Capítulo 1

Evaluación del aprendizaje

Melchor Sánchez Mendiola

Evaluación es un intento de conocer a la persona.
Derek Rowntree, 1987

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

Introducción

¿Qué es evaluación del aprendizaje? Si le preguntamos a un estudiante probablemente nos dirá: ¡exámenes!, si le preguntamos a un profesor podría contestar: ¡es uno de los aspectos más difíciles de la enseñanza, por el que generalmente no me pagan las horas extra que requiere, y del que he recibido muy poco entrenamiento! Vienen a la mente los comentarios de Sigmund Freud sobre las *profesiones imposible*s, aquellas en las que puedes estar seguro de lograr resultados insatisfactorios: el psicoanálisis, gobernar y la educación. Los docentes habitualmente vivimos en una nube de falsas expectativas y premisas en las que creemos que todo lo que enseñamos es aprendido por los estudiantes. Desafortunadamente esto no es así, por ello la única manera de tener mayor claridad sobre el efecto de la educación y su impacto en los estudiantes es llevar a cabo una evaluación técnicamente adecuada, alineada con los currículos y los métodos de enseñanza, que suministre resultados interpretables y útiles para los diferentes actores del proceso educativo.

Una definición de *evaluación* ampliamente utilizada en educación es: "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 a dicho proceso…" (Miller, 2012). Evaluación implica un proceso sistemático de acopio de

información mediante la aplicación de diversos instrumentos, como pueden ser exámenes escritos u orales, para ser analizada con rigor metodológico, fundamentar la toma de decisiones y promover el aprendizaje complejo en los estudiantes. Los docentes debemos internalizar la evaluación educativa desde una perspectiva amplia, como sugirió en 1977 Derek Rowntree, académico australiano: "cuando una persona, con algún tipo de interacción directa o indirecta con otra, obtiene e interpreta información de manera consciente sobre el conocimiento y la comprensión, habilidades y actitudes de la otra persona. Hasta cierto punto *evaluación es un intento de conocer a esa persona*".

Es fundamental tener en cuenta que existen algunos principios generales de la evaluación en educación (Miller, 2012):

1) Es determinante especificar claramente lo que se va a evaluar.
2) La evaluación es un medio para un fin, no un fin en sí mismo.
3) Los métodos de evaluación deben elegirse con base en su relevancia, tomando en cuenta los atributos que se van a evaluar en el estudiante.
4) Para que la evaluación sea útil y efectiva, se requiere una variedad de procedimientos e instrumentos.
5) Su uso adecuado requiere tener conciencia de las bondades y limitaciones de cada método de evaluación.

Tipos de evaluación

Evaluación diagnóstica, formativa y sumativa

Una de las clasificaciones tradicionales de la evaluación educativa, desde el punto de vista de su objetivo, la divide 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 conocimientos, habilidades o actitudes del educando. Esta información puede ser de gran utilidad para el docente, ya que le permite hacer adecuaciones en el contenido y la implementación de las actividades académicas programadas, que correspondan a las características de los alumnos participantes. Un ejemplo de este tipo de evaluación es el Examen Diagnóstico de Ingreso a las Carreras de Licenciatura de la Universidad Nacional Autónoma de

México (UNAM), en el que se valoran conocimientos generales de español y de inglés en los estudiantes de nuevo ingreso. Los resultados se envían a cada facultad o escuela para su uso y difusión. Recientemente, colocamos estos resultados en el portal de la Coordinación de Desarrollo Educativo e Innovación Curricular como material de acceso abierto para cualquier persona que quiera explorar los datos, incluyendo, además del reporte oficial, unas tablas dinámicas de Excel que permiten al usuario realizar comparaciones de varios tipos y visualizarlas gráficamente. Consultar en https://www.codeic.unam.mx/index.php/resultado-de-diagnostico/

La evaluación *sumativa* es aquella compuesta por la suma de valoraciones efectuadas durante un curso o unidad didáctica, a fin de determinar el grado con que los objetivos de la instrucción se alcanzaron, otorgar calificaciones o certificar competencia. Ejemplos de este tipo de evaluación son los exámenes de fin de curso, los exámenes de certificación de individuos y el examen profesional de la carrera. Estos exámenes son eventos de alta trascendencia para la vida del estudiante, quien en ocasiones los percibe como obstáculos a sortear para alcanzar un objetivo, en lugar de visualizarlos como oportunidades para identificar su estado real de aprendizaje.

La evaluación *formativa* es la que se utiliza para monitorizar el progreso del aprendizaje, con la finalidad de proporcionar realimentación al estudiante sobre sus logros, deficiencias y oportunidades de mejora. Esta evaluación debería ocurrir a lo largo de todo el proceso educativo del estudiante —incluso cuando se ha graduado y se encuentra en la práctica profesional—, y puede ser *formal* o *informal*, *positiva* o *negativa*. La evaluación formativa tiene un poderoso efecto en el aprendizaje, ya que, durante las actividades cotidianas, permite identificar aquellas que se llevan a cabo correctamente para continuar realizándolas así, y aquellas que poseen alguna deficiencia, a fin de detectarlas a tiempo y corregirlas (Martínez Rizo, 2009a, 2013b). Este tipo de evaluación forma parte del concepto de evaluación *para* el aprendizaje, el cual se describe en otro capítulo de este libro.

Es importante enfatizar que en las últimas décadas se ha generado una falsa dicotomía entre la evaluación sumativa y formativa. Por un lado, a la sumativa se le ha etiquetado como excesivamente cuantitativa, centrada en los números, punitiva y discriminatoria, usada con fines políticos, de ejercicio del poder o de control, demasiado estandarizada y poco relevante para el aprendizaje individual. Por el contrario, la evaluación formativa ha surgido como la parte bondadosa, positiva, nutritiva educacionalmente, que toma en cuenta los aspectos afectivos y emocionales

de los estudiantes, además de ayudar a los educandos a salir adelante y a aprender mejor, sin importar sus limitaciones personales y de contexto. Este debate ha creado una situación que recuerda la famosa frase de George Orwell en *Rebelión en la Granja: "Cuatro patas bueno, dos patas malo"*. Si bien hay algo de verdad en esta polarización, debemos visualizar estos dos tipos de evaluación como un continuo, o tal vez como una espiral en la que hay un traslape sustancial, ya que todas las evaluaciones pueden tener un componente sumativo y formativo, que depende de su intencionalidad y el uso de los resultados (Man Sze Lau, 2016).

Un examen de ingreso a la universidad tiene un fuerte componente sumativo, pero también puede usarse como evaluación diagnóstica e incluso formativa si se provee la información a los docentes y estudiantes. Una sesión de realimentación durante el curso puede ser predominantemente formativa, pero si dicha información se utiliza como un componente de la calificación, adquiere una dimensión sumativa. Debemos hacer un esfuerzo para lograr cierto balance entre los extremos del continuo de la evaluación, que promueva un aprendizaje más profundo y significativo.

Evaluación referida a norma y criterio

Otra manera de clasificar a la evaluación es de acuerdo con la interpretación de los resultados. Ello puede ser con referencia a norma (relativa), o con referencia a criterio (absoluta). 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 (Miller, 2012; Sánchez, Delgado, Flores, Leenen y Martínez, 2015). Este tipo de evaluación se utiliza para colocar a los alumnos en escalas de rendimiento y puntaje, con la finalidad de asignarles un lugar dentro del grupo. Un ejemplo en México es el Examen Nacional de Aspirantes a Residencias Médicas (ENARM), evaluación sumativa que presentan los médicos graduados que desean realizar una especialidad médica. La puntuación obtenida por el profesional de la salud se evalúa con relación al desempeño del grupo y de su lugar secuencial en la lista, para aspirar a una de las plazas y no a un criterio de nivel de conocimientos previamente definido. Consultar en: https://es.wikipedia.org/wiki/Examen_Nacional_de_Aspirantes_a_Residencias_Médicas

La evaluación con *referencia a criterio* describe el resultado específico que se encontró, de acuerdo con criterios o metas preestablecidos. Este tipo de evaluación busca la comparación del estudiante con relación a un estándar definido previa-

mente. Un ejemplo es el examen de inglés como segunda lengua, TOEFL, en el que hay niveles de desempeño previamente determinados y los resultados se interpretan de acuerdo a dichos estándares, no de acuerdo al desempeño del grupo de sustentantes. Consultar en: https://www.ets.org/toefl

De manera similar a la controversia en evaluación sumativa y formativa, la evaluación normativa versus la evaluación criterial puede verse desde diversas perspectivas. Mientras que los resultados de un examen criterial también pueden utilizarse para jerarquizar a los estudiantes por la puntuación obtenida, en un examen normativo pueden definirse niveles de desempeño.

Instrumentos de evaluación del aprendizaje

Los instrumentos de evaluación son técnicas de medición y recolección de datos que tienen distintos formatos, atendiendo a la naturaleza de la evaluación. Existe una gran variedad de instrumentos para documentar el aprendizaje de los conocimientos, habilidades y destrezas de los estudiantes, con sus respectivas ventajas y limitaciones. Es responsabilidad del profesor y de la institución educativa elegir los métodos más apropiados para el proceso de evaluación, dependiendo del modelo educativo utilizado, la normatividad institucional y las particularidades del contexto.

Los instrumentos pueden clasificarse en las siguientes categorías:

- ***Evaluaciones escritas (de respuesta construida o de selección):*** Ensayos, preguntas directas de respuesta corta, exámenes de opción múltiple, relación de columnas, disertaciones, reportes.
- ***Evaluaciones prácticas:*** Exámenes orales, exámenes prácticos con casos, examen clínico objetivo estructurado (ECOE).
- ***Observación:*** Reporte del profesor, listas de cotejo, rúbricas.
- ***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.

Cada uno de estos métodos tiene ventajas y desventajas, características psicométricas, así como recomendaciones para su implementación. Es responsabilidad de los profesores y responsables de la evaluación en las instituciones educativas, diseñar,

seleccionar, implementar y acumular evidencia de validez de los instrumentos más apropiados para evaluar el aprendizaje de los estudiantes, de acuerdo con el currículo y las características del contexto local.

Criterios para una buena evaluación

La evaluación educativa depende de la metodología utilizada, la calidad del proceso y el uso que se hace de los resultados. Varias organizaciones internacionales han propuesto criterios sobre las *buenas prácticas* en evaluación, se cita las enunciadas por el Grupo de Consenso de la Conferencia de Ottawa, un evento académico dedicado a la evaluación de la competencia clínica en ciencias de la salud, que se ha destacado por promover los aspectos académicos de la evaluación educativa (Norcini, Anderson, Bollela, Burch, Costa, Duvivier, 2011). Estos criterios son: validez, confiabilidad, justicia, equivalencia, factibilidad, efecto educativo y catalítico, y aceptabilidad.

Validez

Uno de los conceptos más importantes para que los resultados de los procesos de evaluación tengan sustento sólido y uso apropiado, es el de *validez*. La validez de un proceso de evaluación es el grado con el que mide lo que se supone que mide. Tradicionalmente, la validez en educación se clasificaba como las tres C: validez de contenido, de criterio y de constructo. En la definición actual de validez, este esquema de tres tipos de validez desaparece. Ahora la validez es un concepto unitario y se considera que toda la validez es validez de constructo (AERA, 2014; Downing, 2003; Kane, 2013). 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: la inteligencia, la timidez, los conocimientos sobre química, las habilidades de comunicación escrita, etc. (Brennan, 2006; Downing, 2003).

Validez es un juicio valorativo holístico e integrador que requiere múltiples fuentes de evidencia para la interpretación del constructo evaluado, ya que intenta responder a la pregunta: *¿qué inferencias pueden hacerse sobre la persona basándose en los resultados del examen?* (Downing, 2003; Mendoza Ramos, 2015). Debemos tener claro que no es el instrumento el que es válido per se, ya que la validez de un examen es específica para un propósito, 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 examen para ingresar a las residencias médicas (ENARM), no deben interpretarse categóricamente como evidencia de la calidad de la enseñanza en 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 la que estudió el aspirante), debe acumularse evidencia de diversas fuentes para sustentar esta interpretación (Sánchez Mendiola, 2016).

En el modelo vigente, las cinco fuentes importantes de validez de constructo en evaluación educativa son (AERA, 2014; Downing, 2003):

1) *Contenido:* En los exámenes escritos la documentación de evidencia de validez de contenido es fundamental. Por ejemplo, la tabla de especificaciones de la prueba y el proceso seguido para elaborarla, definición del contenido temático, 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, entre otros.

2) *Procesos de respuesta:* Es evidencia de la 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 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, el control de calidad del reporte de los resultados del examen, la familiaridad del estudiante con el formato de evaluación (lápiz y papel vs. computadora).

3) *Estructura interna:* Se refiere a las características estadísticas y psicométricas del examen y de las preguntas que lo componen: análisis de reactivos con el grado de dificultad e índices de discriminación, desempeño de los distractores en las preguntas de opción múltiple, confiabilidad del examen, 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 sumativos.

4) *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 otra medición de características conocidas. Este rubro busca evidencia confirmatoria y contradictoria, representando 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 en el examen de admisión a la licenciatura con las obtenidas en los exámenes parciales durante la carrera, se consideraría evidencia de validez para interpretar el resultado de dicho examen como de utilidad predictiva del proceso de admisión.

5) *Consecuencias:* Se refiere al impacto de la evaluación en los sustentantes, de las decisiones que se toman considerando los resultados del examen, y su efecto en la enseñanza y el aprendizaje. Por ejemplo: 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.

El concepto actual de validez de constructo implica 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, pasar o reprobar un curso, certificar o no a un especialista, admitir o no a un estudiante en la universidad (Downing, 2003; Kane, 2013). Bajo estas premisas, probar la validez es un proceso que nunca queda completo, ya que siempre se puede indagar más sobre el significado de los resultados de un examen con diversos grupos de estudiantes 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 (Brennan, 2006). Quienes elaboramos exámenes, tenemos la obligación ética y el imperativo educativo de documentar qué tan defendible es la

interpretación de los resultados, en beneficio de los educandos y de la sociedad en general.

Confiabilidad

La confiabilidad o fiabilidad (*reliability* en inglés) tiene un significado técnico preciso en evaluación educativa, que no debe confundirse con la percepción coloquial del término. Es la capacidad del examen de arrojar un resultado consistente cuando se repite, es decir, su reproducibilidad (Downing, 2004). Se trata de un concepto estadístico, que representa el grado en el cual las puntuaciones de los alumnos serían similares si fueran examinados de nuevo, en el que el instrumento mide el fenómeno de manera consistente. 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.

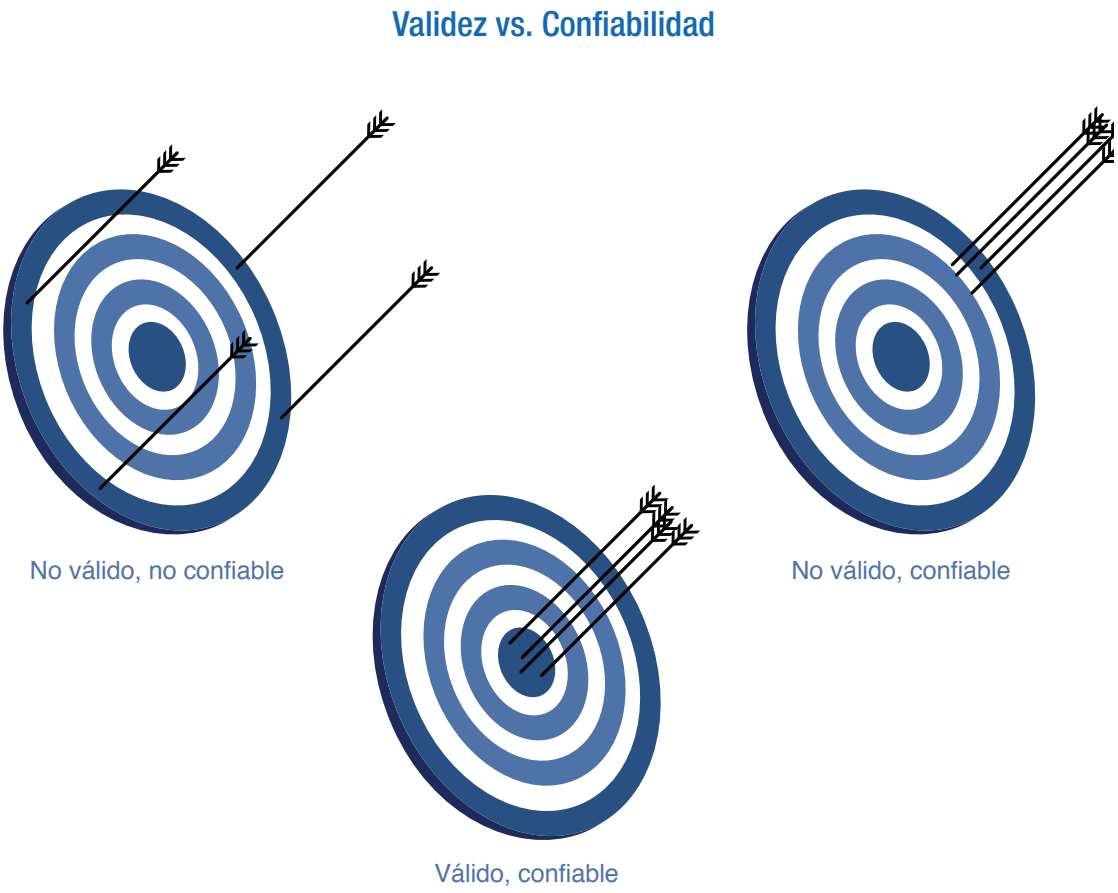
Generalmente se expresa como un coeficiente de correlación, 1.0 es una correlación perfecta y 0 ninguna correlación. Mientras más alta es la cifra de confiabilidad, por lo común, es mayor su peso como evidencia de validez en el apartado de *estructura interna* del examen. La magnitud de la cifra de confiabilidad 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 las consecuencias que tendrá la evaluación sobre los estudiantes.

Para exámenes de muy altas consecuencias, 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 puede afectar de manera importante a los examinados. Para exámenes de consecuencias moderadas, como las evaluaciones sumativas de fin de curso en la escuela, es deseable que la confiabilidad sea de 0.80 a 0.89. En exámenes de menor impacto, como la evaluación formativa o exámenes parciales diagnósticos, es aceptable una confiabilidad de 0.70 a 0.79. Estas cifras no representan rangos absolutos, debido a que hay diferencias de opinión entre los expertos, pero pueden servir de marco de referencia (Downing, 2004).

La confiabilidad de una medición es necesaria para obtener resultados válidos, aunque puede haber resultados confiables sin validez —es decir, la confiabilidad

es necesaria, pero no suficiente para la validez—. La analogía con la diana de un blanco de tiro es útil para entender la relación entre los dos conceptos como se demuestra en la Figura 1.

Figura 1. Esquema visual de los conceptos de validez y confiabilidad, con el símil de un blanco de tiro



Justicia y equidad

En las últimas décadas, las principales organizaciones de evaluación educativa del mundo han hecho énfasis en la necesidad de justicia y equidad en todo el proceso educativo, incluso en la evaluación del aprendizaje, para ser congruentes con el principio social de la educación (AERA, 2014; INEE, 2017). Hay una fuerte controversia sobre el tema, pues los exámenes estandarizados a gran escala, que por necesidad se aplican y analizan en contextos altamente controlados para que cada sustentante se enfrente al mismo reto en igualdad de condiciones, por definición, tratan a todos los estudiantes de la misma manera. Si queremos promover la evaluación formativa para el aprendizaje deberíamos individualizar el uso de los instrumentos para cada caso específico. Esta permanente tensión entre lo ideal y lo real continúa sin resolverse. La tarea de la equidad y justicia integral en evaluación educativa, en la sociedad moderna, es aún una asignatura pendiente.

Equivalencia

La equivalencia se refiere a que los exámenes proporcionen puntuaciones o decisiones equivalentes, cuando se administran en diferentes instituciones o tiempos (AERA, 2014; Norcini et al., 2011). La mayoría de los docentes no conocemos este concepto técnico, a pesar de su importancia para interpretar exámenes aplicados de manera periódica que pretenden evaluar el mismo constructo, o exámenes en diferentes contextos en los que queremos asegurar que sean de la misma dificultad (Carter, 1984; Moreno Olivos, 2010). Para lograr equivalencia se requiere de procedimientos estadísticos de varios grados de sofisticación, de la familia de métodos de equiparación o *igualación* de exámenes. Uno de ellos es el uso de *reactivos ancla* (preguntas con un grado de dificultad similar y comportamiento estadístico bien documentado) en un porcentaje de reactivos de cada versión del examen. Para aplicar estas técnicas se requieren profesionales de evaluación educativa, expertos en estos procedimientos.

Factibilidad y aceptabilidad

Estas características se refieren a que las evaluaciones sean prácticas, realistas y apropiadas a las circunstancias del contexto, incluyendo las instalaciones físicas y disponibilidad de recursos humanos y financieros. Por ejemplo, el método más utilizado en el mundo para evaluar la competencia clínica en medicina es el Exa-

men Clínico Objetivo Estructurado (ECOE), que consiste en una serie de múltiples estaciones estandarizadas en las que en cada una, el sustentante se enfrenta a un reto que requiere la aplicación de competencias específicas, como pueden ser habilidades de comunicación, realizar un diagnóstico o interpretación de radiografías y estudios de laboratorio (Boursicot et al., 2011). Este tipo de examen requiere una gran cantidad de recursos humanos, instalaciones apropiadas y mucha dedicación en términos de disciplina, tiempo y organización. Esta disponibilidad de recursos puede no estar al alcance de varias escuelas, de manera que, aunque el examen sea excelente y tenga mucha evidencia de validez y confiabilidad, si no se puede aplicar hay que buscar otras alternativas. Otros ejemplos podrían ser el uso de exámenes adaptativos por computadora, simuladores de alta fidelidad y tecnología de punta, herramientas que requieren una gran inversión inicial y de mantenimiento. Las evaluaciones también deben ser aceptables, tanto por los estudiantes como por los profesores. Si hay rechazo de la comunidad de docentes o estudiantes a algún método de evaluación, por ejemplo, los exámenes estandarizados de opción múltiple que no sean aceptados por cuestiones socioculturales locales o por el modelo educativo utilizado, ello puede hacer difícil su implementación.

Efecto educativo, efecto catalítico

Todos los métodos de evaluación, sobre todo los de índole sumativa, tienen un efecto en los métodos de estudio y prioridades de aprendizaje de los estudiantes (Newble, 1983). Aunque los profesores les digamos a nuestros alumnos que un tema o concepto es fundamental, la pregunta común es: ¿eso va a venir en el examen? La cultura prevalente en muchas escuelas es que si algo no cuenta para el examen no se le da mucha importancia, se ha documentado que la manera como se aplica la evaluación tiene efectos en la motivación de los estudiantes y en sus métodos de estudio. La evaluación además tiene un efecto *catalítico* en el contexto educativo, debido a que puede tener influencia en los demás docentes, en los departamentos académicos y en la institución misma (Norcini et al., 2011). Por ejemplo, si se privilegian los exámenes escritos de opción múltiple, habrá un efecto en cascada en los diferentes actores del proceso educativo. Si se fomenta la evaluación formativa, de la misma manera habrá influencia en las actitudes hacia la evaluación de los participantes, sobre todo cuando sientan sus efectos positivos.

Amenazas a la validez

Si partimos de la base de que la validez es el concepto más importante en evaluación educativa, es fundamental adquirir conciencia de las amenazas a la misma. Existen diversas amenazas para la validez de un proceso de evaluación, elementos que disminuyen la credibilidad de las inferencias que se pueden hacer de los resultados de un examen. Pueden clasificarse de la siguiente manera (Downing y Haladyna, 2004):

- ***Infrarrepresentación del constructo (IC):*** Se refiere a una representación inapropiada de los dominios del contenido a evaluar por los exámenes, por ejemplo: pocos reactivos en el examen que no muestreen apropiadamente el área de conocimiento explorada; distribución de reactivos que no siga fielmente la tabla de especificaciones del examen, de manera que algunas áreas sean sobreexploradas y otras infraexploradas; uso de muchos ítems que exploren procesos cognoscitivos de bajo nivel, como la memoria o reconocimiento de datos factuales, mientras que las metas de la enseñanza son la aplicación o solución de problemas.

  Otra amenaza a la validez es el fenómeno de *enseñando para la prueba (teaching to the test en inglés)*, en el que se enfatiza demasiado lo que va a venir en el examen, distorsionando la totalidad del currículo y del proceso educativo, generando resultados educativos incompletos que no preparan al estudiante para enfrentarse al ejercicio profesional (Popham, 2001). Esto puede llegar al grado en que algunos profesores utilizan reactivos del examen en clase para aumentar, artificialmente, las calificaciones de sus alumnos y mejorar las evaluaciones de su grupo o escuela. No hay que estudiar con la visión limitada de aprender solamente lo que va a venir en el examen.

- ***Varianza irrelevante al constructo (VIC):*** Se refiere a variables que de manera sistemática 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 los reactivos elaborados con deficiencias y que tienen fallas de acuerdo a las recomendaciones basadas en evidencia educativa; incluir preguntas demasiado difíciles, fáciles o que no permiten identificar a los estudiantes que saben más y los que saben menos; uso de estructuras gramaticales complejas en las preguntas o instrucciones difíciles de entender. Escribir buenos reactivos para exámenes requiere entrenamiento y experiencia, no es algo tan sencillo.

Otro ejemplo de VIC son los problemas en la seguridad del examen y fuga de información, de manera que el resultado del examen no refleja los conocimientos de los estudiantes. Este problema invalida los resultados de los exámenes, tiene implicaciones éticas y de uso de recursos como es repetir el examen con otra versión. La mayoría de las instituciones educativas en países como el nuestro tienen bancos de reactivos limitados, por lo que *quemar* o sobreexponer reactivos, se constituye en un gran problema operacional.

La *astucia* para responder los exámenes (en inglés *testwiseness*) ocurre cuando los estudiantes que se preparan con estrategias para responder exámenes pueden obtener puntajes que no reflejen lo que realmente saben. Se ha creado un mercado gigantesco de organizaciones que dan cursos para pasar exámenes, en los que el elemento principal es adiestrar a los asistentes en métodos para obtener la mayor puntuación posible, como detectar errores gramaticales en los reactivos, entre otros. Las familias de los estudiantes pagan un precio alto por estos cursos que son de efectividad cuestionable y que promueven una competencia desleal.

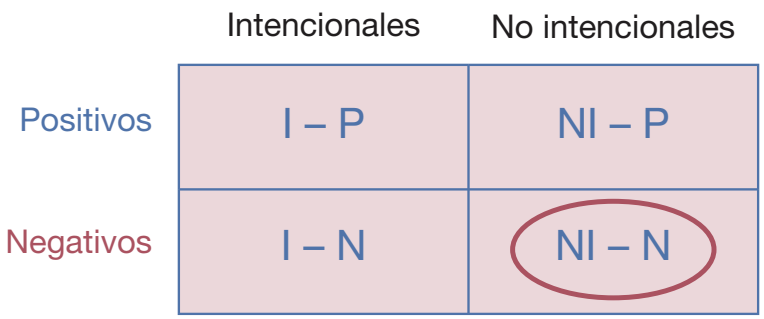
Exámenes de alto impacto

Los exámenes de altas consecuencias o de alto impacto (*high-stakes tests* en inglés) son aquellas pruebas o exámenes que tienen consecuencias serias e importantes para los individuos que los toman, por lo que han sido sujetos de mucha controversia (Márquez Jiménez, 2014; Nichols, 2007; Sánchez Mendiola y Delgado Maldonado, 2017). Ejemplo de ellos son: exámenes de fin de carrera, de admisión a la universidad, finales de curso, de certificación profesional, entre otros. Estos exámenes tienen un impacto poderoso en aspectos políticos, sociales, económicos, humanos y de forma importante, efectos educativos. Estos efectos no son sencillos ni lineales, sino que forman parte de una compleja red de interacciones que han convertido a los exámenes de alto impacto en uno de los "villanos favoritos" de la educación moderna (Cizek, 2001).

Las consecuencias de un examen de alto impacto pueden ser positivas o negativas, así como intencionales y no intencionales, como se muestra en la Figura 2 (adaptado de Brennan 2006 y Cizek, 2001). Por supuesto que lo deseable es tener efectos positivos, intencionales o no. Es poco plausible que existan efectos intencionales negativos, y es fundamental identificar los efectos negativos no intencionales, que ocurren con inusitada frecuencia. La clasificación dicotómica en efectos positivos

y negativos es un tanto artificial para fines didácticos, ya que puede haber ambivalencia en el impacto de un examen, dependiendo del individuo y del contexto. Algo positivo para alguien puede ser negativo para el otro, por ejemplo, la estandarización de los exámenes.

Figura 2. Esquema de los potenciales efectos de un examen de alto impacto (I = intencionales; NI = no intencionales; P = positivos, N = negativos)



Uno de los principales efectos positivos de estos exámenes es la motivación para estudiar. De hecho, la intencionalidad de los exámenes es enviar una señal de qué es lo importante aprender, y si se hacen bien, el efecto es primordialmente positivo. Por otra parte, es obligación de los docentes hacer explícitos los criterios de evaluación a utilizar en los cursos o asignaturas para que los estudiantes tengan claros los parámetros con los que serán evaluados. Diversos factores motivacionales extrínsecos e intrínsecos convergen en los estudiantes, lo que puede mejorar el aprendizaje de los conceptos importantes del curso.

La estandarización de la evaluación es uno de los aspectos más controversiales de los exámenes de alto impacto (Debray et al., 2003; Márquez Jiménez, 2014). Las recomendaciones de las principales organizaciones internacionales de profesionales de evaluación plantean que es importante realizar los exámenes sumativos en condiciones estandarizadas, en ambientes consistentes y con reglas y especificaciones detalladas y predefinidas, para que la situación en que los sustentantes presentan

el examen sea similar y las inferencias que se hagan de los resultados sean válidas y comparables (AERA, 2014; INEE, 2017).

Recientemente se le ha dado a la justicia e imparcialidad de los exámenes, el mismo nivel de importancia que la validez y confiabilidad (AERA, 2014). La evaluación debe ser justa y equitativa para todos los individuos como acto de elemental justicia. Sin embargo, varios activistas y educadores han criticado este aspecto de la evaluación de alto impacto, en virtud de que los seres humanos somos demasiado complejos para que nuestra esencia sea capturada por exámenes escritos estandarizados (principalmente los de opción múltiple). Se argumenta que estos exámenes, principalmente evalúan el conocimiento, y que la riqueza de las personas consiste en que somos mucho más que un cúmulo organizado de conocimientos (Nichols, 2007; Cizek, 2001).

El debate continúa, pero el peso de la evidencia empírica sugiere que los exámenes estandarizados, elaborados y analizados profesionalmente, con un uso apropiado y prudente de los resultados, son una de las herramientas con mayor evidencia de validez y confiabilidad, para identificar de manera justa y equitativa el nivel de conocimiento, capacidad de entender conceptos y resolver problemas, tanto de individuos como de grupos (AERA, 2014; Sánchez Mendiola y Delgado Maldonado, 2017).

Otro de los potenciales efectos positivos es la mejora de la calidad educativa. Este efecto también es controversial. Si se siguen apropiadamente los lineamientos para realizar buenos exámenes y se hace un esfuerzo por alinear la evaluación con el currículo y los métodos de enseñanza, es posible mejorar la calidad educativa. La mejora de la calidad en el sistema educativo de un país depende de una red extremadamente compleja de factores gubernamentales, sociales, económicos y personales de docentes y estudiantes, en los que los exámenes de alto impacto son solo un componente. Cualquier intento por mejorar la calidad educativa debe tener una perspectiva sistémica y tratar de identificar las estrategias que podrían contribuir a dicho proceso en el contexto local (Hattie, 2015).

En los últimos años se ha documentado que realizar exámenes potencia el aprendizaje profundo a largo plazo, más allá de su efecto directo en la solución del instrumento. A dicho concepto se le denomina "aprendizaje potenciado por exámenes" *(test-enhanced learning)*, por lo que es menester analizarlo e incorporarlo en nuestras estrategias educativas (Larsen, 2008).

El uso de exámenes de alto impacto también puede contribuir a homogeneizar diversos componentes de los procesos educativos, como los contenidos a enseñar, las metas educativas a alcanzar, la identificación de un currículo nuclear, el tipo de instrumentos a utilizar en evaluación, entre otros. Este tipo de efectos puede producir rechazo en algunos docentes con el argumento de que se limita la libertad de cátedra, por lo que hay que trabajar con los profesores desde el principio de cualquier cambio curricular (Madaus, 1988).

Un efecto potencialmente negativo importante es el interesante fenómeno de lo que se ha denominado "enseñando para la prueba" *(teaching to the test)* (Popham, 2001). El principal objetivo de las evaluaciones es obtener información que permita realizar inferencias sobre la adquisición de conocimientos y logros de las metas educativas definidas en el currículo, pero cuando los docentes y las instituciones enfatizan sobremanera lo que vendrá en los exámenes de altas consecuencias durante las actividades de enseñanza, entonces el currículo se distorsiona y puede llegarse al grado de enseñar solo lo que vendrá en los exámenes. Incluso hay casos en que los docentes y escuelas enseñan a sus estudiantes con preguntas de exámenes que pueden venir (o que vendrán, en el caso de que exista corrupción) en los exámenes de alto impacto. Enseñar para los exámenes puede ser toda una gama de actividades, desde las muy sutiles, implícitas e inconscientes por parte del docente, hasta las explícitas y dirigidas, principalmente, a subir las puntuaciones en los exámenes.

Otro aspecto negativo es la proliferación de los cursos para preparar a las personas a contestar exámenes de alto impacto. Ante lo importante de las consecuencias de no aprobar un examen de esta naturaleza, no es de extrañar que aparezcan una serie de cursos, libros y aplicaciones informáticas para mejorar las puntuaciones en los exámenes. Estos eventos y recursos se han convertido en un lucrativo negocio en nuestro país y en el resto del mundo, tomando ventaja de la necesidad de los aspirantes a cualquier nivel educativo, de aumentar sus puntuaciones. En Norteamérica McGaghie (2004) realizó una revisión sistemática sobre los cursos comerciales para preparar aspirantes para los exámenes de alto impacto en educación médica, en los que una sola empresa reporta ventas por más de 250 millones de dólares. Encontraron que, prácticamente, no existe evidencia de su utilidad, los pocos estudios que muestran un efecto débil tienen una metodología de investigación deficiente, por lo que concluyen que no está demostrado que los cursos comerciales de este tipo tengan valor real y que el temor a las evaluaciones sumativas, la poca cultura de evaluación de los estudiantes de medicina y las estrategias agresivas de publicidad de las empresas involucradas, son los responsables de su prosperidad financiera.

Existe gran controversia sobre el efecto de los exámenes de alto impacto en el currículo formal, vivido y oculto de las instituciones de educación (Koretz, Linn, Dunbar y Shepard, 1991; Martone, 2009; Mehrens, 1998; Sánchez Cerón, 2013). Hay poca investigación rigurosa, lo que hace difícil tener conclusiones contundentes y claras. Tradicionalmente se utiliza la premisa de que las evaluaciones de alto impacto tienen influencia importante en el currículo, los métodos de enseñanza de los docentes y las estrategias de estudio de los alumnos, con el argumento de que "lo que examinas es lo que obtienes". Existe la percepción global de que los graduados de las universidades tienen deficiencias en varias de las habilidades necesarias para salir adelante en el siglo XXI y que han dedicado demasiado esfuerzo a "saber contestar exámenes", que sirve poco en la vida real. No se ha demostrado de manera contundente que los exámenes estandarizados de alto impacto influyan sustancialmente en el currículo.

En una metasíntesis de 49 estudios cualitativos sobre cómo los exámenes de alto impacto afectan el currículo, los contenidos de conocimiento enseñados y las estrategias pedagógicas de los docentes, se encontró que el efecto principal de este tipo de exámenes es el estrechamiento del currículo que se dirige a los contenidos examinados en las pruebas (Au, 2007). También se encontró que las áreas de conocimiento de los contenidos educativos se fragmentan en piezas relacionadas con los exámenes, y que los docentes incrementan el uso de estrategias pedagógicas centradas en el profesor, como la instrucción directa con conferencias y menor interactividad. Ciertos tipos de exámenes de alto impacto tuvieron, al contrario, efectos positivos en las tres dimensiones arriba citadas, con expansión del currículo, integración del conocimiento y estrategias de enseñanza centradas en el estudiante, por lo que Au (2007) concluye que la naturaleza del control curricular, inducido por los exámenes de alto impacto, es altamente dependiente de la estructura de los exámenes y la forma en que se implementan en el proceso educativo.

El problema de los usos e inferencias inapropiados de los resultados de exámenes de alto impacto es uno de los retos más importantes que enfrenta la comunidad de profesionales de evaluación educativa. Aún hay un largo trecho por caminar en el incremento de una cultura de la evaluación en alumnos, docentes, directivos y funcionarios gubernamentales, así como la sociedad en su conjunto. Uno de los efectos negativos más frecuentes de los exámenes de alto impacto es realizar inferencias de los resultados que no son congruentes con los objetivos iniciales del examen, por lo que dichas conclusiones tienen validez limitada. Con facilidad las declaraciones breves y sensacionalistas en los medios de comunicación generan malentendidos y distorsión de las conclusiones, limitaciones e implicaciones reales

de los exámenes, como ocurre frecuentemente con los exámenes PISA. La comprensión clara del concepto moderno de validez es fundamental para entender las limitaciones de los resultados de los exámenes de alto impacto, ya que extrapolar conclusiones y decisiones más allá de lo académicamente obtenible, es inapropiado e incluso puede ser peligroso. Si un estudiante tiene un desempeño deficiente en una aplicación de un examen sumativo de alto impacto, eso no significa que sea "mala persona", "incompetente", alguien que "no debió estudiar esa carrera", entre otros muchos calificativos que se asignan como etiquetas y que tienen un impacto emocional importante en los sustentantes.

Conclusiones y recomendaciones

La evaluación del aprendizaje es un componente fundamental del proceso educativo, por lo que debemos profundizar nuestros conocimientos y habilidades en sus aspectos metodológicos y aplicativos.

Es necesario que los docentes en la práctica, los responsables institucionales de coordinar la enseñanza y las autoridades universitarias y gubernamentales, adquieran una amplia perspectiva del campo de la evaluación educativa. Se requiere una gama de expertos en diferentes elementos del proceso de evaluación, como: pedagogos, psicólogos, matemáticos, científicos de datos, para avanzar en la profesionalización de esta disciplina y contribuir a mejorar el aprendizaje complejo de nuestros estudiantes.

Una de las principales recomendaciones de los expertos mundiales en evaluación es: *"Los desarrolladores del examen son los candidatos obvios para validar las afirmaciones que hacen sobre la interpretación de los resultados de un examen…"* (Brennan, 2006), por lo que la responsabilidad de realizar buenos instrumentos e informar a la sociedad sobre sus limitaciones recae en nuestras organizaciones y grupos de expertos, en colaboración con las autoridades y los medios de comunicación. La asimetría de poder intrínseca en los procesos de evaluación sumativa conlleva una gran responsabilidad de las autoridades académicas e institucionales.

Los instrumentos de evaluación y el uso que se hace de ellos en las universidades son la declaración pública más importante de "lo que realmente cuenta" para la institución. Los estudiantes están muy alerta a estas señales, que a veces son sutiles y en ocasiones explícitas y visibles sobre lo que deben aprender, lo que realmente aprenden y cómo lo aprenden, por lo que las instancias evaluadoras y

los docentes deben hacer lo posible porque estos procedimientos de evaluación se realicen con profesionalismo educativo en un entorno de calidad y abundante evidencia de validez. Al final del día, el uso de la puntuación de un examen, definitivamente implica consecuencias; de otra manera, "uso" es solo una abstracción. Los exámenes de alto impacto han adquirido un enorme grado de sofisticación técnica y metodológica, y llegaron para quedarse. Lo más importante es encontrar un balance entre este tipo de evaluación y la evaluación para el aprendizaje que analizaremos en otro capítulo.

Como ha dicho un académico mexicano, el Dr. Tiburcio Moreno, la evaluación tiene muchas caras, y en países como el nuestro, ha estado permeada por una visión empirista que descansa en el principio: *"Todos sabemos de evaluación, porque alguna vez hemos sido evaluados"* (Moreno Olivos, 2010). Debemos mejorar nuestros conocimientos y habilidades en evaluación, es una obligación ética y moral de todos los docentes.

Referencias

American Educational Research Association, American Psychological Association, National Council on Measurement in Education, y Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington, D.C.: AERA.

Au, W. (2007). High-Stakes testing and curricular control: A qualitative metasynthesis. *Educational Researcher, 36*(5), 258-267.

Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S., et al. (2011). Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Med Teach*., *33*(5), 370-383. DOI:0.3109/0142159X.2011.565831

Brennan, R. L. (2006). Perspective on the Evolution and Future of Educational Measurement. En R. L. Brennan (Ed.), *Educational Measurement. National Council on Measurement in Education and American Council on Education* (pp. 1-16). Westport, Connecticut: Praeger Publishers.

Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education, 35*(6), 57-60.

Cizek, G. J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice, 20,* 19-27.

Debray, E., Parson, G., y Avila, S. (2003). Internal alignment and external pressure. En M. Carnoy, R. Elmore y L. S. Siskin (Eds.). *The new accountability: High schools and high-stakes testing* (pp. 55-85). Nueva York: Routledge Falmer.

Downing, S. M. (2003a). Validity: on the meaningful interpretation of assessment data. *Med Educ*., *37,* 830-837.

Downing, S. M. (2004b). Reliability: on the reproducibility of assessment data. *Med Educ*., *38,* 1006-1012.

Downing, S. M., y Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*., *38,* 327-333.

Hattie, J. (2015). *What Doesn't Work in Education: The Politics of Distraction*. Londres: Pearson. Recopilado de https://www.pearson.com/content/dam/corporate/global/pearson-dot-com/files/hattie/150602_DistractionWEB_V2.pdf

Instituto Nacional para la Evaluación de la Educación. (2017). *Criterios Técnicos para el Desarrollo y Uso de Instrumentos de Evaluación Educativa 2015-2016*. Ciudad de México: INEE. Recuperado de https://www.inee.edu.mx/index.php/criterios-tecnicos

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *J. Educ. Meas*., *50*(1), 1-73.

Koretz, D. M., Linn, R. L., Dunbar, S. B., y Shepard, L. A. (1991). The Effects of High-Stakes Testing on Achievement: Preliminary Findings About Generalization Across Tests. Presented at

the annual meeting of the American Educational Research Association. En R.L. Linn, *The Effects of High Stakes Testing, annual meeting of the American Educational Research Association and the National Council on Measurement in Education,* Chicago, April 1991. https://bit.ly/2qWl2XD

Larsen, D. P., Butler, A. C., y Roediger, H. L. (2008). Test-enhanced learning in medical education. *Med. Educ.*, *42*(10), 959-966.

Madaus, G. F. (1988). The influence of testing on the curriculum. En L. N. Tanner (Ed.), Critical issues in curriculum: *Eighty-seventh year-book of the national society for the study of education* (pp. 83-121). Chicago: University of Chicago Press.

Man Sze Lau, A. (2016). "Formative good, summative bad?" — A review of the dichotomy in assessment literature. *Journal of Further and Higher Education, 40*(4), 509-525. doi:10.1080/0309877X.2014.984600

Márquez Jiménez, A. (2014). Las pruebas estandarizadas en entredicho. *Perfiles Educativos, 36*(144), 3-9.

Martínez Rizo, F. (2009a). Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. *Revista Electrónica de Investigación Educativa, 11*(2). Recuperado de http://redie.uabc.mx/redie/article/view/231

Martínez Rizo, F. (2013b). Dificultades para implementar la evaluación formativa: Revisión de literatura. *Perfiles Educativos, 35*(139), 128-150. Recuperado de http://www.scielo.org.mx/pdf/peredu/v35n139/v35n139a9.pdf

Martone, A., Sireci, S. G. (2009). Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research, 79*(4), 1332-1361.

McGaghie, W. C., Downing, S. M. y Kubilius, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? Teach Learn Med., *16*(2), 202-211.

Mehrens, W. A. (1998). Consequences of assessment: What is the evidence? *Education Policy Analysis Archives, 6*(13). Recuperado de http://epaa.asu.edu/ojs/article/download/580/703

Mendoza Ramos, A. (2015). La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles Educativos, 37*(149), 169-186.

Miller, M. D., Linn, R. L. y Gronlund, N. E. (2012). Measurement and Assessment in Teaching (11va ed.). Londres: Pearson.

Moreno-Olivos, T. (2010). Lo bueno, lo malo y lo feo: las muchas caras de la evaluación. *Revista Iberoamericana de Educación Superior, I*(2), 84-97.

Newble, D. I., y Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Med Educ., 17*(3), 165-71.

Nichols, S. L., y Berliner, D. C. (2007). *Collateral damage: How high-stakes testing corrupts America's schools*. Cambridge, Massachusetts: Harvard Education Press.

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., (2011). Criteria for

good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med. Teach., 33*(3), 206-214.

Popham, W. J. (2001). Teaching to the Test? *Educational Leadership,* 58(6),16-20. Recuperado de http://www.ascd.org/publications/educational-leadership/mar01/vol58/num06/Teaching-to-the-Test%C2%A2.aspx

Rowntree, D. (1977). *Assessing students: How shall we know them?* Londres: Kogan Page Ltd.

Sánchez Cerón, M. y del Sagrario Corte Cruz, F. M. (2013) Las evaluaciones estandarizadas: sus efectos en tres países latinoamericanos. *Revista Latinoamericana de Estudios Educativos, 43*(1), 97-124.

Sánchez-Mendiola, M., Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. *Inv Ed Med., 6*(21), 52-62. DOI:10.1016/j.riem.2016.12.001

Sánchez-Mendiola, M., Delgado-Maldonado, L., Flores, F., Leenen, I., y Martínez, A. (2015). Evaluación del aprendizaje. En Sánchez M. Sánchez Mendiola, A. Lifshitz Guinzberg, P. Vilar Puig, A. Martínez González, M. Varela Ruiz, y E. Graue Wiechers (Eds.), *Educación Médica: Teoría y Práctica* (pp. 89-95). Ciudad de México: Elsevier.