Algunos métodos de evaluación de las competencias: Escalando la pirámide de Miller

Eduardo Durante

INTRODUCCIÓN

En el artículo *"La evaluación de los conocimientos: lo que parece ser, ¿es realmente lo que es?"* del Número 1 de 2005 de esta misma revista[2], se trataron los fundamentos de la evaluación de las competencias en medicina. Allí se comparó a la "incompetencia médica" con una enfermedad que puede ser diagnosticada a través de pruebas que no son perfectas.

Como fue desarrollado en ese artículo, la competencia es específica según el contenido o el contexto. Esto significa que el hecho de lograr una competencia óptima en un área no es un buen predictor de competencia en otra, aún en el caso de que dichas áreas se encuentren muy relacionadas. Esto tiene que ver con que la adquisición de competencias es específica para contenidos o contextos diferentes aunque parezcan similares (la competencia del examen físico en un paciente con insuficiencia cardíaca no predice el desempeño en el examen físico de un paciente con fibrosis pulmonar). Como consecuencia de este fenómeno, es necesario aumentar el número de casos en las evaluaciones para asegurar una adecuada confiabilidad intercasos[13].

Otra conclusión es que un único método no puede valorar a toda la pirámide de Miller. Se necesita una combinación de diferentes métodos.

Además, se mencionó el papel fundamental de la evaluación como orientadora del aprendizaje, por lo que su diseño debe ser estratégico en función de los efectos de aprendizaje deseados.

En este artículo, se describirán las características de algunas pruebas frecuentemente utilizadas en nuestro medio, y se hará una reflexión final sobre su aplicación en el día a día.

ESCALANDO LA PIRÁMIDE DE MILLER

En ese mismo artículo, se presentó a la pirámide de Miller, del año 1990[4]. Este es un modelo para la evaluación de la competencia profesional organizada como una pirámide de cuatro niveles. En los dos niveles de la base se sitúan los conocimientos (saber) y cómo aplicarlos a casos concretos (saber cómo). En el nivel inmediatamente superior (mostrar cómo), se ubica a la competencia cuando es medida en ambientes "in vitro" (simulados) y donde el profesional debe demostrar todo lo que es capaz de hacer. En la cima se halla el desempeño (hace) o lo que el profesional realmente hace en la práctica real independientemente de lo que demuestre que es capaz de hacer (competencia).

Además, durante la década de los ´90 se volvió más evidente la necesidad de una evaluación del aprendizaje más auténtica e integrada, así como la mayor incorporación de los estudiantes a la evaluación.

Hoy en día se ha vuelto bastante clara la noción de que la evaluación tradicional organizada en la combinación de constructos ya no es sustentable. En Educación, el modelo más conocido es el de **conocimientos teóricos, habilidades y destrezas, actitudes.** Los constructos o conceptos son rasgos (*traits*, en idioma inglés) más o menos estables en el tiempo, que pueden ser medidos en forma separada e independientes entre sí, y son genéricos; es decir, son competencias generales no dependientes del contexto. Sin embargo, repetidamente se ha observado que hay más variación dentro de un mismo instrumento de evaluación (de un caso a otro, de una estación a otra en el Examen Clínico Objetivo y Estructurado [ECOE][3]) que la que hay entre diferentes instrumentos. La correlación entre un examen escrito de conocimientos y una estación del ECOE del mismo contenido puede ser mayor que entre dos estaciones del mismo ECOE. Se concluye que lo importante no es el método, sino el contenido para determinar cuál es la competencia medida[7].

Actualmente, a la luz de estos hallazgos, se ha abandonado la búsqueda del instrumento ideal que mida todos los constructos a la vez. La idea actual es que para completar una determinada tarea, es necesario que diferentes aspectos de la competencia estén juntos e integrados. La pirámide de Miller marca el comienzo de esta forma de pensamiento. Cada nivel usa un verbo o acciones que son observables, por lo que pueden ser valoradas y usadas para la evaluación. De esta manera, se acepta en la actualidad que varios instrumentos deben ser combinados para obtener juicios sobre la competencia de los estudiantes en los distintos niveles[9].

Otro factor importante es la autenticidad[8,9]. Su inclusión

Recibido: 18 ago. 2006 Aceptado: 24 ago. 2006

Servicio de Medicina Familiar y Comunitaria. Hospital Italiano de Buenos Aires. Correspondencia: eduardo.durante@hospitalitaliano.org.ar

debería ser prioritaria cuando se están diseñando programas para la evaluación de las competencias médicas. Esto significa que las situaciones en las cuales es evaluada la competencia de los estudiantes se parezca lo más posible a la situación en la que la competencia deba ser utilizada en la realidad. Varias razones sostienen esta afirmación:

1. Las personas guardan y recuperan información de manera más efectiva cuando es aprendida en un contexto relevante[7].
2. Durante el proceso de aprendizaje, las personas almacenan información contextual, alguna de ella aparentemente irrelevante[7].

En conclusión, cuanto más auténtica sea la aproximación al aprendizaje y la evaluación, más información contextual será incorporada en el proceso.

QUÉ PRUEBAS USAR

La pirámide de Miller, como ya fue comentado, presenta 4 niveles de competencia, definidos como "sabe", "sabe cómo", "demuestra" y "hace". En la Figura 1 se observan los cuatro niveles y algunos de los métodos utilizados para evaluar cada uno de ellos.

Describiremos los métodos mencionados en la Figura 1 y además, usaremos la noción de **utilidad** de un examen en donde se vinculan estas variables y se les da diferente peso[2,14]. Así:

$$\textbf{Utilidad} = \textbf{Confiabilidad} \times \textbf{Validez} \times \textbf{Impacto educacional} \times \textbf{Aceptabilidad} \times \frac{1}{\textbf{Costo}}$$

En la Tabla 1, se enumeran diferentes tipos de pruebas y cuáles son sus características en relación a la aplicación de la fórmula de utilidad:

EXÁMENES ESCRITOS

NIVEL DE "SABE" Y "SABE CÓMO"

En este nivel los exámenes son escritos. Este tipo de evaluaciones pueden ser clasificadas como de formato de respuesta o formato de estímulo, según dónde se ponga el énfasis del *ítem* [10,11].

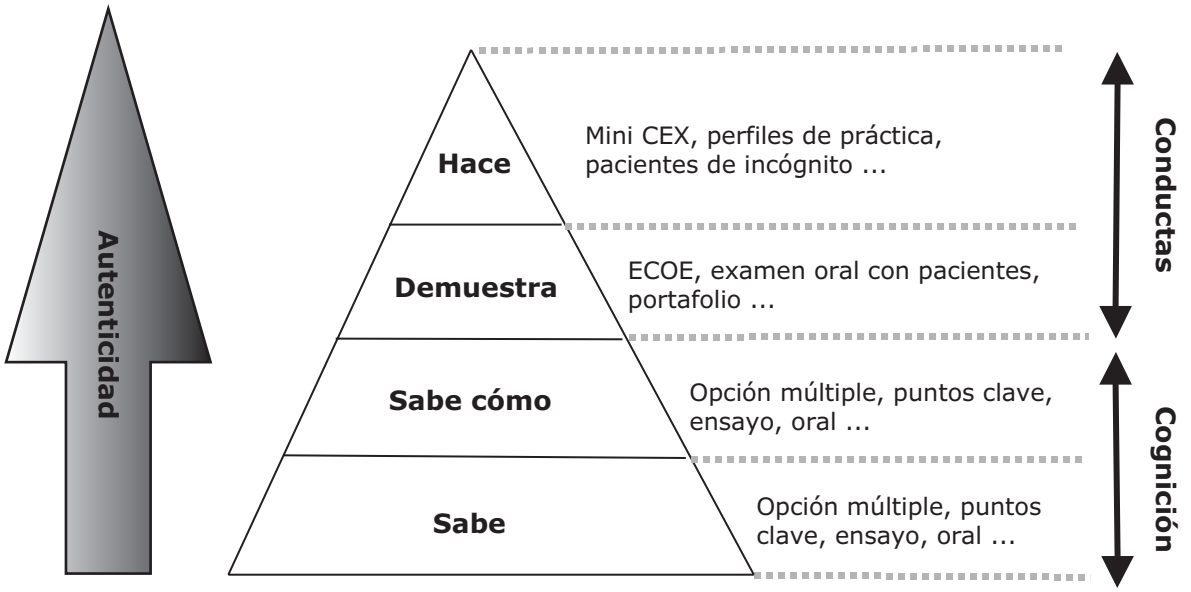
Las pruebas de formato de respuesta incluyen las preguntas de respuesta corta y las de ensayo. Además de la longitud, ambas solicitan al candidato respuestas cognitivas diferentes.

En las preguntas de respuesta de ensayo, se pregunta por conocimiento o procesamiento de la información más que reproducción, requiriendo de los candidatos que establezcan procesos de razonamiento o síntesis de información, o pidiéndoles que apliquen conocimientos en diferentes contextos. La confiabilidad *inter-casos* es baja dada la limitación de presentar muchas preguntas a desarrollar en un tiempo determinado. La confiabilidad *inter-observador* también es baja, dada la disparidad de los criterios de corrección.

Las preguntas de respuesta corta no son mejores que las de elección de opciones múltiples (EOM). Por lo tanto, sólo deberían usarse en las situaciones en las que la generación espontánea de la respuesta sea un aspecto esencial del estímulo (la viñeta o descripción del caso).

Las pruebas de formato de estímulo incluyen las que tie-

Figura 1. la pirámide de Miller y los métodos de evaluación. Mini CEX, *Mini Clinical Evaluation Exercise*; ECOE, Examen Clínico y Objetivo Estructurado.



nen en su enunciado descripciones ricas de claves contextuales, o no. Las que tienen formato rico en contexto evalúan más efectivamente el razonamiento clínico y pertenecen al nivel de "sabe cómo". Las de formato pobre en contexto evalúan conocimiento de tipo memorístico y pertenecen al nivel de "sabe". Estimulan la toma de decisiones más simple, del tipo si / no.
En la Tabla 2, se ejemplifican las diferencias entre formatos.

LA PRUEBA DE ELECCIÓN DE OPCIONES MÚLTIPLES (EOM)

Este tipo de prueba no necesita presentación, ya que cualquier profesional de la salud estuvo expuesto a sus efectos en algún momento de su vida. Su propósito es evaluar conocimiento teórico (nivel de sabe y sabe cómo) [10,11,12].

Sus ventajas logísticas son su mayor fortaleza. Cientos o miles de alumnos pueden ser evaluados al mismo tiempo con mínima participación humana. Sus desventajas más reconocidas son que evalúan (en general, debido a su construcción como de formato de estímulo pobre en contexto), conocimiento de tipo memorístico más que razonamiento más elaborado y la diferencia entre "reconocer" la respuesta correcta en lugar de recuperarla de la memoria. Dado su empleo tan extendido, existe buen grado de evidencia sobre sus características psicométricas.
Sin embargo, como ya se mencionó, es posible diseñar *ítems* con descripciones ricas del contexto (formato de estímulo rico en contexto) que simulan casos reales y evalúan más adecuadamente las competencias del nivel "sabe cómo".
Confiabilidad: los EOM son capaces de tomar muestras de

Tabla 1. Características de las pruebas según la fórmula de utilidad de una evaluación. ECOE, Examen Clínico y Objetivo Estructurado; Mini CEX, *Mini Clinical Evaluation Exercise*; (+), bajo; (++), moderado; (+++), alto.

Tipo de prueba	Confiabilidad	Validez	Impacto educacional	Costo
Opción Múltiple	+++	+++ (de contenido)	+	+
"Puntos Clave"	+++	+++	++	++
Examen oral	+	+	+	+/++
Ensayo	+	+	+	+/++
Caso largo	+	+	+	++/+++
ECOE	++/+++	+++	+++	+++
Mini CEX	++	+++	+++	++/+++
Portafolio	+/++	++	+++	++

Tabla 2. diferencias entre formatos de estímulo rico y pobre en contexto.

Formato pobre en contexto con preguntas abiertas	Formato rico en contexto con preguntas abiertas
Cuando el oxígeno difunde desde el aire alveolar hasta el eritrocito atraviesa diferentes estructuras. Mencione esas estructuras.	El Sr. García es un enfermo terminal. Tiene un mesotelioma en el tórax derecho. Desarrolla disnea súbita CF IV por lo que se interna. Se diagnostica derrame pleural masivo y se decide la evacuación y sellado pleural. Tres días después lo ve en el domicilio. Le refiere que se le hincharon ambos miembros inferiores. Explique los posibles diagnósticos diferenciales y por qué.
Formato pobre en contexto con *ítems* de elección múltiple	**Formato rico en contexto con *ítems* de elección múltiple**
Marque cuáles de las siguientes son medidas de prevención primaria: a. Anticoagulación en un paciente con FA. b. Uso del cinturón de seguridad. c. Rehabilitación cardiovascular. d. Suero hiperinmune antitetánico.	Los padres traen a control a Julián, de 3 años de edad. En el examen usted observa que el testículo izquierdo del niño no se encuentra en la bolsa escrotal, sino en el conducto inguinal, pero que puede llevarlo hasta el escroto, donde permanece. Usted le informa a los padres que se trata de un testículo en ascensor y le indica: a. Cirugía. b. Testosterona. c. Gonadotrofina coriónica. d. Observación y control.

amplios contenidos muy efectivamente, ya que en poco tiempo de examen es posible alcanzar altos índices de confiabilidad, que son impensables con otros métodos. La experiencia y la opinión de los expertos indican que los alumnos podrían responder una pregunta por minuto, por lo que es posible exponerlos a 180 preguntas en tres horas. En estos niveles, la confiabilidad puede ser tan alta como, o superior a 0.90. Esto se explica por el hecho de que al existir tantas preguntas que evalúan diferentes dominios del conocimiento, se controla la especificidad de caso[2]. Una prueba con 40 o 50 preguntas o menos tiene baja confiabilidad para el uso en decisiones importantes [12]. La confiabilidad inter-observador es 1.0, ya que los criterios son previos a la corrección, que es automática.

Validez: existen varios cuestionamientos a la validez del EOM. Entre ellos se destacan los siguientes:

1. Evalúa el reconocimiento de la respuesta correcta, no la memoria: existe evidencia de que esto parece no ser así. Estos estudios muestran una alta correlación entre pruebas que miden la memoria con los ítems de EOM. Estas observaciones se basan en el hecho de que tanto el reconocimiento como la memoria acceden a los mismos sistemas cognitivos[12].
2. Sólo mide conocimiento teórico, no resolución de problemas: cuando los ítems están construidos con formato de estímulo pobre en contexto, sólo miden conocimiento teórico, al igual que cualquier otra prueba. Para evaluar habilidades como resolución de problemas, lo importante es que la raíz de la pregunta (enunciado) exponga un caso contextualizado sobre el que hay que tomar decisiones particulares[12].
3. Sólo mide la capacidad de reconocer la respuesta correcta; no tiene nada que ver con medir competencia o desempeño: esto no parece ser tan así. Existen estudios que indican una correlación aceptable (0,60 - 0,70) entre EOMs y el desempeño en la práctica[12].

Impacto educacional: sin duda, este es el aspecto donde el EOM presenta mayores debilidades. Si es utilizado como evaluación sumativa, su impacto en el estilo de aprendizaje de los alumnos es negativo, ya que orienta el estudio hacia la memorización. Sin embargo, se lo ha utilizado con ese fin como evaluación de progreso (*progress testing*) en instituciones con *curricula* basadas en problemas, como Maastricht o McMaster. En esta forma de evaluar, todos los alumnos de la carrera son evaluados con EOM orientado al que egresa de la carrera. Los alumnos de cada año son percentilados y los que están por fuera del percentilo 95 identificados para ser aconsejados sobre sus estudios. Así, se consigue que los alumnos no estudien exclusivamente para el examen, sino que mantengan un ritmo continuo a lo largo del año[12,14].

PROBLEMAS BASADOS EN "PUNTOS CLAVE" (*KEY FEATURES*)

En nuestro país, también se conocen con el nombre de "simuladores". Consiste en una descripción corta de un escenario o viñeta (caso) en el que se presenta un problema. Por cada viñeta se pueden realizar varias preguntas que están orientadas a evaluar las decisiones importantes solamente[11,12]. Un estudio reciente recomienda hasta tres o cuatro preguntas por caso para asegurar una adecuada confiabilidad[6]. El formato de las preguntas puede variar desde EOM a respuestas abiertas cortas (ver Tabla 3). Otro formato posible es el de selección de un lista larga de opciones (*extended-matching questions*). En este formato, las preguntas ofrecen una lista de opciones de la que el examinado debe elegir las respuestas apropiadas. Es muy flexible y permite construir un puntaje de manera sencilla. También es posible penalizar las decisiones que puedan poner en riesgo la vida del paciente o que sean altamente inadecuadas.

Puede aplicarse tanto en forma escrita como electrónica.

Confiabilidad y validez: de acuerdo a varios estudios, el índice de confiabilidad varía entre 0,6 y 0,80 y se ha demostrado su validez para medir las habilidades de resolución de problemas[11,12].

Impacto educacional: son aceptadas tanto por estudiantes como por docentes, como un buen escenario de "simulación" con lápiz y papel. La construcción de un caso de *puntos clave* puede ser trabajosa e insumir varias horas, ya que son necesarios varias preguntas para producir exámenes con confiabilidad aceptable[11,12].

NIVEL DE "DEMUESTRA CÓMO"

EXAMEN ORAL (EL CASO LARGO)

Incluye una variedad de técnicas que estimulan al alumno a demostrar el razonamiento usado en la práctica profesional, en general como respuesta a las preguntas del docente [14]. Los exámenes orales tienen una larga tradición en medicina. El más tradicional es el llamado caso largo o extenso. Este método ya fue mencionado y descrito por Flexner en su famoso informe como el que mejor evalúa las competencias clínicas de los estudiantes de grado. Prácticamente, desde entonces su estructura no ha variado y es utilizado universalmente, aunque en los últimos años ha sido desplazado por el ECOE [3] como evaluación sumativa final de la carrera de medicina.

El caso largo consiste en que el examinado debe entrevistar y examinar a un paciente, en la mayoría de los casos internado, en general sin ser observado, durante un tiempo que oscila entre 30 y 45 minutos. Luego el examinador le pide que le reporte sus hallazgos y se establece una serie de preguntas basadas en hipotéticos casos cortos u otros contenidos no relacionados con el caso, a criterio del examinador.

Confiabilidad: presenta dos problemas que casi los han inhabilitado para la evaluación sumativa: falta de confiabilidad inter-observador y, sobre todo, falta de confiabilidad inter-casos. Es claro que, debido a que el acuerdo entre dos observadores es diverso, la confiabilidad inter-ob-

servador es baja. Esto se debe sobre todo a lo que se describe como estilos de calificar "duros" o "blandos" (*dove/hawk*, en inglés). El principal problema es que el caso largo evalúa en profundidad un solo caso y amenaza seriamente la confiabilidad inter-casos. Como ya fue descripto, la competencia en un caso no la predice en otros, aún cuando sean similares[6,14]. Por lo tanto, es necesario aumentar el número para mejorar la confiabilidad inter-casos. Un estudio demuestra que para mejorar la confiabilidad inter-casos es necesario que el estudiante sea evaluado con 10 casos largos para alcanzar un índice alfa de Cronbach de 0,8. En este caso, se asumió que diferentes pares de observadores evaluarían a los estudiantes, por lo que también se obtuvo una amplia muestra de los juicios de los observadores. Teóricamente y de acuerdo a las conclusiones de ese trabajo, es posible alcanzar una adecuada confiabilidad inter-casos e inter-observadores cuando el estudiante es expuesto a por lo menos 10 casos, siempre y cuando se cuente con la logística y el número suficiente de pacientes y examinadores[14].

Dadas las restricciones logísticas imperantes, las facultades de medicina tradicionalmente evalúan a sus estudiantes con un solo caso.

Validez e impacto educacional: la aproximación al mundo "real", por lo menos del paciente internado, y su amplia aceptación entre los médicos, le otorgan una alta validez de primera impresión (*face validity*), su principal fortaleza. Aproxima al examinado a las tareas reales con los pacientes. Sin embargo, la validez de constructo no está suficientemente estudiada así como tampoco su impacto educacional. Es posible que la evaluación con pacientes reales a través de los casos largos tenga diferentes consecuencias sobre el aprendizaje y el estudio, comparado con el ECOE.

EL EXAMEN CLÍNICO OBJETIVO Y ESTRUCTURADO (ECOE)

El ECOE fue introducido hace treinta años como una aproximación confiable para la evaluación de las habilidades clínicas [3]. Es una prueba con formato flexible, basado en un circuito de pacientes en las llamadas "estaciones". En cada estación, los examinados interactúan con un paciente simulado o estandarizado, para demostrar habilidades específicas. Los pacientes simulados o estandarizados, son personas entrenadas para representar problemas de los pacientes de una manera real. Este tipo de pacientes son valiosos sobre todo para evaluar las habilidades para entrevistar. Las estaciones pueden ser cortas (5 minutos) o largas (15 minutos), simples (evalúa un solo problema por vez) o dobles (la segunda evalúa otros conocimientos luego de haber entrevistado un paciente en la primera, por ejemplo).

En cada estación, un observador pone un puntaje de acuerdo a una lista de cotejo o escala global previamente diseñada y validada. Los observadores son entrenados en el

Tabla 3. Ejemplo de un problema basado en "puntos clave". TA, tensión arterial; BMI, *body mass index* (índice de masa corporal).

Marta tiene 57 años. Hace 8 años que no menstrúa. Es la primera vez que lo consulta. Viene a hacerse un control. Su marido falleció en un accidente hace 3 años. Vive sola. Ahora hace tres meses que está de novia. Esto la ha motivado para intentar bajar de peso. Hace dos meses ha iniciado una dieta y camina ida y vuelta al colegio en donde da clases (queda a 20 cuadras). Ya bajó 3 kg. Su padre murió de un infarto a los 72 años. Su madre aún vive y es sana. Sus hermanos y sus dos hijos también son sanos. Fumó 10 cigarrillos por día desde los 20 hasta los 40 años.
El registro de enfermería le informa: TA 160/90; BMI 27 kg/m^2.
Trae laboratorios de hace 6 meses, que muestran: glucemia de 200 mg/dL, triglicéridos 150, HDL 40 mg/dL, colesterol total de 240.
Luego del interrogatorio usted le vuelve a controlar la TA y encuentra: 140/85. El resto del examen físico es **normal**.

En relación a este caso seleccione los problemas que presenta la paciente:
(puede seleccionar varias opciones, pero **atención**, las elecciones erróneas pueden disminuir el puntaje).

Elevación de la glucemia en ayunas	❍
Alto riesgo cardiovascular	❍
Moderado riesgo cardiovascular	❍
Bajo riesgo cardiovascular	❍
Hipertensión arterial	❍
Registro de tensión arterial elevado	❍
Dislipemia	❍
Diabetes	❍
BMI elevado	❍
Sedentarismo	❍
Ex tabaquista de 10 años/pack	❍
Ex tabaquista de 20 años/pack	❍

uso de esas escalas, quienes pueden ser profesionales o los mismos pacientes simulados entrenados.

Confiabilidad: la confiabilidad inter-observador es alta y varía entre 0,62 y 0,99 en diferentes publicaciones[12,14]. Este nivel se explica sobre todo por el uso de listas de cotejo y escalas globales y el entrenamiento de los observadores. La confiabilidad intercasos es baja, como en cualquier prueba, si el número de casos es bajo (se mide en número de horas ya que la duración de las estaciones puede variar). Sin embargo, se demostró que mejora cuando se aumenta el tiempo de duración de ECOE. Para un ECOE de 4 horas se reconoce una confiabilidad intercasos de 0,8[12,14].

Validez: sin duda, tiene una alta validez de primera impresión (*face validity*) dada la seducción que la "realidad" simulada ofrece.

La validez de contenido para la evaluación de competencias clínicas ha sido demostrada en varios estudios. El uso de estaciones cortas permite la evaluación de muchos aspectos por hora de examen, pero limita la inclusión de casos complicados. Esto puede atentar contra la validez de contenido.

La validez de constructo también ha sido demostrada a través de estudios que muestran puntajes más altos en estudiantes de medicina de primero a cuarto años. Sin embargo, hay algunos estudios que muestran que la correlación entre los puntajes del ECOE y pruebas escritas de conocimiento es alta (0,72), sugiriendo que, si bien no miden los mismos dominios, el ECOE aportaría poca discriminación cuando se aplican las dos pruebas secuencialmente. La evidencia que soporta la validez de criterio no es tan robusta [12,14].

Impacto educacional: no es sorprendente que la adición de evaluación de habilidades clínicas en forma sistematizada tenga un impacto positivo en el estilo de aprendizaje de los alumnos y en el diseño del currículo.

Costo: sin duda, una de las mayores limitaciones, debido al costo directo así como el invertido en la implementación. Depende del número de estaciones y del tipo de encuentros que se planifiquen: número de pacientes simulados, tipo de observadores, etc.

Fortalezas[12]:

1. Una amplia gama de habilidades para un relativo amplio número de alumnos puede ser evaluada en relativamente poco tiempo.
2. El uso de escalas predeterminadas asegura cierta "objetividad".
3. La variabilidad del paciente y el observador es disminuida al máximo, a diferencia de los casos largos.
4. Puede ser usado para fines formativos o sumativos.
5. El formato es flexible: número y duración de las estaciones, circuitos paralelos, rango de competencias a ser evaluadas etc.

Debilidades[12]:

1. A menudo, las estaciones solicitan que los alumnos demuestren habilidades "aisladas" del encuentro clínico.
2. El ECOE se asienta sobre el uso de listas de cotejo que ponen el énfasis en la evaluación exhaustiva y paso por paso de la habilidad, lo que puede atentar contra la evaluación del resultado del desempeño global y su relevancia.
3. Las limitaciones sobre lo que puede ser simulado acota el tipo de pacientes que puede ser presentado en las estaciones.
4. Logística y costo.

NIVEL DE HACE:

LA OBSERVACIÓN DIRECTA (EL Mini-CEX)[1,5]

Intuitivamente, evaluar a los estudiantes observándolos en "acción" es atractivo. En los ambientes clínicos, los docentes evalúan el progreso de los alumnos observándolos con los pacientes, a menudo, resumiendo sus observaciones en una escala global al final de un período de formación, por otra parte de dudoso valor. La evaluación del desempeño de los alumnos con pacientes reales a través de la observación puede ser realizada de varias maneras: el observador puede estar físicamente presente en el consultorio, observar desde un lugar contiguo a través de un espejo o de cámaras de video u observar un video de la entrevista. Esta decisión depende de los objetivos de la evaluación: presencia en caso de maniobras del examen físico o cirugías, cámaras o video-grabaciones para las habilidades de entrevista clínica. El observador debería recolectar información a través de una lista de cotejo o una escala global que le permita dar *feedback* sobre el desempeño con el propósito de mejorarla. Un problema con este tipo de observaciones es que los estándares usados para los casos pueden variar porque en general un solo experto observa el encuentro entre el examinado y el paciente y los expertos rara vez estudian los casos en profundidad.

Confiabilidad: un estudio concluyó que se necesitan al menos diez observaciones estructuradas al año para obtener resultados reproducibles de competencia clínica con un instrumento estructurado[1,5]. Probablemente, el mayor problema en la evaluación de las habilidades clínicas es la falta de observación por parte de los docentes a los residentes.

Por otra parte, existe en general demora entre la observación y la trascripción de las calificaciones obtenidas en el encuentro. La demora introduce error en las calificaciones. Observaron que cuando los formularios no eran estructurados, una frecuente característica de los formularios de evaluación que se utilizan en la práctica clínica, los docentes detectaban sólo el 30% de las debilidades y fortalezas. Las fortalezas fueron omitidas con mayor frecuencia

que las debilidades. La evaluación de estas fortalezas y debilidades aumentaron un 60% con los docentes que utilizaban formularios estructurados[1,5].

Existen variaciones intraobservador vinculadas a cambios de la atención, de perspectiva, de estándares, de humor o de estado de ánimo. Existen variaciones interobservadores. Diferencias de criterios, de puntos de vista, de rigor son algunas fuentes de este problema.

En conclusión, en base a lo antes expuesto se proponen las siguientes recomendaciones[1,5]:

- Los estudiantes deben ser observados en un amplio espectro de situaciones clínicas y procedimientos y por múltiples evaluadores. La bibliografía en general sugiere al menos entre 7 a 11 observaciones para obtener conclusiones razonables de la competencia clínica global del estudiante o residente[14-16].
- Utilizar formularios cortos y estructurados como el Mini CEX (ver Alves de Lima A. Claves para la evaluación efectiva del residente. Rev Hosp. Ital. B.Aires, 2005; 25(3/4):107-111).
- Definir claramente las consignas.
- Dar tiempo para la evaluación.
- Maximizar el valor de devolución (*feed-back*) como herramienta formativa.
- Solicitar la trascripción inmediata de las calificaciones luego del examen.
- Complementar observaciones formales con informales.
- Considerar trabajar en grupo para tomar decisiones de promoción.
- Entrenar y calibrar a los evaluadores.
- Chequear los instrumentos de evaluación.

CONCLUSIONES

Este artículo pretende resumir algunos de los aspectos más importantes de los métodos de evaluación más frecuentemente utilizados en nuestro medio y dar algunas sugerencias sobre su uso.

En primer lugar, no hay un solo tipo de método de evaluación de las competencias que sea intrínsecamente superior. Esta afirmación tal vez vaya en contra de mucha literatura sobre el tema y sobre lo que a menudo se piensa en los ambientes docentes.

Un segundo punto es que, en el caso de los métodos escritos, los formatos de respuesta tienen menos influencia sobre lo que está siendo medido que lo que estamos inclinados a pensar. En ese sentido, deberíamos poner el foco en el formato del estímulo más que en el formato de respuesta. El tipo de preguntas debería ser seleccionado de acuerdo a sus fortalezas y debilidades. Un buen manual para escribir preguntas puede encontrarse en el sitio de la National Board of Medical Examiners:
http://www.nbme.org/about/itemwriting.asp#spanish.

A modo de recomendación final, para mejorar la calidad de nuestras evaluaciones, es necesario aumentar el número de casos en los exámenes para asegurar una adecuada confiabilidad intercasos.

Otra conclusión es que un único método no puede valorar a toda la pirámide de Miller; se necesita una combinación de diferentes métodos.

Además, no se debe olvidar el papel fundamental de la evaluación como orientadora del aprendizaje, por lo que su diseño debe ser estratégico en función de los efectos de aprendizaje deseados.

Agradecimientos

A la Dra. Alejandrina Lo Sasso por su apoyo y haber ofrecido generosamente los casos de problemas de puntos clave como ejemplos.

BIBLIOGRAFÍA

1. Alves de Lima A. Claves para la evaluación efectiva del residente. Rev Hosp. Ital. B.Aires 2005; 25(3/4):107-11.
2. Durante E. La evaluación de los conocimientos: lo que parece ser, ¿es realmente lo que es? Rev Hosp. Ital. B.Aires 2005;25(1):18-23.
3. Harden RM, Gleeson FA. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979;13(1):41-54.
4. Miller GE. The assessment of clinical skills/competence/performance. Acad Med 1990;65(9 Suppl):S63-7.
5. Norcini JJ, Blank LL, Arnold GK, Kimball HR. The mini-CEX (clinical evaluation exercise): a preliminary investigation. Ann Intern Med 1995;123(10):795-9.
6. Norman G, Bordage G, Page G, Keane D. How specific is case specificity? Med Educ 2006;40(7):618-23.
7. Regehr G, Norman GR. Issues in cognitive psychology: implications for professional education. Acad Med 1996;71(9):988-1001.
8. Schuwirth LW, van der Vleuten CP. The use of clinical simulations in assessment. Med Educ 2003;37 Suppl 1:65-71.
9. Schuwirth LW, van der Vleuten CP. Changing education, changing assessment, changing research? Med Educ 2004;38(8):805-12.
10. Schuwirth LW, van der Vleuten CP. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004;38(9):974-9.
11. Schuwirth LW, et al. How to write short cases for assessing problem-solving skills. Med Teacher 1999;21(2):144-50.
12. Shannon S, Norman G. Evaluation methods: a resource handbook. 3rd ed. Hamilton, Ont: McMaster University. The Program for Educational Development, 1995.
13. Wass V, van der Vleuten C. The long case. Med Educ 2004;38(11):1176-80.
14. Van der Vleuten CP. The assessment of professional competence: development, research and practical implications. Advances in Health Sciences Education 1996;1:41-67.