REVALUE
REVISTA DE EVALUACIÓN EDUCATIVA
Anáhuac México Norte

Para citar este artículo, le recomendamos el siguiente formato:
Sánchez, A. y Backhoff, E. (2015). Generación automática de ítems: una nueva aproximación para evaluar el aprendizaje*, 4* (2). Consultado el día de mes de año en: http://revalue.mx/revista/index.php/revalue/issue/current

Generación automática de ítems: una nueva aproximación para evaluar el aprendizaje

Citlalli Sánchez Álvarez
Universidad Autónoma de Baja California (UABC), México

Eduardo Backhoff Escudero
Instituto Nacional para la Evaluación de la Educación (INEE), México

Resumen

El uso de las evaluaciones estandarizadas de aprendizaje requiere de un recambio constante de sus reactivos debido al desgaste que sufren por su uso frecuente. Esta tarea es altamente compleja, onerosa y poco eficiente. Por ello, desde los años sesenta se han buscado formas diversas para diseñar y construir ítems que agilicen y hagan más barato el proceso de la elaboración de exámenes de logro académico. Con el advenimiento de las ciencias computacionales surge una nueva aproximación para diseñar y construir ítems isomorfos (equivalentes) que midan los mismos constructos educativos. A esta nueva aproximación se le conoce como generación automática de ítems (GAI), campo emergente de la evaluación psicológica y educativa que utiliza los principios de la ingeniería de test, el desarrollo de la psicometría y las nuevas teorías cognitivas. Debido a la importancia de la GAI para el campo de la evaluación del aprendizaje a gran escala, este trabajo tiene tres propósitos: 1) explicar los principios de la GAI a través de su evolución histórica, 2) describir los distintos modelos de ítems que permiten generar reactivos isomorfos y 3) ejemplificar el uso de la GAI en México. A partir de una revisión de la literatura sobre la GAI y los ejemplos presentados a lo largo del texto, se discuten sus grandes ventajas sobre los métodos tradicionales y los retos psicométricos que plantea el estudio de la validez de sus resultados**.**

Palabras clave: generación automática de ítems, modelos de ítems, exámenes computarizados, evaluación del aprendizaje, Excoba.

Fecha de recepción: 31 de agosto 2014
Fecha de aceptación: 23 de diciembre 2015

Abstract

The use of standardized assessment instruments requires the constant renewal of items, due to the wearing effects. This task is highly complex, expensive and inefficient. Therefore, research has been done since the 1960's to look for different forms of item design that permit a more efficient, cost-effective and expedited form of developing achievement tests. Advancements in computer sciences have enabled a new approach to the development and construction of isomorphic items that measure the same educational constructs. This new approach is known as Automatic Item Generation (AIG); an emerging field of assessment that uses Assessment Engineering principles, new psychometric tools and cognitive theories. Because of the importance of AIG to the field of large-scale assessment, this paper has three purposes: (1) to explain the principles of AIG through its historical evolution, (2) to describe the different item models that allow the generation of isomorphic items, and (3) to illustrate the use of AIG in Mexico. By reviewing the literature in the field of AIG and the examples included in this study, we discuss the advantages of this field over traditional methods and the psychometric challenges it poses over the validity of assessment results.

Keywords: Automatic Item Generation, Item models, Computer assisted testing, Learning assessment, Excoba.

Introducción

Los avances de las ciencias computacionales han proporcionado a todos los campos del conocimiento oportunidades de mejoría sustancial. La tecnología digital está presente en prácticamente toda actividad humana y ha permitido la realización de actividades cuyos productos han generado nuevos caminos, propuestas y descubrimientos significativos en materia de salud, ciencias exactas y ciencias sociales. Los campos de la evaluación psicológica y educativa no son la excepción, ya que, junto con los avances tecnológicos, el desarrollo de las ciencias cognitivas y los nuevos modelos psicométricos se han generado cambios y mejoras sustantivas en el campo de la medición, particularmente el del aprendizaje (Gierl, Zhou y Alves, 2008).

La fuerte necesidad de contar con alternativas eficientes para la elaboración de instrumentos de medición ha tenido como consecuencia el desarrollo de nuevas líneas de investigación y desarrollo. Este es el caso de la ingeniería de los test (IT) (Luecht, 2006a, 2006b, 2007a, 2008a, 2008b, 2009, 2010, 2011), que es una aproximación innovadora y sofisticada dentro del campo de la medición.

La IT parte de una óptica distinta a la forma tradicional en que se elaboran las pruebas de aprendizaje, para lo cual se apoya en las teorías de las ciencias cognitivas, los sistemas informáticos, los principios del diseño ingenieril y las herramientas psicométricas. Su finalidad es diseñar una gran cantidad de ítems, tareas evaluativas y pruebas por medio de mecanismos sistemáticos y automatizados que permitan asegurar que los instrumentos midan el constructo deseado (Luecht, 2008a, 2012; Luecht, Burke y Devore, 2009).

Los objetivos que persigue la IT son cuatro: a) diseñar instrumentos que midan consistentemente el mismo constructo, automatizar los procesos de elaboración de ítems que realizan los especialistas, proporcionar métodos eficientes, adaptables y de bajo costo para la producción de grandes cantidades de ítems que no requieran ser piloteados, y reducir la dependencia de los análisis psicométricos para la calibración de reactivos y validación de escalas (Luecht, 2012).

Debido a la importancia del tema para el campo de la evaluación del aprendizaje y al desconocimiento que se tiene en México al respecto, este trabajo tiene el propósito de explicar los principios conceptuales de la IT y su evolución histórica, que permitió el desarrollo actual de lo que hoy se conoce como generación automática de ítems (GAI). Para lograr este objetivo se abordarán los siguientes apartados: 1) bases conceptuales de la IT, 2) surgimiento y etapas históricas de la GAI, 3) concepto de modelo de ítem, 4) modelos de ítems con base en teorías fuerte y débil y 5) experiencia en México en el uso de la GAI. El artículo concluye con una serie de reflexiones respecto a los retos que plantea la GAI para la psicometría.

Ingeniería de test (IT)

De acuerdo con Zhou (2009) y Luecht (2012), la IT se centra en cuatro procesos para el desarrollo y análisis de instrumentos de evaluación. El primero se enfoca en la descripción detallada de los constructos que se pretenden evaluar (llámense conocimientos, habilidades o competencias) y en los niveles de dominio de la escala de puntuación correspondiente. A este proceso, Wilson (2005) le llamó mapeo del *constructo* y lo consideró como un proceso iterativo en el cual se pueden realizar tantas modificaciones como sean necesarias.

El segundo proceso se relaciona con la construcción de modelos de tareas (MT), que implica una elaboración cuidadosa y sistemática de especificaciones para clases o familias de ítems. Su diseño es distinto al modelo convencional de una especificación de reactivos, ya que la intención no es describir la forma en que se n, sino ubicar dentro del mapeo del constructo el nivel de domino en el que se encuentra un examinado, dependiendo de su ejecución en una tarea evaluativa. La diferencia principal entre la especificación tradicional de ítems y la del MT estriba en que su orientación es de tipo cognitiva; es decir, integra de manera sistémica: a) información acerca de los componentes del tipo de conocimiento evaluado y los distintos niveles cognitivos involucrados en su dominio; b) las relaciones que se dan entre dichos niveles y las habilidades cognitivas que los subyacen, y c) los contenidos relevantes, contextos y elementos auxiliares que afectan el nivel de complejidad cognitiva de la tarea que deberá realizar el estudiante.

El tercer proceso que se debe considerar en el diseño de un instrumento es la forma en que se elaboran las plantillas y los reactivos que de ellas se deriven. Estas deben incluir tres aspectos fundamentales. En primer lugar, se requiere un modelo representativo de los ítems, también llamado plantilla de ítems o modelo de ítems (Bejar et al., 2003; Haladyna, 2004), que pueda generar al menos dos ítems. En segundo término, es necesario contar con un sistema de evaluación (preferentemente automatizado) que organice las respuestas y les asigne un valor, que funcione a manera de clave de respuestas. Por último, se debe tener un modelo de información que incluya a todos los componentes fijos y variables que serán manipulados, combinados y ensamblados con la finalidad de construir diferentes reactivos (denominados ítems hijos). Si dicha información se genera mediante algoritmos automatizados (informáticos), entonces se hablaría de la GAI.

El cuarto y último proceso se relaciona con la forma en que se realiza la calibración psicométrica de los reactivos y del instrumento. Bajo los modelos de la teoría clásica de los test (TCT) y la teoría de respuesta al ítem (TRI), el diseño de un instrumento de evaluación implica la elaboración de diversos ítems, su pilotaje y calibración psicométrica. En la IT se establecen controles de calidad muy estrictos desde la fase del diseño de un instrumento, mediante la cuidadosa elaboración de mapas de constructos, mapas de tareas, plantillas y modelos de ítems se detalla con mucha precisión lo que se pretende obtener en términos de

dificultad (y de otras propiedades psicométricas). De esta manera, la perspectiva psicométrica es confirmatoria porque se basa en decisiones de diseño intencionales y sistemáticas, tomadas antes de la construcción de los ítems. Los procesos estadísticos se enfocan en el MT como la unidad central de análisis y se llevan a cabo para identificar y controlar la varianza observada entre los ítems hijos que pertenecen a un mismo modelo (o ítem padre). El objetivo es reducir dicha varianza en la medida de lo posible.

Como se puede apreciar, el proceso de diseño de un instrumento de evaluación bajo las premisas de la IT es un paso fundamental que debe ser vigilado de manera estricta desde un inicio. Será en función de dicho diseño que se podrán elaborar modelos de ítems que contengan los elementos necesarios para generar los ítems que conformarán el instrumento final que será utilizado para evaluar los niveles de dominio de los estudiantes en los constructos delimitados. Asimismo, el uso y aplicación de la informática en el campo de la evaluación del aprendizaje ha permitido atender algunas de las demandas asociadas con la aplicación de evaluaciones a gran escala y así automatizar el proceso de generación de ítems.

Etapas históricas de la generación automática de ítems (GAI)

Todo instrumento de gran escala y alto impacto que se utilice de manera recurrente necesita contar con una gran cantidad de ítems, que deberán ser renovados de manera permanente para garantizar la validez de los resultados den el tiempo. Lo anterior se debe al desgaste natural que sufren los ítems por su uso intensivo. Este proceso además de laborioso, es costoso y poco eficiente debido a la necesidad de pilotear y calibrar los nuevos reactivos, muchos de los cuales se tendrán que desechar por no cumplir con los estándares psicométricos deseados.

La idea de contar con procedimientos y herramientas que permitan automatizar los procesos de diseño y construcción de ítems inició hace más de cincuenta años con técnicas muy rudimentarias para elaborar reactivos semejantes en forma manual. En la actualidad, dichos procedimientos han evolucionado a sistemas más sofisticados que permiten desarrollar ítems isomorfos (equivalentes conceptual y psicométricamente) de manera automática, que se conocen con el nombre de GAI (Gierl y Hollis, 2012; Bezruczko, 2014; Hadyna y Rodríguez, 2013; Bejar et al., 2003).

De acuerdo con Haladyna (2012), la evolución histórica de la GAI ha pasado por cinco grandes momentos (ver figura 1): desde la teoría de facetas de Guttman, a fines de los años cincuenta del, hasta la publicación de un libro sobre la GAI de Irvine y Kyllonen a principios del segundo milenio.

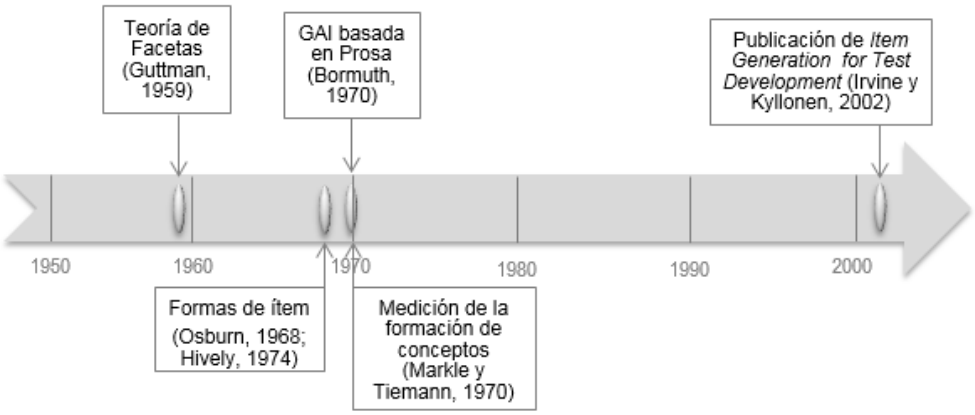


Figura 1. Línea del tiempo en la generación automática de ítems.

El primer momento se le atribuye a Guttman, quien en 1959 creó la teoría de facetas, en la que la generación de reactivos dependía de la cuidadosa elaboración de mapas de enunciados, divididos en diversas secciones que se mantenían fijas y otras que podían variar (llamadas facetas) para modificar su contenido y convertirse en nuevos ítems. Uno de los problemas al operar esta teoría fue que los elaboradores de reactivos no siempre utilizaban el mismo criterio para determinar cómo elaborar los mapas de enunciados.

El segundo momento se dio cuando Osburn (1968), seguido por Hively (1974), propuso lo que llamó formas de ítems, las cuales eran utilizadas para definir dominios de contenido. Tenían una estructura similar a la de los mapas de enunciados de Guttman, porque se trataba de estructuras sintácticas fijas en las que se podía reemplazar una o varias de sus secciones para cambiar los aspectos conceptuales que evaluaban y generar una multiplicidad de reactivos distintos. Esta línea de trabajo fue adoptada durante la década de los años setenta y ochenta y, aunque ya no se utiliza bajo el mismo nombre, algunos de sus principios se siguen aplicando para generar ítems equivalentes.

Posteriormente surgió la GAI basada en prosa, propuesta por Bormuth en 1970, en la que detectó que la forma en que se estaban elaborando los reactivos de las pruebas era subjetiva e ineficiente, debido a la intervención del factor humano en su elaboración. Ante esta situación, propuso que se automatizara el proceso y se centró en la elaboración de ítems cuya base sintáctica se pudiera transformar algorítmicamente de prosa a pregunta. A pesar de que la propuesta de Bormuth no tuvo buenos resultados, después fue retomada por diversos investigadores, quienes transformaron y simplificaron sus algoritmos y la hicieron funcional (Finn, 1975; Roid y Finn, 1977; Roid y Haladyna, 1978). Esta línea de trabajo desapareció debido a: 1) las dificultades en la elección de textos para elaborar ítems y 2) a que evaluaban conocimientos memorísticos.

Un cuarto momento histórico se dio cuando surgieron los trabajos para generar ítems que midieran la formación de conceptos (Markle y Tiemann, 1970), en la que se tomaba un concepto y se identificaban sus principales componentes, atributos y elementos. El trabajo de los elaboradores de reactivos era crear distintas combinaciones de los atributos más importantes de un concepto y elaborar ejemplos y contraejemplos de este; de tal manera que el examinado demostrara su habilidad para discriminar las respuestas correctas y descartar las incorrectas.

El quinto evento importante tuvo lugar cuando Irvine y Kyllonen publicaron un libro en 2002 titulado Item Generation for Test Development, el cual compiló una serie de contribuciones realizadas por investigadores de distintos países en el campo de la GAI, particularmente de Estados Unidos y Gran Bretaña. Dicho libro fue producto de un seminario realizado en 1998 por el Educational Testing Service (ETS), cuyo tema central fue la GAI. En él se menciona la existencia de tres paradigmas de medición utilizados en la GAI: 1) los modelos R, que utilizan programas convencionales para medir el rendimiento escolar o las competencias profesionales; 2) los modelos L o de latencia, que son utilizados para determinar la velocidad de las respuestas de los examinados; y, 3) los modelos D que implican una medición repetida a través del tiempo de tipo predictivo. En general, lo que se busca mediante estos tres modelos es el control de la dificultad de los reactivos.

Como se observa, la evolución de la GAI ha seguido una variedad de líneas de trabajo y, aunque existen puntos de convergencia, también se observa una evolución desarticulada. Sin embargo, todos los esfuerzos realizados han tenido la misma necesidad subyacente: automatizar el proceso para generar grandes cantidades de reactivos que sean equivalentes conceptualmente y semejantes psicométricamente. Actualmente, los trabajos realizados en el campo de la GAI se han centrado en la elaboración de modelos de ítems como el punto medular para la elaboración de reactivos con estructura y propiedades psicométricas similares, pero aún no hay consenso en la terminología utilizada para referirse a ellos.

Modelo de ítems

Un elemento central en la GAI es la noción de Modelo de ítem, que sustituye al de especificación de reactivos. Mientras que este se refiere a las características que debe tener un solo ítem, el primero hace referencia a las de una familia de ítems que mide el mismo constructo de manera equivalente, condición que los hace intercambiables.

Los modelos de ítems han sido el tema central de la GAI y fueron documentados por primera vez en 1968 por Osburn, quien los llamó formas de ítems; sin embargo, fueron Hively, Patterson y Page (1968) quienes desarrollaron de mejor manera el concepto y propusieron un sistema para generar una gran cantidad de problemas de matemáticas, en los que cada forma de ítem contenía un texto fijo con elementos intercambiables y reglas que servían para realizar dichos

intercambios. La figura 2 muestra un ejemplo de una forma de Ítem que contiene estos tres elementos, así como un ejemplo de reactivo.

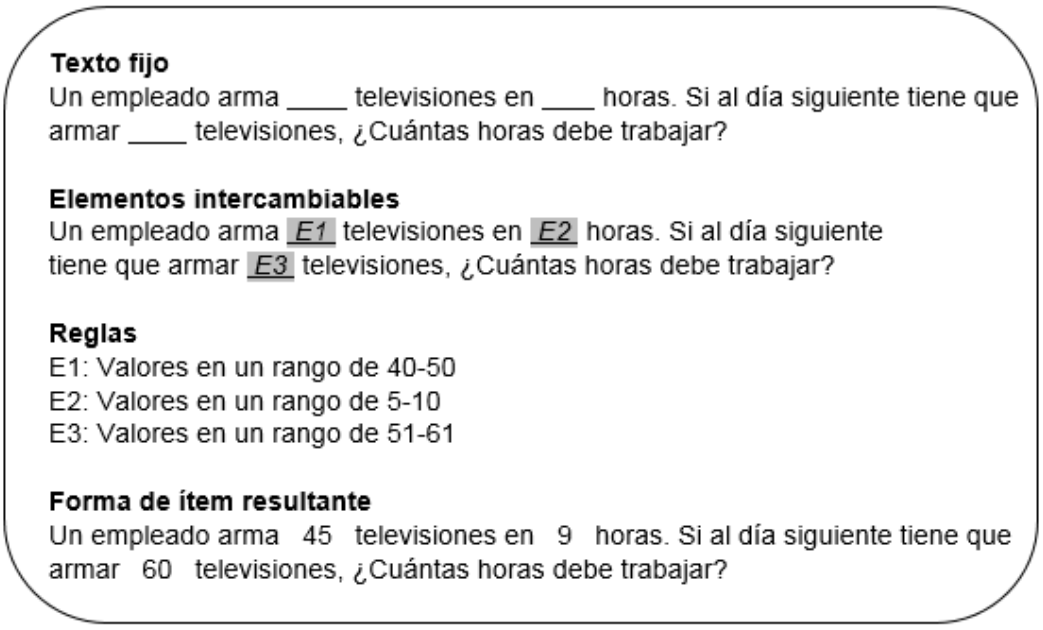
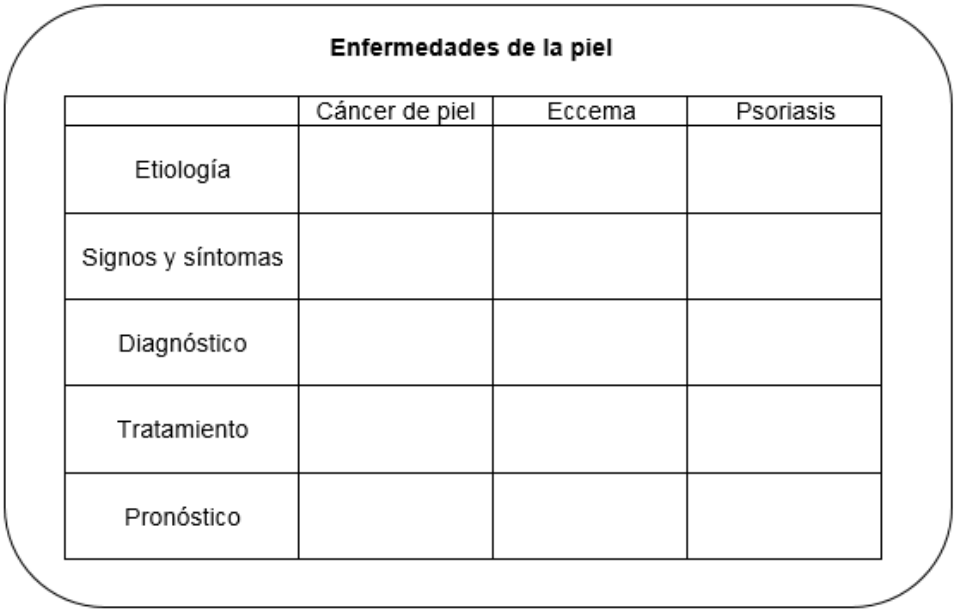


Figura 2. Ejemplo de una forma de ítem, de acuerdo con lo planteado por Osburn (1968), Hively, Patterson y Page (1968).

Posteriormente, en 1974, Minsky llamó marcos (frames) a las estructuras de información que utiliza una persona para representar y dar significado a ciertas experiencias de aprendizaje. Mencionó que, al exponerse a una situación nueva, se genera un marco de referencia que se aloja en la memoria. Este marco se divide en dos niveles: 1) en el más alto se encuentra la información fija, representativa de lo que siempre es verdadero respecto a una experiencia dada; y 2) en el nivel más bajo se encuentran muchas ranuras (slots) que serán rellenadas cuando la persona intente dar significado a situaciones similares a las del nivel más alto, pero que requieran adaptarse debido a que no comparten los detalles específicos. De esta manera, cuando una persona se encuentra ante nueva información, utiliza los niveles de información fijos del marco para situarse y dar significado general a la situación, pero incorpora nuevas ranuras y da pie a la adquisición de nueva información.

Aunque la propuesta de Minsky surgió para explicar la forma en que se organiza la información en la mente de un individuo, ha sido utilizada de forma práctica en contextos escolares para apoyar la enseñanza de textos académicos (Armstrong, Armbruster y Anderson, 1991). Como se muestra en la figura 3, después de revisar en una clase de patología los contenidos teóricos del tema enfermedades de la piel, se puede solicitar al estudiante que complete la información faltante en

los espacios de una tabla (ranuras). De esta manera realizará un ejercicio de organización de la información aprendida.



Enfermedades de la piel

	Cáncer de piel	Eccema	Psoriasis
Etiología			
Signos y síntomas			
Diagnóstico			
Tratamiento			
Pronóstico			

Figura 3. Ejemplo de un marco sobre el tema Enfermedades de la piel, con ranuras para que las complete el estudiante.

El término modelo de ítems fue introducido por primera vez en 1986 por LaDuca, Staples, Templeton y Holzman, quienes lo utilizaron dentro del contexto de la generación de ítems como una herramienta para elaborar ítems isomorfos; es decir, con contenidos y propiedades psicométricas similares (Bejar, 2002). En 1989, Haladyna y Shindoll introdujeron el término molde (shell) para referirse a una técnica creada para responder a las necesidades de certificación de los profesionales de la salud, quienes requerían desarrollar un sistema de evaluación que produjera ítems de opción múltiple de alta calidad y cuya construcción no implicara mucha inversión en tiempo y recursos humanos. Denominaron molde a todo "ítem hueco" que contenía la estructura sintáctica y el contexto de un reactivo, más no su contenido específico. La técnica implica la selección de ítems específicos, pertenecientes a un instrumento de evaluación elaborado y administrado previamente y cuya evidencia empírica indica cuáles cuentan con las propiedades psicométricas necesarias para ser utilizados como moldes. Una vez seleccionados los ítems, se toma su estructura sintáctica básica (base del reactivo) y se elimina su contenido específico, lo que da como resultado un molde en el cual se pueden incrustar distintos elementos que lo convierten en uno nuevo, con estructura similar al original y conservando las propiedades psicométricas del ítem original. De esta forma se cuenta con una estructura gramatical que servirá como punto de partida para que el elaborador de los ítems elija entre los contenidos específicos que desea evaluar, permitiéndole la posibilidad de tener ítems similares en su estructura genérica, pero distintos en sus particularidades

(Haladyna y Shindoll, 1989). La figura 4 presenta un ejemplo de un molde con las características mencionadas.

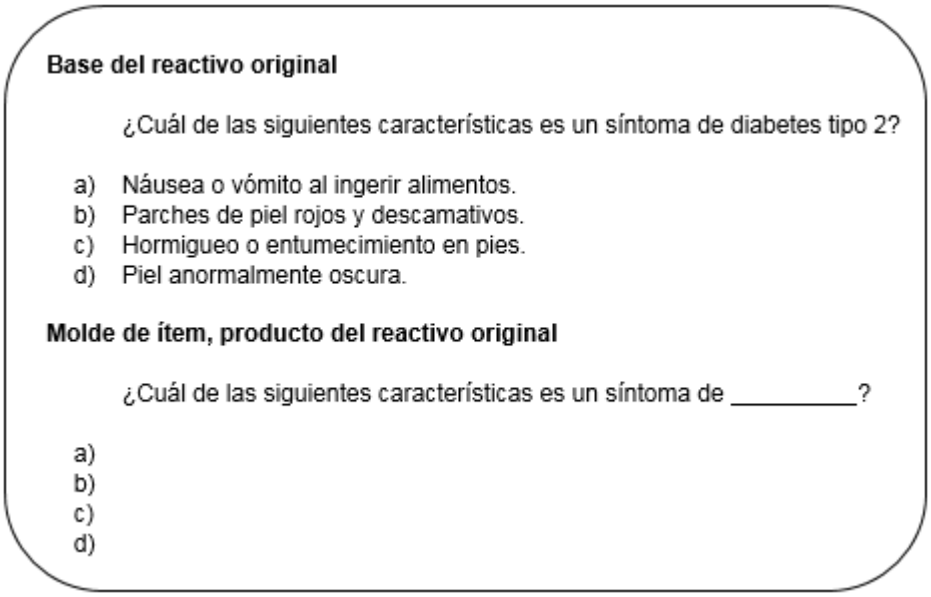


Figura 4. Ejemplo de un molde de ítem en el que se ha eliminado el elemento conceptual de la base del reactivo y conservado la estructura sintáctica básica.

Originalmente, los moldes de ítems fueron creados con la intención de ser utilizados para generar ítems equivalentes para exámenes de opción múltiple en formato lápiz-papel; sin embargo, con el paso del tiempo fueron utilizados exitosamente en distintos campos del conocimiento y con diversas finalidades, incluyendo el desarrollo de sistemas válidos y confiables para evaluar la ejecución de los estudiantes (Draaijer y Hartog, 2007; Enright, Morley y Sheehan, 2002; Haladyna, 1991; Liu y Haertel, 2011; Shea et al., 1992; Simon, 1989; Solano-Flores, 2001; Solano-Flores y Shavelson, 1997). Dentro de las ventajas que se pueden observar en el uso de los moldes de ítems, se encuentra la posibilidad de elegir el nivel de complejidad cognitiva que se desea evaluar mediante la ejecución de los examinados. Este dependerá del tipo de estructura gramatical (redacción) de la cual se parta en la base del reactivo, donde las preguntas cerradas o enunciados incompletos servirán para evaluar habilidades de pensamiento de primer orden, mientras que las estructuras que se utilicen para valorar conceptos de orden superior, presentarán al elaborador de los reactivos una serie de instigadores contextuales que le permitirán flexibilizar la estructura y contar distintas rutas, de las cuales podrá elegir la que más se adecue al tipo de ejecución que desea evaluar. Haladyna y Shindoll (1989) propusieron distintas formas de elaborar moldes de ítems abiertos, en los cuales siempre se utiliza un punto de partida gramatical que sirve para sugerir distintos caminos en la elaboración de la base del reactivo. La figura 5 muestra cómo se puede aplicar el principio anterior a la construcción de un molde de ítems.

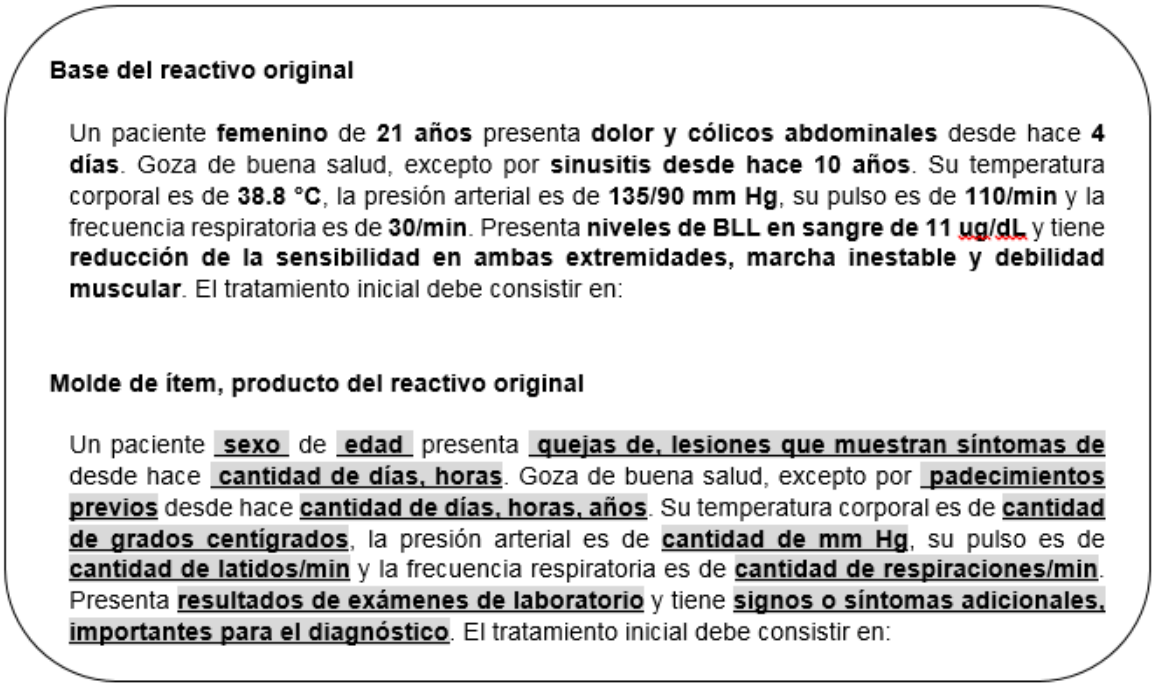
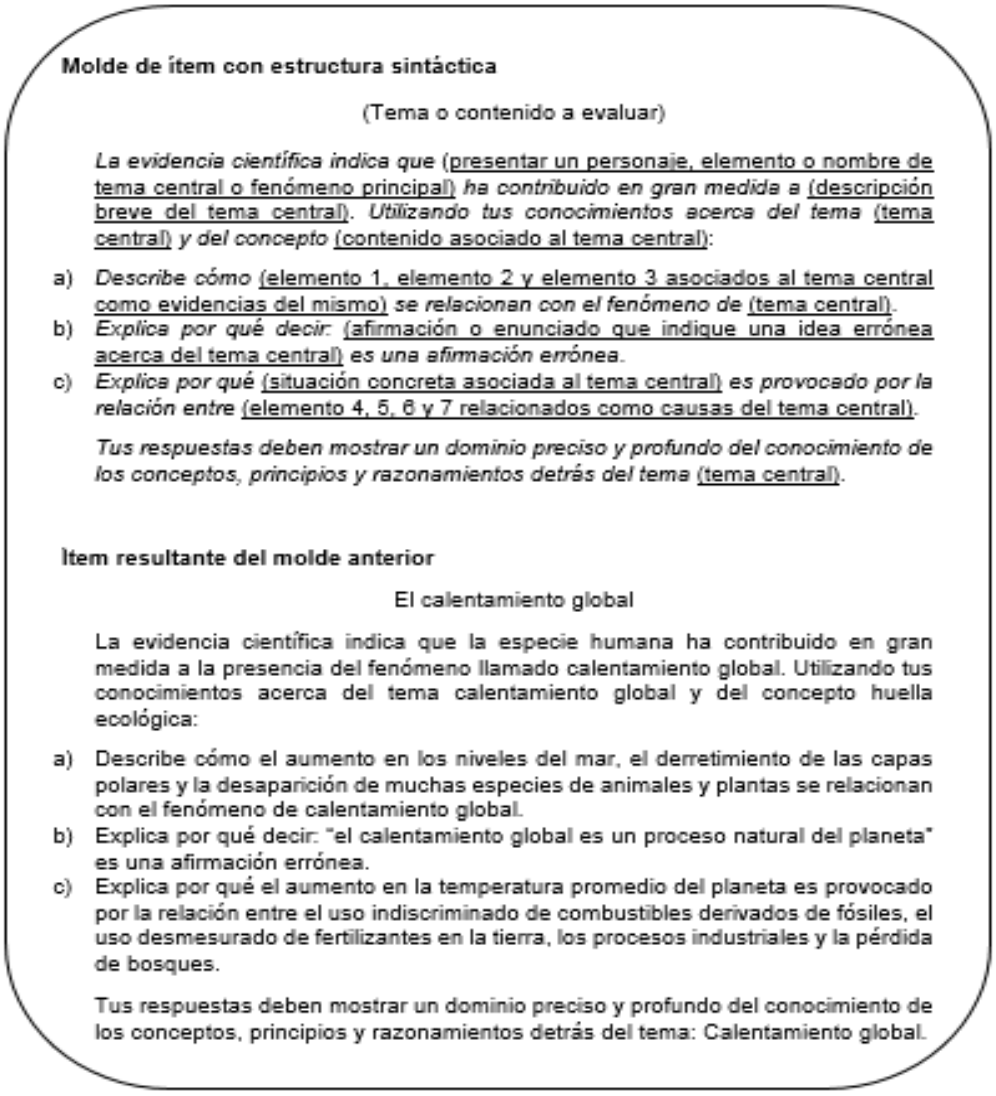


Figura 5. Ejemplo de un molde de ítem abierto para evaluar habilidades mentales de orden superior. Las secciones marcadas en color gris son las que servirán como instigadores contextuales.

Solano-Flores, Shavelson y Schneider (2001) mencionaron que los beneficios de utilizar moldes (término que tradujeron al español como templete) se debían al hecho de que son: a) herramientas para desarrollar pruebas de respuesta construida, b) documentos que formalizan las propiedades estructurales de los ejercicios de evaluación, c) ambientes para la creación de ejercicios de evaluación que permiten estandarizar y simplificar los formatos de respuesta para los estudiantes, y d) herramientas conceptuales que regulan el proceso de desarrollo de exámenes. Debido a que los ítems se construyen a partir de una estructura sintáctica preestablecida, los moldes de ítems permiten cuidar aspectos de redacción gramática y ortografía), que de otra forma podrían pasar inadvertidos cuando distintas personas participan en la elaboración del instrumento, o a que la función primordial del elaborador del ítem es concentrarse en los contenidos a evaluar y no en su redacción. La figura 6 muestra un ejemplo de un molde de ítems en el que se presenta una estructura sintáctica como base, junto con una serie de especificaciones que deben seguir los elaboradores de ítems para rellenar los espacios en blanco. Las letras en itálicas indican que la estructura sintáctica del molde es fija, mientras que los paréntesis con subrayado muestran las características del texto que se debe elaborar.

Los moldes de ítems, siempre y cuando sean desarrollados cuidadosamente y midan dominios específicos del aprendizaje, además de ser buenas herramientas

para desarrollar instrumentos de evaluación, aportan evidencias de validez de contenido ya que valoran contenidos representativos del constructo en cuestión (Solano-Flores et al., 1999).



Molde de ítem con estructura sintáctica

(Tema o contenido a evaluar)

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

a) *Describe cómo* (elemento 1, elemento 2 y elemento 3 asociados al tema central como evidencias del mismo) *se relacionan con el fenómeno de* (tema central).
b) *Explica por qué decir*: (afirmación o enunciado que indique una idea errónea acerca del tema central) *es una afirmación errónea*.
c) *Explica por qué* (situación concreta asociada al tema central) *es provocado por la relación entre* (elemento 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 detrás del tema (tema central).

Ítem resultante del molde anterior

El calentamiento global

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

a) 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 de calentamiento global.
b) Explica por qué decir: "el calentamiento global es un proceso natural del planeta" es una afirmación errónea.
c) Explica por qué el aumento en la temperatura promedio del planeta es provocado por la relación entre el uso indiscriminado de combustibles derivados de fósiles, el uso desmesurado de fertilizantes en la tierra, 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 detrás del tema: Calentamiento global.

Figura 6. Ejemplo de molde de ítem de respuesta construida con estructura sintáctica fija y especificaciones para generar ítems.

Por otro lado, entre las limitaciones encontradas, algunos autores han señalado que los ítems creados mediante un mismo molde son vulnerables al entrenamiento de los examinados (Arendasy y Sommer, 2012; Gierl, 2007), y que la similitud en las propiedades psicométricas de ítems del mismo molde no es señal de isomorfismo, sino de que el constructo no fue debidamente definido en el modelo de ítems (Gierl, Lai y Breithaupt, 2012). Otra desventaja que se ha encontrado al elaborar ítems mediante esta herramienta es la sobreexplotación o "agotamiento" de aspectos específicos del campo o contenido que se mide, en el cual se corre el

riesgo de dejar de lado otros aspectos que son igualmente importantes de medir. Se recomienda el uso de distintos tipos de moldes para elaborar ítems que conforman un mismo instrumento (Haladyna y Shindol, 1989; Solano-Flores, Shavelson y Schneider, 2001). Sin embargo, la elaboración de moldes de ítems no es tarea fácil, ya que implica una serie de procedimientos en los que se definen las características que deberán tener los ítems para tomar la decisión respecto al tipo de molde que se utilizará. Este proceso conlleva complicaciones cuya magnitud dependerá de dos factores: el tipo y nivel taxonómico del campo de conocimiento que se desee evaluar, y el grado de complejidad cognitiva implicado en la tarea que se requerirá para resolver el ítem que se genere a partir de él.

Modelos de ítems con base en teoría fuerte y débil

De los distintos términos que se han empleado para denominar y describir la forma de elaborar ítems y manipular sus componentes, el que más se utiliza dentro del campo de la GAI es el modelo de ítems (Bejar, 1996, 2002; Bejar et al., 2003; LaDuca et al., 1986). Los modelos de ítems se pueden elaborar mediante dos aproximaciones: teoría fuerte y teoría débil. Cuando se utiliza una teoría fuerte, los esfuerzos se centran en develar los mecanismos cognitivos subyacentes al proceso de solución de los ítems generados, así como en estipular cuáles son los elementos específicos que contienen y determinan su nivel de dificultad (Gitomer y Bennett, 2002). Se trata de un proceso mediante el cual se utiliza un modelo cognitivo para no solo identificar estos elementos, sino utilizar el conocimiento teórico que se tiene tanto del contenido evaluado, como de las habilidades y conocimientos que utilizan los examinados para responder al ítem, con la intención de manipular sus propiedades y por ende su nivel de dificultad (Gierl y Lai, 2012). Se trabaja bajo la premisa de que al conocer la dificultad de las demandas cognitivas del contenido que evalúan los ítems se pueden predecir los parámetros de un modelo de respuestas y controlar características psicométricas de los reactivos, como la homogeneidad y dificultad (Bejar, 1993).

Debido a lo anterior, uno de los grandes beneficios del uso de este método es que los ítems generados no requieren ser piloteados previo a su incorporación al instrumento de evaluación, y gracias al sustento teórico subyacente estos se pueden elaborar de manera sistemática atendiendo niveles específicos de complejidad cognitiva con los que cuentan los estudiantes (Lai, Alves y Gierl, 2009). Sin embargo, entre sus limitaciones está el hecho de que no existen suficientes teorías cognitivas para llevar a cabo estos principios a la práctica. Además, en ciertos instrumentos como los exámenes de admisión o de rendimiento académico, en los que se evalúan varios campos del conocimiento no resulta práctico realizar un análisis cognitivo tan exhaustivo (Gitomer y Bennett, 2002). Por ello, la elaboración de instrumentos bajo estos principios se ha limitado a campos específicos del conocimiento en los que ya existen modelos cognitivos establecidos (Gierl, Lai y Breithaupt, 2012).

La figura 7 muestra un modelo cognitivo similar al que propusieron Gierl y Lai (2012) para evaluar el nivel de conocimientos y habilidades requeridas en la solución de problemas para hacer inferencias diagnósticas. Utilizando el caso clínico planteado en la figura 5, se creó un modelo cognitivo para establecer las habilidades y conocimientos que se requieren para llegar a un tratamiento inicial de intoxicación por plomo; también se puede observar el ítem resultante.

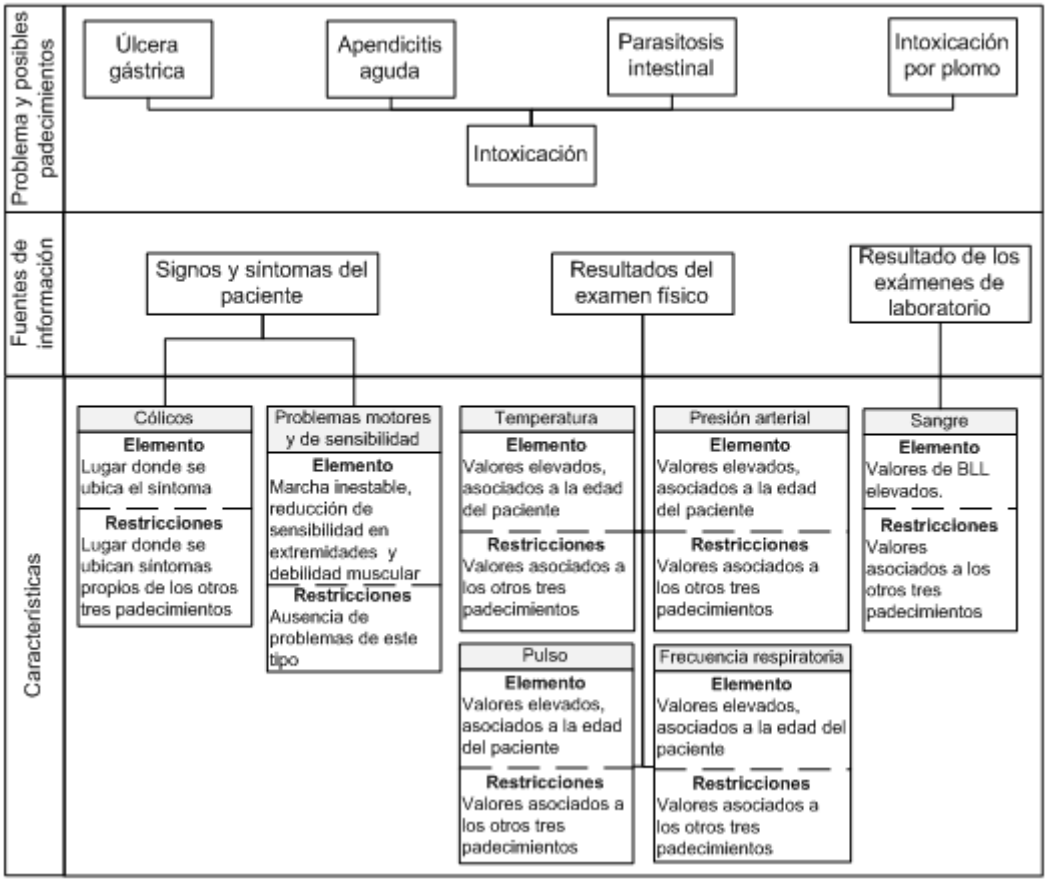


Figura 7. Ejemplo de modelo cognitivo, utilizado para generar ítems mediante teoría fuerte.

Por otro lado, la elaboración de modelos de ítems mediante teoría débil tiene como punto de partida el uso de un ítem padre (Drasgow, Luecht y Bennett, 2006) que se puede crear de varias formas, ya sea mediante la revisión de los ítems de pruebas que ya han sido administrados, eligiéndolo de un inventario de ítems existentes o elaborándolo (Gierl y Lai, 2012). El término teoría débil implica que no es necesario realizar un análisis exhaustivo para conocer y determinar cuáles son los procesos cognitivos base del dominio del contenido evaluado y de las respuestas de los examinados, como sucede cuando se utiliza el método de teoría fuerte. Se trabaja mediante una teoría de invariancia en la que el elaborador del instrumento debe detectar mediante su experiencia, intuición, conocimiento teórico o investigación, las características del ítem padre que no afecten su operación.

Una vez que las identifique, las modificará de manea que se puedan generar variantes del ítem padre (Drasgow, Luecht y Bennett, 2006). Cuando se utiliza la teoría débil con la intención de producir ítems estadísticamente calibrados a partir del ítem padre, la tarea del elaborador será manipular únicamente los elementos que sirvan para producir ítems isomorfos; sin embargo, cuando el objetivo es producir ítems sin calibración estadística, entonces se podrán manipular las características que ayuden a producir grandes cantidades de ítems, independientemente de sus propiedades psicométricas. A diferencia de los ítems generados con base en la teoría fuerte, este tipo de reactivos sí requerirán ser piloteados y calibrados (Gierl y Lai, 2012).

Generalmente, los instrumentos que exploran dominios amplios del conocimiento, tales como los exámenes de admisión utilizan la teoría débil como base para su elaboración, ya que los modelos de ítems construidos bajo esta modalidad muestran todas las propiedades, características y elementos que afectan o no los niveles de dificultad del ítem y permiten hacerle modificaciones para generar grandes cantidades de ítems similares. Aunque el objetivo de la elaboración de instrumentos mediante teoría débil es el mismo que el de teoría fuerte (generar grandes cantidades de ítems calibrados de manera automatizada), el procedimiento es distinto, ya que en la teoría débil se utilizan reglas o directrices para su diseño (Gitomer y Bennet, 2002) y no un análisis o mapeo de los procesos cognitivos subyacentes al contenido evaluado y a las respuestas emitidas por el estudiante.

Los modelos de ítems contienen todas las variables que se incluirán en una tarea evaluativa y serán manipuladas para crear distintas versiones de los ítems; están conformados por tres elementos: 1) la base del reactivo, 2) las opciones de respuesta y 3) la información auxiliar (ver figura 8).

Ítem padre

Juan compró un nuevo refrigerador. Si la temperatura desciende 6ºC cada hora después de que lo conectaron, y la temperatura actual es de 18ºC, ¿cuál será la temperatura dentro de cinco horas?

a) -48º C
b) -19º C
c) -7º C
d) -12º C

Modelo de ítem

Base del reactivo

Juan compró un nuevo refrigerador. Si la temperatura desciende __A__ cada hora después de que lo conectaron, y la temperatura actual es de __B__ , ¿cuál será la temperatura dentro de __C__ horas?

Elementos

A: Rango de 3-9 en valores de 1
B: Rango de 12-22 en valores de 1
C: Rango de 3-9 en valores de 1

Opciones

- Respuesta correcta: -12º C
- Distractores: -48º C, -19º C, -7º C

Información auxiliar: ninguna

Respuesta correcta: d

Figura 8. Ejemplo de modelo de ítem utilizado para generar ítems mediante teoría débil.

La base del reactivo contiene el contexto, contenido, ítem y la pregunta que deberá responder el examinado; las opciones de respuesta incluyen la respuesta correcta y al menos un distractor o respuesta incorrecta, de tal manera que el estudiante elija la que considere pertinente. La información auxiliar se refiere a todos los contenidos adicionales que complementan tanto a la base del reactivo como a las opciones de respuesta, y puede presentarse en forma de texto, imágenes, tablas, diagramas, sonidos o videos (Gierl, Lai y Breithaupt, 2012). No siempre es necesaria la información auxiliar, ya que en algunos casos como el del ejemplo, la base del reactivo contiene todos los elementos necesarios para que el estudiante responda al ítem.

Tanto la base del reactivo como las opciones de respuesta se pueden dividir en elementos de dos tipos: los que contienen información no numérica llamada cadenas (strings), y valores numéricos llamados integrales (integers). Mediante la manipulación sistemática de estos elementos se pueden generar grandes cantidades de ítems, que se pueden clasificar en dos tipos: isomorfos y variantes. En los primeros se busca elaborar reactivos con propiedades psicométricas similares, lo cual se logra mediante la manipulación de los elementos superficiales del ítem llamados incidentales, porque modifican su apariencia más no ejercen influencia significativa sobre su dificultad o propiedades psicométricas. En el caso de los ítems variantes lo que interesa es generar otros con propiedades psicométricas distintas, para lo cual se manipularán elementos llamados radicales,

que son aquellos componentes de la estructura de los ítems relacionados con los aspectos teóricos que les subyacen y funcionan como variables cuasi-independientes, que al ser manipuladas causan modificaciones estadísticamente significativas en las dificultades de los ítems, lo cual se determina mediante la medición del índice de error o tiempo de ejecución (Irvine, 2002).

Aplicaciones de la GAI en México: el caso del Excoba

En México solamente existe una experiencia documentada en el uso de modelos de ítems. Se trata del Examen de Competencias Básicas (Excoba) y el Generador Automático de Exámenes (GenerEx), instrumentos diseñados por Backhoff et al. (2015) para generar exámenes de ingreso a la educación media superior y superior. Este trabajo se basó en la teoría débil de la GAI, debido a que los aprendizajes escolares que evalúa se sustentan en el currículum mexicano de la educación básica y no en un modelo de los procesos cognitivos.

El Excoba/GenerEx busca evaluar las competencias básicas que el estudiante ha adquirido durante su experiencia escolar y que se supone son necesarias para alcanzar aprendizajes significativos de mayor nivel y así cursar con éxito los distintos niveles escolares. El modelo utilizado para su diseño y construcción se basa en la evaluación auténtica, por lo que la mayoría de sus modelos de ítems son de respuesta construida o semiconstruida. En la figura 9 se muestra el ejemplo de un ítem en el que el estudiante debe llevar elementos gráficos hacia una figura para colocarlos y emitir su respuesta. Dos aspectos sobresalen de este: 1) el estudiante debe colocar, mediante la acción de arrastre, los distintos recuadros (conceptos) en el lugar correcto; y, 2) se solicitan cinco respuestas, en vez de una, por lo que el reactivo puede ser calificado con el modelo de crédito parcial.

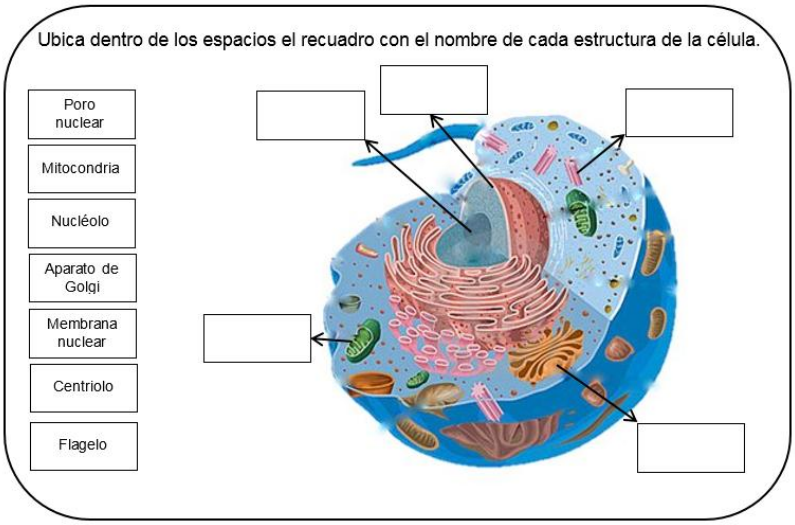


Figura 9. Ejemplo de un ítem de respuesta múltiple semejante a los utilizados por Backhoff et al., 2015).

Todos los modelos de ítems del Excoba/GenerEx incluyen información suficiente para asegurar que al construir un ítem se esté evaluando el aprendizaje que se espera que el estudiante domine en un contenido particular. Esta información se divide en tres grandes secciones: 1) datos que identifican al contenido a evaluar, 2) características del contenido curricular evaluado y 3) plantilla con los elementos indispensables para generar los ítems.

La primera sección contiene la información para identificar el contenido que se va a evaluar: a) la clave que se le asignó, b) el nombre de la asignatura o campo disciplinario al que pertenece, c) el nivel educativo en el que se debe dominar dicho contenido, d) el ámbito al que pertenece en el plan y programa de estudios y e) el nombre del contenido y competencia por evaluar, así como su definición en términos de los aprendizajes esperados del estudiante (ver figura 10).

Clave		**Asignatura o área**	**Nivel educativo**
B-000-I		Biología	Secundaria
Ámbito		**Bloque**	
Estructura celular		I: La biodiversidad: resultado de la evolución	
Contenido	**Nombre**	La célula	
	Definición	Identificación de la célula como unidad fundamental de los organismos vivos, con componentes tridimensionales con anchura, longitud y profundidad posibles de medir.	

Figura 10. Ejemplo de modelo de ítems. Sección de datos de identificación del contenido por evaluar.

La segunda sección contiene las características del contenido curricular evaluado, como los motivos por los cuales se consideró importante incluirlo dentro del instrumento de evaluación, su delimitación conceptual y los conocimientos y habilidades que requiere el estudiante para poder demostrar su dominio sobre él (ver figura 11).

Importancia (justificación) del contenido a evaluar
El conocimiento de la estructura y componentes de una célula es un aprendizaje esencial porque permite que el estudiante la observe como un sistema complejo, organizado, dinámico y auto-dirigido. Representa las bases para la adquisición de nuevos conocimientos como tipos y funciones de los organismos, biodiversidad, nutrición y reproducción.
Delimitación del contenido
El alumno será capaz de identificar los componentes esenciales de la célula.
Conocimientos y habilidades involucrados en la solución correcta del reactivo
Para mostrar dominio en este contenido, el estudiante deberá contar con habilidades de observación, análisis y discriminación de las estructuras básicas de la célula.

Figura 11. Ejemplo de modelo de ítems. Sección de características del contenido por evaluar.

La tercera y última sección del modelo de ítems contiene las especificaciones del modelo para generar diversos ítems. Es en la que se precisan las reglas que se establecieron para combinar los elementos cadena o integrales y así producir ítems. Entre sus apartados se encuentran: a) la estrategia de evaluación que se seguirá, en la cual se deberá mencionar la clasificación del modelo de evaluación del que se trata (según la programación requerida en el GenerEx), así como el tipo de ejecución que realizará el estudiante; b) la base del reactivo; c) los elementos que se utilizarán para generar los ítems (cadenas e integrales), así como las restricciones que se implementarán en caso de ser necesarias; d) las opciones de respuesta, incluyendo la naturaleza de los distractores, como en el caso del ejemplo, en el que todas las opciones son respuestas correcta y al mismo tiempo distractores; e) los elementos auxiliares y f) la respuesta correcta.

En la figura 12 se muestra el modelo de ítem correspondiente para evaluar el conocimiento que tiene el estudiante sobre la estructura y composición de la célula. De este modelo surgió el reactivo que se mostró en la figura 9, que representa uno de tantos reactivos que se pueden generar de manera automática con el Excoba/GenerEx.

III. Generador de ítems

Estrategia de evaluación

Se presentará una lista conformada por siete etiquetas con los nombres de algunos componentes esenciales de una célula. Dicha lista será acompañada de una imagen en la cual se colocarán cinco recuadros vacíos para que el estudiante mediante la acción de arrastre, coloque únicamente cinco de las siete etiquetas en los lugares correctos.

- Clasificación del modelo de ítems en el GenerEx: Elemento imagen.
- Tipo de ejecución implicada: arrastre de elementos.

Base del reactivo

Ubica dentro de los espacios el recuadro con el nombre de cada estructura de la célula.

Elementos

Cadenas:

C1 – C7: Membrana nuclear, nucléolo, citoplasma, flagelo, retículo endoplasmático, ribosoma, lisosoma, mitocondria, vacuola, vesícula, aparato de Golgi, fibras internas, membrana plasmática, poro nuclear, centriolo.

Integrales: ninguno

Restricciones:

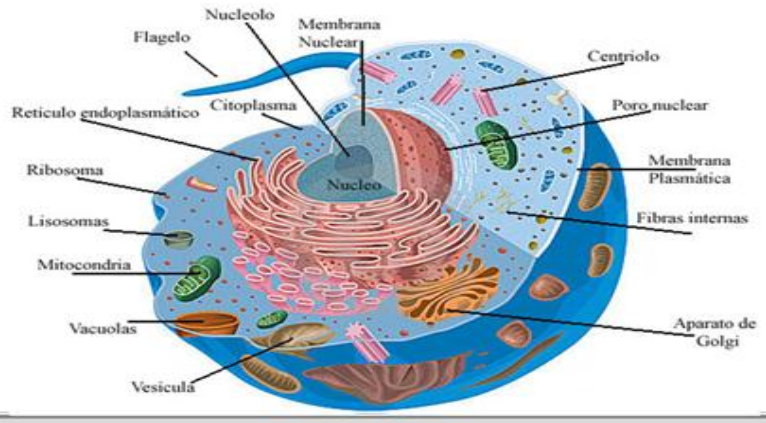
- Seleccionar únicamente 7 elementos cadena para generar el ítem, de los cuales 5 serán respuestas correctas y 2 serán distractores.
- En la figura, ubicar solamente 5 recuadros y sus flechas en los lugares correspondientes, siguiendo la imagen auxiliar.

Opciones

- Respuesta correcta: Selección y ubicación correcta de los cinco recuadros en su espacio.
- Distractores: dos de los elementos presentados en la lista.

Información auxiliar

Imagen con todos los elementos cadena visibles



Respuesta correcta

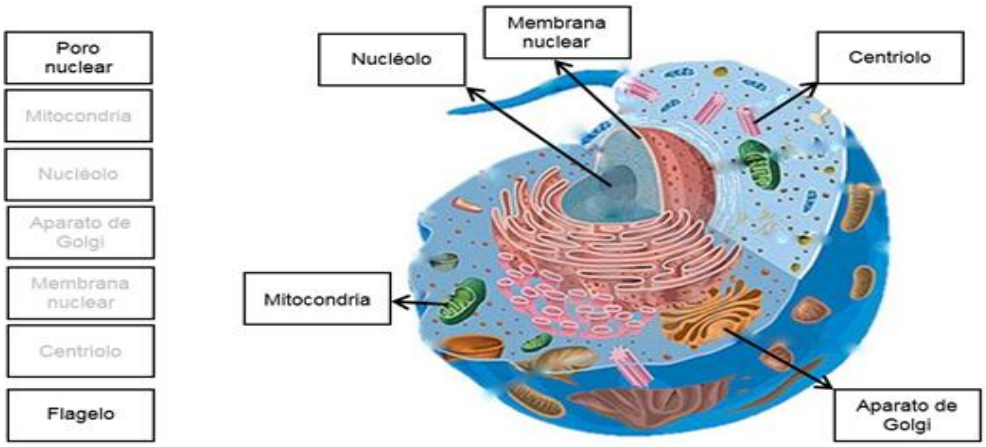


Figura 12. Ejemplo de modelo de ítems del Excoba/GenerEx. Sección que contiene el generador.

Conclusiones

La elaboración de instrumentos válidos para evaluar el aprendizaje de las personas es una tarea técnica de alta complejidad donde confluyen distintas disciplinas. Los inicios de la evaluación del aprendizaje datan de los años veinte del siglo pasado, que se inspiraron en los principios de la evaluación de la inteligencia, desarrollados por Alfred Binet. Así, durante casi 100 años se ha acumulado una vasta experiencia y un cúmulo de conocimientos en el campo de la evaluación del aprendizaje.

En las últimas tres décadas la evaluación del logro educativo ha dado un salto cualitativo con el desarrollo de la informática, que ha permitido la realización de tareas imposibles de lograr sin el uso de las computadoras. Este es el caso de la generación automática de ítems, disciplina que se fundamenta en la ingeniería de test, las teorías cognitivas y los modelos psicométricos, que busca diseñar y construir de manera automática una gran cantidad de reactivos isomorfos y con ellos, pruebas paralelas que midan el mismo constructo. La GAI viene a resolver el problema del recambio de ítems que se requiere realizar de manera constante cuando se hace un uso intensivo de las pruebas, como es el caso de los exámenes de ingreso a las instituciones educativas.

En los ejemplos descritos a lo largo de este trabajo se pudo observar una evolución en la forma en que se han concebido y elaborado los modelos de ítems: transitando desde un molde sencillo en el cual se intercambian segmentos pequeños asociados a un contenido, hasta complejos sistemas y mapas cognitivos que subyacen al ítem; las distintas aproximaciones presentadas muestran cómo el establecimiento de reglas y restricciones, así como la selección y combinación de elementos conceptuales de un generador de ítems requiere de una gran labor creativa por parte de los elaboradores de los modelos de ítems, así como de su experiencia y conocimientos sobre la elaboración de instrumentos.

Como ya se explicó, los modelos de ítems de la GAI se pueden basar en teorías fuertes del aprendizaje y en teorías débiles de las competencias que se desean medir. El primer caso implica que se tiene una teoría cognitiva capaz de precisar la estructura del aprendizaje que subyace a los conocimientos y habilidades de las personas. Con base en una teoría fuerte se diseñan tareas evaluativas que tienen el mismo nivel de complejidad intelectual, con lo que se puede anticipar las propiedades psicométricas de los reactivos. La restricción de esta aproximación es la carencia de teorías cognitivas en los distintos campos del conocimiento que se desean evaluar, razón por la cual su uso actual es muy limitado.

Por su parte, los modelos de ítems que se basan en teorías débiles no presuponen que se puedan predecir las propiedades psicométricas de los reactivos, pero sí que se puedan diseñar ítems conceptualmente semejantes y psicométricamente equivalentes; aunque haya que probarlo ex post facto. El desarrollo del Excoba/GenerEx es un ejemplo claro de la GAI basada en una

teoría débil, ya que su objetivo es medir una diversidad de competencias curriculares de los estudiantes, con propósitos de selección. El desarrollo actual de Excoba/GenerEx permite generar reactivos isomorfos y con ellos construir pruebas paralelas que midan los mismos constructos.

Si bien la GAI resuelve el problema práctico y económico que implica el desgaste de cualquier examen que se utilice frecuentemente, también plantea un nuevo reto al campo de la evaluación del aprendizaje, el cual tiene que ver con la forma de estudiar la validez de los modelos de ítems, con los cuales se pueden generar decenas o cientos de reactivos isomorfos. Ya no se trata de calibrar ítems y formas de pruebas en lo individual, sino de validar familias de reactivos y conjuntos de pruebas que en teoría miden los mismos constructos. Con ello, se abre un nuevo capítulo en el campo de la psicometría.

Referencias

Arendasy, M. & M. Sommer (2012). "Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment". *Learning and individual differences*, 22, 112-117.

Armstrong, J., B. Armbruster y T. Anderson (1991). *Teacher-constructed frames for instruction with content area text* (Reporte técnico núm. 537). Urbana-Champaign, Illinois: Center for the study of reading, College of education, University of Illinois at Urbana-Champaign.

Backhoff, E. E. *et al.* (2015). *Excoba: Examen de Competencias Básicas*. México: Instituto Nacional de Derechos de Autor.

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

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium.* (Reporte de investigación de ETS 96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederikson, R. J. Mislevy e I. I. Bejar (Eds.). *Test theory for a new generation of tests* Mahwah, New Jersey, Erlbaum, pp. 323-359.

Bejar, I. I. *et al*. (2003). "A feasibility study of on-the-fly item generation in adaptive testing", *Journal of Technology*, *Learning and Assessment*, 2(3), 1-29.

Bezruczko, N. (2014). "Automatic item generation implemented for measuring artistic judgment aptitude". *Journal of applied measurement*, 15(1), 1-25.

Bormuth, J. R. (1970). *On a theory of achievement test items*. Chicago, University of Chicago Press.

Draaijer, S. & R. J. M. Hartog (2007). "Design patterns for digital item types in higher education", *E-Journal of instructional science and technology*, 10(1), 1-32.

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

Enright, M. K., M. Morley & K. M. Sheehan (2002). Items by design: The impact of systematic feature variation of item statistical characteristics. *Applied measurement in education*, 15(1), 49-74.

Finn, P. J. (1975). A question writing algorithm. *Journal of reading behavior*, 4, 341-367.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of educational measurement*, 44, 325-340.

Gierl, M. J. & L. Hollis, (2012). "Using weak and strong theory to create item models for automatic item generation. Some practical guidelines with examples". En Mark J. Gierl y Thomas M. Haladyna (eds.), *Automatic Item Generation: Theory and practice* New York, Routledge, pp. 26-39.

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

M. J. Gierl, H. Lai, & K. Breithaupt (2012, abril). Methods for creating and evaluating the item model structure used in automatic item generation. Ponencia presentada en la reunión anual del National Council on Measurement in Education en Vancouver, Canadá.

Gierl, M. J., J. Zhou, & C. Alves (2008). "Developing a taxonomy of item model types to promote assessment engineering". *The Journal of Technology, Learning, and Assessment*, 7(2).

Gitomer, D. H. & Bennett, R. E. (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. (2012). "Automatic item generation: A historical perspective". En M. J. Gierl y T. M. Haladyna (eds.), *Automatic item generation: Theory and practice* Nueva York, Routledge, pp. 13-25.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*, tercera edición. Estados Unidos, Routledge.

Haladyna, T. M. (1991). "Generic questioning strategies for linking teaching and testing". *Educational technology: research and development*, 39, 73-81.

Haladyna, T. M. & Rodríguez, M. (2013). *Developing and validating test items*. New York, Routledge.

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

Hively, W. (1974). “Introduction to domain-referenced testing”. *Educational technology*, 14(6), 5-10.

Hively, W., H. L. Patterson & S. H. Page, (1968). “A ‘universe-defined’ system of arithmetic achievement tests. *Journal of Educational Measurement*, 5, 275-290.

Irvine, S. H. (2002). “The foundations of item generation in mass testing”. En S. H. Irvine y P. C. Kyllonen (eds.), *Item generation for test development* (pp. 3-34). Mahwah, NJ, Lawrence Erlbaum Associates.

Irvine, S.H. & P.C. Kyllonen (eds.). (2002). *Item generation for test development* (pp. 219-250). Mahwah, NJ: Lawrence Erlbaum Associates.

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

Lai, H., C. Alves, & M. J.Gierl (2009, junio). *Using automatic item generation to address item demands for CAT.* Ponencia presentada en el 2009 GMAC Conference on Computerized Adaptive Testing.

Liu, M. & G. Haertel (2011). *Design Patterns: A Tool to Support Assessment Task Authoring* (Reporte técnico 11: Evaluación a gran escala). Menlo Park, CA: SRI International.

Luecht, R. M. (2012). “An introduction to assessment engineering for automated item generation”. En Mark J. Gierl y Thomas M. Haladyna (eds.), *Automatic Item Generation: Theory and practice* New York, Routledge, pp. 59-76.