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

Evaluación estandarizada de logro educativo: contribuciones y retos

Eduardo Backhoff Escudero

Resumen

Las evaluaciones estandarizadas en el ámbito educativo tienen una larga historia, que inicia a principios de siglo XX. El campo de la Psicología y de la Educación ha impactado enormemente el mundo educativo y, últimamente, ha servido para diseñar políticas públicas y para rendir cuentas a la sociedad. La principal característica de las pruebas estandarizadas es que pueden administrarse a una gran cantidad de personas, cuyas respuestas se califican de manera automática con dispositivos electrónicos. Su gran desventaja radica en que utilizan, principalmente, el formato de *selección de respuestas* que hace un tanto artificial la evaluación. A pesar de esta limitación, las evaluaciones estandarizadas se utilizan con una gran variedad de propósitos: desde la admisión a instituciones educativas hasta la evaluación de la calidad educativa de un país. A lo largo de su historia, las evaluaciones estandarizadas han sido objeto de críticas, algunas de ellas justas y otras no. Dada su importancia, el propósito de este texto es precisar lo que se entiende por evaluaciones estandarizadas de aprendizaje, describir su origen y evolución en el ámbito educativo, explicar sus principales usos y puntualizar las limitaciones y retos que tendrán.

Palabras clave: evaluaciones a gran escala, evaluaciones estandarizadas, logro académico, aprendizaje, México.

DOI: http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a3

Universidad Nacional Autónoma de México, Coordinación de Desarrollo Educativo e Innovación Curricular (CODEIC)

Standardized assessment: contributions and challenges

Abstract

Standardized assessments have a long history, which started at the beginning of the 20th century. This field of Psychology and Education has greatly impacted the educational world and, lately, has served to design public policies and as a form of accountability. The main characteristic of standardized tests is that they can be administered to many people, whose responses are automatically qualified with electronic devices. Their great disadvantage is that they use, mainly, the answer's selection format, that makes the testing somewhat artificial. Despite this limitation, standardized tests are used for a variety of purposes: from admission to educational institutions to the evaluation of a country's educational quality. Throughout its history, standardized evaluations have been object of criticisms, some fair and others not. Given its importance, the purpose of this text is to specify what is meant by standardized assessments, to describe its origin and evolution in the educational field, to explain its main uses, and to point out the limitations and challenges that such evaluations will have in the future.

Keywords: large-scale assessment, standardized assessment, educational achievement, learning, Mexico.

Eduardo Backhoff Escudero ebackhoff@gmail.com

Universidad Nacional Autónoma de México. Es licenciado en Psicología, doctor en Educación y miembro del Sistema Nacional de Investigación (desde 1990). Se ha desempeñado como profesor de Psicología en la Universidad Nacional Autónoma de México (UNAM), investigador de la Universidad Autónoma de Baja California (UABC), director del Instituto de Investigación y Desarrollo Educativo de la UABC, Director de Pruebas y Medición del Instituto Nacional para la Evaluación de la Educación (INEE), director de la Revista Electrónica de Investigación Educativa (REDIE), consejero presidente de la Junta de Gobierno del INEE. Actualmente es presidente del Consejo Directivo de Métrica Educativa, A.C. Ha publicado cerca de 120 de artículos de investigación en revistas arbitradas, 30 capítulos y 25 libros en el ámbito educativo, así como de 30 manuales técnicos. Ha participado en más de 170 congresos nacionales y 55 internacionales. Su área de interés y especialidad es el diseño y validación de pruebas psicológicas y educativas, asistidas por computadora, así como la elaboración de instrumentos de evaluación no cognitivos.

Introducción

Tradicionalmente, la evaluación ha ocupado un lugar destacado en el quehacer educativo, como un instrumento para verificar el logro académico de los estudiantes, para retroalimentar su aprendizaje y para certificar los conocimientos adquiridos (Popham, 2002). Por ello, todos los docentes evalúan a sus alumnos, práctica que es fundamental en los procesos de enseñanza-aprendizaje (Anderson, 2018). Sin embargo, lo novedoso en la actualidad es que hacemos referencia a diversas cosas cuando se habla de evaluación. Por ejemplo, la evaluación de alumnos, de la práctica docente, del currículum, de las instituciones o del sistema educativo en su conjunto. Es decir, la evaluación educativa ha ampliado considerablemente sus fronteras, por lo que no se limita al desempeño académico de los estudiantes, ni a las evaluaciones que realizan los profesores en su práctica pedagógica cotidiana (Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura [UNESCO], 2018; Tiana, 1996).

En la historia de la educación mundial, destaca el surgimiento de las evaluaciones estandarizadas, cuyo propósito y formato son distintos a las que realizan los docentes en el aula. Estas evaluaciones, también conocidas como objetivas o de gran escala, rebasan el ámbito del aula para proporcionar resultados que sean confiables, válidos y comparables entre distintas poblaciones de estudiantes. Los exámenes de admisión a las universidades son un ejemplo clásico de una evaluación de esta naturaleza. Igualmente, las evaluaciones para medir el desempeño de los estudiantes de un país (como los casos de Planea[1] y PISA[2]) son un ejemplo más de evaluaciones estandarizadas que se utilizan en el ámbito educativo (ej.: Instituto Nacional para la Evaluación de la Educación [INEE], 2018; Organización para la Cooperación y el Desarrollo Económicos [OCDE], 2016). También lo fue la prueba Evaluación Nacional de Logros Académicos en Centros Escolares (ENLACE), que utilizó la Secretaría de Educación Pública (SEP) hasta 2013.[3] Este tipo de evaluaciones son necesarias para poder estimar objetivamente el desempeño de grandes grupos de individuos y, consecuentemente, tomar las decisiones que correspondan.

A pesar de la utilidad que puedan tener las evaluaciones estandarizadas, muchas personas desconocen sus características, sus bondades y limitaciones; por lo que es común que algunos docentes y especialistas en educación, no sólo las critiquen, sino que estén en contra de su uso, basándose en argumentos que van desde sus limitaciones técnicas hasta sus implicaciones políticas e ideológicas.

Por lo anterior, este trabajo tiene el propósito de precisar lo que se entiende por evaluaciones estandarizadas de aprendizaje, describir su origen y evolución en el ámbito educativo, explicar sus principales usos y puntualizar las limitaciones y retos que dichas evaluaciones tendrán en un futuro próximo.

1. El Plan Nacional para la Evaluación del Aprendizaje es el proyecto del INEE para conocer los niveles de desempeño de los estudiantes mexicanos que terminan distintos grados escolares de la educación obligatoria.

2. *Programme for International Student Assessment* es el proyecto de la OCDE para evaluar y comparar el aprendizaje de los estudiantes de 15 años de edad de distintos países.

3. La prueba se dejó de utilizar debido a que sus resultados se corrompieron (ver, Backhoff y Contreras, 2014).

Características de las evaluaciones estandarizadas de aprendizaje

El propósito de cualquier evaluación define su estructura, sus contenidos y sus usos y, por lo tanto, determina sus alcances y limitaciones. Las evaluaciones de aprendizaje o logro educativo[4] se pueden clasificar en dos grandes categorías: 1) las que diseña el docente para utilizar en su salón de clase, con el objetivo de retroalimentar y calificar a sus estudiantes, y 2) las que desarrollan grupos de especialistas, que se basan en la literatura científica (con marcos de referencia teóricos y metodológicos rigurosos), que tienen como propósito evaluar de manera objetiva, confiable y válida lo que los estudiantes han aprendido en un dominio escolar determinado, independientemente del contexto en que ha ocurrido su aprendizaje (Popham, 2001a; 2002). A este segundo tipo de evaluaciones se les conoce como *estandarizadas*, y se fundamentan en diversas teorías de la medición, como aquellas que dieron origen a la evaluación de la inteligencia, las habilidades numéricas y verbales, o la personalidad. Las tres teorías que han aportado más al campo de la evaluación del aprendizaje a gran escala o estandarizada son la teoría clásica de la medición, la teoría de la generalizabilidad y la teoría de respuestas al ítem. Adicionalmente a estas teorías de la medición, cada prueba debe de tener un marco de referencia de la disciplina que se vaya a evaluar.

4. Por logro educativo se entiende lo que los estudiantes aprenden en el proceso de enseñan-aprendizaje. Estos aprendizajes se refieren a los conocimientos, habilidades o destrezas que el estudiante tuvo la oportunidad de aprender en la escuela.

> **"...estas evaluaciones están diseñadas para evaluar el grado en que los individuos dominan una competencia específica, independientemente de su historia académica..."**

El segundo tipo de evaluaciones, usualmente, se utilizan con grandes poblaciones de estudiantes, razón por la cual requieren que su formato permita calificar las respuestas de manera objetiva y automática; de aquí su nombre de estandarizado (ver figura 1). Una forma de lograrlo es formular preguntas donde se deba identificar y seleccionar la respuesta correcta, entre un conjunto de opciones plausibles; es decir, que pudieran ser verdaderas. De esta manera, se pueden utilizar dispositivos electrónicos (ya sean ópticos o computarizados) capaces de calificar a una gran cantidad de individuos de forma estandarizada y objetiva, en cuestión de minutos. Hay diversos tipos de formatos para seleccionar respuestas. Entre los más utilizados se encuentran los tres siguientes: opción múltiple, falso/verdadero y relación de columnas. Seguramente, el primero de ellos es el formato más utilizado y conocido en las evaluaciones de gran escala (*National Education Association*, s.f.; Fletcher, 2009).

Figura 1. Proceso de admisión a una universidad pública mexicana.

Fuente: reproducido con autorización de Métrica Educativa, A.C.

En los Estados Unidos y en otros países, las evaluaciones estandarizadas se empezaron a utilizar como instrumentos para la selección de personas a diversas instituciones: principalmente a las educativas y a las fuerzas armadas. También se han utilizado para certificar las competencias de algunas profesiones (por ejemplo, Medicina) y otorgar la licencia para ejercer la práctica profesional correspondiente. Igualmente, desde mediados del siglo pasado, las evaluaciones de aprendizaje estandarizadas se empezaron a usar para comparar la calidad educativa de los países (Bloom, 1969). Y, actualmente, una gran cantidad de naciones han creado instituciones evaluativas para medir diversos componentes de sus sistemas educativos y rendir cuentas a la sociedad (Tiana, 1996; Backhoff *et al*. 2017). Por ejemplo, en México se creó el Instituto Nacional para la Evaluación de la Educación (INEE) en agosto de 2002.

Entre las características más sobresalientes de las evaluaciones o pruebas estandarizadas, se encuentran las siguientes:

- Se diseñan de tal manera que las preguntas, las condiciones para su administración, los procedimientos de calificación y la manera de interpretar los resultados son uniformes, consistentes y comparables de una evaluación a otra.
- No necesariamente son pruebas de alto impacto, de tiempo limitado o pruebas cuyo formato de respuesta es la opción múltiple. Las preguntas pueden ser simples o complejas y no se limitan a medir el logro educativo.
- Están diseñadas para administrarse a grandes grupos de personas, como es el caso de las pruebas de admisión a las universidades y las evaluaciones de aprendizaje que se realizan para evaluar la calidad educativa de un país.
- Su desarrollo requiere de personal especializado y capacitado en el desarrollo de instrumentos de evaluación, entre los que destacan: psicólogos expertos en medición y psicometría, especialistas en currículo y docentes de las asignaturas y grados escolares que se evalúan.

- Deben de cumplir con criterios internacionalmente reconocidos por la comunidad académica, como: *The Standards for Educational and Psychological Testing* (American Educational Research Association [AERA], American Psychological Association [APA], National Council on Measurement in Education [NCME], 2014).
- Deben de contar con evidencias que garanticen la validez y confiabilidad de sus resultados.

Las evaluaciones estandarizadas pueden proporcionar dos tipos de resultados: normativos y criteriales. Los primeros sirven para comparar los resultados de un individuo con respecto a una población de referencia, con la cual se normalizan estadísticamente las puntuaciones. Así, los resultados de una persona pueden presentarse en puntuaciones Z,[5] en una escala predefinida (ej.: 200 a 800, con una media de 500 y una desviación estándar de 100 puntos), en puntuaciones percentilares[6] o, bien, en niveles de desempeño (ej.: alto, medio, bajo). Por su parte, los resultados de una evaluación referidos a un criterio hablan de la cantidad o proporción de competencias que un estudiante domina, del total de competencias evaluadas. Los resultados se presentan, por lo general, de dos maneras: porcentaje de respuestas correctas y si se cuenta o no con el nivel de maestría requerido para ejecutar una tarea.

Las pruebas estandarizadas están diseñadas para permitir una comparación confiable de los resultados entre todas las personas examinadas, porque todos responden la misma prueba. Sin embargo, es importante hacer notar que, a menudo, los individuos no han tenido las mismas oportunidades para aprender y adquirir las competencias que evalúa una prueba estandarizada (National Education Association, s.f.). Esto es así, debido a que estas evaluaciones están diseñadas para evaluar el grado en que los individuos dominan una competencia específica, independientemente de su historia académica; de la misma manera que un análisis sanguíneo mide los niveles de colesterol de las personas, independientemente de su historial médico.

5. Las puntuaciones Z se basan en el supuesto de una distribución normal. Usualmente, se presentan en una escala de -3 a +3, donde el 0 es la media (promedio) de la población y la desviación estándar es igual a 1. Es común que estas puntuaciones se transformen a una escala más intuitiva; la más utilizada de todas es la de 200 a 800, con una media de 500 puntos y una desviación estándar de 100 unidades.

6. Las puntuaciones percentilares se refieren a la posición que ocupa una calificación respecto a la totalidad de las puntuaciones, medidas en porcentajes. Estas se presentan en una escala de 0 a 100. Por ejemplo, un alumno que obtiene una puntuación percentilar de 20, significa que por arriba de él está el 80% de alumnos y por debajo el 20%.

Surgimiento de las evaluaciones estandarizadas

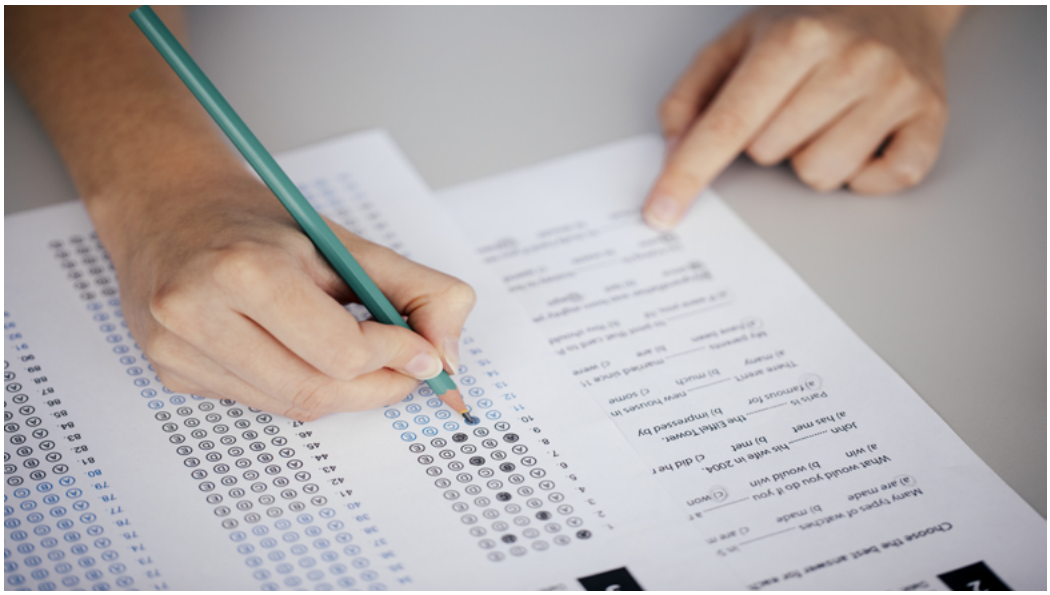
A continuación, se hace una síntesis de algunos eventos históricos que han marcado el rumbo de las evaluaciones estandarizadas en el mundo y en México.

Antecedentes internacionales

La evaluación educativa ha sido producto de dos disciplinas que han convergido históricamente: la Psicología y la Educación. Desde hace más de cien años, la Psicología se ha interesado en evaluar ciertos atributos de los individuos: su personalidad, su inteligencia y sus capacidades cognitivas. Con este interés nació la Psicometría, campo disciplinario cuyo propósito es medir cuantitativamente las características de los individuos. Con nuevas herramientas estadísticas fue

posible diseñar y construir diversos exámenes estandarizados para medir el logro académico de los estudiantes. Uno de estos instrumentos fue la prueba *Stanford Achievement Test* (Anastasi y Urbina, 1998).

Hace cerca de sesenta años, en los Estados Unidos se impulsó fuertemente el uso de pruebas estandarizadas, como consecuencia de la aprobación del *Acta de la Educación Primaria y Secundaria,* cuyo propósito era evaluar la eficacia de todos los programas educativos de este país (Tiana, 1996). Poco después, se publicó el informe Coleman (1966) que tuvo un gran impacto en la sociedad norteamericana, pues mostraba que el nivel socioeconómico de los estudiantes tenía mayor influencia que la escuela en sus niveles de logro académico.



En consecuencia, la sociedad estadounidense empezó a demandar información objetiva y confiable de su sistema educativo, con lo cual se impulsó la evaluación educativa de gran escala (Hanusek, 1986), que dio pie a la creación del programa *Evaluación Nacional de Progreso Educativo* (NAEP), cuya principal función es evaluar los aprendizajes de los estudiantes norteamericanos y darles seguimiento a lo largo del tiempo. La información que genera NAEP sirve para evaluar la calidad de la oferta educativa del país, con lo cual se rinde cuentas a la sociedad (Jones, 1996).

Una década anterior había surgido la Asociación Internacional para la Evaluación del Logro Educativo (IEA, por sus siglas en inglés), con el propósito de comparar los niveles de aprendizaje de los estudiantes de distintos países y con ello poder aprender sobre las buenas prácticas en materia de educación (Ben-Simon y Cohen, 2004). El proyecto más emblemático de esta asociación es el que hoy se conoce, por sus siglas en inglés, como TIMSS (Tendencias de la Medición Internacional de Matemáticas y Ciencias).

En los Estados Unidos, las pruebas a gran escala tal como las conocemos hoy comenzaron a tomar forma con la publicación en 1983 de *A Nation at Risk*. El informe preparado para el Departamento de Educación de los Estados Unidos por la Comisión Nacional de Excelencia en Educación pidió la adopción de normas rigurosas y estándares medibles junto con mayores expectativas para los estudiantes (DePascale, 2013).

La importancia de la evaluación educativa, a través de pruebas estandarizadas, se hizo presente en muchas partes del mundo, por lo que surgieron tanto organismos nacionales como internacionales para medir lo que los estudiantes son capaces de aprender al término de ciertos grados de la educación básica o al cumplir una determinada edad. Estas evaluaciones han tenido dos grandes propósitos: 1) conocer la eficacia de los países en materia educativa y 2) hacer recomendaciones de política pública para mejorar la calidad y equidad de la educación en los países participantes. Entre las organizaciones que destacan en estos esfuerzos evaluativos se encuentran: la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura (UNESCO) y la Organización para la Cooperación y Desarrollo Económico (OCDE). Esta última organización ha destacado por su proyecto PISA, que impactó al mundo desde su primera aplicación en el año 2000 y que ha logrado interesar en la actualidad a más de 70 países.

Antecedentes nacionales

Es hasta la década de los noventa que el campo de la evaluación del aprendizaje fue abordado de manera formal y sistemática en México. Esto sucedió gracias al desarrollo e implementación en las universidades del Examen de Habilidades y Conocimientos Básicos (EXHCOBA) en 1992, con la creación del Centro Nacional de Evaluación para la Educación Superior (Ceneval) en 1994 y con la creación de la Dirección General de Evaluación (DGE) de la SEP en 1992 (Martínez-Rizo, 2001).

En la década de los noventa las autoridades educativas se interesaron en desarrollar exámenes confiables cuyos resultados pudieran servir para diseñar programas y políticas orientados al mejoramiento de la educación. Con este propósito la SEP desarrolló diversas evaluaciones y participó en proyectos de evaluación, de los que destacan: 1) el Factor de Aprovechamiento Escolar, componente del programa de Carrera Magisterial, 2) la prueba IDANIS, (Instrumento de Diagnóstico para Alumnos de Nuevo Ingreso a Secundaria), 3) la participación en el proyecto TIMSS, 4) la coordinación del estudio del Laboratorio Latinoamericano para la Evaluación de la Calidad de la Educación (LLECE) en 1997, y 5) la construcción de las pruebas Estándares Nacionales, para evaluar el logro académico de los alumnos mexicanos de educación básica.

México inicia una fase acelerada de evaluación educativa con la entrada del nuevo milenio: en 2000, participa en el incipiente proyecto de PISA y, en 2002, se crea el INEE con el propósito de evaluar al sistema educativo mexicano y coadyuvar a la rendición de cuentas. Por primera vez, se dan a conocer los resultados de las evaluaciones que se les realizan a estudiantes de educación obligatoria.

Paralelamente, la SEP desarrolla las pruebas ENLACE (Evaluación Nacional de Logros Académicos en Centros Escolares), con las cuales evalúa a los alumnos de primaria (de tercero a sexto grados), de secundaria y de educación media superior (último grado). A partir de 2010, ENLACE se convirtió en un componente esencial del programa de Carrera Magisterial, de tal manera que los resultados de los estudiantes contaban para que sus maestros recibieran estímulos económicos (Santibañez *et al.*, 2006).

Finalmente, la reforma educativa de 2013 le da autonomía al INEE, que lo convierte en autoridad en materia de evaluación educativa y en el coordinador del Sistema Nacional de Evaluación Educativa, donde las evaluaciones estandarizadas del aprendizaje juegan un papel muy importante para medir la calidad educativa del país y de cada una de sus 32 entidades.

Usos de las evaluaciones de gran escala

Sin querer ser exhaustivos, a continuación, se mencionan algunos de los usos de mayor importancia de las evaluaciones estandarizadas en el ámbito educativo, los que divido en dos grandes categorías, de acuerdo con su ámbito de acción: 1) la escuela y 2) el sistema educativo.

Respecto a los usos que le pueden dar las autoridades escolares y los docentes a las evaluaciones estandarizadas, destaco los siguientes:

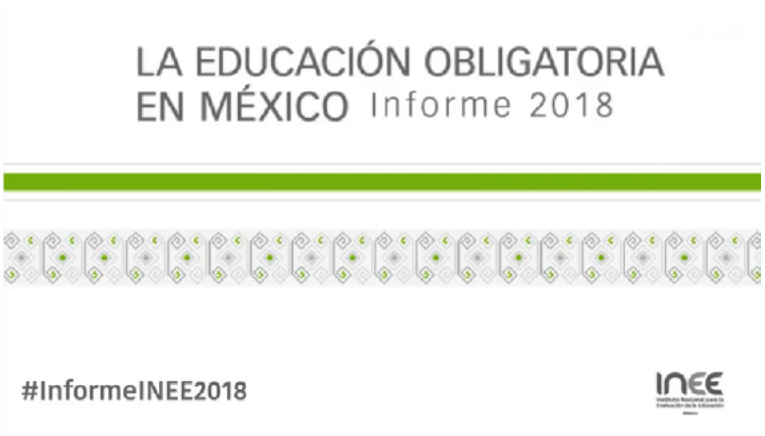
- *Admisión a las instituciones*. Las pruebas estandarizadas son instrumentos muy útiles para que las instituciones puedan realizar procesos confiables, eficientes, transparentes y objetivos para determinar qué estudiantes deben ingresar a una institución cuya demanda rebasa la oferta educativa.

Video 1. Ejemplo de examen de admisión. Revisión del examen de admisión por Notario Público, Universidad Autónoma de Ciudad Juárez. https://youtu.be/82qX5PPORig

- *Ubicación escolar*. En muchos países cuando un estudiante cambia de escuela, de estado o de país se acostumbra evaluar las competencias académicas que posee para decidir en qué nivel educativo lo deben de inscribir y, en su caso, si debe o no recibir ayuda especial en alguna materia (como podría ser la lectura o las matemáticas). Un ejemplo de este tipo de evaluación la lleva a cabo la ITESO con su Examen de Ubicación.
- *Formación*. Algunas pruebas estandarizadas se utilizan para conocer el nivel de dominio que tienen los estudiantes en cada asignatura y grado escolar. Cuando un alumno se retrasa en alguna de estas asignaturas, se utilizan estas pruebas para determinar dónde hay que reforzar su aprendizaje.
- *Retroalimentación institucional.* Algunas evaluaciones estandarizadas sirven para comparar el logro educativo de los estudiantes de una escuela con relación a otra. Esta comparación sirve para definir estrategias de mejora de la calidad de los aprendizajes. Con el paso del tiempo, la escuela que se vuelve a evaluar puede obtener información que retroalimente la eficacia de sus estrategias pedagógicas.
- *Certificación*. Muchas disciplinas requieren certificar las competencias de los profesionistas para otorgarles una licencia para ejercer. Este es el caso de los pilotos aviadores que, además de comprobar la acreditación de sus estudios, deben pasar por un examen de certificación al egreso de su formación. Lo mismo sucede con la carrera de Medicina y otras disciplinas, dependiendo del país del que se trate.

Por otra parte, las autoridades educativas de un país o de un estado pueden utilizar los resultados de las evaluaciones estandarizadas con los siguientes propósitos:

- *Mejorar la calidad y equidad de la educación*. Cada día se reconoce más que los aprendizajes de los estudiantes son la razón de ser de las instituciones educativas. Por ello, es común que los países deseen conocer en qué medida el sistema educativo nacional y el de los estados cumplen con los planes y programas de estudio, para poder diseñar e implementar políticas públicas orientadas al mejoramiento educativo del país.



Video 2. Ejemplo de evaluación a gran escala con objetivos de mejora en la calidad y equidad de la educación. *La educación obligatoria en México. Informe 2018*, INEE. https://www.youtube.com/watch?v=aFvmDc0R4fg

- *Evaluación de programas y políticas educativas*. Las evaluaciones estandarizadas también son útiles para evaluar la eficacia e impacto de los programas y políticas educativas que implementen las autoridades educativas federal y estatales. México cuenta con información de logro escolar la cual se puede utilizar para conocer si algún programa de la SEP tiene o no impacto en el rendimiento académico de los alumnos (ej. Nuevo Modelo Educativo).
- *Rendición de cuentas a la sociedad*. Las evaluaciones estandarizadas de aprendizaje proporcionaron información para conocer cómo se encuentra un país respecto al resto de las naciones y en qué medida sus estudiantes adquieren las competencias escolares básicas (ej.: lectura, matemáticas); información que sirve para que las autoridades rindan cuentas a la sociedad en materia educativa y ésta pueda exigirle al Estado que brinde mejores servicios.

Clic sobre la imagen.

Limitaciones de las evaluaciones estandarizadas

No hay duda de que las evaluaciones estandarizadas han tenido un gran impacto en el ámbito educativo en casi todos los países y México no ha sido la excepción. Sin embargo, es importante reconocer las limitaciones y retos que presentan este tipo de instrumentos en el ámbito educativo. Sin querer ser exhaustivos, menciono primero las siguientes limitaciones de las evaluaciones cuyo formato es de *selección de respuestas*: 1) el formato de opción múltiple no es una forma auténtica, o apegada a la realidad, de evaluar las competencias de

una persona, 2) el formato de *selección de respuestas* fomenta que se aprenda por reconocimiento y, en consecuencia, que se evalúan conocimientos de bajo nivel cognitivo y 3) el formato de *opción múltiple* tiene serias limitaciones para evaluar algunos de los contenidos escolares que son importantes (por ejemplo, expresión escrita y comunicación oral).

Adicionalmente, cuando las evaluaciones estandarizadas son de alto impacto debido a sus consecuencias —por ejemplo, al promover el incremento o reducción del presupuesto de una escuela de acuerdo con los resultados de los estudiantes— presentan otro tipo de problemas. Por un lado, fomentan que los docentes enseñen para la prueba y que los estudiantes aprendan para responderla. Esto ocasiona que el docente se centre en enseñar aquellos contenidos que serán evaluados y deje de atender aquellos que, siendo de importancia curricular, no se evaluarán (por ejemplo, expresión escrita). Lo anterior ocasiona que el currículo implementado se estreche considerablemente (Popham, 2001b). Otro efecto negativo que pueden llegar a tener las evaluaciones de alto impacto es el de corromper la misma evaluación, con el objetivo de mejorar las puntuaciones (Backhoff y Contreras, 2014).

Por otro lado, las evaluaciones de alto impacto pueden fomentar que se sacrifique lo verdaderamente importante de la educación, con tal de obtener resultados que no afecten negativamente a la escuela. Por ejemplo, se deja de apoyar a los estudiantes que están muy lejos de tener buenos resultados en las evaluaciones, para no "desperdiciar" los recursos con aquellos que no podrán mejorar los resultados de las escuelas.

Retos futuros de las evaluaciones estandarizadas

Las evaluaciones estandarizadas tienen un siglo de vida y se han utilizado para una gran cantidad de propósitos educativos. Una de sus bondades es la de poder administrarlas simultáneamente en poblaciones muy grandes de estudiantes. Para ello, se ha tenido que utilizar el formato de selección de respuesta, donde sólo una de las opciones es la correcta. Esto permite que los dispositivos ópticos o electrónicos puedan leer y calificar las respuestas de los estudiantes de manera automática. Sin embargo, como ya se describió, el formato de selección de respuestas impone varias limitaciones que se deben superar.

Afortunadamente, el desarrollo de las ciencias computacionales nos permite superar las limitaciones que impone el formato de selección, permitiendo que las evaluaciones puedan utilizar preguntas cuyas respuestas sean más naturales, "auténticas" (ver figura 2). Por ejemplo, el estudiante puede resolver una ecuación y escribir en la pantalla de la computadora su solución. O, bien, puede balancear una ecuación química, trazar una pendiente, identificar puntos geográficos en un mapa, subrayar las partes importantes de un texto, etcétera. Las ciencias computacionales también permiten elaborar *pruebas adaptativas* que requieren mucho menos tiempo del alumno, sin perder la precisión de la medición.

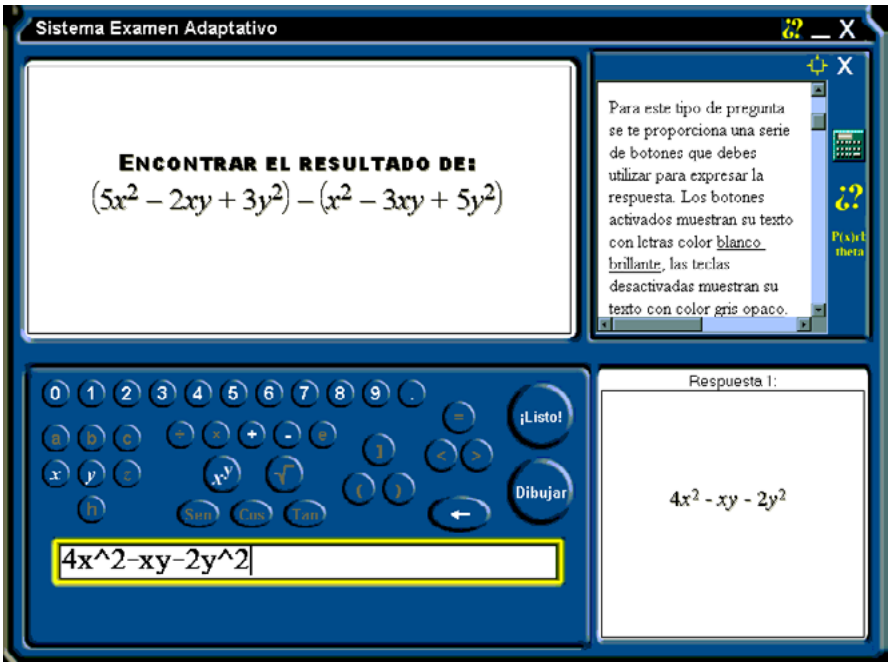


Figura 2. Interfaz del examen de ubicación de matemáticas en el que el estudiante tiene que escribir la respuesta (no seleccionarla). Fuente: reproducido con autorización de Métrica Educativa, A. C.

De igual manera, el desarrollo de las ciencias cognitivas ha permitido mejorar sustancialmente los contenidos y la validez de las evaluaciones de aprendizaje, tanto de pequeña como de gran escala (Pellegrino, Chudowsky y Glaser, 2001). Sin embargo, para lograr mejoras significativas en las prácticas evaluativas en el ámbito de la educación, falta mucho por estudiar, tanto sobre aspectos cognitivos como acerca de su medición.

Por otro lado, las pruebas se desgastan rápidamente con el uso, por lo que es necesario renovarlas constantemente, lo que implica mucho gasto y esfuerzo. Afortunadamente, la *ingeniería de los tests*[7] ha desarrollado lo que se conoce como "generadores automáticos de ítems",[8] que permiten desarrollar una cantidad importante de reactivos isomorfos (conceptual y estadísticamente, equivalentes) y, en consecuencia, pruebas paralelas (para mayor información consulte a Gierl y Haladyna, 2013).

Los problemas ocasionados al establecer consecuencias asociadas a las evaluaciones no son exclusivos de las pruebas estandarizadas, sino que son comunes a cualquier instrumento cuyos resultados tengan consecuencias positivas o negativas, ya sea para los estudiantes, los profesores o para el centro escolar. Siempre va a haber un interés por obtener el mejor resultado al menor costo; condición que opera en contra de los propósitos de la evaluación y que puede hacer que pierda su validez y todo sentido de seguirla utilizando cuando se llega a corromper.

La solución a muchos de los problemas anteriormente expuestos es saber a ciencia cierta cuáles son los alcances y limitaciones de las evaluaciones estandarizadas, para poderlas utilizar de acuerdo con sus propósitos y no "pedirles" más de lo que puedan dar.

7. Ésta es una disciplina bastante novedosa que combina las disciplinas de la psicométrica, las ciencias cognitivas y la computación, con el objetivo de desarrollar pruebas (por ejemplo, de logro educativo) de una manera más eficiente.

8. Los generadores automáticos de ítems (o de reactivos) es una nueva disciplina en el campo de la medición cuyo objetivo es generar una cantidad importante de reactivos de manera automática, que midan los mismos constructos (conocimientos, habilidades, destrezas, competencias, actitudes, etcétera), con lo cual se pueden generar una infinidad de pruebas o exámenes en muy poco tiempo.

Referencias

- American Educational Research Association (AERA), American Psychological Association (APA), National Council on Measurement in Education (NCME), Joint Committee on Standards for Educational and Psychological Testing (Estados Unidos). (2014). *Standards for Educational and Psychological testing*. Washington, DC: AERA.
- Anastasi, A. y Urbina, S. (1998). *Test psicológicos* (7ª edición). México: Prentice Hall.
- Anderson, L.W. (2018). Una crítica a las calificaciones: políticas, prácticas y asuntos técnicos. En, De Ibarrola, M. (Ed.), *Temas clave de la evaluación de la educación básica*. México: FCE.
- Backhoff, E. y Contreras, S. (2014). "Corrupción de la medida" e inflación de resultados de ENLACE. *Revista Mexicana de Investigación Educativa (*RMIE*), 19* (63), 1267-1283.
- Backhoff, E., Vázquez-Lira, R., Contreras-Roldán, S., Caballero-Meneses, J. y Rodríguez-Jiménez, J.G. (2017). *Cambios y tendencias de aprendizaje en México: 2000-2015*. Ciudad de México: Instituto Nacional para la Evaluación de la Educación.
- Ben-Simon, A. y Cohen, Y. (2004). *International assessment: merits and pitfalls*. Trabajo presentado en la 30ª Conferencia Anual de la Asociación Internacional para la Evaluación Educativa, Filadelfia.
- Bloom, B.S. (1969). *Cross-national study of educational attainment: Stage I of the IEA investigation in six subject areas* (Vols. 1-2). Washington, EEUU: Office of Education (DHEW).
- Coleman, J. (1966). *Equality of Educational Opportunity*. Washington, EEUU: Department of Health, Education and Welfare.
- DePascale, Ch. A. (2003). The Ideal Role of Large-Scale Testing in a Comprehensive Assessment System. *Journal of Applied Testing Technonogy, 5* (1), 1-11. Recuperado de: http://www.jattjournal.com/index.php/atp/article/view/48343/39213
- Fletcher, D. (2009). Standardized Testing. *Time*, Diciembre, 11. Recuperado de: http://content.time.com/time/nation/article/0,8599,1947019,00.htm
- Gierl, m. y Haladyna, T.M. (2013). Automatic Item Generation: theory and practice. Nueva York: Routledge.
- Instituto Nacional para la Evaluación de la Educación (INEE). (2018). *La educación obligatoria en México. Informe 2018.* México: INEE.
- Jones, L.V. (1996). A History of the National Assessment of Educational Progress and Some Questions About Its Future. *Educational Researcher, 25* (7): 15-22.
- Martínez-Rizo, F. (2001). La evaluación educativa en México: experiencias, avances y desafíos. Recuperado de: http://www.fmrizo.net/fmrizo_pdfs/capitulos/C%20047%202010%20Evaluacion%20Educativa%20en%20Mexico_FMR-EB%20COLMEX.pdf

- National Education Association (s.f.). Lessons from the Past: A History of Educational Testing in the United States. Recuperado de: https://www.princeton.edu/~ota/disk1/1992/9236/923606.PDF
- Organización para la Cooperación y el Desarrollo Económico (OCDE). (2016). PISA *2015 Results (Volume I). Excellence and Equity in Education*. París: OCDE.
- Pellegrino, J.W., Chudowsky, N. y Glaser, R. (2001). *Knowing what students know. The science and design of educational assessment*. Washington, DC: The National Academies Press. DOI: https://doi.org/10.17226/10019.
- Popham, W.J. (2002). *What Every Teacher Should Know about Educational Assessment*. Boston: Allyn & Bacon.
- Popham, W.J. (2001a). *The Truth About Testing: An educator's call to action*. Alexandria, Virginia: Association for Supervision and Curriculum Development.
- Popham, W. J. (2001b). Teaching to the test. *Educational Leadership, 58* (6), 16-20.
- Tiana, A. (1996). La evaluación de los sistemas educativos. *Revista Iberoamericana de Educación, 10*, 37-61.
- United Nations Educational, Scientific and Cultural Organization (UNESCO). (2018). The impact of large scale learning assessment. París: UNESCO. Recuperado de: http://uis.unesco.org/sites/default/files/documents/impact-large-scale-assessments-2018-en.pdf

Cómo citar este artículo

- Backhoff Escudero, Eduardo (2018). Evaluación estandarizada de logro educativo: contribuciones y retos. *Revista Digital Universitaria* (RDU). Vol. 19, núm. 6 noviembre-diciembre. DOI: http://doi.org/10.22201/codeic.16076079e.2018.v19n6.a3.