Inv Ed Med. 2017;6(21):52–62

Investigación en Educación Médica

http://riem.facmed.unam.mx

ARTÍCULO DE REVISIÓN

Exámenes de alto impacto: implicaciones educativas

Melchor Sánchez-Mendiola[a,*] **y Laura Delgado-Maldonado**[b]

[a] *División de Estudios de Posgrado, Facultad de Medicina, Universidad Nacional Autónoma de México, Ciudad de México, México*
[b] *Dirección General de Medición y Tratamiento de Datos, Instituto Nacional para la Evaluación de la Educación, Ciudad de México, México*

Recibido el 8 de septiembre de 2016; aceptado el 13 de diciembre de 2016

PALABRAS CLAVE
Evaluación sumativa;
Evaluación del aprendizaje;
Medición educativa;
Exámenes;
Calidad de la educación;
México

Resumen

Introducción: Los exámenes de alto impacto o altas consecuencias tienen una larga historia en la educación superior y han contribuido al desarrollo científico de la evaluación educativa como una disciplina sofisticada. A pesar de ello, han surgido reacciones encontradas sobre el tema en diversos sectores de la sociedad y los profesionales de la educación, cuestionando su valor real y enfatizando sus potenciales efectos negativos. Es necesaria una discusión balanceada de esta temática, fundamentada en argumentos académicos con sustento en investigación, específicamente en educación médica.

Objetivo: Proveer un panorama de las implicaciones educativas de la evaluación sumativa con exámenes de alto impacto, con énfasis en la educación médica.

Método: Revisión narrativa de la literatura. Se identificaron publicaciones relevantes al tema en las bases de datos disponibles de literatura académica publicada y gris, sobre los exámenes de alto impacto en educación superior en niveles internacional y nacional. Se enfocó en artículos académicos que reportaran aspectos metodológicos y resultados, principalmente en evaluación de educación médica.

Discusión: Los exámenes de alto impacto han tenido en general efectos positivos en la educación, aunque también se han reportado efectos negativos importantes y sobre los cuales siempre se debe reflexionar. Existe abundante literatura sobre el tema, pero más del 95% no son trabajos formales de investigación, lo que hace difícil tener una discusión razonable usando argumentos con sustento metodológico. La mayoría de los estudios sobre este tema están publicados en el litigioso contexto de Norteamérica, por lo que es necesario realizar investigación original sobre evaluación educativa en el contexto nacional y local, sin perder la perspectiva global.

* Autor para correspondencia. Coordinación de Desarrollo Educativo e Innovación Curricular, UNAM, Edif. De los Consejos Académicos de Área, Circuito Exterior, C.U. Del. Coyoacán, Ciudad de México, 04510, México. Teléfono: (5255) 5622-1509.
Correo electrónico: melchorsm@unam.mx (M. Sánchez-Mendiola).
La revisión por pares es responsabilidad de la Universidad Nacional Autónoma de México.

http://dx.doi.org/10.1016/j.riem.2016.12.001

Conclusión: Los exámenes de alto impacto tienen efectos positivos y negativos en el currículo, los métodos de enseñanza y las estrategias de aprendizaje. Es necesario hacer un uso prudente y profesional de los resultados de estos exámenes, incorporando el concepto moderno interpretativo de validez para obtener inferencias apropiadas de estos datos.

KEYWORDS
Summative assessment;
Learning assessment;
Educational measurement;
Tests;
Quality of education;
México

High-stakes testing: Educational implications

Abstract
Introduction: High-stakes test have a long history in higher education and have contributed to the scientific development of educational assessment as a sophisticated discipline. Despite this, controversial reactions have emerged in different sectors of society and education professionals, questioning its real value and emphasizing its potential negative effects. A balanced discussion of these issues is needed, grounded in academic arguments about the subject, specifically in medical education.
Objective: To provide an overview of the educational implications of summative high-stakes testing, with emphasis on medical education.
Method: Narrative review of the literature. Relevant papers were identified in available databases of published and grey literature, about high-stakes testing in higher education at the international and national levels. The focus was on scholarly papers that reported methodology and results, emphasizing medical education assessment.
Discussion: High-stakes testing has had positive effects on education globally, although important negative effects have also been reported. There is abundant literature on the subject, although more than 95% are not formal published research papers, which makes it difficult to have a reasonable discussion with methodologically grounded academic arguments. The majority of studies about this theme have been published in the litigious Northamerican context, it is necessary to perform educational assessment original research in the national and local contexts, without losing the global perspective.
Conclusion: High-stakes testing have positive and negative effects on the curriculum, teaching methods and learning strategies. There is a need to use professionally and sensibly the results of educational testing, incorporating the modern interpretative concept of validity to obtain appropriate inferences of testing data.

«. . .el uso de la puntuación de un examen definitivamente implica consecuencias; de otra manera uso es solo una abstracción».

Robert L. Brennan. Educational Measurement.

National Council on Measurement in Education. 4.th Ed. 2006.

«No todo lo que puede ser contado cuenta,

y no todo lo que cuenta puede ser contado».

Albert Einstein

Introducción

¿Qué sentimientos despiertan en usted (como docente y estudiante) los siguientes escenarios?

- Una aspirante a entrar en la carrera de medicina logró ingresar en la universidad. Le fue muy bien en el examen de admisión, porque estudió más de un año preparándose en cursos para el examen, que tuvieron un alto costo económico para su familia.
- Un estudiante de primer año de medicina no pasó el examen final de Anatomía, por lo que tendrá que repetir el año. Desarrolló una profunda depresión, que le impide estudiar y ha afectado su vida personal.
- Un médico interno concursó para una beca en el extranjero, pero no fue aceptado porque su puntaje en un examen estandarizado de inglés fue menor que el solicitado por la universidad en la que quería estudiar.
- Una estudiante de medicina no aprobó el examen profesional para titularse como médica, por lo que deberá dedicar un año extra a prepararse para presentar de nuevo el examen. Mientras tanto, no puede ejercer la medicina y busca algún otro trabajo para sostener sus estudios.
- Una médica terminó la carrera hace tres años, y no ha podido ingresar en la especialidad de Pediatría en México

porque su puntuación en el Examen Nacional para Aspirantes a Residencias Médicas ha sido inferior a la requerida. Solo ha encontrado trabajo en un consultorio anexo a una farmacia, en el que recibe un salario bajo y se le presiona para prescribir determinados medicamentos.

- Un especialista en cirugía general presentó el examen del consejo de certificación nacional, sin aprobarlo. A pesar de haber aprobado el curso de su especialidad en un hospital y universidades reconocidos, por no tener la certificación no puede ser contratado en los hospitales del sistema nacional de salud.

Prácticamente todos los que hemos participado en los procesos de ingreso y permanencia en la educación superior, ya sea como profesor o como estudiante, hemos sido testigos del tremendo impacto que pueden tener en nuestras vidas los resultados de los exámenes de alto impacto (EAI). Desde el ingreso a la educación superior, y durante el transcurso de los cursos, asignaturas y rotaciones en la escuela, facultad o institución hospitalaria, somos bombardeados con una serie de instrumentos de medición que pretenden ubicarnos en un nivel de desempeño específico, el suficiente para acreditar un curso, ser admitido en una institución, obtener una beca o tener acceso a estudios de posgrado e incentivos de diversos tipos.

Con frecuencia escuchamos quejas de los estudiantes, docentes y la sociedad sobre este tipo de evaluación sumativa. Estos reclamos generalmente surgen de personas que no lograron obtener una calificación aprobatoria o puntuación suficiente para lograr la meta de admisión, certificación o premio, y que pertenecen a una sociedad poco conocedora de las fortalezas y debilidades de las técnicas de evaluación del aprendizaje. La disciplina de la evaluación educativa ha adquirido en las últimas décadas una profunda sofisticación técnica, a la par que se ha constituido en un área del conocimiento con sus propias líneas de investigación, revistas científicas con arbitraje por pares y asociaciones profesionales[1]. También se escuchan con frecuencia voces de docentes, académicos y activistas sobre las limitaciones y efectos deletéreos de los exámenes estandarizados masivos o de gran escala, en los estudiantes, docentes y los planes de estudio, haciendo necesaria una discusión madura y objetiva del tema, explorando las diferentes aristas de la situación[2,3].

El objetivo de este artículo es revisar algunas de las implicaciones educativas más importantes de los exámenes de alto impacto, la literatura científica que las sustenta o no, y realizar algunas reflexiones sobre la relevancia de estos conceptos en escuelas y facultades de ciencias de la salud. Para ello se realizaron búsquedas en julio-agosto de 2016, en las siguientes bases de datos: Medline, Google Scholar, Psycinfo, ERIC, Latindex, SciELO, Redalyc. Se usaron los siguientes términos en inglés y en español: *high-stakes testing*, *high-stakes assessment*, *standardized testing*, *large scale testing*, *medical education*, exámenes de alto impacto, exámenes de altas consecuencias, educación médica. Los resultados se separaron en artículos de investigación y artículos de opinión/editoriales, encontrando que más del 90% son artículos de opinión y un porcentaje bajo de investigación empírica. De los trabajos de investigación la gran mayoría están publicados en la literatura anglosajona. Los autores revisaron los documentos y analizaron la información, obteniendo las reflexiones descritas en el resto de este escrito.

¿Qué son los exámenes de alto impacto?

Uno de los principales retos cuando se intenta discutir cualquier tema educativo y de evaluación, es la selección y uso de los términos que representan los conceptos a ser analizados. El solo hecho de utilizar definiciones específicas tiene una serie de implicaciones filosóficas, epistemológicas y ontológicas que determinan los supuestos tácitos de la discusión, y que con frecuencia son blanco de los críticos que no están de acuerdo con el abordaje que se haga del tema. En particular los implícitos de la evaluación del aprendizaje y sus efectos sobre los procesos y fines de la evaluación, generan una enorme controversia por las diversas convicciones filosóficas, políticas o sesgos personales y grupales de aquellos que generan los instrumentos, de quienes los usan, de los que son receptores de sus consecuencias, de la sociedad en su conjunto y de los diferentes grupos de interés[4].

Cuando se discute la evaluación de alto impacto, se genera una intensa retórica que frecuentemente nubla la discusión y dificulta el entendimiento entre las partes. Hay una escasez de evidencia publicada convincente tanto positiva como negativa, por lo que los argumentos académicos ceden a los aspectos afectivos y de intereses gremiales, alimentados por una ubicua falta de conocimiento de la metodología moderna de elaboración de exámenes y de los conceptos actuales de validez en evaluación[4]. Por otra parte, aunado a la relativa escasez de conocimiento original empírico sobre los exámenes de alto impacto y sus efectos en el currículo, métodos de enseñanza de los docentes y de estudio de los estudiantes, hay una ausencia casi total de investigación original en educación superior sobre esta temática en países como el nuestro, lo que dificulta aún más el establecimiento de una discusión que arroje resultados contundentes o de consenso.

La necesidad inescapable de tomar decisiones existe. No hay manera a corto plazo de que la totalidad de la población acceda a los espacios de educación superior, por lo que no se vislumbra un futuro cercano en el que no se realicen procesos de selección para entrar a las escuelas de medicina y a los cursos de especialización médica. La sociedad requiere ser protegida de los médicos generales y especialistas poco competentes, por lo que los exámenes de certificación de estos profesionales no deben desaparecer, sino realizarse con mayor profesionalismo educativo[5,6].

En este documento utilizaremos las definiciones aceptadas por las organizaciones más reconocidas en el desarrollo de instrumentos de evaluación del aprendizaje, que han establecido una serie de principios y mejores prácticas para realizarlos con alta calidad[7–9]. La definición de «examen» («*test*» en inglés) que utiliza el Instituto Nacional para la Evaluación de la Educación en México es: «instrumento de evaluación que se emplea para identificar el nivel de dominio de los sustentantes sobre un constructo específico»[8], similar a la que refiere la última edición de los *Standards for Educational and Psychological Testing* de la *American Educational Research Association, American Psychological*

Association y *National Council for Measurement in Education*: «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»[7].

La definición de EAI o de altas consecuencias («*high-stakes testing*» en inglés) de acuerdo al Instituto Nacional para la Evaluación de la Educación en México es la siguiente: «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»[8], y de acuerdo a la edición 2014 de los *Standards*: «pruebas o exámenes cuyos resultados tienen consecuencias importantes y directas para los individuos, programas o instituciones involucrados en el examen»[7].

Como generalmente ocurre cuando se define un concepto, el establecer líneas rígidas y límites claros en los constructos educativos y de ciencias sociales tiene su propia problemática. Un examen puede ser de alto o bajo impacto dependiendo del contexto y de la percepción del individuo que sustenta el examen, de la persona que realiza inferencias de los resultados o de la institución en la que se lleva a cabo, por lo que es inevitable que se forme un continuo interpretativo del concepto. Un examen parcial de Embriología en la escuela de medicina podría ser de alto impacto para un estudiante, si cuenta para su promedio y de ese promedio depende que mantenga la beca que le permite continuar sus estudios, mientras que para el profesor puede ser de menor impacto ya que solo es un componente menor de la calificación que al final determinará si el estudiante acredita el curso.

Algunos ejemplos de exámenes de alto impacto en educación en ciencias de la salud

En las viñetas con las que inicia este documento se describen diversas situaciones que ejemplifican algunos exámenes de alto impacto en nuestras escuelas, facultades de medicina y hospitales.

- Los exámenes de selección como los de ingreso a las licenciaturas de medicina, odontología, veterinaria, enfermería y otras, que inevitablemente dejan fuera a una proporción sustancial de los aspirantes. Por ejemplo, en el ingreso a la carrera de médico cirujano en la Facultad de Medicina de la UNAM en la Ciudad de México, en 2016 los aspirantes admitidos por concurso de selección fueron 192 (1.36%) de un total de 14,069 estudiantes que concursaron para ingresar a esta institución[10].
- El examen de selección para realizar un curso de especialización médica. Una de los retos educativos y de recursos humanos en salud más importantes de México, es el creciente desbalance entre los aspirantes a realizar una residencia médica y los espacios disponibles para ello en los hospitales e instituciones de educación superior[11]. Los datos recientes del Examen Nacional de Aspirantes a Residencias Médicas de México hablan por sí mismos: en el año 2016 para 7,948 plazas concursaron 35,884 aspirantes, lo que significa que la mayoría de los médicos generales mexicanos no pueden realizar su sueño personal de ser cirujanos, pediatras o ginecólogos, en el tiempo y lugar que ellos desean[12].
- Los exámenes finales para acreditar cursos en el transcurso de la formación de los médicos, que determinan que un estudiante sea promovido al siguiente peldaño de la carrera, también son exámenes de alto impacto.
- Los exámenes de graduación como el examen profesional de las licenciaturas o de los programas de maestría y doctorado, también son considerados exámenes de altas consecuencias[13]. En este caso, además del impacto en los individuos, que si no se gradúan no pueden obtener el título y cédula profesional para incorporarse a la sociedad como profesionales independientes, también hay consecuencias para las instituciones y la sociedad. Un estudiante de medicina que no acredita el examen profesional en nuestro país, no puede realizar el servicio social obligatorio, por lo que un poblado marginado y de escasos recursos socioeconómicos puede quedar sin el único profesional de la medicina asignado a esa comunidad.
- En el caso de las especialidades médicas, los exámenes de alto impacto más relevantes son los exámenes de certificación de los Consejos de la especialidad[5,14]. Actualmente en México un especialista que no aprueba estos exámenes está cada vez más restringido para ejercer su especialidad en las instituciones de salud públicas y privadas[15].

Uno de los principios más importantes en evaluación educativa, es que no es recomendable tomar decisiones de muy alto impacto basados en el uso de un solo instrumento, lo que es muy difícil de lograr en un país con tanta heterogeneidad en la calidad de las instituciones educativas públicas y privadas, en niveles básico, medio superior y superior. Un Consejo de certificación de una especialidad médica, debe determinar la aptitud de los graduados de todas las universidades e instituciones de salud del país que egresen especialistas de esa rama del saber, y debe tomar la decisión de acreditar o no a cada uno de los sustentantes, basándose solamente en su desempeño en el examen de certificación[5,14]. Esto tiene muchas implicaciones al no tomar en cuenta su origen geográfico, el contexto de las limitaciones del hospital en que se entrenó, del tipo de pacientes a los que se enfrentó durante sus rotaciones, las diferencias en los programas educativos de las universidades mexicanas y las grandes diferencias interinstitucionales en la manera cómo se abordan diferentes enfermedades y tipos de pacientes. Como comentamos anteriormente, este tipo de decisiones son de gran trascendencia, y está en las instituciones responsables la carga de la prueba de realizar los exámenes de acuerdo a las mejores prácticas de evaluación educativa internacionales[5,6,14–16].

Implicaciones educativas de los exámenes de alto impacto

Es inevitable que los métodos de evaluación utilizados en las instituciones educativas y hospitalarias tengan una serie de efectos en los estudiantes, los profesores, el currículo formal y el oculto, entre otros[3,4,17–20]. Para algunos autores, los instrumentos de evaluación y el uso que se hace de ellos en las escuelas de medicina son la declaración pública más importante de «lo que realmente cuenta» para cada institución, y los estudiantes están muy alertas a estas

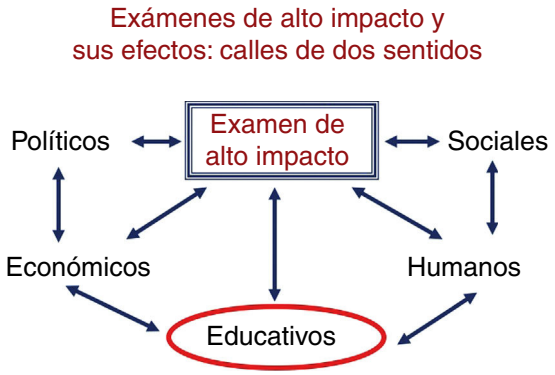

Figura 1 Esquema de las diferentes áreas en las que pueden tener consecuencias los exámenes de alto impacto.

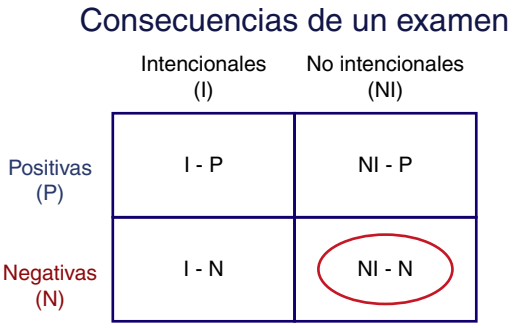

Figura 2 Tipos de consecuencias de los exámenes de alto impacto, de acuerdo a su intencionalidad y su direccionalidad. Adaptada de Brennan[1].

señales, que a veces son sutiles y en ocasiones explícitas y visibles[21–23].

En el caso de los exámenes de alto impacto, las implicaciones educativas y sus efectos son particularmente complejos, ya que puede haber señales encontradas y contradictorias que distorsionen el proceso educativo y las prioridades de estudiantes y profesores[3,17]. Como se muestra en la figura 1, los efectos de los exámenes de altas consecuencias y sus efectos en las diferentes dimensiones del entorno educativo, funcionan como una red multidireccional de complejas interacciones en las esferas política, social, económica, humana y educativa[1,4].

Como ejemplo podemos pensar en una reforma curricular importante de una escuela de medicina, en la que los métodos de evaluación sumativa y formativa propuestos en el cambio del plan de estudios idealmente debieran 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[24–26]. Cualquier cambio importante en los métodos de enseñanza y aprendizaje de una escuela de medicina debe transitar por el proceso complejo de implementación del cambio y adopción de innovaciones, que son procesos más sociales que técnicos, en los que se combinan aspectos políticos de los directivos de la escuela, los gremios de profesores y disciplinas departamentales, las exigencias de los estudiantes y las expectativas de la sociedad[26,27]. Si predominan los exámenes sumativos de opción múltiple, aunque en el plan de estudios se declare que se enseñarán competencias genéricas y específicas, profesionalismo, aspectos éticos y de 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 lunch al currículo formal cuando se trata de exámenes! (en otras palabras, cultura vence a estrategia)[26–28].

Para fines de este documento, describiremos algunos de los efectos educativos de los EAI clasificándolos como positivos y negativos, en el entendido de que esta dicotomía puede ser una simplificación de solo dos caras de una misma moneda (la evaluación), y de que estos efectos pueden convertirse en positivos o negativos dependiendo del contexto específico[4,17,19,29–31] (fig. 2).

Potenciales efectos positivos

Motivación para estudiar. Los estudiantes tienden a identificar rápidamente las evaluaciones que «cuentan» y que formarán parte de la calificación final, y por instinto de supervivencia o anhelo de destacar hacen su mejor esfuerzo en los exámenes[25,28]. Por otra parte, los estudiantes de medicina tienen una enorme motivación intrínseca para aprender lo necesario para ser buenos médicos, ya que durante su formación son expuestos a pacientes con serios problemas de salud y situaciones de vida o muerte. Además los docentes tienen la obligación de 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. En consecuencia, diversos factores motivacionales extrínsecos e intrínsecos convergen en los estudiantes para dedicar mayor esfuerzo a estudiar para los exámenes, lo que puede contribuir a mejorar el aprendizaje de los conceptos importantes del curso[24,28].

Estandarización de la evaluación. Este es uno de los aspectos más controversiales de los exámenes de alto impacto. Las recomendaciones de los estándares de la AERA-APA-NCME 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 sean similares y las inferencias que se hagan de los resultados sean válidas[7]. En la última edición de los estándares se agregó un capítulo de justicia e imparcialidad («*fairness*»), al mismo nivel de importancia que la validez y confiabilidad de los exámenes, ya que actualmente se considera que la evaluación debe ser justa y equitativa para todos los individuos, como acto de elemental justicia[7]. Diversos autores, activistas sociales y educadores han criticado este aspecto de la evaluación sumativa de alto impacto, en virtud de que los seres humanos somos diferentes, cada uno con virtudes y defectos, y que somos demasiado complejos para que nuestra «esencia» pueda ser capturada por exámenes escritos estandarizados (principalmente los de opción múltiple)[2,3,22,32,33]. Se argumenta que estos exámenes principalmente evalúan el conocimiento, y que la riqueza del ser humano consiste precisamente

en que somos mucho más que un cúmulo organizado de conocimientos[21,33,34].

El debate continúa hasta la fecha, pero el peso de la evidencia empírica sugiere fuertemente 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, de individuos y poblaciones[7,16,21,35–37].

En México, el examen estandarizado para ingresar en las residencias médicas se ha constituido en un instrumento aceptado por la sociedad, las universidades, el sistema de salud y los médicos recién graduados para poder aspirar a la especialidad de su preferencia[11,12]. A pesar de las limitaciones del examen, sin un instrumento de esta naturaleza no habría manera válida, confiable y equitativa de seleccionar de la enorme población de aspirantes a los candidatos a cursos de especialización médica en nuestro país.

Mejora de la calidad educativa. Este efecto también es controversial. La comunidad de expertos en evaluación educativa a nivel internacional sugiere que, si se siguen apropiadamente los lineamientos para realizar buenos exámenes, y se hace un esfuerzo importante por alinear la evaluación con el currículo y los métodos de enseñanza, es posible mejorar la calidad educativa[7,8,16,29,38–40]. Es importante hacer notar que la mejora de la calidad educativa 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.

En los últimos años se han publicado varios trabajos de investigación que documentan que el realizar exámenes potencia el aprendizaje, más allá de su efecto directo durante la solución del instrumento. A dicho concepto se le denomina «aprendizaje potenciado por exámenes» («*Test-enhanced learning*»), por lo que es menester incorporarlo en nuestras estrategias evaluativas y estudiarlo en nuestro contexto[41–43].

Unificación de criterios. 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 o «*core curriculum*», el tipo de instrumentos a utilizar en evaluación del aprendizaje, entre otros[7,25]. Este tipo de efectos también es controversial, y puede producir rechazo en algunos docentes con el argumento de que se limita la libertad de cátedra[44].

Consecuencias positivas no intencionales. Gregory Cizek, uno de los investigadores en evaluación educativa más reconocido a nivel internacional, realizó una extensa revisión de la literatura sobre las consecuencias positivas no intencionales de los exámenes de alto impacto, para poner en perspectiva y balancear la gran cantidad de artículos de opinión y anécdotas que enfatizan los aspectos negativos del tema[4]. Identificó los siguientes efectos positivos no intencionales de los exámenes de altas consecuencias[4]:

Desarrollo profesional – las actividades de educación continua de los actores de la educación y evaluación han mejorado en calidad y efectividad.

Acomodación – los exámenes de alto impacto han sido un catalizador para poner más atención a los estudiantes con necesidades especiales.

Conocimiento sobre evaluación – tradicionalmente los docentes han tenido un conocimiento limitado de varios elementos clave de la evaluación educativa, como validez, confiabilidad y diversos aspectos de esta disciplina[45]. Los exámenes de alto impacto han producido una mayor consciencia de la evaluación y su importancia en el proceso educativo en la sociedad y los profesores, p. ej. cada vez más profesores pueden distinguir un examen con referencia a norma vs. uno con referencia a criterio.

Colección de información – como consecuencia de los exámenes de alto impacto, se han fortalecido los mecanismos de colección de datos e información generados en evaluación, y ha mejorado sustancialmente la calidad de los mismos. Lo anterior ofrece grandes oportunidades de análisis de información longitudinal y transversal, con cruces de variables de diversos tipos para informar la planeación de los procesos educativos y la toma de decisiones.

Uso de la información – ligado al efecto anterior, el uso de la información generada en los exámenes de alto impacto puede ser de utilidad para docentes, directivos y la sociedad.

Opciones educativas – se ha incrementado la disponibilidad de datos por áreas geográficas, planteles y niveles educativos para los aspirantes a educación superior, los diferentes niveles de gobernanza educativa y la sociedad.

Sistemas de rendición de cuentas - La rendición de cuentas en educación es uno de los fenómenos recientes que merece especial atención, en el que los exámenes de alto impacto han jugado un relevante papel.

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 especificiaciones con resultados de aprendizaje definidos, participación de expertos en los temas a evaluar, entre otros[46]. El diálogo que resulta de la interacción de los docentes con expertos en evaluación durante el diseño de los exámenes, su análisis y realimentación, conduce a un incremento en la reflexión de los temas de vanguardia de su disciplina y sobre qué es importante evaluar y cómo.

Calidad de los exámenes – la creciente utilización de los exámenes de alto impacto ha llevado a mayor escrutinio en su diseño y análisis, lo que inevitablemente ha producido un incremento en la sofisticación y calidad técnica de los mismos. Existe evidencia publicada sobre la gran diferencia técnica que hay entre los exámenes realizados por los profesores sin ayuda profesional en evaluación, comparados con los exámenes a gran escala realizados por especialistas del ramo[4,45]. 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»[4]. Esto ha sido demostrado también en educación médica, los profesionales de la salud sin entrenamiento apropiado no

elaboramos exámenes de alta calidad que puedan soportar un escrutinio técnico profundo[47].

Potenciales efectos negativos

Como comentamos anteriormente, la clasificación dicotómica en efectos positivos y negativos puede no necesariamente reflejar la compleja y dinámica realidad en la que un efecto puede ser benéfico o dañino dependiendo del contexto y otros factores mediadores, pero creemos que es útil para identificar dichos efectos y analizarlos por separado.

- «Enseñando para la prueba»

  Un efecto potencialmente negativo importante es el interesante fenómeno de lo que se ha denominado «enseñando para la prueba» («*teaching to the test*»)[48–50]. 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 llegar 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[3].

  Como es natural, entonces el aprendizaje se centra en una serie de motivaciones extrínsecas a todos los niveles, que distorsiona el modelo educativo y no genera las habilidades que necesitan los estudiantes para ejercer su profesión en la sociedad moderna. El «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.
- Cursos de preparación para exámenes

  Ante lo importante de las consecuencias de no aprobar un examen de alto impacto, 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 posibilidades de aprobar y subir sus puntuaciones. En Norteamérica McGaghie et al. realizaron una revisión sistemática sobre los cursos comerciales para preparar aspirantes para los exámenes de alto impacto en educación médica de pregrado, en los que una sola empresa reporta ventas por más de 250 millones de dólares[51]. 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[51].

  En las escuelas de medicina de nuestro país, en diversas organizaciones académicas y no académicas, se realizan cursos para «ayudar» a los estudiantes a aprobar asignaturas o mejorar su puntuación en exámenes de promoción, admisión y selección, como el ingreso a la UNAM, el ingreso a las especializaciones médicas, entre otros. No existen estudios publicados de su eficacia, solo la propaganda sesgada y anecdótica de la mercadotecnia y la difusión por comunicación verbal y en redes sociales de estudiantes y familiares. Estos cursos siguen proliferando ante la necesidad de los estudiantes de salir mejor en estos exámenes, en los que hay una enorme competencia[6,10–12].
- Efectos en currículo formal y oculto

  Existe gran controversia en la comunidad de educadores sobre el impacto de los exámenes de altas consecuencias en el currículo formal, vivido y oculto de las instituciones de educación en todos sus niveles[28,36,38,52,53]. Las revisiones sistemáticas sobre esta temática documentan que menos del 5% de la literatura publicada incluye datos empíricos, con resultados obtenidos con metodología de investigación rigurosa, lo que hace difícil tener conclusiones contundentes y claras. Tradicionalmente se maneja 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. 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 de poco en el mundo real. El aumento en las puntuaciones de exámenes no necesariamente significa aumento en el aprendizaje[4].

  Identificamos dos revisiones sistemáticas formales del tema, una de Mehrens de la Universidad del Estado de Michigan, EUA[52] y una metasíntesis cualitativa de Wayne Au de California[53]. No logramos encontrar revisiones sistemáticas de Latinoamérica en las bases de datos exploradas. 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 exámenes estandarizados de alto impacto 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[52].

  En la metasíntesis cualitativa de Wayne Au, en la que se analizaron 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[53]. También encontró que las áreas de conocimiento de los contenidos educativos se fragmentaban 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. De manera interesante, en una minoría

significativa de los estudios revisados por Au, ciertos tipos de exámenes de alto impacto tuvieron 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 concluye que su análisis sugiere que la naturaleza del control curricular inducido por los exámenes de alto impacto es altamente dependiente de la estructura de los mismos exámenes[53,54].

- Inferencias inapropiadas de los resultados de los exámenes

El eterno 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 avanzar 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 el realizar inferencias de los resultados que no son congruentes con los objetivos iniciales del examen, por lo que dichas conclusiones tienen validez limitada[1,55–57]. La elaboración e implementación de exámenes de alto impacto requiere una gran inversión de recursos humanos y materiales, y el público usuario de la información con frecuencia no posee cultura de evaluación en un nivel de sofisticación suficiente para internalizar y aplicar los conceptos de validez y confiabilidad. 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[39,40].

El concepto moderno de validez[7,55–57] 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, aún no ha permeado en la totalidad de la comunidad académica de educación médica[58]. La comprensión clara de dicho concepto 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 un examen sumativo de alto impacto (como el examen profesional al graduarse de medicina), eso no significa que sea una «mala persona», «incompetente», «poco inteligente», alguien que «no debió estudiar medicina», entre otros muchos calificativos que se asignan como etiquetas y que tienen un impacto emocional importante en los sustentantes de estos exámenes. Otro ejemplo son los resultados del Examen Nacional de Aspirantes a las Residencias Médicas, en el que con frecuencia se jerarquizan a las escuelas y facultades de medicina de acuerdo a las puntuaciones logradas por sus egresados en el citado examen, y se realiza el salto conceptual inapropiado a la inferencia de que una escuela es «mejor» o «peor» dependiendo del lugar en el examen que, en promedio, tienen sus egresados[11].

Una de las recomendaciones más importantes 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...»[1,4], 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 consonancia con las autoridades y los medios de comunicación. Cuando Abraham Flexner realizó la evaluación de las escuelas de medicina de Estados Unidos y Canadá en 1910, que tuvo como consecuencia el cierre de muchas de las escuelas de mala calidad que existían más bien con fines de lucro, comentó: «El poder de examinar es el poder de destruir»[59]. La asimetría de poder intrínseca en los procesos de evaluación sumativa conlleva una enorme responsabilidad de las autoridades académicas e institucionales que participan en ellos.

Reflexiones finales y conclusiones

Es imperativo desarrollar y aplicar instrumentos de evaluación siguiendo los principios fundamentales de evaluación educativa[7,60], utilizando las guías de cómo hacer exámenes escritos[61], basados en la mejor evidencia educativa disponible[62], para optimizar los efectos de los exámenes de alto impacto en los individuos (estudiantes y profesores), las instituciones, la sociedad y el desarrollo de las naciones. De no hacerlo así, la posibilidad de afectar a los estudiantes por los resultados obtenidos en estas evaluaciones se incrementa y se convierte en un verdadero problema ético y de equidad[7]. Un examen que viole los principios técnicos de diseño educativo puede ocasionar que estudiantes que no debieran pasar aprueben, y que estudiantes que merezcan aprobar no lo hagan, por lo que el último beneficiario de la profesionalización en evaluación educativa es el sustentante del examen (aunque con frecuencia no tenga consciencia de ello)[63].

El problema de los exámenes de alto impacto en la sociedad contemporánea puede ubicarse en lo que se llama «problemas endiabladamente complejos» («*wicked problems*»), definidos por Rittel y Webber como: «una clase de problemas del sistema social que son inapropiadamente formulados, en los que hay muchos actores y tomadores de decisiones con valores conflictivos, y en los que las ramificaciones en todo el sistema son completamente confusas»[64]. Ejemplos de este tipo de problemas son la planeación urbana, el calentamiento global, los cambios curriculares y reformas educativas, entre otros. Por necesidad el abordaje de este tipo de situaciones requiere un abordaje sistémico que, en el caso de la educación médica, ha sido propuesto recientemente por Eva et al., quienes proponen que ampliemos nuestros horizontes ante los retos de evaluaciones más auténticas y relacionadas al desempeño de los profesionales de la salud[65]. Durning et al. recomendaron recientemente implementar los conceptos de las ciencias de la complejidad y métodos no lineales en evaluación educativa en medicina, utilizando métodos innovadores para lograr la meta de evaluación basada en el trabajo y en el desempeño[32].

Uno de los académicos que más ha contribuido al conocimiento de qué es lo que realmente funciona en educación es John Hattie de la Universidad de Melbourne en Australia, quien recientemente publicó un fascinante ensayo titulado «Lo que no funciona en educación: las políticas de la

distracción»[66]. El Dr. Hattie sugiere que muchas de las políticas educativas y proyectos de evaluación que se realizan en todo el mundo son «buenas intenciones desperdiciadas», ya que se dirigen gran cantidad de recursos financieros y humanos en disminuir la varianza interescuelas (motivados en parte por el gran énfasis en los «rankings» internacionales de las universidades y las comparaciones interpaíses) en lugar de la varianza intraescuela, en la que uno de los factores más importantes es la efectividad de la enseñanza[66]. Por la importancia de la educación se han generado lo que él llama «las políticas de la distracción», remedios rápidos y populares que pretenden «arreglar» el problema educativo y que con frecuencia son irrelevantes o equivocados, desviando la atención de los puntos estratégicos que podrían realmente mejorar la situación educativa. Identifica cinco políticas de la distracción: tranquilizar a los padres, «componer» a la infraestructura, a los estudiantes, a los profesores y a las escuelas[66]. Es crucial no perderse en el laberinto de estas políticas, y mejorar en lo posible la integración de los exámenes de alto impacto con los diferentes elementos del proceso educativo, identificando las estrategias potencialmente más efectivas, tarea nada sencilla. La incipiente investigación sobre los exámenes estandarizados de alto impacto en Latinoamérica debe aumentar y mejorar, de otra manera carecemos de la información necesaria para una toma de decisiones inteligente y de largo plazo[67].

A continuación, anotamos algunas reflexiones a guisa de conclusiones:

- Los exámenes de alto impacto, tal como los conocemos, han llegado para quedarse y han contribuido al desarrollo de las ciencias de la evaluación.
- Estos exámenes tienen tanto efectos positivos como negativos en los actores 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 en medicina y ciencias de la salud tienen que tomarse, es importante conocer las virtudes y limitaciones de los exámenes de alto impacto para incorporar la información generada en la toma 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, no solo a nivel global sino local.
- 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 hay que diseñar estrategias de enseñanza e instrumentos de evaluación apropiados para ello.
- Se requiere incrementar la cultura de evaluación en todas las esferas de la sociedad y en los tomadores de decisiones a nivel de política educativa.
- Es necesario continuar innovando en el diseño, desarrollo y análisis de los exámenes de alto impacto, para mejorar la calidad de la educación.

Responsabilidades éticas

Protección de personas y animales. Los autores declaran que para esta investigación no se han realizado experimentos en seres humanos ni en animales.

Confidencialidad de los datos. Los autores declaran que en este artículo no aparecen datos de pacientes.

Derecho a la privacidad y consentimiento informado. Los autores declaran que en este artículo no aparecen datos de pacientes.

Financiación

Ninguna.

Conflicto de intereses

Los autores declaran no tener conflicto de intereses.

Referencias

1. Brennan RL. Perspective on the evolution and future of educational measurement. En: Brennan RL, editor. Educational Measurement. National Council on Measurement in Education and American Council on Education. 4.th Ed. Westport, CT: Praeger Publishers; 2006. p. 1–16.
2. Márquez Jiménez A. Las pruebas estandarizadas en entredicho. Perf Educ. 2014;36:3–9.
3. Nichols SL, Berliner DC. Collateral damage: How high-stakes testing corrupts America's schools. Cambridge, MA: Harvard Education Press; 2007.
4. Cizek GJ. More unintended consequences of high-stakes testing. Educ Meas. 2001;20:19–27.
5. Sánchez Mendiola M, Delgado Maldonado L. La certificación de médicos especialistas: bases conceptuales. En: Sánchez Mendiola M, Lifshitz Guinzberg A, Vilar Puig P, Martínez González A, Varela Ruiz M, Graue Wiechers E, editores. Educación Médica: Teoría, Práctica. México, D.F.: Elsevier;2015. p. 395-399.
6. Dauphinee WD. Licensure and Certification. En: Norman GR, van der Vleuten CPM, Newble DI, editors. International Handbook of Research in Medical Education., 7. Series: Springer International Handbooks of Education; 2002. p. 835–82.
7. American Educational Research Association, American Psychological Association and National Council on Measurement in Education, and Joint Committee on Standards for Educational and Psychological Testing. Standards for educational and psychological testing. Washington, DC: AERA; 2014.
8. Instituto Nacional para la Evaluación de la Educación. Criterios técnicos para el desarrollo y uso de instrumentos de evaluación educativa 2014-2015. INEE, México. 2014 [consultado Ago 2016]. Disponible en: http://www.inee.edu.mx
9. Case SM, Swanson DB. Cómo construir preguntas de Selección Múltiple para Ciencias Básicas y Ciencias Clínicas. 3.ª ed. Philadelphia, PA, USA.: National Board of Medical Examiners; 2005 [consultado 1 Ago 2016]. Disponible en: http://www.nbme.org/publications/item-writing-manual.html
10. Dirección General de Administración Escolar, UNAM. Cómo ingreso a la UNAM, 2016-2017, pág. 41 [consultado 1 Ago 2016]. Disponible en: https://www.dgae.unam.mx/ingreso_unam/

11. Sánchez Mendiola M. El Seminario de Educación del Plan Único de Especialidades Médicas de la Facultad de Medicina UNAM: una reflexión crítica. En: Lifshitz Guinzberg A, editor. Los retos de la Educación Médica. Ciudad de México; 2012; 1(1): 135-162.
12. Comisión Interinstitucional para la Formación de Recursos Humanos para la Salud, Secretaría de Salud, Ciudad de México. 2016 [consultado 15 Ago 2016]. Disponible en: http://cifrhs.salud.gob.mx/descargas/pdf/enarm_caracteristicas_evolucion.pdf
13. Delgado Maldonado L, Sánchez Mendiola M. Análisis del Examen Profesional de la Facultad de Medicina de la UNAM: Una experiencia de evaluación objetiva del aprendizaje con la Teoría de Respuesta al Ítem. Inv Ed Med. 2012;1:130–9.
14. Porras-Hernandez JD, Mora-Fol JR, Lezama-Del Valle P, Yanowsky-Reyes G, Perez-Lorenzana H, Ortega-Salgado A, et al. Assessment of the mexican board of pediatric surgery certification system. J Surg Educ. 2015;72:829–35.
15. Comité Normativo Nacional de Consejos de Especialidades Médicas, A.C. [consultado 10 Ago 2016]. Disponible en: http://conacem.org.mx/
16. Clauser BE, Margolis MJ, Case SM. Testing for licensure and certification in the professions. En: Brennan RL, editor. Educational Measurement. National Council on Measurement in Education and American Council on Education. 4.th Ed. Westport, CT: Praeger Publishers; 2006. p. 701–31.
17. Madaus GF. The influence of testing on the curriculum. En: Tanner LN, editor. Critical issues in curriculum: Eighty-seventh year-book of the national society for the study of education. Chicago: University of Chicago Press; 1988. p. 83–121.
18. Fickel LH. Paradox of practice: Expanding and contracting curriculum in a high-stakes climate. En: Grant SG, editor. Measuring history: Cases of state-level testing across the United States. Greenwich, CT: Information Age Publishing; 2006. p. 75–103.
19. Koretz DM, Linn RL, Dunbar SB, Shepard LA. 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 RL, editor. 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 [consultado 1 Ago 2016]. Disponible en: https://dash.harvard.edu/bitstream/handle/1/10880553/The%20Effects%20of%20High-Stakes%20Testing%2023%20Dec%2002.pdf?sequence=1
20. Kuhbandner C, Aslan A, Emmerdinger K, Murayama K. Providing extrinsic reward for test performance undermines long-term memory acquisition. Front Psychol. 2016;7:79. Disponible en: https://www.ncbi.nlm.nih. gov/pmc/articles/PMC4740952/
21. Swanson DB, Norman GR, Linn RL. Performance-based assessment: lessons from the health professions. Educ Res. 1995;24, 5-11+35.
22. Moreno-Olivos T. Lo bueno, lo malo y lo feo: las muchas caras de la evaluación. Rev Iberoam Educ Sup. 2010;I:84–97.
23. Sánchez Mendiola M, Delgado Maldonado L, Flores Hernández F, Leenen I, Martínez González A. Evaluación del aprendizaje. En: Sánchez Mendiola M, Lifshitz Guinzberg A, Vilar Puig P, Martínez González A, Varela Ruiz M, Graue Wiechers E, editores. Educación Médica: teoría y práctica. Editorial Elsevier: México D.F; 2015. p. 89–95. Cap. 14.
24. Debray E, Parson G, Avila S. Internal alignment and external pressure. En: Carnoy M, Elmore R, Siskin LS, editores. The new accountability: High schools and high-stakes testing. New York: Routledge Falmer; 2003. p. 55–85.
25. Martone A, Sireci SG. Evaluating alignment between curriculum, assessment, and instruction. Rev Educ Res. 2009;79: 1332–61.
26. Bland C, Starnaman S, Wersal L, Moorehead-Rosenberg L, Zonia S, Henry R. Curricular change in medical schools: How to succeed. Acad Med. 2000;75:575–94.
27. Greenhalgh T, Robert G, Macfarlane F, Bate P, Kyriakidou O. Diffusion of innovations in service organizations: systematic review and recommendations. Milbank Q. 2004;82:581–629.
28. Newble DI, Jaeger K. The effect of assessments and examinations on the learning of medical students. Med Educ. 1983;17:165–71.
29. Yeh SS. Limiting the unintended consequences of high-stakes testing. Education Policy Analysis Archives. 2005;13 [consultado 1 Ago 2016]. Disponible en: http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=1577&context=coedu_pub
30. Sullivan D. A concept analysis of «high stakes testing». Nurse Educ. 2014;39:72–6.
31. Tagher CG, Robinson EM. Critical aspects of stress in a high-stakes testing environment: A phenomenographical approach. J Nurs Educ. 2016;55:160–3.
32. Durning SJ, Lubarsky S, Torre D, Dory V, Holmboe E. Considering ''nonlinearity'' across the continuum in medical education assessment: supporting theory, practice, and future research directions. J Contin Educ Health Prof. 2015;35:232–43.
33. Rethans JJ, Norcini JJ, Barón-Maldonado M. The relationship between competence and performance: implications for assessing practice performance. Med Educ. 2002;36:901–9.
34. Miller GE. The assessment of clinical skills/competence/performance. Acad Med. 1990;65 Suppl.:S63–7.
35. Haladyna TM, Downing SM. Construct-irrelevant variance in high-stakes testing. Educ Meas. 2004;23:17–27.
36. Sackett PR, Borneman MJ, Connelly BS. High stakes testing in higher education and employment: appraising the evidence for validity and fairness. Am Psychol. 2008;63:215–27.
37. Downing SM, Yudkowsky R. Introduction to Assessment in the Health Professions. En: Downing SM, Yudkowsky, editores. Assessment in health professions education. New York, NY: Routledge; 2009. p. 1–21.
38. Martínez Rizo F. Evaluación formativa en aula y evaluación a gran escala: hacia un sistema más equilibrado. Rev Electrón Investig Educ. 2009;11. Disponible en: http://redie.uabc.mx/redie/article/view/231
39. Organización para la Cooperación y el Desarrollo Económicos (OCDE). PISA 2015 Results (Volume I): Excellence and Equity in Education, OECD Publishing, Paris. 2016 [consultado 1 Ago 2016]. Disponible en: http://www.oecd.org/pisa/
40. Organización para la Cooperación y el Desarrollo Económicos (OCDE). PISA 2015 Results (Volume II): Policies and Practices for Successful Schools, OECD Publishing, Paris. 2016 [consultado 1 Ago 2016]. Disponible en: http://www.oecd.org/pisa/
41. McDaniel MA, Roediger HL 3rd, McDermott KB. Generalizing test-enhanced learning from the laboratory to the classroom. Psychon Bull Rev. 2007;14:200–6.
42. Larsen DP, Butler AC, Roediger HL 3rd. Test-enhanced learning in medical education. Med Educ. 2008;42:959–66.
43. Baghdady M, Carnahan H, Lam EW, Woods NN. Test-enhanced learning and its effect on comprehension and diagnostic accuracy. Med Educ. 2014;48:181–8.
44. Aguilar-Tamayo R, Sánchez-Mendiola M, Fortoul van der Goes T. La libertad de cátedra: ¿una libertad malentendida? Inv Ed Med. 2015;4:170–4.
45. Carter K. Do teachers understand principles for writing test? J Teach Educ. 1984;35:57–60.
46. Downing SM. Twelve steps for effective test development. En: Downing SM, Haladyna TM, editores. Handbook of test development. Mahwah, N.J: Lawrence Erlbaum Associates; 2006. p. 3–25.
47. Jozefowicz RF, Koeppen BM, Case S, Galbraith R, Swanson D, Glew RH. The quality of in-house medical school examinations. Acad Med. 2002;77:156–61.

48. Popham WJ. Teaching to the Test? Educational Leadership. 2001;58:16–20. Disponible en: http://www.ascd.org/publications/educational-leadership/mar01/vol58/num06/Teaching-to-the-Test%C2%A2.aspx
49. Downing SM, Haladyna TM. Validity threats: overcoming interference with proposed interpretations of assessment data. Med Educ. 2004;38:327–33.
50. Downing SM. Threats to the validity of locally developed multiple-choice tests in medical education: construct-irrelevant variance and construct underrepresentation. Adv Health Sci Educ Theory Pract. 2002;7:235–41.
51. McGaghie WC, Downing SM, Kubilius R. What is the impact of commercial test preparation courses on medical examination performance? Teach Learn Med. 2004;16:202–11.
52. Mehrens WA. Consequences of assessment: What is the evidence? Education Policy Analysis Archives. 1998;6. Disponible en: epaa.asu.edu/ojs/article/download/580/703.
53. Au W. High-Stakes testing and curricular control: a qualitative metasynthesis. Educational Researcher. 2007;36:258–67.
54. Apple MW. Education and power. 2.nd ed. New York: Routledge; 1995.
55. Downing SM. Validity: on meaningful interpretation of assessment data. Med Educ. 2003;37:830–7.
56. Kane M. Validating the Interpretations and Uses of Test Scores. J Educ Meas. 2013;50:1–73.
57. Mendoza Ramos A. La validez en los exámenes de alto impacto: Un enfoque desde la lógica argumentativa. Perf Educ. 2015;37:169–86.
58. Sánchez-Mendiola M. Mi instrumento es más válido que el tuyo»: ¿Por qué seguimos usando ideas obsoletas? Inv Ed Med. 2016;5:133–5.
59. Flexner A. Medical Education in the United Sates and Canada. Washington, DC: Science and Health Publications, Inc. 1910 [consultado 15 Ago 2016]. Disponible en: http://archive.carnegiefoundation.org/pdfs/elibrary/Carnegie_Flexner_Report.pdf
60. Norcini J, Anderson B, Bollela V, Burch V, Costa MJ, Duvivier R, et al. Criteria for good assessment: consensus statement and recommendations from the Ottawa 2010 Conference. Med Teach. 2011;33:206–14.
61. Haladyna TM, Downing SM, Rodriguez MC. A Review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15:309–34.
62. Sánchez Mendiola M. Educación médica basada en evidencias: ¿Ser o no ser? Inv Ed Med. 2012;1:82–9.
63. Downing SM. The effects of violating standard ítem writing principles on test and students: the consequences of using flawed test ítems on achievement examinations in medical education. Adv Health Sci Educ Theory Pract. 2005;10:133–43.
64. Rittel HWJ, Webber MM. Dilemmas in a general theory of planning. Policy Sciences. 1973;4:155–69.
65. Eva KW, Bordage G, Campbell C, Galbraith R, Ginsburg S, Holmboe E, et al. Towards a program of assessment for health professionals: from training into practice. Adv Health Sci Educ Theory Pract. 2016;21:897–913.
66. Hattie J. What doesn't work in education: The politics of distraction. London: Pearson; 2015 [consultado 1 Ago 2016]. Disponible en: http://visible-learning.org/2015/06/download-john-hattie-politics-distraction/
67. Sánchez Cerón M, del Sagrario Corte Cruz FM. Las evaluaciones estandarizadas: sus efectos en tres países latinoamericanos. Rev Latinoam Estud Educ (México). 2013;43:97–124.