Revista de la Educación Superior
Vol. XXXVII (3), No. 147, Julio-Septiembre de 2008, pp. 31-42. ISSN: 0185-2760.

LA EVALUACIÓN DEL CONOCIMIENTO EN MEDICINA

Rodolfo Rodríguez Carranza*

* Departamento de Farmacología, Facultad de Medicina, UNAM. Correo e: rodcar@servidor.unam.mx
Ingreso: 17/01/08
Aprobación: 12/05/08

Resumen

Se destaca la importancia de la evaluación en medicina y la variedad de formatos que se utilizan para evaluar el aprendizaje de los alumnos y las competencias clínicas de los egresados. Se especifica que el examen de opción múltiple es el formato más frecuentemente utilizado, y se analizan sus características, alcances y limitaciones. Se reconoce que las computadoras y la evolución de los programas de cómputo son dos que factores han renovado el interés en los reactivos de opción múltiple e impulsado el desarrollo de exámenes computarizados, los cuales ahora se aplican regularmente en muchas escuelas de medicina.

Palabras clave: educación médica, evaluación, competencias.

Abstract

This article first discusses some general issues on complexity and relevance of medicine assessment, and describes the wide variety of examination formats that have been developed to evaluate students´ achievements and competence of practicing physicians. Then gives an overview of the multiple-choice test, the most commonly used type of test item, with its major advantages and disadvantages. This article also refers to computer-based testing, which has recently gained popularity as testing modality in most medical schools.

Key words: medical education, evaluation, competencies.

Introducción

Se reconoce que la medicina es una ciencia muy extensa y compleja, cuya enseñanza y aprendizaje representa un reto formidable para los educadores, para los evaluadores y para los estudiantes de medicina. En el campo educativo, según los objetivos que señalan los diversos planes de estudio, el alumno está obligado a adquirir, además de un volumen impresionante de información biomédica y médica, las habilidades y valores establecidos para el ejercicio de la medicina, y los atributos esenciales de la buena instrucción universitaria, como son interés continuo en la superación académica y en el aprendizaje independiente, adaptabilidad para el cambio, y habilidad para pensar de manera crítica, para educar, y para comunicarse claramente. Todo ello con sentido humanista y social. Si se consideran esos objetivos educacionales, el médico debe ser educado y evaluado como clínico, como científico, como humanista, y como docente (Reddy y Vijayakumar, 2000).

Como puede apreciarse, la evaluación en medicina no es un tema menor y cumple funciones personales, institucionales, curriculares y sociales (Epstein y Hundert, 2002). Por ejemplo, sus resultados generan información útil a los estudiantes para sustentar la toma de decisiones, como reforzar o cambiar sus métodos y hábitos de estudio, orientar su estudio independiente y, si es necesario, programar medidas de recuperación; por lo cual, la evaluación, particularmente la de carácter formativo, es un instrumento muy valioso que impulsa el interés en el desarrollo personal. Las instituciones, a través de la evaluación, determinan el grado de avance del aprendizaje, el dominio de una disciplina y el logro de las competencias establecidas en el plan de estudios; información que les permite validar la pertinencia de una promoción, la obtención de un grado y la autorización para el ejercicio de la medicina (evaluación sumativa); asimismo, información que es, o debe ser, fundamento de cambios curriculares. Al final de los estudios, los exámenes de egreso constituyen la principal garantía a la sociedad de la competencia profesional de los médicos generales y de los especialistas.

Por todas esas razones, y porque la evaluación es parte integral de la educación médica (Schuwirth y Van der Vleuten, 2004), los procedimientos de evaluación en medicina han sido motivo de diversos estudios y de un gran número de cuestionamientos y, actualmente, el desarrollo de nuevos y mejores métodos y esquemas de evaluación es tema de numerosas investigaciones, particularmente de las escuelas de medicina que verdaderamente buscan la excelencia académica de sus egresados.

En este artículo se describen algunos de los procedimientos desarrollados para evaluar el conocimiento en medicina, se analizan las características, ventajas y desventajas de los exámenes constituidos por reactivos de opción múltiple, que es el formato de examen más frecuentemente utilizado, y su empleo en los exámenes sustentados en la tecnología informática (computarizados).

La evaluación en medicina

En nuestro campo, el término competencia refiere los conocimientos, habilidades (psicomotoras y cognitivas) y valores que debe tener un estudiante de medicina, y el término desempeño hace referencia a la pericia del médico en la práctica clínica real. En la literatura médica se pueden encontrar varias definiciones de competencia. Destaca la propuesta de Southgate (1999), quien señala que la competencia en medicina "está compuesta por conocimientos, habilidades interpersonales y atributos morales y personales; que es en parte la habilidad y en parte el deseo para seleccionar y efectuar tareas clínicas relevantes en el contexto de un ambiente social con el fin de resolver problemas de salud de los individuos de manera eficiente, económicamente efectiva, y con sentido humanista", y la de Epstein y

Hundert (2002), quienes señalan que la competencia clínica es "el uso habitual y juicioso de la comunicación, conocimiento, habilidades técnicas, razonamiento clínico, emociones, valores y reflexión en la práctica diaria para beneficio de un individuo y de la comunidad a la cual se sirve". Definición que parece referirse más al desempeño profesional que a las competencias, pero que deja claras las dificultades inherentes a su evaluación.

Para evaluar el conocimiento en medicina se han desarrollado diversos procedimientos (Collins y Gamble, 1996; Epstein y Hundert, 2002; Van der Vleuten y Schuwirth, 2005); entre los más frecuentemente utilizados se pueden mencionar: a) examen con reactivos de opción múltiple (Norcini *et al*, 1985; Collins, 2006); examen con respuesta estructurada por el estudiante (Schafer *et al*, 2005); c) examen ante pacientes reales, hospitalizados y externos, con o sin escala de medición (Reddy y Vijayakumar, 2000); d) examen ante pacientes estandarizados (Reddy y Vijayakumar, 2000; Tamblyn *et al*, 1991; Pololi, 1995); e) examen clínico objetivo y estructurado (Harden y Gleeson, 1979; Petrusa *et al*, 1987; Sloan *et al*, 1995); f) examen oral, estructurado y no-estructurado (Kelley *et al*., 1971; Williams *et al*, 1987); g) manejo del problema principal de un paciente (Van der Vleuten y Newble, 1995); y h) portafolio (Friedman Ben-David *et al*, 2001). Todos estos procedimientos tienen ventajas y desventajas. La selección de alguno de ellos depende, precisamente, de un análisis cuidadoso.

Algunos de los criterios a considerar para su selección se derivan de sus propiedades psicométricas, en particular validez y confiabilidad. Por definición, el término validez indica la medida en que el examen mide la competencia que se propone evaluar, y específicamente hace referencia: a) al contenido (validez de contenido), que señala el grado de relación entre lo que se pregunta y los objetivos educacionales del plan y/o del programa de estudios; b) a la construcción (validez de construcción), que hace referencia al grado en que legítimamente se pueden hacer inferencias del contenido del reactivo al concepto que se desea medir; y c) a la predicción (validez de predicción), que refiere la capacidad de predecir el desempeño académico o profesional futuro. Mientras que el término confiabilidad indica la medida en que el puntaje del examen es consistente y puede ser generalizado (Schuwirth y Van der Vleuten, 2004). Como es bien sabido, los reactivos incluidos en un examen sólo suelen representar una muestra pequeña del conjunto de posibles preguntas relevantes que pueden ser incorporadas; por ello es importante que el puntaje en una prueba sea indicativo del valor que el mismo estudiante puede sacar en otro grupo de preguntas relevantes.

En la selección del procedimiento de evaluación también deben considerarse otros aspectos: a) aceptabilidad, que se refiere a la opinión que alumnos y profesores tienen sobre el tipo de examen y a su disposición para aceptarlo y, en el caso de los profesores, para elaborar los reactivos correspondientes; b) capacidad discriminatoria, que hace referencia a la capacidad del examen para diferenciar un nivel de conocimientos de otro; y c) costos, que considera los recursos económicos y tiempo necesarios para la elaboración, aplicación y calificación de exámenes (Schuwirth y Van der Vleuten, 2004). Los costos en tiempo y recursos varían considerablemente de un tipo de examen a otro.

En general, ninguno de los formatos de examen arriba descritos es superior a otro y, de manera aislada, no evalúan satisfactoriamente todas las competencias de una ciencia tan compleja como la medicina (Wass *et al*, 2001a). Por ello, la mayoría de los autores concluyen que se requieren varios formatos de evaluación para establecer con certeza el grado de aprendizaje de los estudiantes y la competencia clínica global de los egresados (Schwartz *et al*, 1992; Schuwartz *et al*, 1994; Hull *et al*, 1995; Collins y Gamble, 1996; Schuwirth y Van der Vleuten, 2004; Carr, 2004; Van der Vleuten y Schuwirth, 2005). Asimismo, para una evaluación útil y certera, es muy importante que los planes y programas de estudio describan

claramente lo que los estudiantes deben aprender, ya que ello define qué y cómo evaluar y, en consecuencia, la combinación de procedimientos idóneos a los objetivos educacionales de cada nivel educativo (Wass *et al,* 2001b).

En general, el dominio de los conocimientos de ciencias básicas puede ser más fácilmente evaluado con exámenes de opción múltiple, exámenes orales y ensayos, pero se requieren procedimientos más sofisticados para evaluar las diferentes facetas de la competencia clínica; entre ellos se encuentran los formatos que utilizan pacientes estandarizados, reales (hospitalizados y externos), y el denominado examen clínico objetivo y estructurado (ECOE).

En el ECOE, formato que se utiliza cada vez con mayor frecuencia, los estudiantes pasan a través de una serie de estaciones (8 a más de 20) diseñadas para evaluar habilidades clínicas predeterminadas (historia clínica, exploración, interpretación de resultados de laboratorio, diagnóstico diferencial, manejo integral del paciente, comunicación, etc.) en un tiempo también predeterminado (5 a 30 min). A una señal, el examinando pasa a la siguiente estación y así continúa hasta que se termina el número de estaciones programadas (Harden y Gleeson, 1979; Smee, 2003). En este formato se utilizan pacientes estandarizados; es decir, personas que han sido entrenadas para simular problemas de salud de una manera consistente, confiable y relativamente realista. El ECOE (OSCE, por sus siglas en inglés) ha ganado muchos adeptos, pero se trata de un procedimiento costoso en tiempo y recursos (Barman, 2005).

En la literatura médica se aprecia cierta tendencia a la combinación de por lo menos tres de los formatos de examen arriba descritos; entre ellos: a) de opción múltiple, incluido en la totalidad de las propuestas; b) clínico objetivo y estructurado, que ha cobrado fuerza durante los últimos años; c) con pacientes reales, que representa la situación más cercana al ejercicio real de la medicina; y d) oral, que continúa siendo uno de los más favorecidos. Cabe mencionar que en algunas instituciones, a lo largo de la educación médica se utiliza una combinación de seis diferentes tipos de examen (Davis, 2003).

Por ser los más frecuentemente utilizados y por su relación con los exámenes computarizados, los cuales han estado desplazando a los exámenes escritos, en este trabajo sólo se analizan con detalle las características principales de los exámenes de opción múltiple.

Exámenes de opción múltiple

Como ya se mencionó, los exámenes constituidos por reactivos de opción múltiple (ROM) han sido y son los instrumentos más frecuentemente utilizados a lo largo de la educación médica para evaluar formalmente el aprendizaje y el progreso académico de los alumnos de licenciatura y de posgrado; también se les utiliza en los exámenes profesionales, de ingreso a las residencias médicas, y de certificación (Norcini *et al*, 1985; Reddy y Vijayakumar, 2000; Collins, 2006). La gran aceptación de los ROM para la evaluación del conocimiento médico depende de la posibilidad de usar un número alto de reactivos, lo que permite incluir reactivos sobre los temas representados por los contenidos y objetivos más importantes de una disciplina, de un ciclo escolar o de un plan de estudios. Su aceptación también depende de que se pueden aplicar en un tiempo relativamente corto y de que pueden ser analizados por la computadora (Epstein, 2007). Estas características determinan que los exámenes de opción múltiple se puedan aplicar simultáneamente a un número alto de estudiantes.

Los exámenes de opción múltiple surgieron por primera vez en el campo educativo como parte de un esfuerzo para ordenar la enseñanza formal (elemental y media) y como auxiliares importantes de la evaluación del aprovechamiento escolar, en especial para contender con una población estudiantil cada vez más numerosa (Viniegra, 1979). Por su eficiencia y caracterís-

ticas psicométricas (confiabilidad y validez), su uso se extendió rápidamente a todas las áreas de educación formal, incluida la medicina.

En general, este tipo de reactivos está constituido por dos partes, el enunciado (cuerpo o base del reactivo) que expresa una proposición (imperativa, incompleta, interrogativa) y varias alternativas (tres o más opciones) de respuesta (Garcia *et al*, 2006); las cuales son posibles respuestas o soluciones. Las opciones incluyen la(s) respuesta(s) correcta(s) y una serie de respuestas incompletas o incorrectas, denominadas distractores. A este tipo de reactivos se les clasifica por la forma de respuesta, por su estructura (formato), y por el tipo de aprendizaje que pueden medir. Los formatos de reactivos de opción múltiple más frecuentemente utilizados son los denominados de formato simple. En este caso el enunciado tiene una sola opción de respuesta correcta (única verdadera) y el resto son distractores verosímiles; se trata del formato de examen más versátil y el que regularmente se utiliza en las asignaturas preclínicas.

Otros formatos conocidos, cuyo uso es más frecuente para evaluar el conocimiento clínico, son los de la mejor respuesta, de apareamiento, de omisión, de falso-verdadero, de ítem interpretativo, de falso-verdadero múltiple, y de formato dependiente de contexto (Norcini *et al*, 1985; Schuwirth y Van der Vleuten, 2004; Carr, 2004; García *et al*, 2006).

También se debe mencionar el formato de apareamiento extendido, el cual se utiliza frecuentemente en los exámenes profesionales, de ingreso a las residencias médicas y de certificación, ya que requiere la aplicación del conocimiento médico sobre casos clínicos específicos (reales) y permite que un mayor número de preguntas se resuelva en corto de tiempo, lo que aumenta su confiabilidad y validez. Este formato está constituido por un tema o problema, una lista de opciones o respuestas posibles (siete a 26), una conducción de entrada, y la base del enunciado que refiere un caso clínico (Beullens *et al*, 2002).

La mayoría de los autores acepta que los exámenes de opción múltiple, al reducir al mínimo la intervención humana, eliminan la subjetividad de la evaluación y la inconsistencia de las puntuaciones que se observa en los ensayos, en el examen oral y en el examen ante pacientes; asimismo, el riesgo de sesgo (raza, género) en ese tipo de exámenes. También se acepta ampliamente que cuando los reactivos que los conforman son pertinentes, están claramente escritos y bien construidos, son útiles para medir procesos cognitivos superiores, como interpretación, síntesis y aplicación del conocimiento (capacidad para resolver problemas), y que estos exámenes son válidos, discriminatorios, económicos y confiables (Anderson, 1979; Norcini *et al*, 1985; McCoubrie, 2004; Anderson, 2004; Collins, 2006). De hecho, la confiabilidad es la cualidad principal, y más ampliamente aceptada, de los exámenes de opción múltiple

Diversos autores han cuestionado la validez de los exámenes de opción múltiple para evaluar el aprendizaje en medicina, particularmente la clínica (Newble *et al*, 1979; Viniegra, 1979; Darling-Hammond y Lieberman, 1992). Entre otras críticas, destacan: a) que miden esencialmente niveles cognitivos de orden bajo (memoria), como recordar hechos aislados y/o triviales y que, por lo tanto, no son ideales para evaluar habilidades cognitivas de nivel alto, como razonamiento, síntesis y evaluación; b) que tienen escasa relación con la acción que se requiere en la práctica clínica; c) que no revelan la actitud del estudiante ni su capacidad para integrar conocimientos, para resolver problemas o para comunicarse, y que su aprobación no es garantía de competencia profesional; y d) que evalúan el reconocimiento de la respuesta correcta, no la capacidad de generar espontáneamente una respuesta correcta.

No hay duda de que los reactivos de opción múltiple pobremente construidos inducen en los estudiantes la estrategia errónea de aprender hechos aislados, pero es claro que cuando los reactivos de opción múltiple se preparan cuidadosamente también exploran, como ya se men-

cionó, habilidades cognitivas complejas (Glaser, 1984; Maguire *et al*, 1992; Skakun *et al*, 1994; Van der Vleuten y Newble, 1995; Coderre *et al*, 2004). Además, algunas variantes del examen de opción múltiple han sido desarrolladas con el fin específico de evaluar mejor el razonamiento clínico y la aplicación del conocimiento médico (Aeschlimann *et al*, 2001; Beullens *et al*, 2002; Beullens *et al*, 2005). Y cabe destacar que la construcción de la base del reactivo determina la calidad y el nivel cognitivo de un examen, no el formato o el número de alternativas (Coderre *et al*, 2004); asimismo, la escasa importancia que algunos críticos de los reactivos de opción múltiple otorgan a la memoria, componente fundamental del pensamiento y del razonamiento.

En este trabajo se considera que el examen de opción múltiple es mucho más que un procedimiento confiable para medir la capacidad de recordar y reconocer el conocimiento establecido, y se acepta que los exámenes de opción múltiple bien construidos pueden medir habilidades cognitivas superiores, además de que estimulan el aprendizaje de los alumnos. Tampoco se trata, desde luego, de un procedimiento ideal ni completo para la evaluación del conocimiento clínico. Se les acepta como uno de los componentes más útiles de los sistemas de evaluación del conocimiento en medicina, sistemas que deben incluir otros formatos, como arriba se indica.

Conviene agregar que el diseño de reactivos de opción múltiple requiere dominio de la disciplina motivo de la evaluación, así como de la técnica para elaborar un reactivo útil y relevante (Collins, 2006; Garcia *et al*, 2006). De hecho, la creación de un banco de reactivos de opción múltiple bien balanceado, no es un ejercicio trivial. Su desarrollo requiere de mucho tiempo, ya que es necesario asegurar la precisión, relevancia e importancia del contenido, además de que la estructura de cada reactivo debe ser aceptable en términos de un enunciado apropiado y estar libre de errores y ambigüedades. También se requiere que, antes de ser utilizados, cada uno de ellos sea revisado y aprobado por los profesores encargados de los cursos y por los comités de evaluación (Palmer y Devitt, 2006).

Aplicación de exámenes computarizados

La introducción y evolución de las computadoras y de la tecnología son dos factores que han impactado de manera determinante en el proceso educativo en todas las áreas del conocimiento. Además del apoyo invaluable al proceso enseñanza-aprendizaje (Greenhalgh, 2006), los avances tecnológicos en ese campo han orientado, en gran medida, la forma en que se evalúan los conocimientos, se aplican y se cuantifican los exámenes, y se reportan los resultados. Estos cambios son particularmente evidentes en los exámenes sustentados en la tecnología informática, que aquí denominamos exámenes computarizados, donde dichos avances han determinado el desarrollo de un gran número de variantes de examen (Zenisky y Sireci, 2002). A la fecha, se pueden distinguir dos tipos básicos de exámenes computarizados. En el primero, se pide a los examinandos marcar la respuesta correcta, al igual que en los exámenes escritos, marcas que reconoce la computadora. En el segundo, la computadora genera una interfase, los examinandos ingresan su respuesta, y son retroalimentados vía la computadora.

La evaluación computarizada es un procedimiento que se utiliza desde hace más de 40 años (Swets y Feurzeig, 1965) y que paulatinamente ha estado desplazando, o complementa, la evaluación escrita en medicina (Doull y Walaszek, 1978; Money *et al*, 1998; Cantillon, *et al*, 2004). El interés inicial en los exámenes computarizados y su aceptación creciente se puede atribuir a su mayor eficiencia (Wise y Kingsbury, 2000), ya que algunos de ellos pueden reducir de manera significativa los tiempos de examen, manteniendo así la calidad de los exámenes escritos. Sus ventajas generales se pueden resumir de la

siguiente manera: a) rapidez; b) no afectan las puntuaciones; c) no requieren programación especial; d) ensamblado automático de exámenes; e) empleo de imágenes de alta calidad; f) mayor aceptación por parte de los estudiantes; g) retroalimentación inmediata a los examinandos; h) almacenamiento automático de los datos y la opción de análisis estadístico; e i) opción para que los estudiantes determinen por sí mismos (autoevaluación) sus niveles de aprendizaje y progreso académico (Cantillon e*t al*, 2004; Vrabel, 2004; Lim *et al*, 2006).

Por todo lo anterior, la evaluación computarizada ahora se aplica de manera sistemática en las escuelas de medicina y se le utiliza en evaluaciones de tipo formativo, sumativo, diagnóstico, y para autoevaluación. Su validez para exámenes con un número alto de reactivos está ampliamente documentada (Russell y Haney, 1997; Wolfson *et al,* 2001). Sus principales desventajas se refieren a la alteración emocional (ansiedad) que genera este tipo de exámenes en los alumnos no familiarizados con el manejo de la computadora (Olea *et al*, 2000), y a los costos de operación (Booth, 1998). También es importante tener en cuenta que la implementación de las evaluaciones computarizadas es difícil y prolongada (Cantillon *et al*, 2004); asimismo, que se requieren medidas especiales de vigilancia y seguridad. Entre ellas, cuando el mismo examen se aplica a varios estudiantes, el uso de pantallas opacas o distribución al azar del orden de los reactivos. Si la evaluación es en línea (Internet), el examen debe efectuarse en un ambiente donde el acceso a las fuentes de información y al correo electrónico esté controlado.

Para los exámenes computarizados se utilizan los formatos de reactivos de opción múltiple previamente descritos; además, los avances tecnológicos han impulsado el desarrollo de otros formatos más sofisticados, muchos de ellos interactivos, con los cuales se pretende mejorar la calidad de la evaluación y aumentar la eficiencia de los procedimientos tradicionales. Huff y Sireci (2001) y Zenisky y Sireci (2002), refieren un total de 21 formatos diferentes, entre los que destacan los de selección y colocación, de análisis de casos breves, y los de solución secuencial de problemas. También se han descrito formatos para evaluar el desempeño técnico, como habilidades quirúrgicas básicas (Datta *et al,* 2002).

Por sus características, conviene destacar los exámenes computarizados adaptativos (CATs, por sus siglas en inglés), también denominados tests adaptativos informatizados (TAIs). Se trata de una forma especial de examen "a la medida" (Kreiter *et al*, 1999; Ponsoda, 2000; Roex y Degryse, 2004; Gershon, 2005), cuyo objetivo básico es seleccionar el grupo de reactivos que mejor informa sobre el nivel de preparación de cada examinando (examen personalizado). El CAT se deriva de las pruebas adaptativas desarrolladas en 1905 por Bidet y Simon (citado por Weiss, 2004) y sus métodos están basados en la Teoría de Respuesta al Item (TRI); teoría estadística que genera en una serie modelos matemáticos que permiten estimar la probabilidad de una respuesta particular, en una escala de ítems, como función de los atributos del examinando y ciertas características (parámetros) del ítem (Lord y Novick, 1968; Green *et al*, 1984; Wise y Kingsbury, 2000). Ello implica que la probabilidad de que un examinando responda correctamente ante un ítem dado depende básicamente de los conocimientos y habilidades del examinando. Además, el principio invariable de la TRI facilita la generación de diferentes grupos de reactivos para diferentes examinandos y la estimación de su nivel de preparación en la misma escala de medición, lo cual permite comparar los puntajes de estudiantes que presentaron examen con diferentes juegos de reactivos obtenidos del mismo banco (Chang y Reeve, 2005). Los elementos básicos del CAT son dos: a) un banco robusto (2000 para una licenciatura) de reactivos calibrados; y b) los algoritmos interactivos necesarios para presentar un reactivo, estimar el nivel de aprovechamiento del examinando basándose en la respuesta al ítem previo, seleccionar del banco el siguiente mejor ítem, y así suce-

sivamente hasta que se alcance una condición predeterminada para el examen (error estándar mínimo, contenido, máximo de ítems) y para determinar cuándo se debe terminar la prueba de cada examinando (Wise y Kingsbury, 2000; Chang y Reeve, 2005).

De manera práctica, se puede señalar que el CAT opera y utiliza el principio de que los reactivos demasiado fáciles o demasiado difíciles para un estudiante contribuyen muy poco para conocer el nivel académico real de ese estudiante (Green *et al*, 1984). Así, una vez que el examinando ingresa al CAT, el sistema revisa los antecedentes del estudiante (haber tomado previamente el examen, nivel de estudios) y, con base en ello, selecciona el primer reactivo; si no hay antecedentes, el sistema selecciona el primer reactivo en función de un nivel preestablecido (medio). Una vez que se emite la primera respuesta, y en conjunción con las especificaciones iniciales del examen, un algoritmo selecciona el mejor reactivo disponible para continuar el examen de ese estudiante. En cada ocasión que se emite una respuesta, ésta es calificada inmediatamente y el programa decide el grado de dificultad y discriminación del siguiente reactivo. Si el estudiante responde correctamente, los reactivos subsecuentes son más difíciles, mientras que las respuestas incorrectas conducen a reactivos cada vez más fáciles. De esta manera, los reactivos que están más allá del nivel de conocimientos del estudiante (demasiado fáciles o difíciles) son evitados; lo que personaliza el examen.

Con este procedimiento, el número de reactivos (y tiempo) necesario para evaluar el nivel de aprovechamiento del estudiante es considerablemente menor que el requerido en exámenes escritos (Krieter *et al*, 1999). Esta ventaja es particularmente atractiva para exámenes muy extensos, ya que en este caso el problema principal que enfrenta el examinando es la fatiga (por ejemplo, examen de ingreso a las residencias médicas, constituido habitualmente por 700 reactivos). Además, este tipo de exámenes parece reducir la angustia y frustración de los examinandos (Olea *et al,* 2000). En todo caso, la calidad de este tipo de exámenes depende de la disponibilidad de un banco de reactivos bien construidos, pertinentes (válidos), y calibrados. Más que una colección de reactivos, los bancos para CATs deben contener ítems que desarrollen, definan y cuantifiquen un tema común, con ello se proporciona una definición operativa de sus características (Green *et al*, 1984).

Por otra parte, este sistema de evaluación no está libre de problemas y riesgos. Por ejemplo, el empleo de criterios estadísticos para la selección de los reactivos no asegura una prueba con contenidos totalmente válidos (Hambleton *et al*, 1991). Asimismo, y por las características del sistema, algunos reactivos se utilizan de manera repetitiva (alta exposición), mientras que otros son ocasionalmente seleccionados o nunca tocados por el algoritmo correspondiente; ello altera el nivel de discriminación (Chang y Van der Linden, 2003). También debe tomarse en cuenta que este sistema es muy vulnerable cuando está en línea (Internet) y que requiere de medidas adicionales de vigilancia y seguridad (Wise y Kingsbury, 2000). Los reactivos pueden ser extraídos por el sustentante del examen y compartidos con futuros examinandos. Asimismo, las medidas de seguridad del banco de reactivos pueden ser vulneradas por personas u organizaciones (*hackers*) ilegalmente interesadas en el banco. La seguridad de la información en el Internet es siempre cuestionable; ello debilita seriamente el valor del sistema en línea para la toma de decisiones importantes (Butcher *et al*, 2004).

Se puede concluir que los programas computarizados adaptables ofrecen una alternativa de evaluación muy interesante, que merece atención y estudios especiales. Aun cuando ya se les utiliza en exámenes de grado y de certificación (Roex y Degryse, 2004; Bergstrom y Lunz, 1999), su verdadera utilidad en las evaluaciones formales del conocimiento médico está por determinarse.

Finalmente, se puede agregar que los resultados de nuestra búsqueda bibliográfica revelan

que los exámenes computarizados no se utilizan formalmente en nuestras escuelas de medicina. Asimismo, que se ha puesto en marcha un proyecto piloto (experimental) que pretende establecer, en nuestro medio, las ventajas y desventajas del uso de las computadoras en la evaluación del aprendizaje de los alumnos de medicina. Su desarrollo está planeado en varias etapas. La primera de ellas, ya terminada, y la más costosa en tiempo, se refiere a la elaboración de un banco de reactivos útiles (clasificados y calificados) y al diseño de un programa de cómputo que permite la selección al azar de los reactivos del banco y la elaboración y aplicación simultánea de exámenes diferentes. En su segunda etapa (en proceso), este proyecto sólo contempla una de las asignaturas básicas en la tercera fase, el estudio se extenderá, también de manera experimental, a otras asignaturas preclínicas y a las asignaturas clínicas. Si los resultados de este proyecto son satisfactorios, se puede considerar la pertinencia de los exámenes computarizados en otro tipo de evaluaciones (exámenes profesionales, de ingreso a las residencias médicas, de certificación).

Referencias

Aeschlimann, A. *et al.* (2001). "Multiple choice question quiz: a valid test for needs assessment in CME in rheumatology and for self assessment", en *Ann Rheumatol Dis,* 60.

Anderson, J. (1979). "For multiple choice questions", en *Med Teach*, 1.

Anderson, J. (2004). "Multiple-choice questions revisited", en *Med Teach*, 26.

Barman, A. (2005). "Critiques on the Objective Structured Clinical Examination", en *Ann Acad Med,* Singapore, 34.

Bergstrom, B. A. & M. E. Lunz (1999). "CAT for certification and licensure", en F. Drasgow & J. B. Olson-Buchanan (Eds.), *Innovations in Computarized Assessment*. Hillsdate, N.J., Lawrence Erlbaum.

Beullens J, *et al.* (2002). "Are extended-matching multiple-choice items appropriate for a final test in medical education?", en *Med Teach*, 24.

Beullens, J. *et al.* (2005). "Do extended matching multiple-choice questions measure clinical reasoning?", en *Med Educ*, 39.

Booth, J. F. (1998). "Guest Editorial: Uses of PC technology in selection and assessment*",* en *Int J Selec Assess*, 6.

Butcher, J. N. *et al.* (2004). "Computers in clinical assessment: historical developments, present status, and future challenges", en *J Clin Psychol*, 60.

Cantillon, P. *et al.* (2004). "Using computers for assessment in medicine", en *Brit Med J*, 329.

Carr, S. J. (2004). "Assessing clinical competency in medical senior house officers: how and why should we do it?", en *Postgrad Med J*, 80.

Coderre, S. P. *et al.* (2004). "The impact of two-choice formats on the problem-solving strategies used by novices and experts", en *BMC Med Educ*, 4.

Collins, J. P. & G. D. Gamble (1996). "A multi-format interdisciplinary final examination", en *Med Educ*, 30.

Collins, J. P. (2006). "Educational techniques for lifelong learning: writing multiple-choice questions for continuing medical education activities and self-assessment modules", en *Radiographics*, 26.

Chang, H. H. & W. J. van der Linden (2003). "Optimal stratification of item pools in alpha-stratified computerized adaptive testing", en *Appl Psychol Meas*, 27.

Chang, C.H. & B.B. Reeve (2005). "Item response theory and its application to patient-reported outcomes measurement", en *Eval Health Prof*, 28.

Darling-Hammond, L. & A. Lieberman (1992). "The shortcomings of standardized tests", en *Chron Higher Educ*, 29.

Datta, V. *et al* (2002). "The relationship between motion analysis and surgical technical assessments", en *Am J Surg*, 184.

Davis, M. H. (2003)."OSCE: the Dundee experience", en *Med Teach*, 25.

Doull, J. & E. J. Walaszek (1978). "The use of computer-assisted teaching systems (CATS) in pharmacology", en E. C. DeLand (Ed.), *Information Technology in Health Science Education (CATS)*, New York, Plenum Pub Corp.

Epstein, R. M. & E. M. Hundert (2002). "Defining and assessing professional competence", en *J Am Med Assoc*, 287.

Epstein, R. M. (2007). "Assessment in medical education", en *New Engl J Med*, 356.

Friedman Ben-David, M. *et al.* (2001). "AMEE medical education guide No. 24: portfolios as a method of student assessment", en *Med Teach*, 23.

Garcia, R. *et al.* (2006). "Elaboración de ítems objetivos" en Castañeda S. (Coord.). *Evaluación de aprendizaje universitario. Elaboración de exámenes y reactivos*, México, Universidad Nacional Autónoma de México.

Gershon, R. C. (2005). "Understanding rasch measurement: computer adaptive testing", en *J Appl Meas*, 6.

Glaser, R. (1984). "Education and thinking: the role of knowledge", en *Amer Psychol*, 19.

Green, B. F. *et al.* (1984). "Technical guidelines for assessing computerized adaptive tests", en *J Educ Meas*, 21.

Greenhalgh, T. (2006). "Computer assisted learning in undergraduate medical education", *Brit Med J*, 322.

Hambleton, R.K. *et al.* (1991). *Fundamentals of item response theory*, Newbury Park, CA, Sage.

Harden, R. M. & F. A. Gleeson (1979). "Assessment of clinical competence using an objective structured clinical examination (OSCE)", en *Med Educ*, 13.

Huff, K. L. & S. G. Sireci (2001). "Validity issues in computer-based testing", en *Educ Meas: Issues and Practice*, 20.

Hull, A. L. *et al* (1995). "Validity of three clinical performance assessments of internal medicine clerks", *Acad Med*, 70.

Kelley, P. R. Jr, *et al.* (1971). "Analysis of the oral examination of the American Board of Anesthesiology", en *J Med Educ*, 46.

Kreiter, C. D. *et al.* (1999). "Evaluating the usefulness of computerized adaptive testing for medical in-course assessment", en *Acad Med*, 74.

Lim, E. C. H. *et al* (2006). "Computer-based versus pen-and-pencil testing: students' perception", en *Ann Acad Med, Singapore*, 35.

Lord, F. M. & M. R. Novick (1968). *Statistical theory of mental test scores*, Reading, MA, Adisson-Wesley.

Maguire, T. *et al.* (1992). "Setting standards for multiple choice items in clinical reasoning", en *Eval Health Prof*, 15.

McCoubrie, P. (2004). "Improving the fairness of multiple-choice questions: a literature review", en *Med Teach*, 26.

Mooney, G. A. *et al.* (1998). "Some techniques for computer-based assessment in medical education", en *Med Teach*, 20.

Newble, D. *et al.* (1979). "A comparison of multiple-choice questions and free-response tests in examination of clinical competence", en *Med Educ*, 13.

Norcini, J. J. *et al.* (1985). "Reliability, validity and efficiency of multiple choice questions and patient management problem items formats in assessment of clinical competence", en *Med Educ*, 19.

Olea, J. *et al.* (2000). "Psychometric and psychological effects of review on computerized fixed and adaptive tests", en *Psicológica*, 21.

Palmer, E. & P. Devitt (2006). "Constructing multiple choice questions as a method for learning", en *Ann Acad Med,* Singapore, 35.

Petrusa, E. R. *et al.* (1987). "An objective measure of clinical performance", en *Am J Med*, 83.

Pololi, L. H. (1995). "Standardised patients: as we evaluate, so shall we reap", en *Lancet*, 25.

Ponsoda, V. (2000). "Overview of the computerized adaptive testing special section", en *Psicológica*, 21.

Reddy, S. & S. Vijayakumar (2000). "Evaluating clinical skills of radiation oncology residents: parts I and II", en *Int J Cancer*, 90.

Roex, A. & J. Degryse (2004). "A computerized adaptive knowledge test as an assessment tool in general practice", en *Med Teach*, 26.

Russell, M. & W. Haney (1997). "Testing writing on computers: an experiment comparing student performance on tests conducted via computer and via paper-and pencil", en *Educ Pol Anal Arch*, 5.

Schafer, W. D. *et al.* (2005). "Resistance to confounding style and content in scoring constructed-response items", en *Educ Meas: Issues and Practice*, 24.

Schwartz, R. W. *et al.* (1992). "Undergraduate surgical education for the twenty-first century", en *Ann Surg*, 216.

Schwartz, R. W. *et al.* (1994). "Assessing senior residents' knowledge and performance: an integrated evaluation program", en *Surgery*, 116.

Schuwirth, L. W. T. & C. P. M. van der Vleuten (2004). "Different written assessment methods: what can be said about their strengths and weakness?", en *Med Educ*, 38.

Skakun, E. *et al.* (1994). "Strategy choices in multiple-choice items", en Acad Med, Supl 10.

Sloan, D. A. *et al.* (1995). "The Objective Structured Clinical Examination. The new gold standard for evaluating postgraduate clinical performance", en *Ann Surg*, 222.

Smee, S. (2003). "ABC of learning and teaching in medicine: skill based assessment", en *Brit Med J*, 29.

Southgate, L. (1999). "Professional competence in medicine", en *Hosp Med*, 60.

Swets, J. A. & W. Feurzeig (1965). "Computer-aided instruction", en *Science*, 15.

Tamblyn, R. M. *et al.* (1991). "The accuracy of standardized patient presentations", en *Med Educ*, 25.

Van der Vleuten, C. P. M. & D. I. Newble (1995) "How can we test clinical reasoning?", en *Lancet*, 345.

Van der Vleuten, C. P. M. & L. W. T. Schuwirth (2005). "Assessing professional competence: from methods to programmes", en *Med Educ*, 39.

Viniegra, L. (1979). "Exámenes de opción múltiple", en *Gac Med Mex,* 115.

Vrabel, M.(2004). "Computarized *versus* paper-and-pencil testing methods for a nursing certification examination: a review of the literature", en *Comput Inform Nurs*, 22.

Wass, V. *et al.* (2001a). "Assessment of clinical competence", en *Lancet*, 357.

Wass, V. *et al.* (2001b). "Composite undergraduate clinical examinations: how should the components be combined to maximize reliability?", en *Med Educ*, 35.

Weiss, D. J. (2004). "Computarized adaptive testing for effective and efficient measurement in counseling and education", en *Meas Eval Counsel Dev*, 37.

Williams, R. G. *et al.* (1987). "Direct, standardized assessment of clinical competence", en *Med Educ*, 21.

Wise, S. L. & G. G. Kingsbury (2000). "Practical issues in developing and maintaining a computerized adaptive testing program", en *Psicológica*, 21.

Wolfson, P. J. *et al.* (2001). "Administration of open-ended test questions by computer in a clerkship final examination", en *Acad Med*, 76.

Zenisky, A. L. & S. G. Sireci (2002). "Technological innovations in large-scale assessment", en *Appl Meas Educ*, 15.