Revista Digital Universitaria
Vol. 19, Núm. 6, noviembre-diciembre 2018

La evaluación del aprendizaje de los estudiantes: ¿es realmente tan complicada?

Melchor Sánchez Mendiola

Resumen

El proceso educativo incluye diversos elementos como los métodos de enseñanza, el aprendizaje y la evaluación. Frecuentemente se privilegian los métodos de enseñanza en la formación de los profesores, y la adquisición de conocimientos y habilidades sobre evaluación del aprendizaje en educación es limitada o se lleva a cabo de manera informal. Es importante que los participantes del proceso educativo, estudiantes, docentes, autoridades, así como la sociedad en general, conozcan algunos de los conceptos básicos de evaluación del aprendizaje, ya que como cualquier área técnico-científica, posee una terminología propia. Los actores de la educación debemos adquirir conciencia de los alcances y limitaciones de los instrumentos de evaluación del aprendizaje. Este escrito presenta algunas de las definiciones relevantes en el área de la evaluación del aprendizaje.

Palabras clave: evaluación educativa, evaluación del aprendizaje, evaluación para el aprendizaje, validez, amenazas a la validez.

DOI: http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a1

The assessment of learning in students: is it really so complicated?

Abstract

The educational process comprises several important elements, including teaching methods, learning and assessment. Teaching methods are frequently emphasized in faculty development activities, while the acquisition of knowledge and skills in educational assessment is limited or undertaken informally. It is important that all stakeholders of the educational process, students, teachers, authorities and society have a basic understanding of some educational assessment basic concepts. This area, like any other technical-scientific discipline, has its own terminology. We need to develop conscience about the virtues and limitations of the educational assessment tools. This paper presents some of the relevant definitions in the area of educational assessment.

Keywords: educational assessment, assessment of learning, assessment for learning, validity, threats to validity.

Melchor Sánchez Mendiola melchorsm@unam.mx

Universidad Nacional Autónoma de México (UNAM), Coordinador de Desarrollo Educativo e Innovación Curricular. Es Médico pediatra, UDEFA. Fellow en Investigación Clínica por el Hospital General de Massachusetts, Boston y Centro de Investigación Clínica, Instituto Tecnológico de Massachusetts, Cambridge, EUA. Maestro en Educación en Profesiones de la Salud, Universidad de Illinois en Chicago, EUA. Doctor en Ciencias, Educación en Ciencias de la Salud, UNAM. Profesor de Carrera Titular C de Tiempo Completo Definitivo, División de Estudios de Posgrado de la Facultad de Medicina, UNAM.

"El aprendizaje no es una calificación"

Roman Nowak

"Evaluación es un intento de conocer a la persona"

Derek Rowntree

"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

¿Qué es la evaluación del aprendizaje?

Si le preguntamos a un estudiante probablemente nos dirá: "exámenes", y si le preguntamos a un profesor podría contestarnos: "es algo difícil que toma tiempo y experiencia, por lo que generalmente no me pagan, y para lo que no fui capacitado". Pensamos que la mayor parte de lo que enseñamos es aprendido por los estudiantes, aunque la única manera de conocer los efectos de la enseñanza es realizar una evaluación continua y técnicamente adecuada, alineada con los planes de estudio y métodos de enseñanza, que incluya al estudiante como actor activo en el proceso. Esta evaluación debe idealmente arrojar resultados interpretables y utilizables por el mismo estudiante, el docente, la institución educativa y la sociedad.

Existen varias definiciones de evaluación, una de las más utilizadas es: "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" (Miller, 2012). Dichos juicios necesitan algún referente, como puede ser el plan de estudios. Evaluación implica obtener información de diferentes fuentes como realimentación, exámenes, tareas y diversas interacciones con el educando. Los profesores que interactuamos con estudiantes debemos incorporarla desde una visión más profunda, como sugirió Derek Rowntree: "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*" (Rowntree, 1977). No debemos olvidar que a quienes evaluamos son seres humanos, con todo lo que ello implica.

Las siguientes son algunas recomendaciones para que la evaluación del aprendizaje se lleve a cabo de forma apropiada (Miller, 2012):

1. Especificar claramente lo que se va a evaluar es fundamental.
2. La evaluación es un medio para un fin, no un fin en sí mismo.
3. Los métodos de evaluación del aprendizaje deben elegirse por su relevancia para las características que se van a evaluar del estudiante.
4. Requiere de una variedad de procedimientos e instrumentos.
5. Su uso adecuado requiere tener conciencia de su propósito y de las bondades y limitaciones de cada método.

Tipos de evaluación del aprendizaje

Evaluación diagnóstica, formativa y sumativa

Una de las clasificaciones tradicionales de la evaluación educativa es desde el punto de vista de su objetivo: 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 puede ser de utilidad para el docente, ya que le permite hacer adecuaciones en el contenido y en la implementación de las actividades académicas programadas. Un ejemplo de este tipo de evaluación es el Examen Diagnóstico de Ingreso en las licenciaturas de la Universidad Nacional

Autónoma de México (UNAM), en el que se valoran los conocimientos generales de Español y de Inglés de los estudiantes de nuevo ingreso. Los resultados se envían a cada facultad o escuela, para su uso y difusión. Recientemente colocamos estos resultados en la página de la Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC), como material de acceso abierto para cualquier persona que quiera explorar los datos, incluyendo además del reporte oficial, unas tablas dinámicas que permiten al usuario realizar comparaciones y visualizarlas.

La *evaluación sumativa* es aquella compuesta por la suma de valoraciones efectuadas durante un curso, para determinar, al final del mismo, el grado con que los objetivos de la enseñanza se alcanzaron y así otorgar calificaciones. Ejemplos de esta evaluación son los exámenes de fin de curso, los exámenes de certificación de profesionistas, el examen profesional de fin de 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 oportunidades para identificar su estado real de aprendizaje. Un tipo de exámenes sumativos que merece atención especial, son los llamados "exámenes de altas consecuencias o de alto impacto" (*high-stakes testing*, en inglés), que han generado una intensa controversia en las últimas décadas (Sánchez Mendiola, 2017).

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. Es un proceso mediante el cual se recaba

información sobre el proceso de enseñanza aprendizaje, que los maestros pueden usar para tomar decisiones sobre cómo enseñan y los alumnos para mejorar su propio desempeño, convirtiéndose en una fuente de motivación para ellos. Esta evaluación idealmente debería ocurrir a lo largo de todo el proceso educativo del estudiante. Puede ser *formal* si está oficialmente programada y es esperada en determinados momentos del proceso, o *informal* si ocurre de manera espontánea, no programada. Si se reconoce un logro del estudiante para estimularlo y reforzar su conducta se le llama *positiva*, y si critica de manera explícita algo que se hizo mal o que se puede mejorar se le llama *negativa*. La evaluación formativa tiene un poderoso componente educativo, ya que durante las actividades del día a día permite identificar aquellas que se hacen bien, así como aquellas que tienen alguna deficiencia, para detectarlas a tiempo y corregirlas (Martínez Rizo, 2009 y 2013). Este tipo de evaluación forma parte de la llamada "evaluación para el aprendizaje", en la que el enfoque no es verificar, sino apoyar y motivar al estudiante, al mismo tiempo que proporciona al profesor información sobre el aprendizaje del educando.

> **La evaluación educativa es tan buena como la metodología utilizada y el uso que se hace de los resultados.**

Desafortunadamente, se ha creado una diferencia artificial entre la evaluación sumativa y formativa, que ha generado mucha controversia. 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 e inaplicable en los seres humanos que somos individualmente diferentes. Por el contrario, la evaluación formativa ha surgido como la heroína de la película, la parte buena, positiva, nutritiva educacionalmente, que toma en cuenta los aspectos afectivos y emocionales de los estudiantes, y que ayuda a los educandos a salir adelante y aprender mejor, sin importar sus limitaciones personales y de contexto. Este debate ha creado una situación que recuerda la frase de George Orwell en *Rebelión en la Granja*: "Cuatro patas bueno, dos patas malo". Creo que debemos ver a estos dos tipos de evaluación como un continuo, ya que todas las evaluaciones pueden tener un componente sumativo y formativo, que dependerá del uso de los resultados (Man Sze Lau, 2016).

Por ejemplo, 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 de alguna manera la información a los docentes y estudiantes. En cambio, una sesión de realimentación durante el curso puede

ser principalmente formativa, pero si esta información cuenta para la calificación, adquiere una dimensión sumativa. Debemos hacer un esfuerzo por lograr un balance razonable, que promueva una mayor integración de la evaluación con el proceso de enseñanza y aprendizaje.

Evaluación referida a norma y criterio

Otra manera de clasificar la evaluación es de acuerdo con la interpretación de los resultados. 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 Mendiola *et al.*, 2015). Este tipo de evaluación se utiliza para colocar a los alumnos en listas de rendimiento y puntaje, para asignarles un lugar en el 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. La puntuación obtenida por el aspirante se evalúa en relación a lo que obtuvieron los demás 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 definido.

En cambio, 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 con relación a un nivel o

estándar establecido previamente. Un ejemplo es el examen de inglés como segundo lenguaje, *Test of English as a Foreign Language* (TOEFL), en que hay niveles de desempeño previamente determinados y los resultados se interpretan de acuerdo con dichos estándares, no de acuerdo al desempeño del grupo de sustentantes.

Uno de los retos de la evaluación criterial es que si el nivel exigido es muy alto para la población que toma el examen pueden fracasar todos los aspirantes, por lo que este tipo de exámenes deben "calibrarse" para plantear metas de evaluación congruentes con la realidad. Además, la evaluación criterial nos permite tener mayor claridad sobre nuestra situación educativa real, ya que no depende del desempeño del grupo sino de la meta a lograr. En cambio, en la evaluación por norma o relativa si tenemos un grupo de estudiantes con muy baja preparación, de cualquier manera aprobarán el examen o serán seleccionados los que tengan las puntuaciones más altas, dando una imagen arbitraria del nivel de aprendizaje de los estudiantes.

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 con diversas ventajas y limitaciones para documentar el aprendizaje de los conocimientos, habilidades y destrezas de los estudiantes. Los instrumentos de evaluación del aprendizaje pueden clasificarse en estas categorías:

- *Evaluaciones escritas*: ensayos, preguntas directas de respuesta corta, exámenes de opción múltiple, relación de columnas, disertaciones, reportes.
- *Evaluaciones prácticas*: exámenes orales, prácticas con casos, examen clínico objetivo estructurado (ECOE).
- *Observación*: reporte del profesor, listas de cotejo, rúbricas.
- *Registros del desempeño*: libretas de registro, portafolios, registros de procedimientos.
- *Autoevaluación y evaluación por pares*: reporte del educando y de los compañeros.

Cada uno de estos métodos tiene sus ventajas y desventajas, así como recomendaciones para su implementación. Es responsabilidad de los profesores y responsables de la evaluación en las instituciones educativas diseñar, seleccionar y utilizar los instrumentos más apropiados para evaluar el aprendizaje de los estudiantes, de acuerdo al plan de estudios y las características del contexto local.

Criterios para una buena evaluación

La evaluación educativa es tan buena como la metodología utilizada y el uso que se hace de los resultados. Varias organizaciones internacionales han propuesto criterios sobre las "buenas prácticas" en evaluación (American Educational Research Association [AERA], American Psychological Assocation [APA] y National Council on Measurement in Education [NCME], 2014; Norcini *et al.*, 2011). Estos criterios son: validez, confiabilidad, justicia, equivalencia, factibilidad, efecto educativo 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. La validez es un concepto unitario, y actualmente se considera que toda la validez es validez de constructo (AERA, APA y NCME, 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: inteligencia, timidez, conocimientos sobre química) (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). No es el examen 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 de los exámenes, 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 como evidencia de la calidad de las escuelas de medicina de donde provienen, ya que el examen no está diseñado con ese propósito.

Las cinco fuentes importantes de validez en evaluación del aprendizaje son (AERA, APA y NCME, 2014; Downing, 2003):

1. *Contenido.* Debe utilizarse una tabla de especificaciones de la prueba y el proceso seguido para elaborarla, la definición de los temas, la congruencia del contenido de las preguntas con las especificaciones del examen, la representatividad de las preguntas de las diferentes áreas a examinar, la calidad de las preguntas, las credenciales de las personas que elaboran las preguntas, entre otros.

2. *Procesos de respuesta.* Se requiere evidencia de integridad de los datos, de manera que las fuentes de error que se pueden asociar con la administración del examen hayan 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 (lápiz y papel o computadora).

3. *Estructura interna.* Se refiere a las características estadísticas del examen y de las preguntas que lo componen, como son el análisis estadístico de reactivos, el funcionamiento de los distractores en las preguntas de opción múltiple, la confiabilidad 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 alto impacto.

4. *Relación con otras variables.* La relación de los resultados en el examen con otras variables se refiere a la correlación estadística entre los resultados obtenidos por medio de una prueba con otra medición de características conocidas. Por ejemplo, la correlación entre el examen de admisión a la licenciatura y las calificaciones obtenidas en los exámenes parciales durante la carrera y el examen profesional.

5. *Consecuencias.* Se refiere al impacto en los estudiantes de las puntuaciones de la evaluación, de las decisiones que se toman como resultado 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.

Validez 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 rechazar 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). 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 estudiantes y de la sociedad en general.

Confiabilidad

La *confiabilidad* o *fiabilidad* tiene un significado técnico en evaluación educativa, que no debe confundirse con el significado coloquial de la palabra. La confiabilidad de un examen se refiere a la consistencia de las puntuaciones obtenidas por las mismas personas en ocasiones diferentes o con diferentes conjuntos de preguntas equivalentes, es decir, la reproducibilidad de la prueba (Downing, 2004). Es un concepto estadístico, que representa el grado en el cual las puntuaciones de los alumnos serían similares si fueran examinados de nuevo. 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. 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 alto impacto, la confiabilidad debe ser alta para que las inferencias de los resultados del examen sean defendibles. Varios expertos recomiendan una confiabilidad de por lo menos 0.90 para evaluaciones de muy altas consecuencias. 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 menores consecuencias, como la evaluación formativa o exámenes parciales diagnósticos, es aceptable una confiabilidad de 0.70 a 0.79. Estas cifras no representan rangos absolutos, ya 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 muestra en la figura 1. Si las flechas están muy dispersas entre sí y lejos de la diana, la medición es poco confiable y no es válida; si las flechas están muy juntas pero lejos del centro la medición es reproducible (confiable) pero no es válida; y si las flechas están juntas en la diana, la medición es confiable y válida.

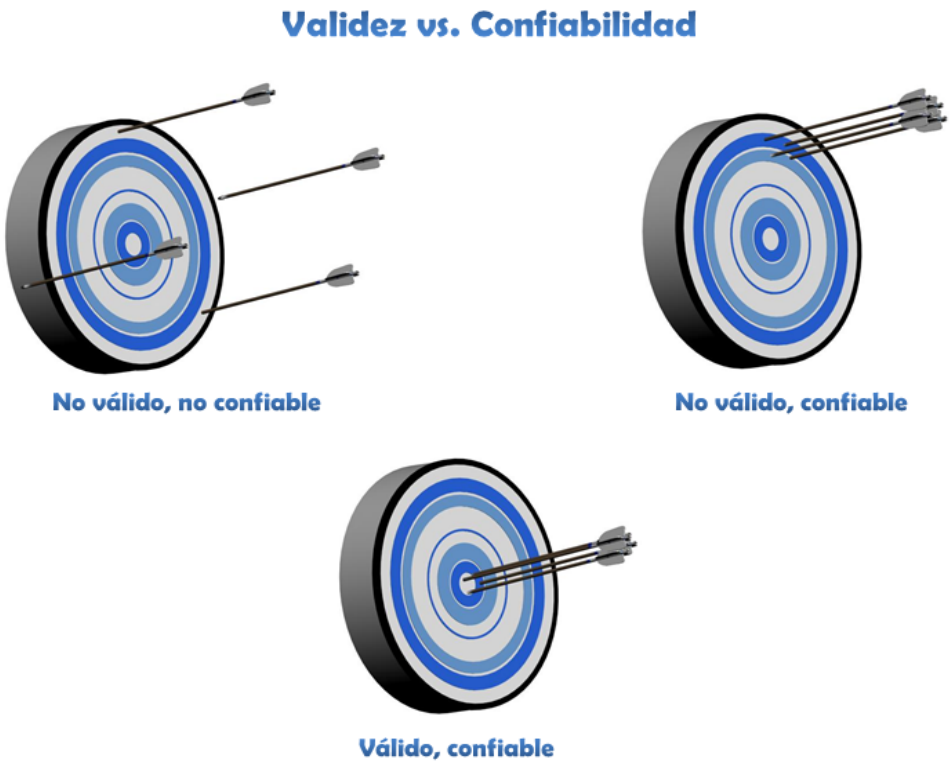

Figura 1. Esquema visual de los conceptos de validez y confiabilidad, con el símil de un blanco de tiro.

Justicia y equidad

En las últimas décadas las principales organizaciones de evaluación educativa del mundo han hecho mucho énfasis en la necesidad de justicia y equidad en todo el proceso educativo, incluyendo la evaluación del aprendizaje, para ser congruentes con el sentido social de la educación (AERA, APA y NCME, 2014; Instituto

Nacional para la Evaluación de la Educación [INEE], 2017). Existe controversia sobre el tema, ya que los exámenes estandarizados en gran escala –que por necesidad se aplican y analizan en contextos altamente controlados para que cada estudiante 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 de evaluación para cada caso específico. Esta permanente tensión entre lo ideal y lo real continúa sin resolverse. Podríamos ampliar el abanico de estrategias de evaluación educativa que utilizamos en la práctica, y capacitar a los profesores en el uso de diversos instrumentos de evaluación para promover el aprendizaje a lo largo del proceso.

Equivalencia

La *equivalencia* se refiere a que los exámenes proporcionen puntuaciones o decisiones equivalentes, cuando se administran en diferentes lugares o tiempos (AERA, APA y NCME, 2014; Norcini *et al.*, 2011). La mayoría de los docentes y estudiantes no conocemos este concepto, a pesar de su importancia para interpretar exámenes aplicados de manera periódica que pretenden evaluar lo mismo, o exámenes en diferentes contextos en los que queremos asegurar que sean de la misma dificultad, sobre todo en evaluación sumativa de alto impacto (Carter, 1984; Moreno Olivos, 2010). Para lograr equivalencia se requiere de procedimientos estadísticos sofisticados, que caen en la familia de métodos de equiparación o "igualación" de exámenes. Uno de estos métodos 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 estas técnicas se requieren profesionales en dichos procedimientos.

Factibilidad y aceptabilidad

Las evaluaciones deben ser prácticas, realistas y apropiadas a las circunstancias del contexto, incluyendo las instalaciones físicas y los recursos humanos y financieros disponibles. Por ejemplo, el método más utilizado en el mundo para evaluar la competencia clínica en medicina es el Examen Clínico Objetivo Estructurado (ECOE), que consiste en una serie de múltiples estaciones estandarizadas, en las que cada estudiante se enfrenta a un reto que requiere que aplique algunas competencias específicas, como pueden ser las habilidades de comunicación, el hacer un diagnóstico, o interpretar radiografías y estudios de laboratorio (Boursicot *et al.*, 2011). Este tipo de examen requiere gran cantidad de recursos humanos, instalaciones apropiadas y mucha dedicación en disciplina, tiempo y organización. Esta disponibilidad de recursos puede no estar al alcance de algunas escuelas, de manera que, aunque el examen sea excelente y se use en muchas partes del mundo, si no se puede hacer en

una institución hay que buscar 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 un rechazo de la comunidad a algún tipo de evaluación –por ejemplo, la evaluación por pares que implica ser evaluado por sus compañeros–, se hace difícil su implementación.

Efecto educativo, efecto catalítico

Todos los métodos de evaluación, sobre todo los sumativos, pueden tener efectos 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: "¿y eso va a venir en el examen?". La cultura de algunas escuelas es que si algo no cuenta para el examen no se le da mucha importancia, así que la manera cómo se aplica la evaluación tiene consecuencias en la motivación de los estudiantes y en sus métodos de estudio. También la evaluación puede tener un efecto "catalítico" en el contexto educativo, ya que puede influir en los demás docentes, en los departamentos académicos y en la institución misma (Norcini *et al.*, 2011). Si se privilegian los exámenes escritos de opción múltiple, habrá un efecto en cascada en los diferentes participantes 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 vivan sus efectos positivos.

Amenazas a la validez

Existen diversas "amenazas" para la validez de un proceso de evaluación del aprendizaje, que disminuyen la credibilidad de las inferencias que se pueden hacer de los resultados de un examen. Al ser la validez uno de los principales elementos de una buena evaluación, todo lo que ponga en riesgo la veracidad de las conclusiones que podamos tener sobre los resultados de una prueba o examen debe identificarse y, en la medida de lo posible, evitarse o corregirse. Pueden clasificarse de la siguiente manera (Downing y Haladyna, 2004):

- *Infrarrepresentación del constructo (*IC*).* Se refiere a una representación inapropiada del contenido a evaluar por los exámenes, teniendo en mente que el constructo es aquello que queremos investigar (como los conocimientos de química en el bachillerato). Son ejemplos de esta amenaza: muy pocas preguntas en el examen, que no exploren apropiadamente el área de conocimiento; uso de preguntas que exploren principalmente memoria o reconocimiento de datos, cuando 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 el plan de estudios y el proceso educativo, y generando resultados incompletos que no preparan al estudiante para enfrentarse al ejercicio profesional (Popham, 2001). A veces ocurre al grado 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.

- *Varianza irrelevante al constructo (*VIC*).* Se refiere a elementos que interfieren con la capacidad de interpretar los resultados de la evaluación de una manera significativa, y que causan "ruido" en la evaluación. Por ejemplo, las preguntas elaboradas con fallas, gramaticales o de otro tipo; y las que dan pistas al estudiante sobre cuál es la respuesta correcta, aunque no sepa el concepto explorado en la pregunta. Recordemos que escribir buenas preguntas de examen requiere entrenamiento y experiencia. Otro ejemplo son los problemas de 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, con diversas implicaciones éticas y de uso de recursos, como es repetir el examen con otra versión.

  La "astucia" o habilidad para responder los exámenes (en inglés, *testwiseness*) ocurre cuando los estudiantes se preparan con estrategias para responder exámenes y pueden obtener puntajes que no reflejen lo que realmente saben. Se ha creado un mercado de organizaciones que dan cursos para pasar exámenes, en los que el objetivo es adiestrar a los asistentes en métodos para obtener la mayor puntuación posible. Las familias de los estudiantes pagan un precio alto por estos cursos, que son de efectividad cuestionable y que además promueven una competencia poco sana.

Algunas reflexiones y conclusiones

El eterno problema de los usos e inferencias inapropiados de los resultados de la evaluación de los aprendizajes de los estudiantes 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 de la sociedad en su conjunto. Uno de los efectos negativos más frecuentes de los exámenes es afirmar y diseminar conclusiones de los resultados que no son congruentes con los objetivos iniciales del mismo, por lo que dichas conclusiones carecen de validez. Con facilidad, las declaraciones breves y sensacionalistas se propagan en los medios de comunicación, generando malentendidos y distorsión sobre las conclusiones, limitaciones e implicaciones reales de los exámenes.

La comprensión clara del concepto moderno de validez es fundamental para entender las limitaciones de los resultados de los exámenes, 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.

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 exámenes 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 conlleva una enorme responsabilidad de las autoridades académicas e institucionales.

Los instrumentos de evaluación y el uso que se hace de ellos en las universidades y otras instituciones son la declaración pública más importante de "lo que realmente cuenta" para la institución. Los estudiantes están muy alertas a estas señales, que a veces son sutiles y en ocasiones explícitas y visibles, sobre lo que deben aprender y cómo lo deben aprender, por lo que las instancias evaluadoras deben hacer lo posible para que estos procedimientos de evaluación se realicen con profesionalismo educativo en un entorno de calidad y atención a las facetas humanas y sociales de los estudiantes. Al final del día, el uso de la puntuación de un examen definitivamente implica consecuencias; de otra manera "uso" es sólo una abstracción. Los exámenes han adquirido un enorme grado de sofisticación técnica y metodológica, y llegaron para quedarse. Tal vez lo más importante es encontrar un balance entre este tipo de evaluación

y la evaluación formativa. Por otra parte, es relevante tener conciencia de que aún existen grandes retos para evaluar de forma adecuada varios atributos fundamentales de los profesionistas que requiere la sociedad moderna, como empatía, liderazgo, asertividad, creatividad, trabajo en equipo, entre otros muchos, por lo que el campo de estudio de la evaluación educativa debe seguir modernizándose para enfrentar los constantes cambios de nuestra sociedad.

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, como una obligación ética y moral de todos los docentes, e informar al resto de la sociedad sobre las virtudes, alcances y limitaciones de este fascinante y controversial tema.

Referencias

- American Educational Research Association (AERA), American Psychological Assocation (APA) y National Council on Measurement in Education (NCME) (2014). Standards for educational and psychological testing. Washington, DC: AERA.
- Boursicot, K., Etheridge, L., Setna, Z., Sturrock, A., Ker, J., Smee, S. y Sambandam, E. (2011). Performance in assessment: consensus statement and recommendations from the Ottawa conference. *Med Teach, 33*(5), 370-83. DOI: https://doi.org/10.3109/0142159X.2011.565831.
- Brennan, R. L. (2006). Perspective on the Evolution and Future of Educational Measurement. En Brennan, R. L., (ed.), *Educational Measurement. National Council on Measurement in Education and American Council on Education* (4a ed., pp. 1-16). Westport, CT: Praeger Publishers.
- Carter, K. (1984). Do teachers understand principles for writing tests? *Journal of Teacher Education, 35*(6), 57-60.
- Downing, S. M. (2003). Validity: on the meaningful interpretation of assessment data. Med Educ., 37, 830-837.
- Downing, S. M. y Haladyna, T. M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ., 38*, 327-333.
- 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: HTTPS://DOI.ORG/10.1080/0309877X.2014.984600.
- Márquez Jiménez, A. (2014). Las pruebas estandarizadas en entredicho. *Perfiles Educativos, 36*(144), 3-9. Recuperado de: http://www.redalyc.org/pdf/132/13230751001.pdf.

- Martínez Rizo, F. (2009). 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. (2013). 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.
- Mendoza Ramos, A. (2015). La validez en los exámenes de alto impacto: un enfoque desde la lógica argumentativa. *Perfiles Educativos*, 37(149), 169-186. Recuperado de: http://www.scielo.org.mx/pdf/peredu/v37n149/v37n149a10.pdf.
- Miller, M. D., Linn, R. L. y Gronlund, N. E. (2012). *Measurement and Assessment in Teaching* (11a ed.). USA: 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.
- Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., ... Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Med Teach., 33*(3), 206-14.
- 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?* London: Kogan Page.
- Sánchez-Mendiola, M., Delgado-Maldonado, L. (2017). Exámenes de alto impacto: Implicaciones educativas. *Inv Ed Med., 6*(21), 52-62. DOI: http://dx.doi.org/10.1016/j.riem.2016.12.001
- Sánchez Mendiola, M., Delgado Maldonado, L., Flores Hernández, F., Leenen, I., Martínez González, A. (2015). Evaluación del aprendizaje. En Sánchez Mendiola, M., Lifshitz Guinzberg, A., Vilar Puig, P., Martínez González, A., Varela Ruiz, M., Graue Wiechers, E. (Eds.), *Educación Médica: Teoría y Práctica* (cap. 14, pp. 89-95). México: Elsevier.

Cómo citar este artículo

- Sánchez Mendiola, Melchor (2018). La evaluación del aprendizaje de los estudiantes: ¿es realmente tan complicada? *Revista Digital Universitaria* (RDU). Vol. 19, núm. 6 noviembre-diciembre. DOI: http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a1.