# Capítulo 5

# EVALUACIÓN SUMATIVA Y EXÁMENES DE ALTO IMPACTO

Melchor Sánchez Mendiola, Laura Delgado Maldonado

*"…el uso de la puntuación de un examen definitivamente implica consecuencias; de otra manera uso es solo una abstracción."*
Robert L. Brennan

*"No todo lo que puede ser contado cuenta,*
*y no todo lo que cuenta puede ser contado."*
Albert Einstein

## EVALUACIÓN SUMATIVA

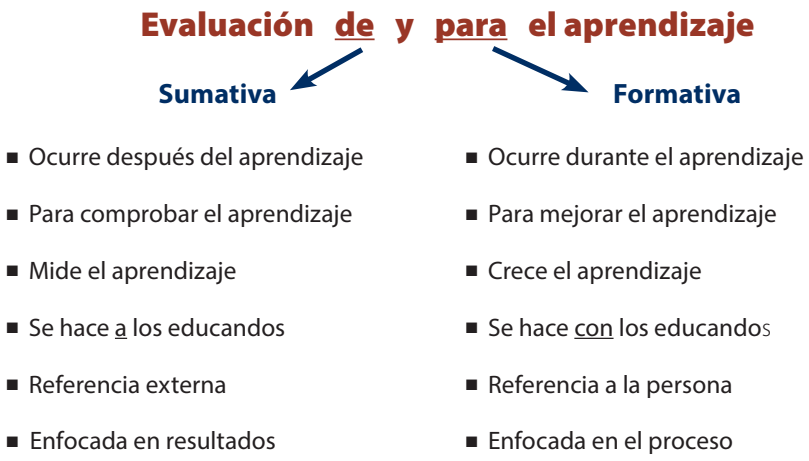
Como se ha descrito en otros capítulos de este libro, la evaluación puede clasificarse por su objetivo en diagnóstica, formativa y sumativa (capítulo 1). En este capítulo se describirán los principales conceptos de evaluación sumativa y de exámenes de alto impacto (EAI), también llamados de altas consecuencias (AERA, APA y NCME, 2014; INEE, 2017). La **evaluación sumativa** es aquella que se utiliza para determinar el aprendizaje del estudiante, adquisición de habilidades y logro académico al concluir una actividad instruccional específica, que puede ser el final de una unidad, curso, programa, año académico o semestre (Bennett, 2015; Miller et al, 2013; UNESCO, 2021). Este tipo de evaluación está compuesto por la suma de valoraciones efectuadas durante un curso, para determinar al final del mismo el nivel con que los objetivos de la enseñanza se alcanzaron y se utiliza para asignar calificaciones. Generalmente se define con tres criterios principales:

- Se determina a través de exámenes, pruebas, tareas o proyectos, que se usan para establecer que los estudiantes han aprendido, lo que deberían haber aprendido de acuerdo a las metas del curso y del plan de estudios.
- Se realizan al concluir periodos instruccionales específicos, y cubren el temario de ese periodo de tiempo.
- Se registran como puntuaciones o calificaciones que se incluyen en el registro académico del estudiante (que puede ser individual por parte del profesor o institucional por la universidad).

Lo que hace a una evaluación "sumativa" no es el diseño del examen o su contenido, sino la forma en que es utilizada para documentar cuánto han aprendido los estudiantes, para efec-

tos de calificación parcial, final, promoción, certificación, admisión o graduación (Miller et al, 2013). Este tipo de evaluación forma parte de la llamada "evaluación **del** aprendizaje", en contraste con la evaluación formativa (**para** el aprendizaje), que se describe en otro capítulo de esta obra. Una forma elegante e ingeniosa para entender los dos tipos de evaluación fue descrita por Paul Black: "*…cuando el cocinero prueba la sopa, esa es evaluación formativa. Cuando el cliente prueba la sopa, esa es evaluación sumativa*" (Black y Wiliam, 1998). En la Figura 1 se mencionan algunas de las diferencias entre estos dos tipos de evaluación.

**Figura 1. Características de la evaluación sumativa y formativa**




Desafortunadamente, con el paso del tiempo, se ha creado una separación artificial entre la evaluación sumativa y formativa, dando la impresión generalizada de que se trata de una dicotomía absoluta (Lau, 2016). A la evaluación sumativa se le ha etiquetado como meramente cuantitativa, que solo mide y no evalúa, que se centra solo en los números, que es punitiva y discriminatoria, usada con fines políticos, de ejercicio del poder o de control, y que es demasiado estandarizada para ayudar a los estudiantes en lo individual. Por el contrario, a la evaluación formativa para el aprendizaje se le describe como positiva, educativamente efectiva, 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 creó una situación similar a la frase "*cuatro patas bueno, dos patas malo*", de George Orwell en "Rebelión en la Granja", que raya en lo absurdo al plantear límites arbitrarios que fragmentan innecesariamente estos complejos conceptos y generan divisiones epistemológicas que se convierten en irreconciliables (Lau, 2016).

Los dos tipos de evaluación deben visualizarse como un continuo, en el que las evaluaciones pueden tener componentes sumativos y formativos, dependiendo del uso que se dé

a los resultados (Bennett, 2015; Lau, 2016). Por ejemplo, un examen de ingreso a la universidad tiene un fuerte componente sumativo, aunque también puede usarse con fines diagnósticos para la institución y los docentes, así como formativos si se provee la información de los resultados a los estudiantes. Por otra parte, una reunión de retroalimentación con un estudiante durante el curso puede ser primordialmente formativa, aunque si la información obtenida en la entrevista de alguna forma cuenta para la calificación final, esta evaluación adquiere una dimensión sumativa. Ambos tipos de evaluación tienen virtudes y limitaciones, lo más razonable es utilizarlas con sensatez y sensibilidad de forma complementaria, de acuerdo a cada situación específica. Ejemplos de evaluación sumativa son los exámenes parciales o finales de curso, exámenes de certificación de profesionistas, exámenes de fin de carrera, exámenes de grado de maestría o doctorado, exámenes de admisión o de colocación. Estos exámenes tienen alta trascendencia para la vida del estudiantado, quienes a veces los perciben como obstáculos a sortear para alcanzar una meta, en lugar de oportunidades para progresar e identificar el nivel de su aprendizaje o competencia en un momento dado. Un tipo de exámenes sumativos que merece atención especial, son los llamados **"exámenes de alto impacto o de altas consecuencias"** ("*high-stakes testing*" en inglés), que ameritan una descripción aparte en este capítulo debido a sus implicaciones educativas (Sánchez y Delgado, 2017).

## EXÁMENES DE ALTO IMPACTO (EAI)

Los exámenes de alto impacto o altas consecuencias tienen una larga historia en la educación superior a nivel internacional, y han contribuido en varias formas al desarrollo científico de la evaluación educativa como disciplina. A pesar de ello, el tema genera respuestas encontradas en varios sectores de la sociedad, docentes, estudiantes y profesionales de la educación, enfatizando con frecuencia sus potenciales efectos negativos. Todos los que hemos participado en procesos de ingreso y permanencia en educación superior, como profesores o estudiantes, hemos experimentado el impacto que pueden tener este tipo de exámenes. La aplicación de exámenes para ingreso a la universidad, para aprobar cursos, asignaturas y programas educativos, han sido parte integral del paisaje académico por mucho tiempo; el uso de instrumentos de medición que nos ubican en un nivel de desempeño específico, el suficiente para ser admitido en una institución, obtener una beca o tener acceso a estudios de posgrado e incentivos de diversos tipos, se han convertido en rutinas de evaluación que merecen reexaminarse a la luz de las investigaciones recientes y los efectos globales de la pandemia por COVID-19 (Cairns, 2021; Cizek, 2001; Tan et al, 2021).

Existen varias definiciones aceptadas por organizaciones internacionales sobre estos temas, a continuación, se describen algunas:

- **Examen o prueba** ("*test*" en inglés): "instrumento de evaluación que se emplea para identificar el nivel de dominio de los sustentantes sobre un constructo específico" (INEE, 2017). Los Estándares de AERA-APA-NCME lo definen como "recurso o procedimiento

en el que una muestra sistemática de una conducta del sustentante del examen en un dominio específico es obtenida y calificada utilizando un proceso estandarizado" (AERA, APA y NCME, 2014).

- **Exámenes de alto impacto (EAI) o de altas consecuencias** ("*high-stakes testing*" en inglés): "se indica cuando los resultados del instrumento tienen consecuencias importantes para las personas o las instituciones; por ejemplo, en los procesos de admisión o certificación" (INEE, 2017). En los Estándares de AERA-APA-NCME se definen como: "pruebas o exámenes cuyos resultados tienen consecuencias importantes y directas para los individuos, programas o instituciones involucrados en el examen" (AERA, APA y NCME, 2014). De acuerdo a la UNESCO: "evaluaciones con consecuencias importantes para los sustentantes, con base en su desempeño. Pasar el examen tiene beneficios importantes, como progresar al grado siguiente, obtención de un diploma o de una beca, ingresar al mercado laboral u obtener una licencia para practicar una profesión" (UNESCO, 2021).
- **Evaluaciones estandarizadas a gran escala** ("*large scale standardized testing*" en inglés): "son evaluaciones a nivel del sistema que proporcionan un resumen de los resultados de aprendizaje de un grupo de alumnos determinado, en un año académico determinado y en un número limitado de ámbitos. Suelen clasificarse como evaluaciones nacionales o transnacionales (regionales/ internacionales)" (UNESCO, 2021).

Como generalmente ocurre cuando se definen conceptos, establecer límites claros en constructos educativos y de ciencias sociales es problemático. Un examen puede ser de alto o bajo impacto dependiendo del contexto y de la percepción de la persona que presenta el examen, de quien establece inferencias de los resultados o de la misma institución que efectúa la evaluación. Un examen parcial de Estadística en una escuela de matemáticas podría ser de alto impacto si cuenta para la calificación final del estudiante y su beca depende del promedio académico, mientras que para un estudiante de filosofía que lleve Estadística como una asignatura opcional, el mismo examen puede ser de menor impacto.

## IMPLICACIONES EDUCATIVAS DE LA EVALUACIÓN SUMATIVA Y LOS EAI

La discusión de la evaluación sumativa en educación es compleja y tiene varias aristas filosóficas, políticas, económicas, sociales. En particular los aspectos implícitos de la evaluación del aprendizaje y sus efectos sobre los procesos y fines de la evaluación, generan controversia por las convicciones o sesgos personales y grupales de quienes diseñan los instrumentos, de quienes los usan, de quienes son receptores de sus consecuencias, de la sociedad en su conjunto y de diferentes grupos de interés (Cizek, 2001; Márquez, 2014; Martínez Rizo, 2009; Nichols y Berliner, 2007; Sánchez Cerón et al, 2013). Cuando se discute la evaluación sumativa y los exámenes de alto impacto, se crea una retórica intensa que puede obstaculizar la discusión y dificultar el entendimiento.

La relativa escasez de conocimiento original empírico sobre los exámenes de alto impacto y sus efectos en el currículo, métodos de enseñanza y de estudio, dificulta una discusión

balanceada y de consenso, ya que la mayoría de las publicaciones sobre estos temas son artículos de opinión o textos que no han pasado el tamiz del arbitraje por pares, o son estudios con características muy locales y de contexto que hacen difícil su generalización (Sánchez y Delgado, 2017). La mayoría de los estudios sobre este tema están publicados en el litigioso contexto de Norteamérica, por lo que sigue siendo asignatura pendiente realizar investigación original sobre evaluación educativa en el contexto nacional y local, con perspectiva global.

De cualquier forma, existe una necesidad inescapable de tomar decisiones educativas. Por ejemplo, es prácticamente imposible que el 100% de la población acceda a la universidad en todos los países del mundo, y que realice la carrera profesional que cada individuo decida por preferencia personal, por lo que es poco probable vislumbrar un futuro cercano en el que no se realicen procesos de selección que incluyan exámenes sumativos de alto impacto. La sociedad requiere ser protegida de profesionistas y especialistas poco competentes, por lo que es difícil que desaparezcan totalmente los exámenes de certificación profesional, sobre todo en algunas disciplinas como la medicina y sus especialidades (Dauphinee, 2002; Brennan, 2006).

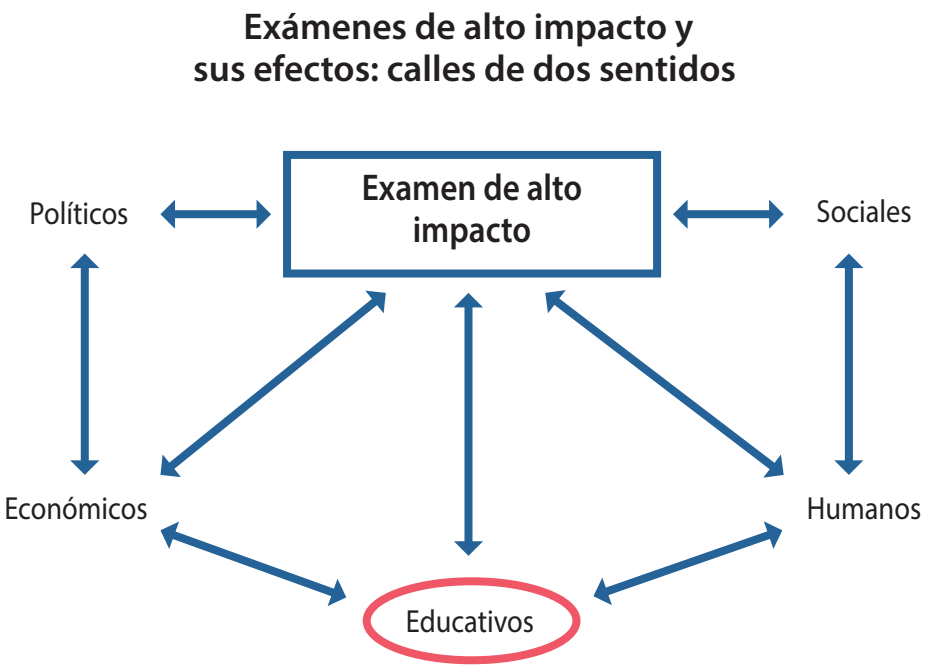
Uno de los principios de la evaluación educativa, es que no es recomendable tomar decisiones de alto impacto basadas en los resultados obtenidos con la aplicación de un solo instrumento. En este tipo de decisiones de gran trascendencia, es compromiso de las instituciones responsables realizar los exámenes de acuerdo a las buenas prácticas internacionales de evaluación educativa (AERA, APA y NCME, 2014; Brennan, 2006). Es inevitable que los métodos de evaluación utilizados en las universidades tengan efectos en estudiantes, profesores, el currículo formal y el oculto, entre otros (Brennan, 2006; Cizek, 2001; Haladyna y Downing, 2004). 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 y la cultura de su cuerpo docente, y los estudiantes están alertas a ello.

Las implicaciones educativas y efectos de los exámenes de alto impacto son particularmente complejos, ya que generan señales a veces contradictorias que pueden distorsionar el proceso educativo y las prioridades del estudiantado y del profesorado. Como se muestra en la Figura 2, los efectos de los exámenes de alto impacto y sus consecuencias en las diferentes dimensiones del entorno educativo, funcionan como una red multidireccional de complejas interacciones en las esferas política, social, económica, humana y educativa.

Un ejemplo puede ser una reforma curricular de una escuela de medicina, en la que los métodos de evaluación propuestos en el plan de estudios deben estar alineados con el currículo, los métodos de enseñanza de los profesores, los resultados educativos esperados al final de la carrera y el contexto local y nacional de atención de la salud (Martone y Sireci, 2009). Cualquier cambio importante en los métodos de enseñanza de la escuela debe transitar por el complejo camino de implementación del cambio y adopción de innovaciones, procesos más sociales que técnicos, en los que se combinan aspectos políticos de las autoridades universitarias, el profesorado y los gremios disciplinarios, las necesidades de los estudiantes y las expectativas de la sociedad. Si predominan en el currículo los exámenes sumativos de

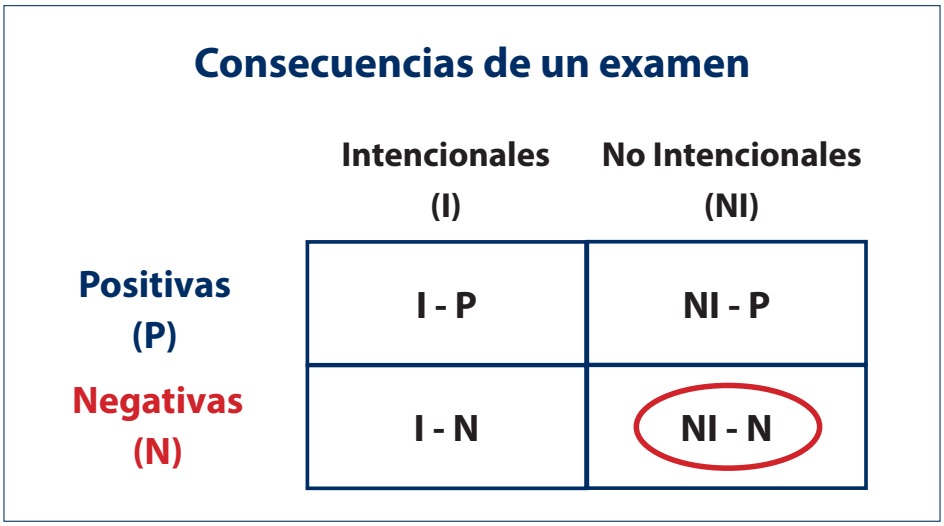
opción múltiple, aunque en el plan de estudios se declare que se enseñarán competencias genéricas, aspectos éticos, habilidades de comunicación, pensamiento crítico y creatividad, la motivación extrínseca de este tipo de exámenes y la sociología de su implementación puede influir dramáticamente en los métodos de estudio de los estudiantes, los contenidos enseñados en el aula y en el hospital, y la aparición de toda una industria de cursos para "preparar" a los estudiantes a obtener las puntuaciones más altas posibles en los exámenes. El currículo oculto se come de almuerzo al currículo formal cuando se trata de exámenes, es decir, la cultura vence a la estrategia (Newble y Jaeger, 1983; Madaus, 1988).

**Figura 2. Esquema de las diferentes áreas en las que pueden tener consecuencias los exámenes de alto impacto (elaboración propia)**




Para fines de este capítulo, describiremos algunos de los potenciales efectos educativos de los exámenes de alto impacto clasificándolos como positivos y negativos, en el entendido de que esta dicotomía es una simplificación de la realidad, y que estos efectos pueden convertirse en positivos o negativos dependiendo del contexto específico (Brennan, 2006; Sánchez y Delgado, 2017) (Figura 3).

**Figura 3. Tipos de consecuencias de los exámenes de alto impacto, de acuerdo a su intencionalidad y direccionalidad (adaptado de Brennan RL, 2006)**




## A. Potenciales efectos positivos

- **Motivación para estudiar.** Los estudiantes identifican con eficacia cuáles evaluaciones "cuentan" para la calificación final, por lo que hacen su mejor esfuerzo en los exámenes parciales y finales (Martone y Sireci, 2009; Newble y Jaeger, 1983). Generalmente el estudiantado tiene una alta motivación intrínseca para aprender lo necesario para ser profesionistas exitosos, y los docentes deben ayudarlos haciendo explícitos los criterios de evaluación a utilizar en los cursos, para que los estudiantes tengan claridad sobre cómo serán evaluados. Diversos factores motivacionales extrínsecos e intrínsecos convergen en los estudiantes para dedicar esfuerzo a estudiar para los exámenes, lo que puede contribuir a mejorar el aprendizaje de los conceptos importantes del curso (Debray et al, 2003).
- **Estandarización de la evaluación.** Este es uno de los aspectos más controversiales de los exámenes de alto impacto. Los estándares de la AERA-APA-NCME afirman que es importante realizar los exámenes sumativos en condiciones estandarizadas, en ambientes consistentes, con reglas y especificaciones predefinidas detalladas, para que los contextos en que los sustentantes presentan el examen sean similares y las inferencias que se hagan de los resultados sean válidas (AERA, APA y NCME, 2014). En la edición más reciente de estos estándares se agregó el capítulo de justicia e imparcialidad ("*fairness*"), dándole el mismo nivel de importancia que a la validez y confiabilidad de los exámenes (AERA, APA y NCME, 2014). Varios autores han criticado la estandarización de la evaluación sumativa de alto impacto, en virtud de que los estudiantes constituyen

una población heterogénea y que los individuos son demasiado complejos para que su aprendizaje pueda ser realmente evaluado por este tipo de exámenes escritos de opción múltiple; uno de los argumentos es que estas pruebas evalúan primordialmente el conocimiento, excluyendo muchos otros aspectos importantes de las habilidades humanas para la vida y la práctica profesional (Márquez, 2014; Nichols y Berliner, 2007; Moreno-Olivos, 2010).

El debate persiste, aunque el peso de la evidencia empírica sugiere que los exámenes estandarizados, si son realizados profesionalmente, con uso apropiado y juicioso de los resultados, son una de las herramientas con mayor evidencia de validez y confiabilidad para identificar de manera justa el nivel de conocimiento, capacidad de entender conceptos y resolver problemas (AERA, APA y NCME, 2014; Brennan, 2006; Norcini et al, 2011; Sackett et al, 2008).

- **Mejora de la calidad educativa.** Si se siguen 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 (AERA, APA y NCME, 2014; Martínez Rizo, 2009; Yeh, 2005). La mejora de la calidad en el sistema educativo de un país depende de una red compleja de factores gubernamentales, sociales, económicos y personales de universidades, docentes y estudiantes, de los que los EAI son solo un componente. Los esfuerzos por mejorar la calidad educativa deben tener una perspectiva sistémica e identificar estrategias que, en el contexto local, incluyan a la evaluación del aprendizaje como un elemento fundamental.
- **Aprendizaje potenciado por exámenes** (capítulo 5 de este libro). Existe abundante investigación que documenta que realizar exámenes potencia el aprendizaje, más allá de su efecto motivacional directo para estudiar. A dicho concepto se le denomina "aprendizaje potenciado por exámenes" ("*Test-enhanced learning*"), es importante incorporarlo en nuestras estrategias evaluativas (Larsen et al, 2008).
- **Unificación de criterios.** El uso de EAI también puede contribuir a establecer consensos sobre diversos componentes de los procesos educativos, como los contenidos y metas educativas, la identificación de un currículo nuclear, el tipo de instrumentos a utilizar en evaluación, entre otros (Madaus, 1988; Martone et al, 2009). Este aspecto también es controversial, y puede generar rechazo en algunos docentes con el argumento de que se limita la libertad de cátedra.
- **Consecuencias positivas no intencionales.** Gregory Cizek, reconocido investigador en evaluación educativa, realizó una revisión de la literatura sobre las consecuencias positivas no intencionales de los exámenes de alto impacto, para poner en perspectiva la gran cantidad de artículos de opinión y anécdotas que enfatizan los aspectos negativos del tema (Cizek, 2001). Identificó los siguientes efectos positivos no intencionales de los exámenes de alto impacto:

  1) Desarrollo profesional. Las actividades de educación continua de los actores de la educación y evaluación han mejorado en calidad y efectividad.

2) Acomodación. Los exámenes de alto impacto han sido un catalizador para poner más atención a los estudiantes con necesidades especiales.
3) Conocimiento sobre evaluación. Los EAI han producido una mayor consciencia de la evaluación y su importancia.
4) Colección de información. Se han fortalecido los mecanismos de colección de datos e información generados en evaluación, y ha mejorado sustancialmente su calidad. Esto ofrece oportunidades para analizar información de forma longitudinal y transversal, con cruces de variables para informar la planeación de los procesos educativos y la toma de decisiones.
5) Uso de la información. Ligado al efecto anterior, el uso de la información generada en los EAI puede ser de utilidad para docentes y universidades.
6) Sistemas de rendición de cuentas. La rendición de cuentas en educación es un fenómeno que merece especial atención, en el que los EAI han jugado un papel relevante.
7) Familiaridad de los docentes con sus disciplinas. Las etapas para elaborar un examen de alto impacto implican varios pasos sucesivos e interdependientes, como la definición del perfil de referencia, tabla de especificaciones con resultados de aprendizaje definidos, participación de expertos en los temas a evaluar, entre otros. El diálogo que resulta de la interacción de docentes con expertos en evaluación durante el diseño de los exámenes, su análisis y retroalimentación, conduce a un incremento en el conocimiento sobre los temas actuales de su disciplina y cuáles son importantes para ser evaluados.
8) Calidad de los exámenes. La creciente utilización de exámenes de alto impacto ha llevado a mayor escrutinio en su diseño y análisis, lo que ha producido un incremento en la calidad técnica de los mismos. En palabras de Cizek: "por lo menos en términos de calidad técnica, el examen típico obligatorio de alto impacto que tome un estudiante será –por mucho– la mejor evaluación que el estudiante verá en todo el año" (Cizek, 2001).

## B. Potenciales efectos negativos

La clasificación en efectos positivos y negativos de los exámenes de alto impacto no refleja necesariamente la compleja y dinámica realidad en la que un efecto puede ser benéfico o dañino dependiendo del contexto y otros factores mediadores. Sin embargo, creemos que es útil este esquema para tener en mente dichos efectos y reflexionar sobre su potencial impacto en los estudiantes.

- **Enseñando para la prueba.** Un efecto potencialmente negativo importante es el fenómeno denominado "enseñando para la prueba" ("*teaching to the test*") (Downing y Haladyna, 2004; 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 docentes e instituciones enfatizan durante las actividades de enseñanza lo que vendrá en los exámenes,

entonces el proceso se distorsiona y puede llegar al grado de enseñar principalmente lo que vendrá en los exámenes. Incluso hay casos en que docentes y escuelas enseñan a sus estudiantes con preguntas que pueden venir en los exámenes, para mejorar las puntuaciones de grupos, escuelas y los mismos profesores. El aprendizaje se centra entonces en motivaciones extrínsecas, alterando el proceso educativo. El "enseñar para los exámenes" puede incluir varias actividades, desde las sutiles e inconscientes por parte del profesor, hasta las explícitas y dirigidas principalmente a subir las puntuaciones en los exámenes.

- **Cursos de preparación para exámenes.** Ante lo importante de las consecuencias de los EAI, han aparecido una gran cantidad de cursos, libros y plataformas digitales para mejorar las puntuaciones en los exámenes. Esto se ha convertido en un lucrativo negocio en México y el resto del mundo, aprovechando la necesidad de los aspirantes de aumentar sus posibilidades de aprobar y mejorar sus puntuaciones. En Estados Unidos, McGaghie y colaboradores realizaron 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 reportó ingresos por más de 250 millones de dólares (McGaghie et al, 2004). 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.
- **Efectos en los currículos formal y oculto.** Existe controversia sobre el impacto de los exámenes de alta consecuencia en el currículo formal, vivido y oculto de las instituciones educativas (Au, 2007; Sackett et al, 2008). Las revisiones sistemáticas sobre el tema documentan que menos del 5% de la literatura publicada incluye datos empíricos, con metodología de investigación rigurosa, lo que hace difícil tener conclusiones contundentes. Tradicionalmente, se dice que las evaluaciones de alto impacto tienen influencia importante en el currículo, los métodos de enseñanza del profesorado y las estrategias de estudio de los estudiantes. Existe la percepción de que los graduados de las universidades tienen deficiencias en las habilidades necesarias para tener éxito en el mercado laboral, y que dedican demasiado tiempo y esfuerzo a contestar exámenes. Identificamos dos revisiones sistemáticas del tema, una de la Universidad de Michigan, EUA (Mehrens, 1998) y una meta-síntesis cualitativa de la Universidad de California (Au, 2007). Mehrens afirma que la totalidad de la evidencia no es clara y que depende del nivel del impacto o consecuencia del examen específico, y que no se ha logrado demostrar de manera contundente que los EAI influyan sustancialmente en el currículo, por lo menos en los trabajos de investigación cuantitativa, sin embargo, la comunidad docente persiste en la creencia de que los exámenes influyen en los planes de estudio, métodos de enseñanza y estrategias de estudio de los alumnos (Mehrens, 1998). En la meta-síntesis cualitativa de Au, se analizaron 49 estudios cualitativos sobre cómo los EAI 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 encontró que las áreas de

conocimiento de los contenidos educativos se fragmentaban en piezas relacionadas con los exámenes, y que los docentes incrementaban el uso de estrategias pedagógicas centradas en el profesor, como la instrucción directa con conferencias y menor interactividad. Curiosamente, en una minoría de los estudios revisados por Au, algunos EAI tuvieron efectos positivos en las tres dimensiones citadas, con expansión del currículo, integración del conocimiento y estrategias de enseñanza centradas en el estudiante, por lo que el análisis sugiere que la naturaleza del control curricular inducido por los exámenes de alto impacto es dependiente de la estructura de los exámenes y del contexto (Au, 2007).

- **Inferencias inapropiadas de los resultados de los exámenes.** 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 (AERA, APA y NCME, 2014; Mendoza, 2015; Sánchez y Delgado, 2017). La elaboración e implementación de exámenes de alto impacto requiere gran inversión de recursos humanos y materiales, y el público usuario de la información con frecuencia no posee una cultura de evaluación suficiente para aplicar los conceptos de validez y confiabilidad. Con facilidad declaraciones breves y sensacionalistas en los medios de comunicación, generan malentendidos y distorsión de las conclusiones, limitaciones e implicaciones reales de los exámenes.

  El concepto moderno de validez es descrito en otro capítulo de este libro, como un modelo holístico en el que toda la validez es validez de constructo que se alimenta de diferentes fuentes, y que requiere de una cadena argumentativa para realizar inferencias apropiadas de los resultados; dicho concepto aún no ha permeado en la totalidad de la comunidad académica (AERA, APA y NCME, 2014; Downing y Haladyna, 2004). La comprensión de la validez es fundamental para entender las limitaciones de los resultados de los EAI, ya que extrapolar conclusiones y decisiones más allá de lo técnicamente correcto es inapropiado y puede ser peligroso y causar daño. Si un estudiante tiene un desempeño deficiente en un examen sumativo de alto impacto (como el examen profesional al graduarse de la carrera), eso no significa que sea una "mala persona", "incompetente", entre otros adjetivos que se asignan como etiquetas y que tienen un impacto emocional importante en los sustentantes. Una recomendación importante en evaluación educativa 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 elaborar buenos instrumentos e informar a la sociedad sobre sus limitaciones recae en las instituciones responsables del examen.

## EVALUACIÓN SUMATIVA Y EXÁMENES DE ALTO IMPACTO EN LA ERA PANDÉMICA

El impacto brutal de la pandemia por COVID-19 en la educación superior global ha sido particularmente fuerte en la evaluación sumativa y los exámenes de alto impacto, ya que muchos de ellos se tuvieron que suspender temporal o definitivamente en el año 2020, y, por otra parte, varios migraron a la modalidad de exámenes en línea (Cairns, 2021; Clark et al,

2020; Isbell y Kremmel, 2020; Luna-Bazaldua et al, 2020; UNESCO, 2020). Como ejemplo de lo profundo de la disrupción causada por la pandemia, uno de los exámenes más importantes en educación médica en Estados Unidos, el Paso 2 de Habilidades Clínicas del *United States Medical Licensing Examination* (Step 2 CS USMLE), examen clínico objetivo estructurado con estaciones estandarizadas y pacientes simulados para evaluar la competencia clínica, que es requisito para ejercer la profesión, así como para ingresar a una especialidad en EUA, fue temporalmente cancelado en un inicio y después suspendido definitivamente (https://www.usmle.org/work-relaunch-usmle-step-2-cs-discontinued), al no ser posible efectuarlo en las condiciones sanitarias vigentes.

Otro ejemplo espectacular ha sido la decisión de algunas de las universidades más famosas de los Estados Unidos como Harvard, Yale y la Universidad de California, de eliminar como requisito de admisión puntuaciones elevadas en los exámenes estandarizados como el SAT o el ACT (https://www.cnbc.com/2020/06/17/7-ivy-league-schools-will-not-require-sats-or-acts-next-year.html), como resultado de un prolongado debate, batallas legales y hallazgos de investigación que, aunado a los efectos de la pandemia en los estudiantes, parece anunciar un cambio profundo del paradigma del uso tradicional de los resultados de este tipo de EAI. Muchas organizaciones, que estaban en una relativa zona de comodidad en sus procesos de aplicación de exámenes estandarizados a gran escala, han tenido que repensar profundamente sus premisas, para iniciar una transformación de sus procesos y planes a corto y largo plazo, con el fin de enfrentar con éxito la nueva realidad trans y post-pandémica, así como aprovechar las herramientas de evaluación en línea que han madurado significativamente en los últimos años (Cairns, 2021; Clark et al, 2020; Isbell y Kremmel, 2020; Luna-Bazaldua et al, 2020; Tan et al, 2021; UNESCO, 2020).

Los retos de las brechas digitales y las dificultades para vigilar la administración de exámenes en línea en contextos menos controlados (como los exámenes en casa vigilados por herramientas digitales de video y algoritmos de inteligencia artificial) han generado un ambiente de rápidos avances y complejidades éticas y tecnológicas que no se han resuelto del todo; como se menciona en un documento de la UNESCO sobre el tema: "…los exámenes en línea solo se deben considerar cuando se hayan examinado exhaustivamente los temas de acceso igualitario a la infraestructura y la conectividad, la seguridad y los métodos de supervisión en línea, la transparencia, y las habilidades y brechas digitales de los estudiantes" (UNESCO, 2020). Luna-Bazaldúa y colaboradores recomiendan tener en mente los siguientes cinco puntos al mover los EAI a la modalidad en línea (Luna-Bazaldua et al, 2020):

1) No todo el estudiantado tiene acceso en casa a dispositivos y conexión a Internet adecuados para realizar un examen en línea, además de la situación personal de cada familia y su vivienda, que puede complicar responder adecuadamente este tipo de exámenes.
2) Pueden ocurrir problemas con el software, la electricidad y otras situaciones imprevistas. Se requiere tiempo y esfuerzo para probar la compatibilidad de los dispositivos con las plataformas de exámenes, así como apoyo técnico eficiente en tiempo real antes y durante el examen.

3) La vigilancia a distancia ("*remote proctoring*") en EAI con seres humanos en tiempo real, con programas de inteligencia artificial, o con grabaciones de video y revisión posterior de las mismas, tiene una gran cantidad de implicaciones éticas, legales, de invasión de la privacidad y manejo de datos personales, que no han sido resueltas en su totalidad (Coghlan et al., 2021).
4) No es suficiente convertir un examen previamente administrado en papel a un formato en línea, sin considerar todos los aspectos que implica la aplicación y análisis de un EAI (estudios piloto de los ítems, validez de contenido y de proceso de respuesta, comportamiento psicométrico de los reactivos, etc.) Debe tomarse en cuenta el formato en línea del examen durante todo el proceso de planeación e implementación.
5) Los EAI en línea deben seguir los principios de diseño universal, tarea nada sencilla en países como México, para permitir que el estudiantado con discapacidades tenga igualdad de oportunidades para demostrar sus habilidades.

La era moderna nos conduce a una oportunidad sin precedentes para mejorar nuestras estrategias de evaluación, y adecuar la evaluación sumativa y los exámenes de alto impacto a condiciones difíciles, dinámicas y complejas (Tan et al, 2021; Fuller et al, 2020).

## CONCLUSIONES

Es imperativo desarrollar y aplicar instrumentos de evaluación siguiendo los principios fundamentales de evaluación educativa, utilizando las guías de cómo hacer exámenes escritos, basados en la mejor evidencia educativa disponible, para optimizar los efectos de los EAI en los individuos (estudiantado y profesorado), las instituciones, la sociedad y el desarrollo de los países. La posibilidad de afectar a los estudiantes por los resultados obtenidos en este tipo de evaluaciones es real y se ha convertido en un verdadero problema técnico, ético y de equidad. Un examen que no sigue los principios técnicos de diseño educativo puede ocasionar que aprueben estudiantes que no debieran pasar, y que estudiantes que merezcan aprobar no lo hagan, por lo que los beneficiarios de la profesionalización en evaluación educativa son múltiples: sustentantes, instituciones, docentes y la sociedad.

Debemos ampliar nuestros horizontes ante los retos de evaluaciones más auténticas y relacionadas al desempeño. Durning y colaboradores recomendaron recientemente implementar los conceptos de las ciencias de la complejidad y métodos no lineales en evaluación educativa, utilizando métodos innovadores para lograr la meta de evaluación basada en el trabajo y en el desempeño (Durning et al, 2015).

A continuación, anotamos algunas conclusiones:

- Los exámenes de alto impacto han contribuido al desarrollo de las ciencias de la evaluación.

- Estos exámenes tienen efectos positivos y negativos en los participantes de la educación superior, que son de naturaleza compleja y variable, dependiendo del contexto y de la naturaleza del examen.
- Las decisiones en los procesos de selección y de promoción tienen que tomarse, y es importante conocer las virtudes y limitaciones de los exámenes de alto impacto para incorporar la información generada en la toma sensata de estas ineludibles decisiones.
- La mayoría de las publicaciones sobre exámenes de alto impacto son opiniones, anécdotas o datos obtenidos sin metodología apropiada, por lo que es indispensable dedicar esfuerzos de investigación en esta temática, a nivel local y global.
- Los exámenes escritos tradicionales tienen utilidad limitada para explorar algunas de las habilidades necesarias para la vida (curiosidad, creatividad, empatía, compasión, resiliencia, entre otras) por lo que es necesario diseñar estrategias de enseñanza e instrumentos de evaluación apropiados para ello.
- Es necesario continuar innovando en el diseño, desarrollo y análisis de los exámenes de alto impacto, integrar la evaluación sumativa del aprendizaje con la evaluación formativa para el aprendizaje, para mejorar la calidad de la educación.

## REFERENCIAS


American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (Eds.). (2014). Standards for educational and psychological testing. American Educational Research Association. https://www.testingstandards.net/open-access-files.html

Au W. (2007). High-Stakes. testing and curricular control: a qualitative metasynthesis. *Educational Researcher. 36*(5):258-267. https://journals.sagepub.com/doi/10.3102/0013189X07306523

Bennett, R. E. (2015). The Changing Nature of Educational Assessment. *Review of Research in Education*, *39*(1), 370–407. https://doi.org/10.3102/0091732X14554179

Black, P., Wiliam, D. (1998) Assessment and Classroom Learning, Assessment in Education: Principles, Policy & Practice, *5*(1), 7-74, DOI: 10.1080/0969595980050102

Brennan, R.L. (2006). Perspective on the Evolution and Future of Educational Measurement. En: Brennan, R.L., Ed. Educational Measurement. National Council on Measurement en Education and American Council on Education. 4th Ed. Westport, CT: Praeger Publishers, pág. 1-16.

Cairns, R. (2021). Exams tested by Covid-19: An opportunity to rethink standardized senior secondary examinations. *Prospects 51*(1-3), 331–345. https://doi.org/10.1007/s11125-020-09515-9

Cizek, G.J. (2001). More unintended consequences of high-stakes testing. *Educational Measurement: Issues and Practice. 20*(4):19-27.

Clark, T.M., Christopher S. Callam, Noel M. Paul, Matthew W. Stoltzfus, and Daniel Turner Testing in the Time of COVID-19: A Sudden Transition to Unproctored Online Exams *J. Chem. Educ.* 2020, *97*(9), 3413–3417. https://doi.org/10.1021/acs.jchemed.0c00546

Coghlan, S., Miller, T. & Paterson, J. (2021). Good Proctor or "Big Brother"? Ethics of Online Exam Supervision Technologies. *Philos Technol*. 34, 1581–1606. https://doi.org/10.1007/s13347-021-00476-1

Dauphinee, W.D. (2002). Licensure and Certification. En: Norman, G.R., van der Vleuten, C.P.M., Newble, D.I. *International Handbook of Research in Medical Education*. Series: Springer International Handbooks of Education, *7*, 835-882.

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

Downing, S.M., Haladyna, T.M. (2004). Validity threats: overcoming interference with proposed interpretations of assessment data. *Med Educ*. *38*(3):327-33. https://doi.org/10.1046/j.1365-2923.2004.01777.x

Durning, S.J., Lubarsky, S., Torre, D., Dory, V., Holmboe, E. (2015). Considering "nonlinearity" across the continuum in medical education assessment: supporting theory, practice, and future research directions. *J Contin Educ Health Prof*. *35*(3):232-243. https://doi.org/10.1002/chp.21298

Fuller, F., Joynes, V., Cooper, J., Boursicot, K., & Roberts, T. (2020) Could COVID-19 be our 'There is no alternative' (TINA) opportunity to enhance assessment? *Medical Teacher*, *42*(7), 781-786, DOI: 10.1080/0142159X.2020.1779206

Haladyna, T.M., Downing, S.M. (2004). Construct-irrelevant variance in high-stakes testing. *Educational Measurement: Issues and Practice*. *23*(1):17-27. https://doi.org/10.1111/j.1745-3992.2004.tb00149.x

Instituto Nacional para la Evaluación de la Educación. Criterios Técnicos para el Desarrollo y Uso de Instrumentos de Evaluación Educativa 2014-2015. INEE, México. 2017. https://www.inee.edu.mx/wp-content/uploads/2019/02/CRITERIOS_TECNICOS_PARA_EL_DESARROLLO_Y_USO_DE_INSTRUMENTOS_10_ABRIL_2014.pdf

Isbell, D. R., & Kremmel, B. (2020). Test Review: Current options in at-home language proficiency tests for making high-stakes decisions. *Language Testing*, *37*(4), 600–619. https://doi.org/10.1177/0265532220943483

Koretz, D.M., Linn, R.L., Dunbar, S.B., 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: Linn, L.R. *The Effects of High Stakes Testing, annual meeting of the American Educational Research Association and the National Council on Measurement in Education*, Chicago, IL; USA. https://dash.harvard.edu/bitstream/handle/1/10880553/The%20Effects%20of%20High-Stakes%20Testing%2023%20Dec%2002.pdf?sequence=1

Lau, A.M.S. (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

Larsen, D.P., Butler, A.C., Roediger, H.L. 3rd. (2008). Test-enhanced learning in medical education. *Med Educ*. *42*(10):959-66. https://doi.org/10.1111/j.1365-2923.2008.03124.x

Luna-Bazaldua, D., Liberman, J., and Levin, V. 2020. Moving high-stakes exams online: Five points to consider. Education for Global Development. https://blogs.worldbank.org/education/moving-high-stakes-exams-online-five-points-consider

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

Márquez Jiménez A. Las pruebas estandarizadas en entredicho. *Perfiles Educativos*. 2014; *36*(144):3-9. https://www.sciencedirect.com/sdfe/reader/pii/S0185269814706208/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). http://redie.uabc.mx/redie/article/view/231

Martone, A., Sireci, S.G. (2009), Evaluating Alignment Between Curriculum, Assessment, and Instruction. *Review of Educational Research*. *79*(4):1332–1361. https://doi.org/10.3102/0034654309341375

McGaghie, W.C., Downing, S.M., Kubilius, R. (2004). What is the impact of commercial test preparation courses on medical examination performance? *Teach Learn Med. 16*(2):202-11. https://doi.org/10.1207/s15328015tlm1602_14

Mehrens, W. A. (1998). Consequences of Assessment: What is the Evidence? *Education Policy Analysis Archives*, *6*, 13. https://doi.org/10.14507/epaa.v6n13.1998

Mendoza Ramos, A. (2015). La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. *Perfiles Educativos*. *37*(149):169-186. https://doi.org/10.22201/iisue.24486167e.2015.149.53132

Miller, M.D., Linn, R.L., Gronlund, N.E. (2013). *Measurement and Assessment in Teaching*. Pearson: USA. 11th Ed.

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. https://doi.org/10.22201/iisue.20072872e.2010.2.6

Newble, D.I., Jaeger, K. (1983). The effect of assessments and examinations on the learning of medical students. *Med Educ*. *17*(3):165-71. https://doi.org/10.1111/j.1365-2923.1983.tb00657.x

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

Norcini, J., Anderson, B., Bollela, V., Burch, V., Costa, M. J., Duvivier, R., Galbraith, R., Hays, R., Kent, A., Perrott, V., & Roberts, T. (2011). Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. *Medical Teacher*, *33*(3), 206–214. https://doi.org/10.3109/0142159X.2011.551559

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

Sackett, P.R., Borneman, M.J., Connelly, B.S. (2008). High stakes testing in higher education and employment: appraising the evidence for validity and fairness. *Am Psychol*. *63*(4):215-27. https://doi.org/10.1037/0003-066X.63.4.215

Sánchez Cerón, M., del Sagrario Corte Cruz, F.M. (2013). Las evaluaciones estandarizadas: sus efectos en tres países latinoamericanos. *Revista Latinoamericana de Estudios Educativos (México). 43*(1):97-124. https://rlee.ibero.mx/index.php/rlee/article/view/285

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

Tan C, Chua W, Vu CK, *et al*. (2021). High-stakes examinations during the COVID-19 pandemic: to proceed or not to proceed, that is the question. *Postgraduate Medical Journal. 97*:427-431. https://pmj.bmj.com/content/97/1149/427

UNESCO (2021). Glossary. Paris: UNESCO. https://learningportal.iiep.unesco.org/es/glossary/e

UNESCO (2020). COVID-19. A glance of national coping strategies on high-stakes examinations and assessments. Paris: UNESCO. https://en.unesco.org/sites/default/files/unesco_review_of_high-stakes_exams_and_assessments_during_covid-19_en.pdf

Yeh, S.S. (2005). Limiting the unintended consequences of high-stakes testing. *Education Policy Analysis Archives. 13*(43). http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1577&context=coedu_pub