno de los grupos (A y B), cuidando que cada uno de ellos contara con la representación de residentes de las 22 sedes de la especialidad. Los grupos quedaron constituidos con 65 residentes respectivamente, sin embargo, el día de la aplicación asistieron al grupo A (examen impreso) 67 sustentantes y al grupo B (examen en computadora) 63. A ambos grupos se les aplicó el mismo examen, el cual fue impreso en dos cuadernillos, I y II, cada uno con 105 reactivos para 21 casos.

El examen se aplicó en la Facultad de Medicina a la misma hora y con la misma distribución de tiempo para ambos cuadernillos del examen, el grupo A en un aula y el grupo B en el aula de cómputo.

Para los alumnos que contestaron el examen en la computadora, se utilizó un programa desarrollado en el Departamento de Cómputo de la Facultad de Medicina, estructurado en Visual Fox 5. La plataforma se sustentó sobre dos bases de datos. La primera para los casos clínicos y los reactivos de opción múltiple, la segunda para los datos del residente y las respuestas a cada una de las preguntas.

Se utilizaron 63 computadoras; el software fue instalado en cada una de ellas mediante un programa de instalación en disquete, media hora antes del inicio del examen. Las bases de datos que contenían las respuestas de cada uno de los residentes fueron procesadas por medio de un programa diseñado ex profeso, para integrarlas todas en una sola base que incluyera los dos cuadernillos del examen. A los residentes de este grupo se les aplicó un cuestionario de opinión respecto a la modalidad del examen, ventajas y desventajas de la misma y sugerencias. Los alumnos que contestaron el examen por el método impreso, vaciaron su respuesta en hojas de lectura óptica, las cuales fueron leídas en el lector OPSCAN 5.

Las dos bases se integraron en una, para fines del análisis y se procesaron por medio de dos programas calificadores; SICAM versión 1.1. y KALT versión 4.1. Los cuales presentan puntajes e informes estadísticos.

Para ambos grupos se llevó un control individual del tiempo de inicio y terminación de cada una de las partes. Para el grupo A, el control fue manual y para el grupo B, el mismo programa registró los tiempos correspondientes.

Para la evaluación del rendimiento académico de los residentes en el examen, se calcularon los siguientes indicadores: media, desviación estándar, mínimo y máximo de aciertos.

En el análisis del comportamiento del examen se calcularon: el grado de dificultad y el poder de discriminación del examen, la confiabilidad del instrumento se obtuvo por medio del alfa de Cronbach. Para el análisis del comportamiento de los reactivos se clasificaron y calcularon el número de reactivos: muy difíciles, difíciles, fáciles, muy fáciles; así como los reactivos con discriminación positiva y negativa. Todo ello lo dan ambos programas calificadores.

El grado de dificultad: se refiere a la complejidad que representa para cada alumno la respuesta de cada reactivo. Mientras más alumnos contesten correctamente el reactivo, será clasificado como fácil o muy fácil; por el contrario mientras más alumnos lo contesten erróneamente, el reactivo será considerado difícil o muy difícil. El gado de dificultad clasifica a los reactivos en:

Reactivos	Grado de dificultad (%)
Muy difíciles	< 10
Difíciles	10<27
Aceptables	27<73
Fáciles	73< 90
Muy fáciles	>90

El poder de discriminación: se refiere a la capacidad de cada reactivo para que a través de su respuesta correcta pueda diferenciar entre los alumnos de mayor y menor rendimiento en el examen, entiéndase como rendimiento en el examen, el número de respuestas correctas obtenidas por cada alumno. Si el mayor número de alumnos que contestaron correctamente el reactivo tienen un alto rendimiento en el examen, el reactivo discrimina positivamente; si por el contrario fue contestado por un mayor número de alumnos con bajo rendimiento en el examen, el reactivo discrimina negativamente y deberá ser eliminado.

La confiabilidad: se refiere a la capacidad de un instrumento de evaluación para valorar los conocimientos de un grupo de alumnos con características específicas y que los resultados obtenidos se repitan en aplicaciones posteriores, ya sea al mismo grupo o a otros grupos de características semejantes.

Para el análisis estadístico, se utilizó, la *t* de student para comparar el rendimiento académico y el comportamiento de los reactivos de los dos grupos y la χ^2 para el análisis comparativo del comportamiento del examen.

Resultados

El análisis del rendimiento académico de los residentes, por grupo, se presenta en el cuadro I. En él, se puede observar una diferencia significativa en el promedio de aciertos entre ambos grupos, siendo mayor para el grupo A (p<0.001), el máximo y mínimo de aciertos también fue mayor en éste, lo cual indica una mayor amplitud en el rango de respuesta del grupo B.

Cuadro I. Rendimiento académico de los residentes

Grupo	Residentes (n)	Aciertos Media ± DE	Máximo	Mínimo
A	67	118±22*	152	81
B	63	91±15	144	47
Ambos	130	105±20	152	47

*p> 0.001 en comparación con grupo B

El comportamiento del examen (cuadro II) muestra un grado de dificultad del examen del 56% en el grupo A y 43% en el grupo B. Respecto al poder de discriminación, 85% de los reactivos del grupo A discriminaron positivamente en tanto que en el grupo B lo hicieron el 78% de ellos. Las diferencias entre ambos indicadores no fueron significativas, lo que demuestra que el comportamiento general del examen y de los reactivos fue similar en ambos grupos. La confiabilidad fue de 0.903 en el grupo B, y de 0.853 para el A, ambos valores caen en un rango aceptable de confiabilidad.

Cuadro II. Grado de dificultad, poder de discriminación y confiabilidad del examen

Grupo	Dificultad (%)	Discriminación (%)	Confiabilidad
A	56*	85**	0. 853
B	43	78	0.903
Ambos	50	84	0.920

*p< 0.11; **p < 0.28 en comparación con grupo B

Ante estos resultados se decidió analizar en forma independiente los cuadernillos I y II del examen, para cada grupo. Los resultados aparecen en el cuadro III, para el cuadernillo I se observa , una diferencia estadísticamente significativa en las medias de aciertos,mayor en el grupo A. En el cuadernillo II, la diferencia es solo de dos aciertos más en el grupo A.

Cuadro III. Aciertos por cuadernillo

Grupo	Aciertos (Media ± DE) Cuadernillo I	Cuadernillo II
A	62±8*	57±10**
B	36±10	55±16

*P< 0.001; **P< 0.39 en comparación con grupo B

El análisis de los reactivos por grado de dificultad (cuadro IV) ratifica lo mencionado en los párrafos anteriores, el grado de dificultad fue menor para el grupo A, donde el número de reactivos fáciles y muy fáciles fue de 65, en tanto que para el grupo B fue de 32.

Cuadro IV. Grado de dificultad de los reactivos del examen

Grupo	Reactivos (n) Muy difíciles	Difíciles	Aceptables	Fáciles	Muy fáciles
A	3	31	111	43	22
B	12	55	111	30	2
Ambos	5	30	135	38	2

En el cuadro V se observa que el número de reactivos con discriminación positiva fue mayor en el grupo A (16 reactivos más).

Del cuestionario de opinión aplicado a los residentes que respondieron el examen en computadora, el 30% consideró entre excelente y muy buena la aplicación del examen a través de esta tecnología; 61% la consideró buena y 9% expresó aspectos de deficiencia del programa; como que era necesario mejorarlo, incrementar el tamaño de la letra, facilitar el manejo del examen, hacer los comandos más ágiles y sencillos.

Entre las ventajas: 37% refirió que era más rápido, 16% más fácil, 13% más cómodo, 5% que se requiere menos tiempo de aplicación y 11% expresó diversos calificativos positivos. Solamente 5% refirió no encontrar ventaja alguna, les resultó irrelevante.

Entre las desventajas anotadas destacan: 25% el riesgo de fallas en el sistema, 9% el que las personas con problemas visuales pueden ver afectados sus resultados por falta de claridad en la lectura, 5% falta de experiencia o desconocimiento en el uso de la computadora, 54% no mencionó desventajas.

Cuadro V. Poder de discriminación de los reactivos del examen

Grupo	Reactivos (n)	
	Discrimanción positiva	Discriminación negativa
A	179	31
B	163	47
Ambos	176	34

Sólo 21% hizo sugerencias, de estás 46% fueron dirigidas a continuar con el uso de la computadora en la aplicación de los exámenes, 23% sugirieron que al término del examen automáticamente se les proporcionara la calificación. Un residente propuso que la elección del método de examen fuera voluntaria.

Cuadro VI. Duración del examen

Grupo	Tiempo (min)	
	Menor	Mayor
A	59	115
B	53	114

Para finalizar, en este análisis se comparó el tiempo que los alumnos dedicaron a dar respuesta al examen. En ello se consideró que después de concluido el cuadernillo I los residentes tuvieron un descanso de quince minutos para iniciar el cuadernillo II (simultáneamente ambos grupos), el control de tiempo se llevó durante la aplicación del primer cuadernillo. La hora de inicio fue igual para todos, se cuantificaron los tiempos en que terminaron el primero y el último de los alumnos de cada grupo y el promedio de duración del examen por grupo (Cuadro VI). En el grupo A, el primer alumno terminó en 59 minutos y en el grupo B en 53 minutos, con una diferencia de seis minutos a favor del grupo B. El alumno del grupo A que lo realizó en mayor tiempo lo concluyó en 115 minutos y en el grupo B en 114 minutos, con una diferencia de un minuto más para el grupo A.

Discusión

Al comparar los resultados del presente estudio con las opiniones proporcionadas por los residentes de tercer año de Medicina Interna, encontramos cierta discrepancia, ya que a pesar de la opinión muy favorable hacia el examen en computadora, el mayor rendimiento académico se dio en el examen impreso. Por lo que si se tratara de una evaluación en donde el rendimiento académico obtenido repercutiera en la obtención de un grado o bien en la acreditación o no de una asignatura, su uso estaría en duda.

Por ello decidimos analizar el comportamiento del rendimiento de los alumnos por cada uno de los dos cuadernillos. Se encontró que la disminución en la media de aciertos se daba sólo en el primero lo cual, al recordar lo revisado en la literatura respecto a la necesidad de capacitar a los alumnos en el manejo de la computadora, sugería que el factor determinante de esta disminución, pudo ser el desconocimiento en el manejo de la máquina o el temor a enfrentar un examen a través de ese medio. Este efecto disminuyó después de la práctica obtenida con el primer cuadernillo, por lo que en el segundo lograron un resultado similar al obtenido por el grupo A.

La mayoría de los residentes de tercer año, que contestaron el examen, ingresaron a la licenciatura entre 1992 y 1994, época en que en la mayoría de las escuelas de medicina aún no incluían la enseñanza de la computación como parte de su plan de estudios, aún más, no se contaba con equipos accesibles a los alumnos, por lo que es muy factible que ese sea un factor determinante del problema encontrado.

Otro aspecto relevante de este estudio, es el hecho de que la confiabilidad del examen, siendo el mismo, fue mayor en el examen aplicado en la computadora (0.90) que en el impreso (0.85), lo cual habla a favor del primer método.

Después de este análisis es conveniente dirigir nuestros esfuerzos a mejorar los programas de aplicación de exámenes en computadora, hacerlos más amigables a los usuarios y mejorar la calidad de la presentación; especialmente cuando, como en este estudio, se utilizan casos clínicos y reactivos de opción múltiple en donde es conveniente que el alumno tenga en la pantalla el caso y los reactivos completos, sin necesidad de estar cambiando la pantalla. Otra recomendación es la de incluir en el programa una serie de ejercicios de práctica para que el alumno se familiarice con su manejo y adquiera la confianza para responder el examen con mayor seguridad.

Es importante obtener el mayor y mejor beneficio de la tecnología, por lo cual es recomendable continuar realizando este tipo de estudios y favorecer el desarrollo de mejores exámenes en los cuales se incorporen estudios de gabinete, imágenes macro y microscópicas,

sonido, video y dinámicas interactivas. Así como favorecer la aplicación de exámenes en la computadora desde los primeros años de la formación del alumno y fomentar el uso de programas de autoformación con los que vayan adquiriendo seguridad en el manejo de la computadora.

Agradecimiento

Los autores expresan su agradecimiento a la Dra. Guadalupe García de la Torre, por su apoyo técnico para el manejo del paquete estadístico.

Referencias

1. **Lee G, Weerakoon P.** The role of computer-aided assessment in health professional education: a comparison of student performance in computer-based and paper-and-pen multiple choice tests. Med Teach 2001;23:152-157.
2. **Morán AC, Cruz LV.** Uso de la computadora en estudiantes de medicina. Rev Fac Med UNAM; 2001;44:195-197.
3. **Herskovic P, Vásquez A, Herskovic J, Herskovic V, Roizen A, Urrutia MT, Miranda C, Beytía M.** Ownership of computers and abilities for their use in a sample of Chilean medical students. Med Teach 2000;22:197-199.
4. **Butzin DW, Friedman CP, Brownlee RC.** A pilot study of microcomputer testing in paediatrics. Med Educ 1984;18:339-342.
5. **Legler JD, Realini JP.** Computerized student testing as a learning tool in a family practice clerkship. Fam Med 1994;26:14-17.
6. **Ogilvie RW, Trusk TC, Blue AV.** Students' attitudes towards computer testing in a basic science course. Med Educ 1999; 33:828-831.
7. **Whorthley LI.** A computer program to prepare and answer multiple choice questions. Anaesth Intens Care 1985;13:417-419.
8. **Herskovic P.** Reutilization of multiple-choice questions. Med Teach 1999;21:430-431.
9. **Ram P, Van Der Vleuten C, Rethans JJ, Schouten B, Hobma S, Grol R.** Assessment in general practice: the predictive value of written-knowledge tests and a multiple-station examination for actual medical performance in daily practice. Med Educ 1999;33:197-203.
10. **Dugdale AE.** The Pathway MCQ: a method for teaching and testing deeper knowledge. Med Teach 1998;20:250-253.
11. **Norcini JJ, Swanson DB, Grosso LJ, Webster GD.** Reliability, validity and efficiency of multiple choice question and patient management problem item formats in assessment of clinical competence. Med Educ 1985;19:238-247.
12. **Scheuneman JD, Van Fan Y, Clyman SG.** An investigation of the difficulty of computer-based case simulations. Med Educ 1998;32:150-158.
13. **Schuwirth LWT, Blackmore DE, Mom E, Van Den Wildenberg F, Stoffers HEJH, Van Der Vleuten CPM.** How to write short cases for assessing problem-solving skills. Med Teach 1999;21:144-150.
14. **Miller AP, Haden P, Scwartz PL, Loten EG.** Pilot studies of in-course assessment for a revised medical curriculum: II. Computer-based, individual. Acad Med 1997; 72: 1113-1115.
15. **Ram P, Van Der Vleuten C, Rethans JJ, Grol R, Aretz K.** Assessment of practicing family physicians: comparison of observation in a multiple-station examination using standardized patients with observation of consultations in daily practice. Acad Med 1999;74:62-69.
16. **Lowe D.** Set a multiple choice question (MCQ) examination. Br Med J 1991;302:780-782.
17. **Ram P, Van Der Vleuten C, Tethans JJ, Schouten B, Hobma S.** Assessment in general practice: the predictive value of griten-knowledge test and multiple-station examination for actual medical performance in daily practice. Med Educ 1999;33:197-203