Capítulo 19

INTRODUCCIÓN A LA GENERACIÓN AUTOMÁTICA DE ÍTEMS

Eduardo Backhoff Escudero

"El propósito de la educación es formar individuos que se puedan realizar como seres humanos y que se conviertan en ciudadanos capaces de contribuir a construir un país justo y próspero; el objetivo de la evaluación es verificar y coadyuvar a que este propósito se cumpla."

INTRODUCCIÓN

Todo proceso educativo busca alcanzar la meta final de formar a los individuos para que desarrollen sus capacidades, se realicen como seres humanos y se conviertan en ciudadanos productivos que contribuyan a mejorar el país donde viven. Independientemente del modelo pedagógico que se utilice, la evaluación es un componente consustancial del proceso educativo, pues representa la forma idónea de comprobar en qué medida las metas de aprendizaje se han alcanzado por cada uno de los estudiantes y, con base en esta información, retroalimentar su ejecución durante el curso escolar y certificar la adquisición de sus habilidades y conocimientos al final de este.

Aunque los exámenes escolares no son la única manera de evaluar las competencias de un estudiante, sí representan la forma más común y práctica de hacerlo. En el caso de la educación básica, los docentes se preparan en las escuelas normales para enseñar y evaluar a sus estudiantes (aunque con muchas limitaciones). En los niveles de educación media superior (EMS) y superior (ES) los profesores se forman en la práctica de la docencia; ya que se trata de profesionistas que no se formaron para ejercer la pedagogía. De esta manera, por lo general, sus actividades de enseñanza y de evaluación no se fundamentan en ninguna teoría del aprendizaje; se trata de prácticas didácticas que, por intuición o imitación, se ejercitan y se mejoran a través del método de ensayo y error. En algunos casos se obtienen buenos resultados y en muchos otros solo se logran prácticas pedagógicas mediocres.

Independientemente de su preparación o experiencia pedagógica, todos los docentes tienen la necesidad de preparar clases, elaborar materiales didácticos y utilizar exámenes para evaluar el logro académico de sus estudiantes. En algunas instituciones educativas se utilizan exámenes departamentales o, bien, exámenes de egreso para certificar las competencias adquiridas, como es el caso del examen de las competencias médicas (UNAM, 2021). En

estos casos, donde se evalúa a una cantidad importante de estudiantes de manera repetida, es común que las instituciones elaboren bancos de preguntas (ítems o reactivos) para construir diferentes versiones de una prueba y así evitar que los estudiantes o egresados conozcan las preguntas de exámenes previos; en cuyo caso, los resultados de los exámenes perderían su validez.

La elaboración de exámenes de gran escala[1] requiere del trabajo colegiado de expertos formados por docentes del nivel educativo de que se trate, especialistas en enseñanza de la disciplina correspondiente, expertos en evaluación del aprendizaje y estadísticos especializados en instrumentos de evaluación (Tiana, 1996). La limitación que presenta este trabajo es que los reactivos elaborados se desgastan una vez que se han utilizado en forma masiva o en repetidas ocasiones, por lo que se necesita construir nuevas versiones de preguntas y, en el mejor de los casos, combinarlas con versiones anteriores para hacer rendir el alto costo de la elaboración de este tipo de exámenes. Ante la necesidad de contar con grandes bancos de reactivos, desde los años sesenta, se han hecho diferentes intentos para poder automatizar (y, así, abaratar los costos y facilitar los procesos) la manera de construir una gran cantidad de preguntas; campo al que hoy se le conoce como Generación Automática de Reactivos (GAI) (Irvine y Kyllonen, 2002).

Por su relevancia para el buen funcionamiento de las instituciones educativas, el propósito de este capítulo es introducir al profesor universitario y a las autoridades encargadas de las evaluaciones institucionales en el tema de la GAI, a fin de que evalúen su posible utilización en el quehacer de su competencia. Para lograr este objetivo, se narrará brevemente los antecedentes históricos del nacimiento y evolución de la GAI, se describirán y ejemplificarán los principales modelos de estos generadores de reactivos, se explicará la forma en que el advenimiento de las computadoras en el sector educativo ha hecho posible avanzar sustancialmente en el campo de la GAI y se ejemplificará el uso de esta tecnología con el Generador Automático de Exámenes (GenerEx) (Ferreyra y Backhoff, 2016; Sánchez y Backhoff, 2015); sistema que fue desarrollado en México por Métrica Educativa A.C. (www.metrica.edu.mx) y que se utiliza para generar los reactivos de varios exámenes nacionales, entre los que se encuentra el Examen de Competencias Básicas (Excoba) (Backhoff, Larrazolo, Ramírez y col, 2015).

ANTECEDENTES DE LA GAI

Los orígenes para evaluar las diferencias individuales de las personas se pueden ubicar en los inicios del siglo pasado, con los trabajos de Alfred Binet sobre la medición de la inteligencia que se utilizó, en un principio en Francia, para identificar a los estudiantes que podrían presentar dificultades para aprender en la escuela (Beltrán-Llera y Pérez-Sánchez, 2011). Muy pronto, en la Primera Guerra Mundial, los tests de inteligencia se empezaron a utilizar en el

[1] También conocidas como estandarizadas, son aquellas que se utilizan en cientos o miles de estudiantes, como es el caso de los exámenes de admisión o certificación.

sector militar para identificar aquellos reclutas que tuvieran las capacidades necesarias para incorporarse a los trabajos de la milicia (Ben-Simon y Cohen, 2004). Igualmente, estas pruebas se empezaron a utilizar masivamente en el sector educativo de los Estados Unidos, con el objeto de poder ubicar a los estudiantes en los distintos cursos escolares de acuerdo con sus capacidades intelectuales (Ben-Simon y Cohen, 2004).

Las evaluaciones de gran escala para medir el logro escolar de los estudiantes se popularizaron a partir de los años cincuenta del siglo pasado, lo que impulsó el nacimiento y crecimiento exponencial de empresas dedicadas a la elaboración de todo tipo de pruebas de aprendizaje, ya sea para seleccionar a los aspirantes de una institución, diagnosticar sus habilidades y conocimientos escolares o, bien, certificar las competencias profesionales (Ben-Simon y Cohen, 2004).

Algunas instituciones educativas empezaron a utilizar pruebas departamentales estandarizadas, al darse cuenta que los estudiantes acreditaban los cursos de distintas asignaturas con un dominio de estas muy diferente. En otras instituciones se implementaron pruebas de egreso de ciertos niveles educativos para poder obtener el certificado de terminación de estudios correspondiente; tal es el caso del examen terminal del bachillerato en Francia, *Baccalauréat* (Wikipedia, 2021). En todos los casos, fue notorio que para que las evaluaciones cumplieran su función estas deberían ser desarrolladas con altos estándares de calidad, para lo cual era necesario contratar a especialistas en el desarrollo de pruebas de aprendizaje. Sin embargo, una vez que se utilizaban las pruebas se desgastaban rápidamente, dado que no había manera de evitar que sus contenidos se filtraran y se dieran a conocer entre los estudiantes que aún no se evaluaban.

La solución a este problema fue, inicialmente, la elaboración de bancos de reactivos para formar distintas versiones de un examen. Sin embargo, si bien se solucionaba el problema de contar con preguntas distintas de una versión a otra, se creaba un problema nuevo: el tener que utilizar exámenes con distintos niveles de dificultad, pues es imposible garantizar *a priori* que dos preguntas que evalúan un mismo contenido escolar tengan el mismo nivel de dificultad; condición que se extiende al examen en su totalidad. Por ello, era necesario utilizar distintos procedimientos para igualar el nivel de dificultad de dos versiones de una misma prueba, lo que se conoce en la literatura especializada como equiparación (*equating*, en inglés). Lo anterior, también abonó a que se incrementaran los costos de la elaboración de exámenes. Teniendo la necesidad de elaborar una cantidad de exámenes y versiones distintas de cada uno de ellos, surgió la idea de poder automatizar la elaboración de sus reactivos; lo que resolvería el problema, al menos, en forma parcial. Posiblemente, el primer intento publicado en este sentido fueron los trabajos de Osburn (1968) y Hively, Patterson y Page (1968), quienes propusieron el uso de *formas de ítems*. Este mecanismo para generar decenas de reactivos consistía en la redacción de la estructura sintáctica de una pregunta, que permitía poder reemplazar algunos de sus elementos, con la finalidad de contar con varias versiones para evaluar un mismo contenido curricular. El Recuadro 1 presenta un ejemplo típico de esta técnica.

Recuadro 1. Ejemplo de una forma de ítem, de acuerdo con Osburn (1968) y Hively, Patterson y Page (1968).

Texto fijo de la pregunta
Una costurera borda _____ manteles en _____ horas. Si tiene que bordar un juego de_____ manteles, ¿cuántas horas tiene que trabajar para terminarlo?

Elementos intercambiables (A1, A2 y A3)
Una costurera borda __ **A1**___ manteles en __ **A2**__ horas. Si tiene que bordar un juego de___ **A3**___ manteles, ¿cuántas horas tiene que trabajar para terminarlo?

Reglas de los elementos intercambiables
A1: la variable puede tomar valores en un rango de 5 a 50
A2: la variable puede tomar valores en un rango de 25 a 250
A3: la variable puede tomar valores en un rango de 3 a 30

Fuente: Adaptación de Sánchez y Backhoff (2015).

Como se aprecia en el recuadro, para generar versiones equivalentes de reactivos se establecían reglas que limitaban la dificultad de las distintas versiones de las preguntas. Estas reglas se referían a los valores mínimos y máximos de cada uno de los elementos intercambiables del reactivo, o bien del tipo de operación matemática que debería de realizar el estudiante para resolver los problemas planteados (Hively, Patterson y Page, 1968).

Veinte años después, Haladyna y Shindoll (1989) propusieron la elaboración de "moldes de reactivos" (*Item Shell*, en inglés*)* con la idea de facilitarles el trabajo a los redactores de ítems y poder realizar su tarea con mayor eficiencia y rapidez. Estos moldes contienen la estructura semántica y otros elementos de los ítems convencionales. Es decir, se trata de ítems "huecos" cuya base del reactivo ya estaba redactada, con excepción de algunos elementos relacionados con su contenido, como se muestra en el recuadro 2.

Recuadro 2. Ejemplo de un molde de ítem siguiendo la propuesta de Haladyna y Shindoll (1989)

Base del reactivo original

¿Cuál es el objeto de estudio de la Ecología?

- Las relaciones de los organismos y su medio.
- La estructura y función de los seres vivos.
- La contaminación atmosférica.
- Los deterioros del medio ambiente.

Molde de reactivo

¿Cuál es el objeto de estudio de _____?

Fuente: Adaptación de Sánchez y Backhoff (2015).

Los moldes de ítems, por lo general, se seleccionan de bancos de reactivos que han sido probados y validados con anterioridad y de los cuales se conoce su funcionamiento métrico (o estadístico), como son sus niveles de dificultad. La idea central de estos moldes es ahorrarles tiempo a los elaboradores de reactivos, quienes deben de concentrarse en las competencias a evaluar y no en la estructura semántica, gramatical o sintáctica que requiere una pregunta. De esta manera se pueden crear múltiples versiones de un ítem con el que se desea evaluar un mismo conocimiento o habilidad.

Solano-Flores, Shavelson y Schneider (2001) avanzaron en el diseño de moldes de ítems acuñando el término "plantilla" (en inglés, *template*), para referirse a un conjunto de instrucciones que sirven para desarrollar ejercicios evaluativos de bajo costo y en poco tiempo. Esto permite cuidar los aspectos de redacción, gramática y ortografía a partir de una estructura sintáctica preestablecida, cuya función principal es facilitar que el elaborador del ítem atienda los contenidos a evaluar y no se distraiga en su redacción. Sin embargo, como los autores lo anticipan, este procedimiento no garantiza que los reactivos que se elaboren sean métricamente equivalentes. El recuadro 3 muestra un ejemplo de un molde de ítem con una estructura sintáctica base (fija) y una serie de especificaciones para rellenar los espacios en el texto (que se señalan entre paréntesis y en itálicas). En la parte inferior del recuadro se muestra un ítem generado con base en este modelo, donde en subrayado se identifican los elementos que conforman el nuevo ítem.

Recuadro 3. Ejemplo de un molde de ítem y del reactivo resultante

Molde de reactivo con estructura sintáctica

(Tema o contenido a evaluar)

La evidencia científica indica que (*presentar un personaje, elemento o nombre de un tema central o fenómeno principal*) ha contribuido en gran medida a (descripción breve del tema central). Utilizando tus conocimientos acerca de (tema central) y del concepto (contenido asociado al tema central):

- Describe cómo (*elementos 1, 2 y 3 asociados al tema central como evidencia del mismo*) se relacionan con el fenómeno de (tema central).
- Explica por qué decir: *(afirmación o enunciado que indique una idea errónea acerca del tema central*), es una afirmación errónea.
- Explica por qué (*situación concreta asociada al tema central*) es provocado por la relación entre (*elementos 4, 5, 6 y 7 relacionados como causas del tema central*).

Tus respuestas deben mostrar un dominio preciso y profundo del conocimiento de los conceptos, principios y razonamientos relacionados con el (tema central).

Ítem resultante del molde del reactivo

El calentamiento global

La evidencia científica indica que la especie humana ha contribuido en gran medida a la presencia del calentamiento global. Utilizando los conocimientos del tema calentamiento global y del concepto huella ecológica:

- Describe cómo el aumento en los niveles del mar, el derretimiento de las capas polares y la desaparición de muchas especies de animales y plantas se relacionan con el fenómeno del calentamiento global.
- Explica por qué decir: "el calentamiento global es un proceso natural del planeta", es una afirmación errónea.
- Explica por qué el aumento en la temperatura promedio del planeta es provocado por la relación entre el uso indiscriminado de combustibles fósiles, el uso desmesurado de fertilizantes, los procesos industriales y la pérdida de bosques.

Tus respuestas deben mostrar un dominio preciso y profundo del conocimiento de los conceptos, principios y razonamientos relacionados con el calentamiento global.

Nota: En paréntesis y en itálicas se señalan los elementos del molde del reactivo que se deben cambiar y en subrayado los elementos que conforman un reactivo, de acuerdo con dicho molde.

Fuente: Adaptado de Sánchez y Backhoff (2015).

El advenimiento de las computadoras y su impacto en la GAI

En las últimas décadas, pocos temas han sido tan relevantes en el mundo como el arribo de la informática, disciplina que ha impactado en todos los ámbitos de la humanidad, entre ellos la educación. Muestra de este impacto es el uso de las tecnologías de la información para atender los efectos del COVID-19, que obligó a las escuelas a cerrar sus puertas y adoptar la modalidad de educación a distancia; lo que requirió que los estudiantes aprendieran a aprender en casa.

Específicamente, en el campo de la educación, los recursos digitales han posibilitado la innovación de nuevas formas de evaluar el aprendizaje, tales como: los exámenes asistidos por computadora, las pruebas adaptativas, los simuladores para evaluar competencias profesionales y el desarrollo de la GAI. Las ventajas que ofrece la evaluación por medios digitales, en comparación con las de lápiz y papel, son numerosas. Entre ellas podemos destacar las siguientes:

- Se generan preguntas y versiones de exámenes de manera automática.
- Se mejora la presentación de los exámenes al utilizar diversos medios digitales, tales como ilustraciones, fotografías, audios, animaciones y videos.
- Se califican las respuestas de los estudiantes de forma automática e inmediata.
- Se generan reportes individualizados de resultados de manera eficiente.
- Se mejora la seguridad de los contenidos de los exámenes.
- Se transparentan los procesos de evaluación y se mejora la imagen institucional.

Una ventaja especial de la evaluación computarizada es que permite utilizar preguntas distintas a las de opción múltiple, cuyas respuestas son más naturales o "auténticas", como sería la escritura de ecuaciones algebraicas, la ubicación de puntos cartesianos en un plano, el balance de fórmulas químicas, el llenado de tablas numéricas, el subrayado de oraciones en un párrafo, la categorización de conceptos, etcétera.

Por estas y otras ventajas, prácticamente, todas las pruebas más prestigiadas del mundo se administran a través de plataformas digitales, ya sea local o remotamente, como son los casos de las pruebas internacionales de PISA (Programme for International Student Assessment), TOEFL (Test of English as a Foreign Language), SAT (Scholastic Assessment Test) y GRE (Graduate Record Examinations).

Por otra parte, el uso de la tecnología digital también ha ayudado a que se avance sustancialmente en el desarrollo de la GAI que, hasta fines del siglo pasado, era muy limitado. La publicación del libro *Item Generation for Test Development* (Irvine y Kyllonen, 2002), que presenta una compilación de las aportaciones más importantes de la GAI a lo largo de su historia, representa un parteaguas en este campo de la evaluación. En este libro se documenta que solo con el apoyo de las computadoras personales se puede aspirar a desarrollar generadores de reactivos de una manera más creativa, sofisticada y eficiente, que responda a las necesidades de quienes se dedican a elaborar instrumentos de evaluación de gran escala.

Las nuevas aproximaciones de la GAI requieren que se desarrollen modelos de ítems que permitan elaborar y manipular los distintos componentes de un reactivo, de tal manera que sea posible medir la competencia de las personas de manera equivalente e intercambiable (Bejar, 2002; Bejar, Lawless, Morley et al., 2003; LaDuca, Staples, Templeton et al., 1986). Un problema de la GAI es que la modificación de cualquier elemento o componente de un ítem, potencialmente, puede cambiar la esencia de la competencia que se pretende medir (conocimiento, habilidad o actitud) o, bien, la dificultad con la que se le mide. Por ello, es importante saber que hay dos aproximaciones teóricas para elaborar instrumentos de evaluación mediante la GAI: la teoría fuerte y la teoría débil. Quienes utilizan la teoría fuerte parten de los principios de teorías psicológicas capaces de explicar los procesos cognitivos que utilizan las personas para poder responder correctamente a una pregunta (Gitomer y Bennett, 2002), con lo que se pueden manipular las propiedades y dificultad de los reactivos generados (Gierl y Lai, 2012) gracias al sustento teórico que los soporta (Lai, Alves y Gierl, 2009). El problema de esta aproximación es que no existen, por ahora, suficientes teorías cognoscitivas para llevar a cabo estos principios a la práctica escolar en los distintos ámbitos del conocimiento (Gitomer y Bennett, 2002).

Por su parte, el uso de la teoría débil de la GAI no requiere que se precisen los procesos cognoscitivos necesarios para responder un reactivo. Pero sí es importante asegurar que las versiones de los ítems que se generen sean invariantes –es decir, que no cambien sus características fundamentales de una población a otra– para poder producir reactivos equivalentes, tanto conceptualmente como métricamente (Drasgow, Luecht y Bennett, 2006). A esta propiedad se le conoce en la literatura especializada con el nombre de isomorfismo, propiedad que se tiene que probar empíricamente (Gierl y Lai, 2012).

Por lo general, los instrumentos que se utilizan en el ámbito educativo, en los que se evalúan dominios amplios del conocimiento –como en los exámenes de admisión, departamentales o de certificación– no requieren que se conozcan los procesos cognoscitivos exactos (como en el caso de la teoría fuerte), pero sí es necesario generar grandes cantidades de reactivos isomorfos. Esto requiere que los modelos de ítems contengan los elementos bien definidos que se incluirán en una tarea evaluativa y que serán intercambiados para elaborar distintas versiones de los ítems, a saber: 1) la base del reactivo, 2) los elementos sustituibles y 3) las reglas para generar ítems equivalentes (como se describirá y ejemplificará en el siguiente apartado).

GENERADOR AUTOMÁTICO DE EXÁMENES (GENEREX)

El GenerEx[2] se puede considerar como un prototipo de la GAI, ya que su principal función es generar de manera automática una gran cantidad de reactivos isomorfos, para poder construir una diversidad de versiones de exámenes equivalentes. Dado que este sistema informático se fundamenta en la teoría débil de los generadores, hay que comprobar que las distintas versiones de reactivos que genera, no solo sean conceptualmente equivalentes –lo que se puede hacer a través de la opinión de expertos– sino que, además, sus características métricas (al menos, su dificultad) sean semejantes, lo que solo se puede comprobar a través de estudios empíricos (Ferreyra y Backhoff, 2016).

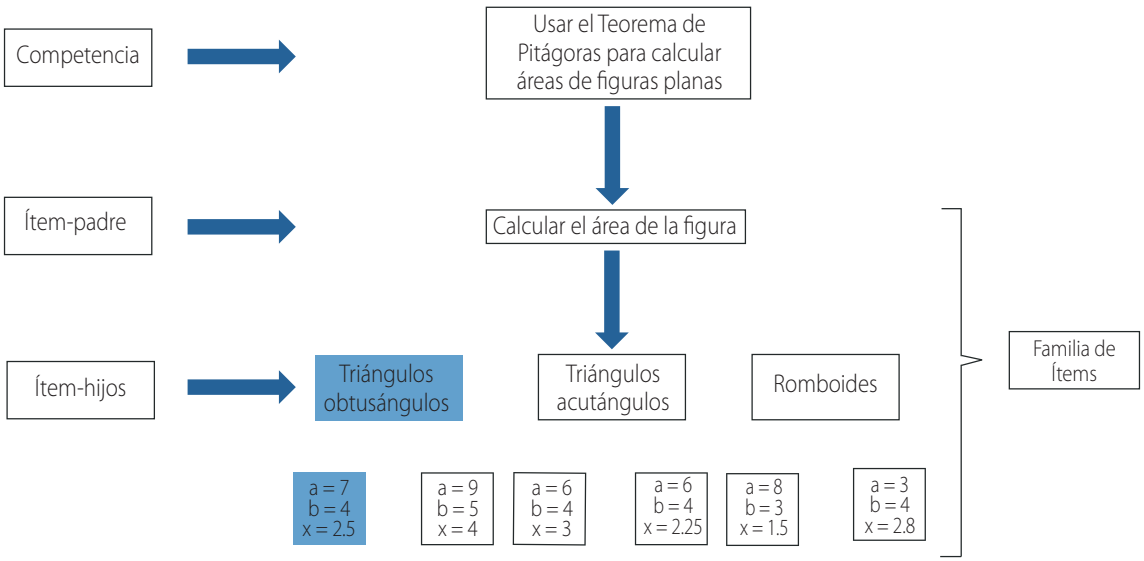
El GenerEx se terminó de desarrollar en 2013, cinco años después de que inició su diseño (Backhoff, Larrazolo y Tirado, 2013). Su antecedente inmediato es el Sistema Computarizado de Exámenes (SICODEX) que se desarrolló, en 1993, con el propósito de administrar exámenes de opción múltiple por computadora (Backhoff, Ibarra y Rosas, 1996), como fueron los casos del Examen de Habilidades y Conocimientos Básicos (EXHCOBA) y del Examen de Egreso del Idioma Inglés del nivel intermedio (EXEDII) (Velazco, Anguiano y Larrazolo, 2007).

Para superar las limitaciones del formato de opción múltiple, avanzar en el uso de la evaluación asistida por computadora y facilitar la generación automática de reactivos, en 2008 se inició el desarrollo de manera conjunta del GenerEx y del Examen de Competencias Básicas (Excoba). Con estos nuevos desarrollos se buscó alcanzar tres metas: evaluar las competencias básicas que se definen en los planes y programas de estudio de la educación obligatoria mexicana; evaluar estas competencias de la manera lo más auténtica posible, es decir, procurando que los estudiantes, en vez de seleccionar las respuestas, las construyeran; y, generar una infinidad de reactivos isomorfos y, en consecuencia, de versiones de exámenes equivalentes.

Para entender cómo funciona el GenerEx, hay que conocer en qué consiste una familia de ítems. Para ello, se muestra el siguiente ejemplo. Supóngase que se desea evaluar el dominio de la competencia de utilizar el Teorema de Pitágoras para calcular áreas de figuras planas (por ejemplo, triángulos). Esta competencia (del nivel de educación básica) puede evaluarse de muchas maneras, pero se requiere especificar las reglas para construir los reactivos con los que se evaluará. Es decir, hay que especificar la familia de reactivos correspondiente, que consta de tres componentes: la competencia a evaluar, así como las características del ítem-padre y de los ítems-hijo. La Figura 1 muestra un ejemplo de estos tres componentes.

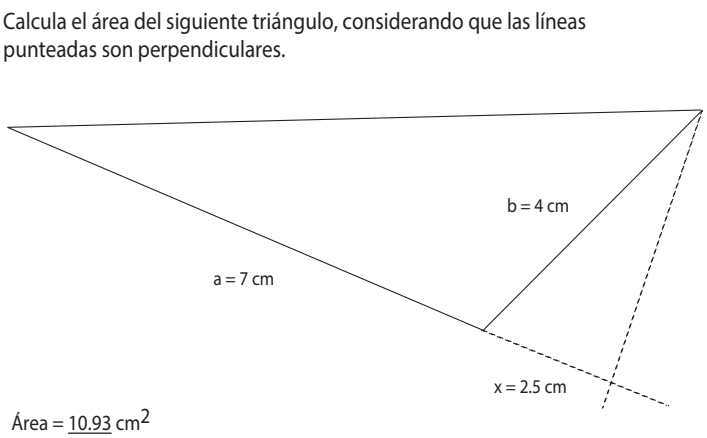
[2] Con registro en el Instituto Nacional de Derechos de Autor.

Figura 1. Familia de reactivos de la competencia uso del Teorema de Pitágoras para calcular perímetros y áreas de figuras planas



En la Figura 2 se muestra el ejemplo de un reactivo generado con la información del modelo de ítem descrito (que en la Figura 1 se resalta sombreándolo). Utilizando la información proporcionada en la Figura 2 (variables a, b y x), el estudiante debe calcular el área del triángulo, para lo cual tiene que calcular, primero, la longitud de la línea punteada con el uso del Teorema de Pitágoras.

Figura 2. Ejemplo de un ítem-hijo para evaluar la competencia de calcular el área de una figura plana regular, utilizando el Teorema de Pitágoras



Es importante mencionar que una familia de reactivos forma parte de un modelo de ítem, que contiene los siguientes componentes: la definición de la competencia a evaluar, la estrategia para evaluarla (en este caso, presentar triángulos o romboides con información incompleta), las reglas para combinar los elementos de la familia de reactivos (en este caso, los valores que pueden tomar las variables a, b y x), el tipo de ejecución que se le solicitará al estudiante (en este caso, escribir la cifra del cálculo correspondiente) y la forma en que se calificarán las respuestas (como podría ser un margen de error aceptable, que en este caso no está especificado). Finalmente, un modelo de ítem puede solicitar más de una respuesta al estudiante; en el ejemplo anterior, se podría haber solicitado también el área del triángulo formado por las líneas punteadas.

A la fecha, el GenerEx se ha utilizado, junto con el Excoba, en procesos de admisión de varias instituciones de educación media superior y superior mexicanas y del extranjero. Igualmente, se ha probado con diversos exámenes diagnósticos de inglés y de matemáticas. La experiencia ha mostrado que el GenerEx es un sistema muy eficiente para generar reactivos y exámenes equivalentes, que se administran y se califican por medios computacionales. Como ya se señaló anteriormente, es necesario verificar que las distintas versiones de reactivos sean isomorfos o métricamente equivalentes, como lo muestran algunos estudios realizados con el Excoba (ver, por ejemplo, Ferreyra y Backhoff, 2016).

REFLEXIONES FINALES

El propósito de este capítulo fue explicar a los docentes, así como a los responsables de los departamentos de evaluación de las instituciones educativas, en qué consisten los generadores automáticos de reactivos, de dónde y desde cuándo proviene el interés por utilizarlos, cómo han ido evolucionando a lo largo del tiempo y qué impacto ha tenido el advenimiento de la tecnología digital en el campo de la evaluación educativa y en el área de la GAI, en particular. Para lograr el propósito del capítulo, se describe el funcionamiento de un generador automático de ítems desarrollado en México: el GenerEx.

Dos aspectos que se deben resaltar de la GAI son sus ventajas y limitaciones. Entre las primeras destacan la eficiencia con que se pueden generar reactivos y exámenes equivalentes, que evita el desgaste que tienen las pruebas de gran escala, debido a su uso intensivo y a la facilidad con que se pueden socializar sus contenidos y, por ello, invalidar sus resultados. Una aportación específica del GenerEx al campo de la evaluación radica en la posibilidad de utilizar reactivos distintos a los de opción múltiple y, con ello, hacer más auténtica la evaluación del aprendizaje; como sería la escritura de una ecuación matemática, el balanceo de una ecuación química o la ubicación de coordenadas geográficas en un mapa.

Por otro lado, dos de las grandes limitaciones que presenta la GAI tienen que ver con su dificultad para desarrollar y probar el isomorfismo de los reactivos que generan. Respecto a la primera limitación, es importante subrayar que la elaboración de modelos de ítems que requiere la GAI es una tarea compleja, en la que se debe tener cuidado en el nivel de detalle con que se precisan las instrucciones para generar ítems. Es necesario contar con

procedimientos rigurosos que garanticen que la forma en que se generan los reactivos a partir de un mismo modelo son lo suficientemente rigurosos teórica y técnicamente como para garantizar, al menos, su equivalencia conceptual. En este sentido una segunda limitación de los generadores como el GenerEx, que se basan en la teoría débil, se refiere a que la semejanza conceptual entre dos reactivos no garantiza su equivalencia métrica, por lo que es necesario que cada ítem se pilotee para conocer su nivel de dificultad (y otras propiedades métricas) y poder sustentar su isomorfismo. Una solución a este problema es equiparar[3] las versiones de un examen, partiendo de la premisa de que las versiones son semejantes conceptualmente, pero no idénticas psicométricamente.

Con todas sus fortalezas y limitaciones, el desarrollo del GenerEx y su uso, a partir de 2013, en diversas instituciones públicas y privadas de educación superior y media superior, marca un parteaguas de la evaluación asistida por computadora en México y en una de sus innovaciones más recientes: la GAI.

En resumen, la GAI es un campo emergente de la evaluación asistida por computadora que promete revolucionar la evaluación del aprendizaje a gran escala. Esto es especialmente cierto para las instituciones educativas que requieren elaborar o utilizar frecuentemente distintas evaluaciones con propósitos de ingreso, diagnóstico y certificación o, bien, que desean implementar exámenes departamentales en distintas áreas del conocimiento. Sería deseable que en México se impulsara el desarrollo e innovación en este campo de la evaluación educativa en aquellas instituciones educativas que utilicen exámenes estandarizados o de gran escala.

REFERENCIAS

Backhoff, E., Ibarra, M. A. y Rosas, M. (1996). Desarrollo y validación del sistema computarizado de exámenes (SICODEX). *Revista de la Educación Superior*, Vol. XXXV, No. *1*(97), pp. 41-54.

Backhoff, E., Larrazolo, N., Ramírez, J.L., Rosas, M., y Tirado, F. (2015). Excoba: Examen de Competencias Básicas. México: Instituto Nacional de Derechos de Autor.

Backhoff, E., Larrazolo, N., Tirado, F., (2013 noviembre). Desarrollo y validación de un Generador Automático de Reactivos de respuesta construida para elaborar exámenes computarizados de ingreso a la educación superior. Memoria III Clabes Conferencia Latinoamericana sobre el Abandono en la Educación Superior.

Beltrán-Llera, J. y Pérez-Sánchez, L. (2011). Más de un siglo de psicología educativa. Valoración general y perspectivas de futuro. *Papeles del Psicólogo*, *3*2(3), 204-231.

Bejar, I. (2002). Generative testing: from conception to implementation. En S. H. Irvine y P. C. Kyllonen (eds.), Item generation for test development (pp. 199-218). Mahwah, NJ: Lawrence Erlbaum Associates.

[3] La equiparación de dos versiones de un examen consiste en ajustar sus puntuaciones para que tengan el mismo nivel de dificultad.

Bejar, I., Lawless, R., Morley, M., Wagner, M., Bennett, R., y Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment, 2*(3), 1-29.

Ben-Simon, A., y Cohen, Y. (2004). International assessments: Merits and pitfalls. Trabajo presentado en la 30ª Conferencia Anual de la Asociación Internacional de Medición Educativa. Filadelfia, PA.

Drasgow, F., Luecht, R., y Bennett, R. (2006). Technology and testing. En R. L. Brennan (e.), Educational measurement (pp. 471-516). Washington, DC: American Council on Education.

Ferreyra, F., y Backhoff, E. (2016). Validez del Generador Automático de Ítems del Examen de Competencias Básicas (Excoba). RELIEVE, *Revista Electrónica de Investigación y Evaluación Educativa, 22*(1) 1-16. Recuperado de http://www.uv.es/RELIEVE/v22n1/RELIEVEv22n1_2.htm

Gierl, M., y Lai, H. (2012). Using weak and strong theory to create item models for automatic item generation: Some practical guidelines with examples. En M. J. Gierl & T. Haladyna (eds.). Automatic item generation: Theory and practice. New York: Routledge.

Gitomer, D., y Bennett, R. (2002). Unmasking Constructs Through New Technology, Measurement Theory, and Cognitive Science (Memorandum de investigación, febrero de 2002, RM-02-01). Educational Testing Service. Statistics and research division. Princeton, NJ.

Haladyna, T. M. y Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. Evaluation and the Health Professions, 12, 97-106.

Hively, W., Patterson, H., y Page, S. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement,* (5), 275-290.

Irvine, S., y Kyllonen P. (eds.). (2002). Item generation for test development. Mahwah, NJ: Lawrence Erlbaum Associates.

LaDuca, A., Staples, W., Templeton, B., y Holzman, G. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical Education,* (20), 53-56.

Osburn, H. G. (1968). Item sampling for achievement testing. Educational and psychological measurement, 28, 95-104.

Sánchez, C., y Backhoff, E. (2015). Generación automática de ítems: una nueva aproximación para evaluar el aprendizaje (Una revisión). REVALUE, *Revista de Evaluación Educativa,* 4(2), 1-25

Solano-Flores, G., Jovanovic, J. Shavelson, R. J. y Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessment. *International Journal of Science Education, 21*(3), pp. 293-315.

Solano-Flores, G., Shavelson, R. J. y Schneider, S. A. (2001). Expanding the notion of assessment shell: from task development tool to instrument for guiding the process of science assessment development. *Revista Electrónica de Investigación Educativa, 3*(1). Recuperado de http://redie.uabc.mx/vol3no1/contents-solano.html

Tiana, A. (1996). La evaluación de los sistemas educativos. *Revista Iberoamericana de Educación,* 10, 37-61.

Universidad Nacional Autónoma de México [UNAM] (2021). Examen de Competencia Académica Ciclo 2021-2022. Recuperado de: https://www.fmposgrado.unam.mx/index.php/examen-de-competencia-academica-ciclo-2021-2022

Velazco, V., Anguiano, M.L. y Larrazolo, N. (2007). Propuesta metodológica para la formulación de una conceptualización del constructo de un examen de certificación del inglés como lengua extranjera. *Revista Electrónica de Investigación Educativa, 6*(2), 1-22. Recuperado de: http://redie.uabc.mx/vol9no2/contenido-velasco.html

Wikipedia (2021). *Baccalauréat* Recuperado de: https://en.wikipedia.org/wiki/Baccalaur%C3%A9at