

Un motivo de controversia para la certificación en las especialidades quirúrgicas es la inclusión de procedimientos operatorios en el examen práctico, que permitirían evaluar la destreza del sustentante. Solamente cinco consejos lo aplican, haciéndose en pacientes. A quienes sostienen su utilidad para avalar este aspecto se oponen aquellos que argumentan que las situaciones propias del examen, el estrés ante los sinodales, la intervención quirúrgica en un medio desconocido, etc., no solo no evalúan la destreza sino que, además, pone en riesgo al sujeto de la cirugía. En el futuro cuando sean más accesibles los simuladores electrónicos podrán ser una buena herramienta para esos menesteres.

## LA CALIDAD DEL PROCESO DE EVALUACIÓN PARA LA CERTIFICACIÓN DEL MÉDICO ESPECIALISTA

Dr. Melchor Sánchez Mendiola

"Colectar datos para evaluación es como recoger la basura.  
Más te vale saber lo que vas a hacer con ella antes de que la recojas".  
Mark Twain

"El poder de examinar es el poder de destruir"  
Abraham Flexner

### I. INTRODUCCIÓN

El profesional de la medicina se desenvuelve en un entorno social y económico extraordinariamente complejo en la actualidad, y son pocas las oportunidades durante la ajetreada vida cotidiana del médico para reflexionar sobre uno de los aspectos más importantes del proceso vitalicio de enseñanza-aprendizaje en el que estamos inmersos desde el ingreso a la escuela de medicina hasta el final de nuestra vida profesional: **la evaluación del proceso educativo (1)**. A pesar de que el médico es sujeto de evaluación desde que solicita su ingreso a la escuela de medicina, y que realiza exámenes de todo tipo durante su entrenamiento, es excepcional que reciba un mínimo adiestramiento formal en los diferentes aspectos de la educación médica, principalmente en los aspectos técnicos de los procesos de evaluación. Lo anterior se debe a diversos factores: la natural tendencia por privilegiar el conocimiento técnico de la medicina en la educación del médico, el cual es cada vez más amplio; el hecho de que la educación es una ciencia social y como tal, tiene un gran bagaje de aspectos teóricos y conceptuales que son ajenos al entrenamiento médico tradicional que enfatiza las llamadas ciencias "duras"; y a la habitual aceptación del médico de los diferentes procesos de evaluación a que es sujeto, en que supone que quienes diseñan, implementan y analizan los exámenes saben lo que hacen y que no es necesario profundizar en esos temas para aceptar el resultado de la evaluación.

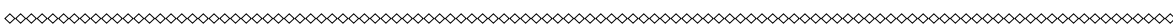
El objetivo de este documento es proveer un panorama de algunos aspectos relevantes del proceso de evaluación en educación médica, en el contexto de los exámenes de altas consecuencias, para que los profesionales de la salud adquieran un mejor entendimiento de los conceptos vigentes en esta área.

## II. DEFINICIÓN DE EVALUACIÓN Y EXÁMENES DE ALTAS CONSECUENCIAS

La palabra "*evaluación*" como muchos vocablos que utilizamos para comunicarnos, tiene un significado coloquial y uno técnico, en este documento nos centraremos en la faceta técnica del término. En uno de los textos más reconocidos sobre medición y evaluación en educación, se define a la evaluación como "*un término genérico que incluye un rango de procedimientos para adquirir información sobre el aprendizaje del estudiante, y la formación de juicios de valor respecto al proceso de aprendizaje...*" (2). Lo anterior implica un proceso sistemático de acopio de información a través de la aplicación de diversos instrumentos (como son los exámenes escritos u orales), para ser analizada objetivamente y así poder fundamentar la toma de decisiones sobre el proceso educativo. A un *examen o prueba* se le define como el instrumento o procedimiento sistemático para medir una muestra de conducta, planteando un conjunto de preguntas o interrogantes de una manera uniforme. Es importante recalcar que no es lo mismo evaluación que medición, *medición* es "el proceso de obtener una descripción numérica del grado al cual un individuo posee una característica particular" (2), es decir el medir es sólo la obtención de datos sin realizar juicios de valor (como medir la temperatura corporal o el cociente intelectual), mientras que la evaluación es un proceso más amplio y valorativo del fenómeno (p.ej. el paciente tiene 39 grados de temperatura, entonces tiene fiebre probablemente secundaria a un padecimiento infeccioso). La medición es parte de la evaluación, pero no lo es todo.

Existen algunos *principios generales de la evaluación en educación*, que siempre debemos tomar en cuenta durante la misma (2):

- 1) Especificar claramente lo que se va a evaluar debe ser una prioridad.
- 2) Los métodos de evaluación deben elegirse por su relevancia para las características que se van a evaluar.
- 3) Se requiere de una variedad de procedimientos para que sea útil y efectiva.
- 4) Su uso adecuado requiere tener conciencia de las limitaciones de cada método en particular.
- 5) La evaluación es un medio para un fin, no un fin en sí mismo.



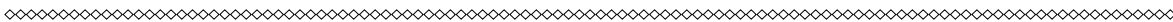
Un tipo de exámenes en particular merecen atención especial, los llamados “**exámenes de altas consecuencias**” (“high-stakes tests” en inglés). Se les denomina así a aquellas pruebas o exámenes que tienen consecuencias serias e importantes para los individuos que los toman, por ejemplo los exámenes de graduación, selección, promoción y certificación. Estos exámenes (como el Examen Nacional de Aspirantes a Residencias Médicas en nuestro país) tiene una gran cantidad de efectos positivos y negativos sobre lo que aprenden los estudiantes y cómo lo aprenden, por lo que las instancias evaluadoras deben hacer lo posible porque estos procedimientos de evaluación se realicen en un marco conceptual técnico apropiado, con profesionalismo educativo (3).

### III. TIPOS DE EVALUACIÓN

Desde el punto de vista de su objetivo, la evaluación se puede clasificar en **diagnóstica, sumativa y formativa**. La **evaluación diagnóstica** se realiza al principio de un curso o actividad académica con la finalidad de determinar el nivel de conocimiento, habilidad o actitud del educando. Esta información es de gran utilidad para el docente ya que le permite hacer adecuaciones en el contenido y en la implementación de las actividades académicas programadas para corresponder a las características del alumno que participará en la actividad educativa.

**La evaluación sumativa** es aquella compuesta por la suma de valoraciones efectuadas durante un curso o unidad didáctica, para determinar el grado con que los objetivos de la instrucción fueron alcanzados, otorgar calificaciones, o certificar competencia clínica (2). Ejemplos de este tipo de evaluación son los exámenes de fin de curso, los exámenes de certificación de los Consejos de especialidad, el examen profesional de la carrera de medicina; eventos de alta trascendencia para la vida del educando, quien suele percibirlos como obstáculos a sortear para alcanzar un objetivo en lugar de oportunidades para identificar su estado real de aprendizaje. Estos exámenes representan la totalidad del concepto de evaluación en educación médica para muchos estudiantes. La evaluación sumativa idealmente provee evidencia a la sociedad de que el individuo ha aprendido lo que tenía que aprender para graduarse como médico o certificarse como especialista. Desafortunadamente los instrumentos que utilizamos para evaluar en educación médica con frecuencia adolecen de defectos que limitan las inferencias que podemos hacer de los resultados, exponiendo a la sociedad a médicos que pudieran no ser competentes en todas las áreas necesarias para una atención de salud de calidad.

**La evaluación formativa** es la que se utiliza para monitorear el progreso del aprendizaje, y proporcionar realimentación al estudiante sobre sus logros, deficiencias y oportunidades de mejora (2,4). Esta evaluación debiera ocurrir a lo largo del proceso educativo del médico, y puede ser **formal o informal**, ambas de importancia para la formación del médico general y especialista. La evaluación formativa tiene un fuerte componente educativo, ya que durante las actividades cotidianas permite identificar aquellas que se hacen bien para continuar haciéndolas así, y aquellas que tienen alguna deficiencia, para detectarlas a tiempo y corregirlas.



La evaluación también puede clasificarse de acuerdo a la interpretación de los resultados, en *con referencia a norma* o *con referencia a criterio* (2,4). Cuando la evaluación se interpreta con *referencia a norma*, el resultado se describe en términos del desempeño del grupo y de la posición relativa de cada uno de los estudiantes evaluados. Este tipo de evaluación se utiliza para colocar a los alumnos en escalas de rendimiento y puntaje, y asignarles un lugar dentro del grupo. Un ejemplo en nuestro medio es el Examen Nacional de Aspirantes a Residencias Médicas (ENARM), en el que la puntuación obtenida por el médico se evalúa en relación

al desempeño del grupo y de su lugar secuencial en la lista para aspirar a una de las plazas, y no en un criterio de nivel de conocimientos previamente establecido.

La evaluación con *referencia a criterio* describe el resultado específico que se encontró, de acuerdo a criterios o metas preestablecidos. Este tipo de evaluación busca la comparación del estudiante en relación a un estándar fijado de antemano. Un ejemplo es el examen de certificación realizado por los Consejos de las especialidades médicas, en que se debe acreditar la correcta solución de diferentes problemas clínicos que idealmente deben tener un estándar de pase prefijado de acuerdo a las competencias establecidas para esa especialidad. La tendencia actual hacia una educación basada en competencias requiere que los procedimientos de evaluación sean con referencia a criterio, ya que uno de los inconvenientes de la evaluación con referencia a norma es que parte del grupo evaluado puede carecer de las competencias mínimas requeridas para ser un buen médico, y terminar siendo aprobados porque se aplicó una curva de distribución normal y se reprobaban solo a los que están dos desviaciones estándar debajo de la media.

Existen gran variedad de instrumentos que tienen diversas ventajas y limitaciones para documentar el aprendizaje de los conocimientos, habilidades y destrezas médicas (1,4). Es responsabilidad del profesor y de la organización evaluadora elegir los métodos apropiados para el proceso de evaluación, que pueden clasificarse en cinco categorías:

- **Evaluaciones escritas:** ensayos, preguntas directas de respuesta corta, exámenes de opción múltiple, relación de columnas, disertaciones, reportes.
- **Evaluaciones clínicas/prácticas:** exámenes prácticos con casos clínicos, examen clínico objetivo estructurado (ECO).E).
- **Observación:** reporte del profesor, listas de cotejo, reporte de pacientes.
- **Portafolios y otros registros del desempeño:** libretas de registro, portafolios, registros de procedimientos.
- **Autoevaluación y evaluación por pares:** reporte del educando, reporte de los compañeros.

El recurso más utilizado en los Estados Unidos es la "caja de herramientas de evaluación" desarrollada por el Accreditation Council for Graduate Medical Education (ACGME) de los E.U.A. y disponible de manera gratuita en la página de Internet de la citada organización (<http://www.acgme.org/Outcome/>), en donde se describen diversos métodos para evaluar a los médicos, sus ventajas y desventajas, características psicométricas, así como sugerencias para su implementación.

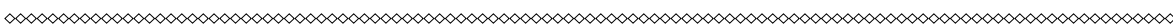
#### IV. VALIDEZ

Uno de los conceptos más importantes para que los resultados de los procesos de evaluación tengan un sustento sólido y un uso apropiado, es el de *validez*. Tradicionalmente se dice que la *validez* de un proceso de evaluación es el grado con el que mide lo que se supone que mide (p.ej. es menos válido medir la presión arterial con un manómetro de mercurio que hacerlo con una medición directa con un catéter intraarterial conectado a un transductor). El concepto de validez en educación ha evolucionado en las últimas décadas, y actualmente se considera que toda la validez es *validez de constructo* y que requiere múltiples fuentes de evidencia para su interpretación, ya intenta responder a la pregunta "¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen?" (5). No es el instrumento el que es válido, ya que la validez de un examen es específica para un propósito y se refiere más bien a lo apropiado de la interpretación de los resultados. En otras palabras, la validez no es una propiedad intrínseca del examen, sino del significado de los resultados en el entorno educativo específico y las inferencias que pueden hacerse de los mismos. Por ejemplo los resultados de los médicos que sustentan el ENARM no deben interpretarse como evidencia de la calidad de las escuelas de medicina de donde provienen, ya que el examen no está diseñado con ese propósito. Si se desea realizar este tipo de inferencia (determinar la calidad de la escuela de medicina en que estudió el aspirante), debe acumularse evidencia de diversas fuentes para sustentar esta interpretación (es decir, para que sea válido decir que los resultados del ENARM discriminan la actuación de las diferentes escuelas), proceso que hasta la fecha no se ha realizado de manera explícita.

Tradicionalmente la validez en educación se clasificaba como "las 3 Cs": validez de contenido, de criterio y de constructo. Con la nueva definición de validez esta distinción desaparece, ya que todas las pruebas de validez responden a la misma pregunta: "¿qué me dice de la persona su puntuación en el examen?". Las diferentes fuentes de evidencia pueden arrojar luz sobre distintos aspectos de la validez, pero no reflejan diferentes tipos de la misma: *la validez es un concepto unitario, y actualmente se considera que toda la validez es validez de constructo* (5). La palabra constructo significa colecciones de conceptos abstractos y principios, inferidos de la conducta y explicados por una teoría educativa o psicológica, es decir, atributos o características que no pueden observarse directamente (por ejemplo, inteligencia, timidez, conocimientos sobre neuroanatomía).

Las cinco fuentes importantes de validez de constructo en evaluación educativa son (3,5):

- **Contenido:** En los exámenes escritos la documentación de evidencia de validez de contenido es fundamental. Los siguientes son algunas fuentes de validez de contenido: la tabla de especificaciones de la prueba y el proceso seguido para elaborarla, el contenido temático definido, la congruencia del contenido de las preguntas con las especificaciones del examen, la representatividad de las preguntas de los diferentes dominios del área a examinar, la calidad de las preguntas, las credenciales de las personas que elaboran los reactivos.
- **Procesos de respuesta:** Se definen como evidencia de integridad de los datos de tal manera que las fuentes de error que se pueden asociar con la administración del examen han sido controladas en la medida de lo posible. Por ejemplo: el control de calidad de la elaboración del examen, la validación de la clave de la hoja de respuestas utilizada, el control de calidad del reporte de los resultados del examen, la familiaridad del estudiante con el formato de evaluación.
- **Estructura interna:** Esta fuente de evidencia de validez se refiere a las características estadísticas y psicométricas de las preguntas del examen, como son: el análisis de reactivos con el grado de dificultad e índices de discriminación de las preguntas, el desempeño de cada distractor en las preguntas de opción múltiple, la confiabilidad del examen, el error estándar de medición, el modelo psicométrico utilizado para asignar la puntuación del examen entre otros. Muchos de estos datos debieran obtenerse de rutina como parte del proceso de control de calidad del examen, principalmente en los exámenes de altas consecuencias, en virtud de la relevancia de sus resultados para los educandos.
- **Relación con otras variables:** La relación de los resultados en el examen con otras variables es intuitivamente atractiva, y se refiere a la correlación estadística entre los resultados obtenidos por medio de un instrumento con una medición de características conocidas. Este rubro busca evidencia confirmatoria y contradictoria, representando claramente el concepto actual de validez como la demostración de una hipótesis. Puede investigarse la correlación positiva de los resultados con exámenes similares que midan el mismo constructo (evidencia convergente), y la falta de correlación con pruebas que midan otros atributos (evidencia divergente). Si se documenta correlación entre las calificaciones obtenidas durante la residencia de especialidad con las obtenidas en el examen de certificación del Consejo respectivo, se consideraría evidencia de validez para interpretar el resultado del examen del Consejo como documentación de las competencias adquiridas durante la residencia.
- **Consecuencias:** Se refiere al impacto en los educandos de las puntuaciones de la evaluación, de las decisiones que se toman como resultado de los resultados en el examen, y su efecto en la enseñanza y el aprendizaje. Las consecuencias pueden ser positivas o negativas, intencionales o no intencionales, las cuales son muy importantes en los exámenes de altas consecuencias. Son ejemplos de este aspecto de la validez: el método de establecimiento del punto de corte para aprobar o reprobar un examen, las consecuencias para el estudiante y la sociedad, las consecuencias para los profesores y las instituciones educativas.



La validez de constructo involucra una aproximación científica a la interpretación de los resultados de los exámenes, es decir, probar hipótesis sobre los conceptos evaluados en el examen. La información proporcionada por un instrumento de evaluación no es válida o inválida, sino que los resultados del examen tienen más o menos evidencia de las diferentes fuentes para apoyar (o refutar) una interpretación específica (por ejemplo, el pasar o reprobar un curso, el certificar o no a un especialista, o el admitir o no a un residente a un curso

de especialidad). Bajo estas premisas, la mala noticia es que el probar la validez es una aventura que nunca queda completa, ya que siempre se puede aprender más sobre el significado de los resultados de un examen con diversos grupos y en diferentes circunstancias. Un aspecto muy importante de la obtención de evidencia de validez en los exámenes de altas consecuencias es que las organizaciones que elaboran e implementan el examen (entidades gubernamentales, instituciones educativas, Consejos de certificación) son los candidatos obvios para validar las afirmaciones que hacen sobre la interpretación de los resultados de un examen, ya que generalmente son quienes tienen los elementos y recursos para hacerlo. Quienes elaboran el examen tienen una obligación ética y un imperativo educativo para documentar qué tan defendible es la interpretación de los resultados, en beneficio de los educandos y de la sociedad en general (3).

Existen diversas amenazas para la validez de un proceso de evaluación, es decir elementos que disminuyen la credibilidad de las inferencias que se pueden hacer de los resultados de un examen. De acuerdo a Downing y Haladyna, se pueden clasificar de la siguiente manera (6):

- **Infrarrepresentación del constructo (IC):** Se refiere a una representación inapropiada de los dominios del contenido a evaluar por el examen, por ejemplo: pocos reactivos en el examen que no muestren apropiadamente el área de conocimiento explorada; proporción inadecuada de reactivos que no siga fielmente la tabla de especificaciones, de tal manera que algunas áreas son sobre-exploradas y otras infra-exploradas; uso de muchos ítemes (preguntas o reactivos) que exploren procesos cognoscitivos de bajo nivel, como la memoria o reconocimiento de datos factuales, mientras que los objetivos de la enseñanza son de mayor nivel como la aplicación o solución de problemas.
- **Varianza irrelevante al constructo (VIC):** Se refiere a variables que de manera sistemática (no al azar) interfieren con la capacidad de interpretar los resultados de la evaluación de una manera significativa, y que causan "ruido" en los datos de medición. Ejemplos de VIC son: reactivos elaborados con deficiencias y que tienen "fallas" de acuerdo a las recomendaciones basadas en evidencia educativa, introducen este tipo de variación (7); problemas con la seguridad del examen y fuga de información, de tal manera que el resultado del examen no refleja los conocimientos de los estudiantes; preguntas demasiado difíciles, fáciles o que no discriminan los estudiantes que saben más de los que saben menos; astucia para responder los exámenes ("testwiseness"), los estudiantes que se preparan con estrategias para responder exámenes pueden obtener puntajes que no reflejen lo que saben; el uso de estructuras gramaticales complejas en las preguntas o instrucciones difíciles de entender; el fenómeno de "enseñando a la prueba" ("teaching-to-the-test") en el que se enfatiza la enseñanza de

lo que será incluido en el examen en lugar de los objetivos del proceso educativo, al grado que algunos profesores pueden utilizar reactivos del examen para exagerar las calificaciones de sus alumnos (6).

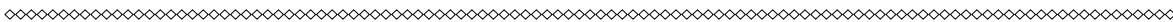
## V. CONFIABILIDAD

La **confiabilidad** ("reliability" en inglés) tiene un significado técnico preciso en evaluación educativa, que no debe confundirse con la percepción coloquial del término. **Confiabilidad** es la capacidad del examen de arrojar un resultado consistente cuando se repite, es decir, es la reproducibilidad del examen. Es un concepto estadístico, que representa el grado en el cual las puntuaciones de los alumnos serían similares si fueran examinados de nuevo, y en el que el instrumento mide el fenómeno de manera consistente en el tiempo (2,8). Si la prueba se repite a lo largo del tiempo, los nuevos resultados deberían ser similares a los iniciales para el mismo instrumento de evaluación y la misma población de estudiantes, suponiendo que no hubiera ocurrido aprendizaje en ese intervalo.

Generalmente se expresa como un coeficiente de correlación, siendo 1.0 una correlación perfecta y cero ninguna correlación. Mientras más alta es la cifra de confiabilidad, generalmente es mayor su peso como evidencia de validez en el rubro de "estructura interna" del examen. Es importante recalcar que la magnitud de la cifra de confiabilidad suficiente para aceptar los resultados de un proceso de evaluación depende del propósito de la misma, el uso que se hará de los resultados del examen y de las consecuencias que tendrá la evaluación sobre los estudiantes. Para exámenes de muy altas consecuencias (p.ej. exámenes profesionales, de certificación, el ENARM), la confiabilidad debe ser alta para que aporte evidencia suficiente de que las inferencias de los resultados del examen son defendibles. Varios expertos en medición educativa recomiendan una confiabilidad de por lo menos 0.90 para evaluaciones de muy altas consecuencias, ya que el resultado de las mismas puede

afectar de manera importante a los examinados y a la sociedad. Para exámenes de consecuencias moderadas, como las evaluaciones sumativas de fin de curso en las escuelas de medicina, es deseable que la confiabilidad sea de 0.80 a 0.89. En exámenes de menores consecuencias, como la evaluación formativa o exámenes parciales diagnósticos, es aceptable una confiabilidad de 0.70 a 0.79 (8). Estas cifras no representan rangos absolutos, ya que hay diferencias de opinión entre los expertos, pero pueden servir de marco de referencia para evaluar nuestros instrumentos de evaluación.

La confiabilidad de una medición es necesaria para obtener resultados válidos, pero puede haber resultados confiables sin validez (es decir la confiabilidad es necesaria, pero no suficiente para la validez), recordemos que la confiabilidad es un elemento de validez de constructo en el terreno de "estructura interna" del examen. La analogía con un blanco de tiro es útil para entender la relación entre los dos conceptos, como se demuestra en la **Figura 1**.



## Validez vs. Confiabilidad

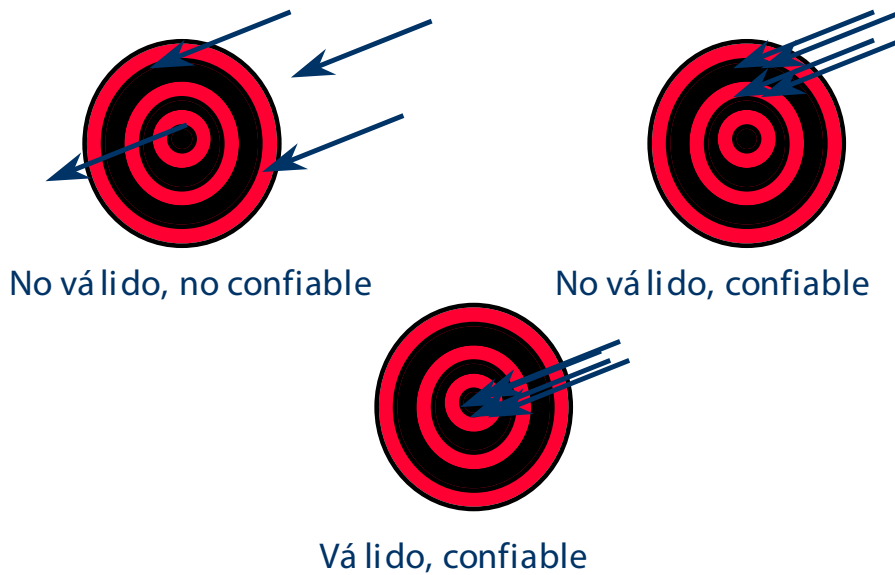


Figura 1. Ilustración de los conceptos de validez y confiabilidad de un instrumento de evaluación, usando la analogía de un blanco de tiro.

El tipo específico de medición de confiabilidad depende del tipo de evaluación y su propósito. El enfoque comúnmente utilizado para medir la confiabilidad en los exámenes escritos (como los exámenes de opción múltiple) utiliza mediciones llamadas de "consistencia interna", como el coeficiente alfa de Cronbach o la fórmula de Kuder-Richardson 20 (KR-20). Las evaluaciones de la competencia y el desempeño clínicos o los exámenes orales necesitan otro tipo de medición de confiabilidad, ya que la principal amenaza a la reproducibilidad de este tipo de evaluaciones es la inconsistencia entre los evaluadores (p.ej. el jurado de un examen profesional oral), por lo que la confiabilidad inter-observadores es la más importante en este contexto. Existen varios métodos para determinar la confiabilidad inter-evaluador con diferentes estrategias estadísticas, como son: el porcentaje de acuerdo, el índice kappa, el coeficiente de correlación intraclase, y el uso del análisis con la "teoría de la generalizabilidad" (8). El uso de estas herramientas dependerá del propósito y consecuencias del examen, así como de los requerimientos técnicos del educador.

Algunas estrategias para incrementar la confiabilidad de los exámenes son las siguientes: incrementar el número de preguntas, utilizar principalmente reactivos de dificultad media, y en general incrementar el nivel de profesionalización desde el punto de vista técnico educativo de las personas que elaboran los exámenes.

---

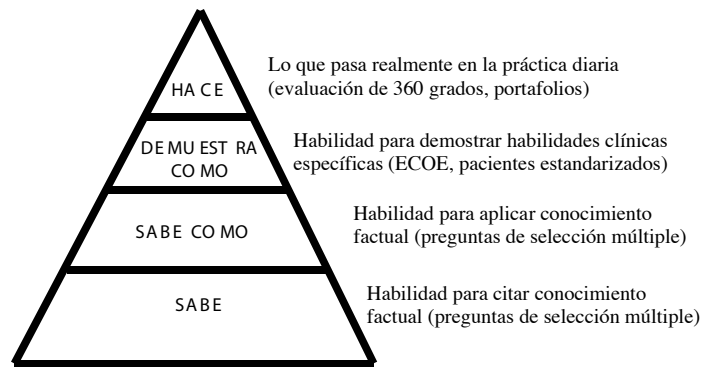
## VI. COMPETENCIA VS. DESEMPEÑO

La importancia de efectuar una evaluación educativa en medicina clínica no puede soslayarse, ya que es la única forma de contar con un referente para establecer si el educando es poseedor de las competencias necesarias que la institución educativa pretende lograr. Lo anterior evaluando no solo conocimientos, sino también formas de comportamiento, valores, afectos y sus formas de expresión, habilidades, destrezas, actitudes y comportamiento ético, lo que indudablemente es complejo por el tipo de variables que se pretende objetivar. En el caso particular de la profesión médica, el consenso de los expertos es que ningún método individual de evaluación puede proveer todos los elementos que se requieren para juzgar algo tan complejo como la atención médica de calidad en la práctica clínica, de tal manera que el reto de la evaluación en la educación médica es de extrema complejidad y exige que lo abordemos con responsabilidad y profesionalismo (9).

Un modelo aceptado en la comunidad de educadores médicos es el de la **pirámide de Miller (Figura 2)**, en el que se muestran de manera escalonada de autenticidad profesional, las características del saber y quehacer del médico, comenzando con la cognición del educando y subiendo hacia la conducta profesional (4,9). En la base, el primer escalón de la pirámide se refiere al conocimiento, el "saber" o recordar, que puede evaluarse por escrito con exámenes de selección múltiple; el segundo escalón se refiere al conocimiento aplicado, el "saber cómo" o integrar, que también se puede evaluar en forma escrita; a partir del tercer escalón las evaluaciones escritas pierden legitimidad, ya que se refiere a la **competencia clínica**, el "*mostrar cómo*" lo hace. Para ello, se requiere un examen práctico clínico en un entorno controlado y estandarizado con pacientes o simuladores, como el examen clínico objetivo estructurado (ECO); el cuarto escalón y punta de la pirámide se refiere al desempeño del médico en la práctica, el "*hacer*" durante el trabajo cotidiano, cuya evaluación requiere de otros métodos como observación directa, portafolios educativos, evaluación por pares, registro de resultados en los pacientes (4,9).

Es importante integrar de manera lógica y planeada los diferentes métodos de evaluación en educación médica, teniendo en cuenta los objetivos a ser evaluados por cada instrumento de acuerdo a su situación en la pirámide de Miller. La educación basada en competencias plantea la documentación de la competencia y el desempeño del educando en las partes altas de la pirámide, para asegurar que el médico está listo para hacer lo que la sociedad y la comunidad médica suponen que debe poder hacer para una práctica clínica efectiva. Los educandos en los programas educativos de residencias médicas tienen un intenso componente de trabajo directo con pacientes bajo la supervisión de los especialistas, por lo que es de particular importancia utilizar métodos de evaluación orientados a los dos últimos escalones de la pirámide (competencia y desempeño), haciendo un esfuerzo para que sean lo más válidos y confiables posible en la medida de las limitaciones de nuestro entorno.

Figura 2. Modelo de la Pirámide de Miller para evaluar las habilidades y competencias del médico

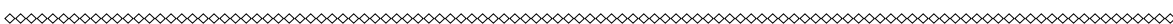


Pirámide de Miller

El reto en los exámenes de altas consecuencias en medicina es el combinar el rigor técnico de los procedimientos de evaluación de acuerdo a las mejores prácticas internacionales, con las realidades y limitaciones técnicas, de tiempo y recursos prevalentes en nuestro medio. Una parte importante de los exámenes profesionales, de certificación y de selección en nuestro país exploran principalmente los dos primeros escalones de la pirámide de Miller, y de manera limitada el tercer escalón (como el uso del ECOE en algunas escuelas de medicina). Pocas instituciones intentan evaluar el vértice de la pirámide, por las características propias del desempeño clínico en la vida real que revela múltiples obstáculos y dificultades para realizarlo de manera precisa. Es importante hacer notar que en varios países la tendencia es a documentar que los médicos tengan un desempeño adecuado y de calidad, más allá de determinado nivel de conocimientos medido con exámenes escritos, como el Foundation Programme en Inglaterra ( [www.mmc.nhs.uk](http://www.mmc.nhs.uk) ) y las estrategias actuales del American Board of Medical Specialties en Estados Unidos ( [www.abms.org](http://www.abms.org) ).

## VII. ESTABLECIMIENTO DE ESTÁNDARES DE PASE

Uno de los pasos más importantes en el proceso de evaluación en educación médica es el establecimiento de "*estándares de pase*" ("passing standards"), que desafortunadamente no se realiza en nuestro medio con la frecuencia que debería, ya que es un elemento indispensable para la credibilidad, validez y confiabilidad de los exámenes de altas consecuencias. Un estándar de pase es una declaración explícita sobre si el desempeño del educando en un examen es lo suficientemente bueno para un propósito en particular, de esta manera estableciendo una calificación específica que sirve como límite para separar a los que aprueban el examen de los que no lo hacen (11). Citando al Dr. John Norcini, uno de los principales expertos mundiales en el tema, "*...es la respuesta numérica a la pregunta: ¿cuánto es suficiente?, ¿qué tan alto es el gigante más chaparro?*".



La definición de estándares de pase tiene un sentido amplio, ya que refleja los valores profesionales en el contexto educativo y social en que se lleva a cabo la evaluación, así como el desempeño de los educandos en el examen. En nuestro medio los docentes y educandos tendemos a reflexionar poco sobre el punto de pase, ya que con frecuencia es establecido de manera tradicional como 6.0, sin que los profesores realicen un proceso definido para determinar cuándo un educando debe ser aprobado o reprobado. Los pasos recomendados como metodología para establecer un estándar de pase para un proceso de evaluación son los siguientes:

- **Paso 1 – Decidir el tipo de estándar.** Existen dos tipos de estándares, los **relativos** y los **absolutos**. Los estándares de pase relativos son basados en una comparación entre el desempeño de los examinados y se expresan como un número o porcentaje de los mismos (p.ej. los que se encuentran arriba del promedio, o los que se encuentran en el 20% superior de las calificaciones). Los estándares absolutos son aquellos que se basan en lo que saben o son capaces de hacer los examinados (p.ej. deben contestar correctamente el 70% de las preguntas, o realizar el 80% de determinadas destrezas). Los absolutos son más apropiados para exámenes de competencia o desempeño, en donde el propósito es establecer que los examinados saben lo suficiente para un propósito específico (como el examen de certificación o el examen profesional), mientras que los estándares relativos se utilizan para seleccionar los más altos o los más bajos para procesos de selección o exclusión en situaciones en donde hay un número limitado de plazas (como el ENARM o los exámenes de admisión a las escuelas de medicina).
- **Paso 2 – Elegir el método para establecer el estándar.** Este paso debe determinar de manera explícita cómo se establece el punto de corte, ser consistente con el propósito del examen y sustentarse en investigación científica y juicio de expertos. Existen varios métodos, como el de Angoff, el de porcentaje fijo, el de grupos contrastantes, el de Hofstee entre otros. Se remite al lector a la literatura especializada para una descripción detallada de cada uno de estos métodos, aunque lo importante es que el que se seleccione debe realizarse con la metodología establecida para tener credibilidad (11).
- **Paso 3 – Selección de los jueces.** En virtud que este proceso es una expresión de juicios de valor, es muy importante que se determine el número y calidad de las personas que funcionarán como jueces para elegir el punto de corte. Mientras de más altas consecuencias sea un examen, más importante la selección y adiestramiento adecuados de los jueces.
- **Paso 4 – Realizar la reunión para establecer el estándar.** Todo el proceso debe estar documentado, con realimentación de los jueces durante el mismo.
- **Paso 5 – Cálculo del estándar.** El procedimiento exacto dependerá del método utilizado, es importante determinar la confiabilidad del proceso.
- **Paso 6 – Después del examen.** Es crucial asegurar que el proceso de establecer el estándar de pase produzca resultados razonables, y que estos sean aceptados por los profesores y

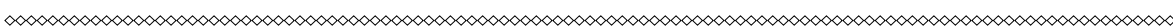
educandos. Si el punto de pase es demasiado alto (lo que ocurre con frecuencia con grupos de jueces que no están calibrados o que no han recibido capacitación sobre el procedimiento) y no aprueban el examen la mayoría de los sustentantes, tal situación no será aceptada por los sustentantes, la sociedad, las instituciones educativas o las instancias certificadoras.

## VIII. LA CERTIFICACIÓN DE MÉDICOS ESPECIALISTAS Y LA CALIDAD DE LA ATENCIÓN MÉDICA

En años recientes el movimiento de la calidad de la atención en salud ha ocupado los reflectores en diversos escenarios de la medicina, con un énfasis importante en el modelo propuesto por Avedis Donabedian que incluye la estructura, los procesos y los resultados de la atención médica como los componentes que determinan la calidad de la atención del sistema y que son susceptibles de ser evaluados (12). Por diversas razones el papel del médico individual (su competencia y desempeño clínicos) no ha sido tomado en cuenta de manera integral en el proceso, e incluso se ha conceptualizado a los médicos como obstáculos en el camino organizacional hacia la mejoría de la calidad. Es importante analizar el papel que juega la certificación del médico en el marco conceptual vigente de la mejoría de calidad, revisando la evidencia existente sobre el status de certificación de un profesional de la salud y su impacto medible en la calidad de la atención.

En un trabajo de investigación reciente, se analizó la literatura existente sobre el papel de la certificación de los médicos por los Consejos de especialistas y su relevancia en la calidad de la atención (13). Los autores identificaron tres cuerpos de evidencia para ligar la certificación de los especialistas con el concepto de calidad en la atención de la salud:

- **La validez del examen de certificación como medida de la calidad del médico.** Los exámenes de certificación, como exámenes de altas consecuencias, deben acreditar de manera explícita y técnicamente apropiada los procesos que intervienen en la elaboración, implementación y análisis del examen (p.ej. tener una tabla de especificaciones definida, establecer un estándar de pase). Por otra parte, es importante documentar una correlación entre la puntuación en el examen de certificación con otras medidas de competencia clínica (p.ej. la calificación durante la residencia de especialidad), con el objeto de proporcionar credibilidad al proceso. Y por último, es crucial documentar una relación entre el status de certificación y recertificación de un especialista y los resultados clínicos en los pacientes, ya que esta sería la prueba de fuego de que la certificación documenta competencia clínica y predice el desempeño en la práctica. En la única revisión sistemática publicada sobre el tema, en pacientes y médicos de Estados Unidos que tenían certificación por las diferentes instancias del American Board of Medical Specialties, se seleccionaron 13 artículos que cubrieron los criterios de inclusión y que analizaban 33 hallazgos identificables (14). Se encontró una correlación positiva entre el status de certificación y buenos resultados clínicos en sólo 16 de los 33 resultados analizados, no se encontró asociación en 14 de ellos, y en 3 la correlación fue negativa entre la certificación y los resultados clínicos en pacientes. Un hallazgo inquietante de esta revisión es que sólo el 5% de los estudios tenía una metodología de investigación apropiada para la pregunta del estudio, lo que revela la necesidad de investigación de calidad metodológica en este campo de la educación médica (14).



- La certificación y la seguridad del paciente. Es primordial adoptar la teoría de prevención de errores utilizada en el enfoque de sistemas del movimiento de la calidad, los programas de entrenamiento de licenciatura y especialidad en medicina deben incluir estos conceptos para que los médicos participen activamente en la solución de este problema que enfrenta la sociedad.
- La percepción social de la certificación. Si bien algunas encuestas revelan que los pacientes creen que es importante que los médicos estén certificados, en los hechos la desinformación es tal que es excepcional que un enfermo le pregunte a su médico si cuenta con certificación vigente en su especialidad. La comunidad médica debe difundir de manera más efectiva el papel de la certificación del médico en la atención de la salud, para que la comunidad tenga plena conciencia de sus virtudes y problemática.

## **IX. RETOS DEL PROCESO DE CERTIFICACIÓN EN MÉXICO**

Los médicos viven una serie de experiencias relacionadas con la evaluación durante su proceso formativo, que los condicionan a aprender teniendo como objetivo principal el pasar el examen. Esta situación puede distorsionar su aprendizaje con consecuencias negativas en los aspectos éticos, de comunicación y emocionales de la relación médico-paciente, ya que la preocupación del educando se centra más en los aspectos puramente técnicos de la profesión. Un ejemplo de esta situación son los exámenes escritos, particularmente los de opción múltiple, que a pesar de sus ventajas psicométricas se encuentran "en el corazón del currículo oculto" (4). Los estudiantes aprenden la información para pasar el examen en lugar de interiorizarla como un todo coherente con el resto del conocimiento que ya poseen, con énfasis en la memorización de datos y no en la aplicación del conocimiento para la solución de problemas.

Cuando se decide utilizar un instrumento de evaluación, deben considerarse los elementos de validez y confiabilidad, para tener credibilidad ante la comunidad científica y la sociedad. En ocasiones se requieren más recursos financieros y humanos para su implementación (p.ej. el ECOE, el uso de pacientes estandarizados), por lo que la factibilidad de su uso en el entorno local y el costo de su adquisición e implementación pueden ser factores determinantes en la decisión de la selección de los métodos a utilizar. Lo anterior no debe evitar que se documente la validez de constructo y confiabilidad del método de evaluación que se esté utilizando, de otra manera es difícil afirmar que estamos midiendo lo que decimos de una manera consistente.

La situación legal de los Consejos de certificación sigue siendo problemática en nuestro país, a pesar de que en general la comunidad médica considera a estas instancias como las más apropiadas para vigilar la competencia clínica de los médicos especialistas, en virtud de su autoridad moral y legitimidad técnica ( [www.conacem.org.mx](http://www.conacem.org.mx) ). La gran responsabilidad social y ante la comunidad médica de los Consejos de Certificación, por la importancia que tiene el documentar que un especialista es competente para ejercer en la práctica, debe ser un motor permanente de capacitación continua y profesionalización de la estructura de los

Consejos, que en última instancia eleve el nivel de la calidad de la atención médica que recibe la población.

