RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa
E-ISSN: 1134-4032
relieve@uv.es
Universitat de València
España

Ferreyra, María Fabiana; Backhoff-Escudero, Eduardo
Validez del Generador Automático de Ítems del Examen de Competencias Básicas (Excoba)
RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa, vol. 22, núm. 1, 2016, pp. 1-16
Universitat de València
Valencia, España

Revista ELectrónica de Investigación y EValuación Educativa

RELIEVE

ISSN: 1134-4032

e-Journal of Educational Research, Assessment and Evaluation

Validez del Generador Automático de Ítems del Examen de Competencias Básicas (Excoba)

Validity for Automatic Generation of Items for the Basic Competences Exam (Excoba)

Ferreyra, María Fabiana[(1)] & Backhoff-Escudero, Eduardo[(2)]

(1) Métrica Educativa, A.C., México (2) Universidad Nacional Autónoma de México

Resumen

La Generación Automática de Ítems (GAI) es el proceso con el cual se diseñan y elaboran reactivos de una prueba, así como versiones completas de exámenes conceptual y estadísticamente equivalentes. Los Generadores Automáticos de Ítems se desarrollan con el apoyo de sistemas informáticos, que los hacen sumamente eficientes. Con esta idea se creó el generador automático de reactivos GenerEx del Examen de Competencias Básicas (Excoba). Si bien la GAI representa un gran avance en el desarrollo de la evaluación psicológica y educativa, validar la gran cantidad de reactivos y exámenes que se generan de manera automática es un reto metodológico para la psicometría. Este trabajo tuvo el propósito de describir una propuesta para analizar la estructura interna y equivalencia psicométrica de los exámenes generados con el GenerEx, así como describir el tipo de resultados que se obtienen para lograr este propósito. La propuesta se fundamenta en la forma de seleccionar las muestras de reactivos, partiendo del principio de que los ítems y exámenes obtenidos deben ser equivalentes psicométricamente. El estudio se basa en tres tipos de análisis con marcos conceptuales diferentes y complementarios: la Teoría Clásica de los Test, la Teoría de Respuestas al Ítem y el Análisis Factorial Confirmatorio. Los resultados indican que el GenerEx produce exámenes psicométricamente similares, aunque con ciertos problemas en algunas áreas temáticas. La metodología permitió obtener una buena descripción del funcionamiento psicométrico del GenerEx y de la validez interna de dos versiones generadas al azar. Los análisis se pueden complementar con un estudio cualitativo de las deficiencias detectadas.

Fecha de recepción
25 de Octubre de 2015

Fecha de aprobación
26 de Octubre de 2015

Fecha de publicación
1 de Febrero de 2016

Palabras clave:

Generación Automática de Ítems, tests educativos, validez de constructo, estructura factorial, análisis de ítems

Abstract

Automatic Item Generation (AIG) is the process of designing and producing items for a test, as well as generating different versions of exams that are conceptually and statistically equivalent. Automatic Item Generation tools are developed with the assistance of information systems, which make these tools very efficient. Under this aim, GenerEx, an automatic item generation tool, was developed. GenerEx is used to automatically generate different versions of the Basic Competences Exam (Excoba). Even though AIG represents a great advance for the development of psychological and educational assessment, it is a methodological challenge to obtain evidence of validity of the enormous quantity of possible items and tests generated in an automatic process. This paper has the purpose of describing an approach to analyze the internal structure and the psychometric equivalence of exams generated by GenerEx and, additionally, to describe kinds of results obtained to reach this objective. The approach is based on the process for selecting samples from the generation tool, founded on the assumption that items and exams must be psychometrically equivalent. This work includes three kinds of conceptually different and complementary analysis: the Classical Test Theory, Item Response Theory and Confirmatory Factor Analysis. Results show that GenerEx produces psychometrically similar exams; however there are problems in some learning areas. The methodology was useful for obtaining a description about GenerEx's psychometric functioning

Reception Date
2015 October 25

Approval Date
2015 October 26

Publication Date:
2015 February 1

Autor de contacto / Corresponding author

Ferreyra, Maria Fabiana. Métrica Educativa, Alvarado 921, Zona Centro. Ensenada, Baja California, C.P. 22800 (México). ferreira@metrica.edu.mx

and the internal structure of two randomly generated versions of Excoba. Analysis can be complemented by a qualitative study of this item deficiencies.

Keywords:
Automatic Item Generation, Educational Testing, Construct Validity, Factor Structure, Item Analysis

La Generación Automática de Ítems (GAI) se define como el proceso para diseñar y elaborar reactivos de una prueba que son conceptual y estadísticamente equivalentes y que se desarrollan con el apoyo de sistemas informáticos (Gierl & Lai, 2012). Este procedimiento requiere de la participación de especialistas que diseñan los modelos de ítems, así como de métodos estadísticos complejos para validar la calidad y equivalencia de los ítems generados.

Las raíces conceptuales de la GAI se pueden ubicar en los trabajos de Hively, Patterson y Page (1968). Estos autores señalaron que los reactivos se podían generar a través de formatos o plantillas de ítems que contenían elementos fijos y variables. Los elementos variables podían cambiar de acuerdo con reglas explícitas, con lo cual se generaban ítems que medían las mismas habilidades cognitivas, pero no necesariamente presentaban las mismas propiedades psicométricas, como es el caso de su dificultad y su discriminación.

El desarrollo de la GAI progresó con el surgimiento de métodos cognitivos para la instrucción y la evaluación diagnóstica. Sin embargo, estos métodos se concentraron en la enseñanza y no en los tests; por lo tanto se desarrollaron modelos cognitivos, pero no se exploraron implicaciones psicométricas, como la equivalencia entre las diferentes pruebas.

Un tercer paso se dio cuando se integraron las perspectivas de la psicometría y de los modelos cognitivos, que dio pie a dos planteamientos teóricos: la Teoría Fuerte (Embretson, 1999) y la Teoría Débil (Bejar, 1993). La Teoría Fuerte se basa en *modelos cognitivos de tareas*, donde se especifican y manipulan los elementos que afectan el nivel de complejidad (o dificultad) de los ítems generados, de acuerdo con el marco teórico correspondiente. Cada modelo cognitivo de tarea constituye la base para crear múltiples *modelos de ítems*[1] que, a su vez, pueden generar una diversidad de ítems equivalentes. Según Embretson (1999), es posible predecir y controlar las propiedades psicométricas de los reactivos cuando se utiliza un modelamiento cognitivo robusto. Gierl y Lai (2012) señalan que la GAI que se sustenta en la Teoría Fuerte posee poco desarrollo en el diseño de pruebas educativas debido a que se ha centrado, principalmente, en procesos psicológicos básicos, por lo que hay pocas teorías cognitivas desarrolladas en las cuales fundamentarse para diseñar una variedad de modelos de reactivos que atiendan las necesidades de evaluación de los distintos dominioseducativos.

Por su parte, la Teoría Débil utiliza plantillas (o moldes) para diseñar *modelos de ítems* que generen reactivos equivalentes o isomorfos. Una plantilla es una especie de molde conceptual compuesto por una estructura sintáctica básica (tarea que debe realizar un estudiante), con elementos fijos y variables, que al completarse, con reglas preestablecidas, permite generar conjunto de reactivos semejantes (Haladyna & Shindoll, 1989). Para el caso de reactivos de opción múltiple, un modelo de ítem debe incluir los siguientes elementos: una base del reactivo, las opciones de respuesta e información auxiliar (Gierl & Lai, 2011). La base del reactivo contiene el contexto, el contenido y la pregunta que el examinado debe responder. Las opciones deben incluir la respuesta correcta y uno o más distractores. La información auxiliar incluye cualquier material adicional necesario para la generación de los ítems (textos, imágenes, tablas, diagramas). Tanto la base del reactivo como las opciones de respuesta pueden subdividirse en *elementos* (frases, palabras, letras, símbolos, números, etc.). Los ítems que se generen con una plantilla, se denominan ítems-hijo. Si los reactivos generados con el modelo de ítem miden un contenido con niveles de dificultad similares, se dice que los ítems son *isomorfos*. En este caso, los desarrolladores de ítems manipulan aquellos elementos que son

características superficiales del reactivo y que no alteran su dificultad, para producir los ítems isomorfos.

Si bien esta propuesta no requiere de una teoría cognitiva y resulta muy apropiada para la diversidad de exámenes educativos, también tiene sus limitaciones. Una de ellas es que los elaboradores de los modelos de ítems deben predecir las propiedades psicométricas de los reactivos para que estas sean semejantes; sin embargo este objetivo no siempre se logra. Otro inconveniente es que, en muchas ocasiones, para elaborar ítems-hijo isomorfos se hacen cambios muy superficiales en las plantillas, con lo que se obtienen ítems demasiado parecidos o prácticamenteidénticos.

Si bien la GAI lleva más de 40 años de existencia, su desarrollo no había progresado por un problema esencial: el de su validez. La estrategia más simple para analizar las respuestas a los ítems-hijo generados por la GAI es estudiar cada uno de ellos de manera individual como una entidad independiente. Sin embargo, si consideramos que un generador de reactivos produce cientos o miles de ítems, esta solución se convierte en un trabajo monumental e ineficiente. Por lo tanto, se necesitan modelos alternativos e innovadores para analizar los miles de reactivos que se obtienen a través de la GAI.

Sinharay y Johnson (2012) describieron tres modelos para analizar y calibrar los reactivos producidos por la GAI. El primero consiste en predecir las propiedades psicométricas de los reactivos, en particular la dificultad, de acuerdo con las características de los modelos de tareas que se utilizan para generar los ítems. El segundo modelo considera la dependencia entre parámetros que pertenecen a una misma familia de reactivos. El tercer modelo combina los dos anteriores.

En cuanto a la primera aproximación, investigadores como Embretson (1999) y Holling, Bertling y Zeuch (2009) utilizaron el Modelo de Test Logístico Lineal o Modelo Logístico Lineal del Rasgo Latente (LLTM, por su nombre en inglés**,** *Linear Logistic Test Model***,** propuesto por Fischer en 1973) que es una extensión del modelo de Rasch. Para ello, se requiere un modelo cognitivo que dé soporte a cada contenido y, por lo tanto, a los ítems generados. Es decir, el modelo se sustenta en una Teoría Fuerte.

El segundo modelo se basa en que los ítems de una plantilla se agrupan en familias, con el objetivo de estimar los parámetros del modelo a nivel familia. Los procedimientos más desarrollados son dos: el modelo de hermanos idénticos (*Identical Siblings Model*, ISM) y el modelo de hermanos relacionados (*Related Siblings Model*, RSM). El ISM, de Hombo y Dresher (2001), asume una función de respuesta única para todos los ítems-hijo de una misma familia. Este modelo contiene ciertas limitaciones porque no considera las variaciones dentro de una misma familia. Glass y Van der Linden (2003) propusieron el RSM con el objeto de resolver este problema, por medio de la incorporación de una estructura asociada entre los ítems de una misma familia. El RSM se aplica, fundamentalmente para pruebas adaptativas, donde la habilidad de los examinados juega un rol primordial. Este análisis está enfocado en el estudio de los ítems-hijo como entidades isomorfas de una misma familia y no a la estructura del examen completo, con un número establecido de ítems.

El tercer modelo combina las aproximaciones de LLTM y RSM en otra denominada Modelo Lineal de Clonación de Ítems (LICM, por sus siglas en inglés), desarrollado por Geerlings, Glas y Van der Linden (2011). Los autores utilizaron un modelo de ojiva normal de tres parámetros para especificar la probabilidad de responder correctamente a un ítem. Por las razones anteriores, se infiere que esta metodología también debe utilizarse para una GAI de Teoría Fuerte.

Examen de Competencias Básicas (Excoba)

El Examen de Competencias Básicas (Excoba) es un examen estandarizado de alto impacto que se utiliza para seleccionar a los estudiantes que aspiran ingresar a la Educación Media Superior (EMS) y Educación Superior

(ES) en México. Esta prueba tiene sus antecedentes en el Examen de Habilidades y Conocimientos Básicos (Exhcoba), examen de gran escala, de opción múltiple y que se administra vía computadora (ver Backhoff y Tirado, 1992; Backhoff, Ibarra y Rosas, 1995). Aunque el Excoba conserva la filosofía de su antecesor, en el sentido de evaluar los aprendizajes básicos y esenciales que los estudiantes adquieren durante su trayecto escolar, su estructura y conformación es totalmente distinta e innovadora.

En cuanto a su estructura, el Excoba está alineado al currículo nacional, por lo que evalúa competencias académicas básicas que se precisan en los planes de estudio de la educación obligatoria. Asimismo, presenta una propuesta innovadora en relación con la forma en que se deben evaluar dichas competencias escolares, pues se aleja del formato de opción múltiple y se acerca a formas más "auténticas o naturales" de evaluar los aprendizajes.

La versión del Excoba para el ingreso a la EMS (Excoba/MS) mide los aprendizajes esperados de los planes y programas de estudio nacionales de la educación básica, a través de una cantidad fija de ítems: 120 en total (40 del nivel de primaria y 80 de secundaria). La tabla 1 muestra la relación y número de competencias que evalúa el examen. En el nivel de primaria se incluyen las competencias matemáticas y de lenguaje, y en el nivel de secundaria, las competencias de matemáticas, lenguaje, ciencias naturales (biología, física y química) y ciencias sociales (historia, geografía y civismo).

Tabla 1 - *Número de reactivos de las competencias académicas básicas que conforman la estructura del Excoba/MS*

Competencias	Primaria	Secundaria	Total
Matemáticas	20	20	40
Lenguaje	20	20	40
Ciencias Naturales	-	20	20
Ciencias Sociales	-	20	20
Total	40	80	120

Nota: La competencia de Matemáticas de primaria se denomina Habilidades matemáticas y en secundaria, Matemáticas. La competencia de Lenguaje en primaria se denomina Habilidades del lenguaje y en secundaria, Español.

Como se mencionó, los reactivos del Excoba/MS no son de opción múltiple, al menos en el sentido tradicional de entender este formato, donde el estudiante tiene que seleccionar una opción de varias alternativas que se le presentan. Entre los distintos tipos de ítems del Excoba, destacan los siguientes: (1) respuesta construida: se escribe literalmente una solución numérica o algebraica; (2) respuesta semiconstruida: se ubican o reubican elementos gráficos o conceptuales en mapas, gráficas, esquemas, planos o formatos (e.g.: ubicar coordenadas geográficas en un plano); y (3) selección múltiple-múltiple: se construye la respuesta seleccionando tres o más opciones (e.g.: seleccionar elementos que forman una categoría).

Algunos de estos reactivos del Excoba se califican de manera dicotómica (correcto-incorrecto), este es el caso de los ítems de respuesta construida. Otros se califican con el modelo de crédito parcial, según el número de respuestas que exige el reactivo; este es el caso de los reactivos de respuesta semiconstruida y de selección múltiple-múltiple. La máxima puntuación que se le otorga a un reactivo es de una unidad. En el caso del crédito parcial, la unidad se divide proporcionalmente entre el número de elementos que contiene el reactivo; por ejemplo, si el reactivo solicita ubicar cuatro fracciones en una recta numérica, cada fracción bien ubicada tiene un valor de 0.25 puntos.

Generador Automático de Ítems del Excoba

Para elaborar los reactivos del Excoba se desarrolló el GenerEx, de acuerdo con las necesidades evaluativas de los distintos contenidos curriculares de la educación básica. Por consiguiente, el GenerEx pertenece a los generadores de la Teoría Débil, ya que no se sustenta en una teoría cognitiva particular que explique de manera detallada los procesos cognitivos de cada competencia académica que se evalúa.

Como se comentó, la GAI requiere que se elaboren *modelos de ítems* en los que se precisen, al menos, los siguientes elementos: 1) una definición de la competencia a evaluar, 2) una estrategia específica para evaluar dicha competencia, y 3) una *plantilla*, con las reglas y elementos (conceptuales o gráficos). Estas plantillas permitan producir una familia de reactivos, que contiene al menos un *ítem-padre*, de donde se deriven diferentes *ítems-hijo*. La idea central es que el modelo de ítem sea capaz de generar de manera automática una gran cantidad de ítems-hijo con los cuales evaluar a los estudiantes, de manera consistente, una misma competencia escolar. Para ello se requiere que los ítems-hijo tengan propiedades conceptuales y métricas equivalentes.

La figura 1 muestra, a través de un ejemplo, cómo se estructura una familia de ítems con el GenerEx. En esta figura se presenta, parcialmente, la competencia de *Representación de fracciones* con un solo ítem-padre (*Seleccionar las partes de una figura geométrica, que indica la fracción*). De este ítem padre se derivan varios ítems-hijo, compuestos de diversas fracciones de distintas figuras geométricas (cuadrado, pentágono y triángulo).

En cada modelo de ítem se debe especificar el contexto en que se presenta el reactivo, la acción que debe realizar el estudiante para poderlo responder y las reglas para generar de manera automática cada reactivo-hijo. En el ejemplo anterior, todos los ítems-hijo solicitan al estudiante *seleccionar las partes de una figura geométrica que corresponden a una fracción dada*. Este modelo de ítem posibilita seleccionar distintas figuras geométricas y, a su vez, para cada figura se pueden elegir distintas fracciones. Esta selección puede ser aleatoria o fija, según se desee.

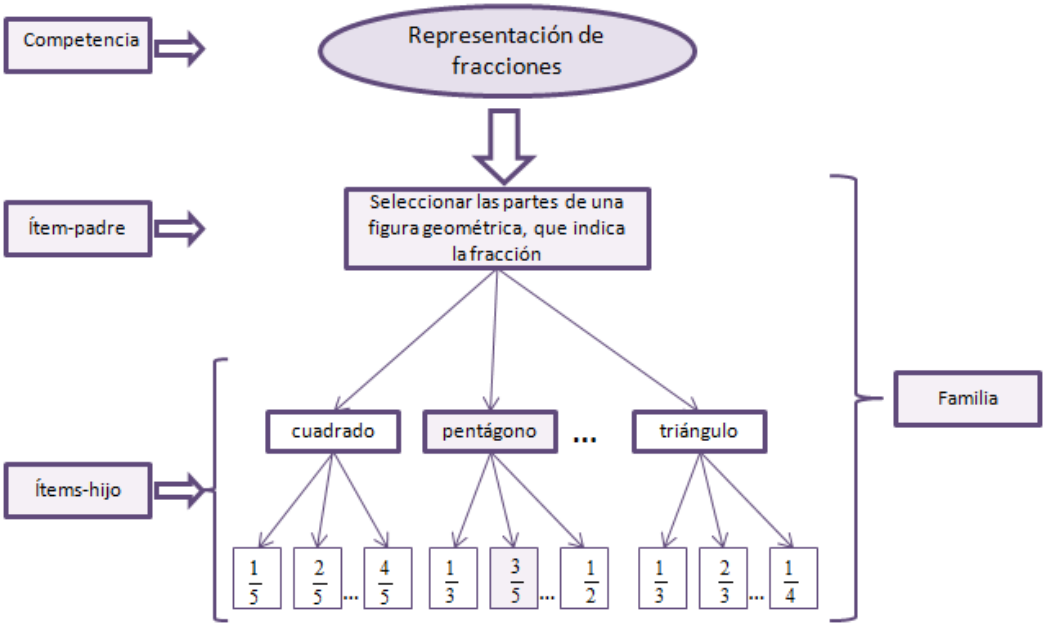

Figura 1. Familia de reactivos de la competencia *Representación de fracciones*, del área de Matemáticas

La figura 2 muestra gráficamente un reactivo-hijo que se puede producir con el GenerEx. En este caso se presenta un pentágono subdividido en partes iguales y se le pide al estudiante que marque las partes que representan la fracción 3/5.

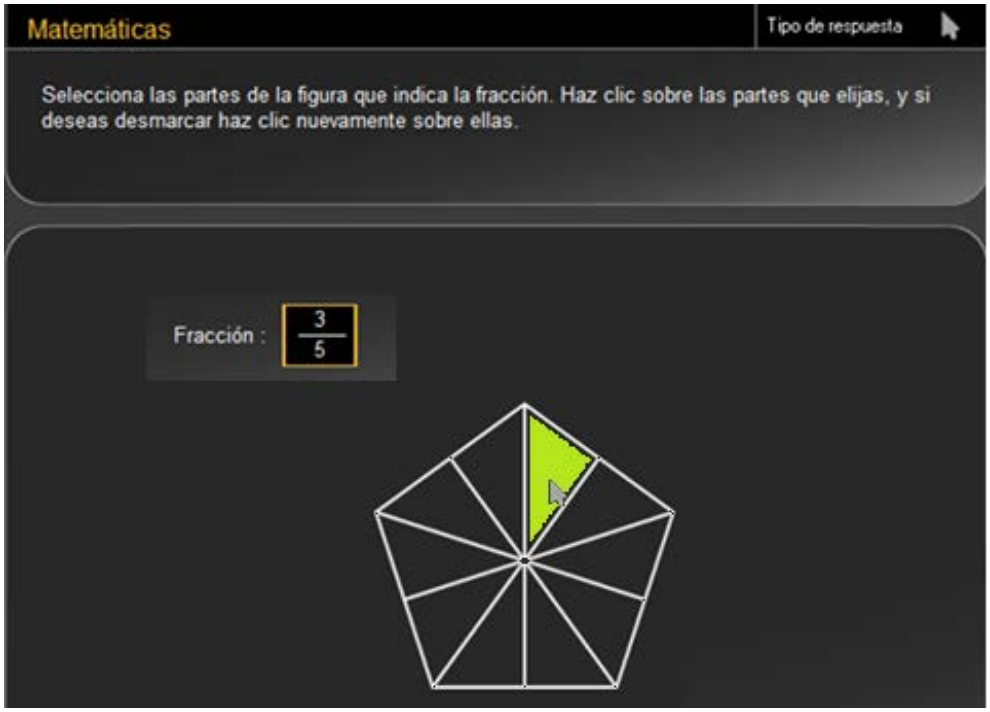

Figura 2. Ejemplo de ítem-hijo para la competencia *Representación de fracciones*, del área de matemáticas

Planteamiento del problema

Los generadores automáticos de ítems, como es el caso del GenerEx, producen decenas, o centenas, de reactivos conceptualmente equivalentes, con los cuales se abre la posibilidad de construir cientos o miles de exámenes paralelos. Si los modelos de reactivos están bien diseñados, poseerán propiedades psicométricas equivalentes, y los exámenes que se construyan con dichos reactivos tendrán estructuras internas similares.

Como se mencionó, la GAI plantea nuevos retos a la psicometría y, seguramente, el más importante es la forma de asegurar su validez, ya que sería imposible conocer las propiedades psicométricas de todos los ítems-hijo que es factible generar con el GenerEx y menos aún conocer la estructura interna de todas las versiones posibles de construir con la combinación de los 120 modelos de ítems, para el caso de la educación media superior.

Por consiguiente, este trabajo tiene el propósito de proponer una forma de estudiar la validez del GenerEx y de mostrar ejemplos de los resultados que se obtienen con esta metodología. Particularmente, nos propusimos aportar evidencias de validez de este generador a dos niveles: para los exámenes que se generan y para cada uno de los modelos de reactivos (con los que se construyen los ítems).

Método

La aproximación metodológica para estudiar la validez del GenerEx consistió en efectuar una serie de estudios comparativos de los reactivos, a tres niveles: 1) *nivel examen*, con el objeto de indagar sobre las medidas de tendencia central y dispersión de la prueba, su confiabilidad, así como comparar diferentes versiones del examen y estudiar el comportamiento de las seis áreas temáticas (que componen el examen), 2) *nivel familia de reactivos*, para analizar y comparar los ítems de una misma competencia, estudiar sus propiedades psicométricas y observar si se agrupan en el constructo o rasgo latente correspondiente, y 3) *nivel de elementos*, donde se estudian los componentes de los ítems para decidir acerca de su calidad.

Lo anterior se hizo con las herramientas que aportan la Teoría Clásica de los Tests (TCT), la Teoría de Respuesta al Ítem (TRI) y el Análisis Factorial Confirmatorio (AFC). En particular, los análisis basados en la TRI se realizaron con el modelo de Rasch clásico para datos

dicotómicos (1961) y con el modelo de Rasch para ítems de crédito parcial (Masters, 1982).

Dada la extensión que implica abordar los tres tipos de análisis referidos, este documento se limita al primer nivel, que corresponde a la validez del examen en general.

Muestra de reactivos y estudiantes

Para realizar el estudio comparativo del GenerEx se generaron dos exámenes paralelos, compuestos por seis áreas temáticas y 120 reactivos, que denominaremos versión A (VA) y versión B (VB). Los reactivos de las dos versiones del Excoba/MS se obtuvieron al azar procurando, en la medida de lo posible, que los ítems que pertenecieran a la misma familia de reactivos no presentaran elementos comunes.

La figura 3 muestra la composición general del examen, para el ingreso a bachillerato, que en ambas versiones fue la misma, pero con ítems-hijo diferentes para cada competencia. En esta figura se puede apreciar que el Excoba/MS mide dos grandes competencias, que se aprenden en el nivel de primaria (Habilidades del lenguaje y Habilidades matemáticas) y cuatro del nivel de secundaria (Español, Matemáticas, Ciencias Naturales y Ciencias Sociales). A su vez, cada competencia está compuesta de 20 modelos de reactivos (por ejemplo, HV01, HV02…HV20) [2] y cada modelo de reactivo puede generar una cantidad muy grande de ítems conceptualmente equivalentes (e.g.: de HV1 surgen I1, I2…In). Para este estudio, solo se generó un ítem por modelo de reactivo, con lo cual se formó un examen de 120 ítems (20 ítems por dominio).

Composición del Excoba /MS

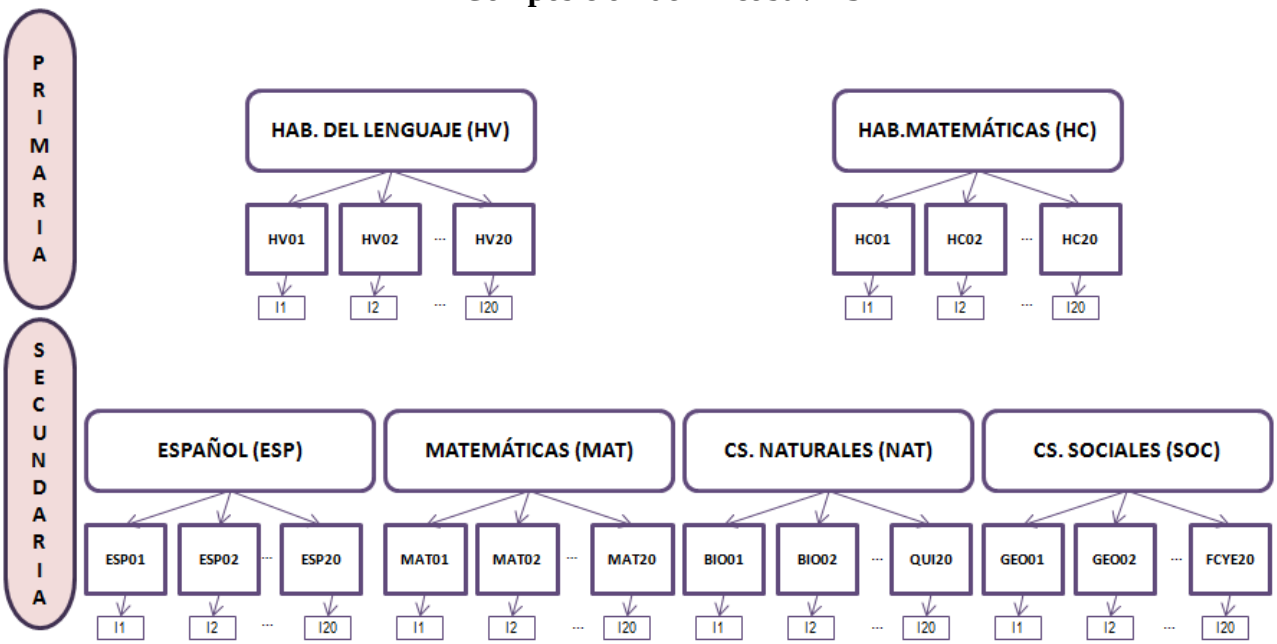

Figura 3. Composición del Excoba/MS

Las dos versiones del examen se aplicaron a grupos de estudiantes aspirantes a ingresar a la Preparatoria Federal Lázaro Cárdenas (PFLC), ubicada en la ciudad de Tijuana, Baja California. La PFLC surgió en 1946, como la primera institución de educación media superior de Tijuana, actualmente cuenta con más de 4,000 estudiantes y es una de las escuelas más prestigiosas del estado, con los mejores resultados en la Evaluación Nacional del Logro Académico en Centros Escolares (ENLACE), según lo atestigua la información del portal de la escuela (http://dir.lazarocardenas.edu.mx/). El número de alumnos que respondió la versión A fue de 401, mientras que 299 contestaron la versión B. La edad de los estudiantes fluctuó entre 15 y 16 años; 59% eran mujeres y 41% hombres. El

promedio de calificaciones de la educación secundaria (en una escala del 5 al 10) de estos jóvenes fue de 9.13 con una desviación estándar de 0.57. La participación fue voluntaria y alentada por la institución, a la cual se le devolvieron los resultados para fines propedéuticos. La selección de quiénes resolvían una u otra versión del examen, fue al azar.

Análisis de resultados

Las versiones A y B del Excoba/MS se analizaron y se compararon a dos niveles.

1) Para el examen completo (conformado por 120 reactivos) se realizaron los siguientes cálculos: distribución de frecuencias y normalidad; medidas de tendencia central y dispersión; sesgo y curtosis. Asimismo, para analizar la consistencia interna se obtuvieron los índices de correlación punto-biserial y el coeficiente de Alpha de Cronbach. Con el modelo Rasch se calcularon los índices de ajuste, de correlación ítem medida y de discriminación, así como el Mapa de Wright.
2) Para cada una de las seis áreas temáticas (20 reactivos) que componen el examen se obtuvieron los siguientes indicadores: dificultad, correlación punto-biserial y confiabilidad; medida, nivel de ajuste (interno y externo), correlación punto-medida y discriminación; además los índices y las cargas factoriales correspondientes a la agrupación de ítems para cada área de la prueba.

Los análisis estadísticos se realizaron con la ayuda de los programas: SPSS 17.0 (SPSS, 2008), Winsteps (Linacre, 2010) y EQS 6.1 (Bentler, 2006). En la tabla 2 se muestran los criterios y límites establecidos para evaluar la calidad del examen en su conjunto y de los ítems en lo individual, de acuerdo con el modelo psicométrico utilizado. Por ejemplo, la correlación punto-biserial mínima aceptable de los reactivos se estableció en 0.2; la confiabilidad (a) mínima de las áreas temáticas del examen (con 20 reactivos) debió ser al menos de 0.6, mientras que la confiabilidad de toda la prueba (120 reactivos) tuvo que ser igual o mayor a 0.9.

Tabla 2 - *Criterios asumidos para los análisis estadísticos de los ítems de las muestras del Excoba/MS*

Modelos psicométricos	**Estadísticos**	**Número variables**	**Criterio**	
			Aceptable	Bueno
Teoría Clásica de los Test	Correlación punto biserial		≥ 0.2	
	Alfa (a)	20	≥ 0.6	
		120	≥ 0.9	
Teoría de Respuestas al Ítem	Correlación punto medida		≥ 0.2	
	Infit-Outfit MNSQ		≥ 0.8 y ≤ 1.3	
	Discriminación		≥ 0.8	
Análisis Factorial Confirmatorio	Carga factorial		≥ 0.20	≥ 0.30
	c^2		≥ 0.01	≥ 0.05
	NNFI		≥ 0.90	≥ 0.95
	CFI		≥ 0.90	≥ 0.95
	RMSEA		< 0.08	< 0.05

Resultados

El total de modelos de ítems evaluados del Excoba/MS fue de 117, de los 120 originales. Esta reducción obedeció a que se descartaron tres modelos de ítems: uno de lenguaje (HV19), otro de historia (HIS06), los cuales presentaron un problema de diseño, y el modelo restante, de matemáticas (MAT14), se suprimió debido a inconvenientes para decodificar las respuestas de los estudiantes.

Versiones completas del examen

A continuación se describen y se comparan las propiedades métricas de las dos versiones completas del Excoba/MS. Primero, se presentan las medidas de tendencia central, de dispersión, normalidad y confiabilidad, para

terminar con el mapa de Wright (donde se incluye el empate que tienen las habilidades de los estudiantes con la dificultad de los reactivos), el porcentaje de varianza que explica cada prueba, los índices de ajuste, así como de correlación ítem-medida y de discriminación.

La tabla 3 muestra que ambas versiones tienen indicadores muy similares[3]. Las medias de respuestas correctas (dificultades) resultan similares, para VA es de 60.9 (p = 0.52) y para VB, de 58.1 (p = 0.50); sus dispersiones son prácticamente iguales, lo mismo que su simetría. En cuanto a la curtosis, la distribución de VA es ligeramente leptocúrtica (0.18), mientras que VB tiende a ser platocúrtica (-0.25). Sin embargo, los errores de medida son grandes, lo cual indica que las diferencias entre estos valores no son significativas. Por lo tanto, estos datos inducen a afirmar que se trata de distribuciones normales. Adicionalmente, el promedio de correlación punto-biserial de los reactivos fue de 0.26 para la VA y 0.25 para la VB. La confiabilidad de ambas versiones de la prueba es la misma (a = 0.90), lo que refuerza la premisa de semejanza entre ambas versiones.

Tabla 3 - *Indicadores de tendencia central, dispersión, normalidad y confiabilidad de VA y VB del Excoba/MS*

Indicador	Versión A	Versión B
N	163	119
Media	60.9	58.1
Desviación Estándar	10.6	10.4
Rango	35 - 95	35 - 90
Simetría	0.31	0.33
Curtosis	0.18	-0.25
Correlación punto biserial	0.26	0.25
Confiabilidad	0.90	0.90

La figura 4 muestra el mapa de Wright de las dos versiones completas del Excoba/MS. En este mapa se aprecian las distribuciones de las dificultades de los reactivos versus las habilidades de los estudiantes. Se observa que para ambas versiones la media de las dificultades de los ítems es mayor (casi en una desviación estándar) que la media de las habilidades de los examinados. El área más compleja fue Matemáticas del nivel de secundaria (MAT); el área más fácil fue Español, también de secundaria (ESP); le siguió Habilidades del lenguaje de primaria (HV). En los dos exámenes, Habilidades Matemáticas (HC) recorrió la gama de dificultades que va de -2 a 3 lógitos; mientras que Ciencias Sociales (GEO, HIS, FCYE) y Ciencias Naturales (BIO, FIS, QUI) se ubicaron en el rango de -1 a 1. En general, si bien existen reactivos que superan las habilidades de los estudiantes evaluados, la mayoría de las dificultades los ítems se empatan con las habilidades.

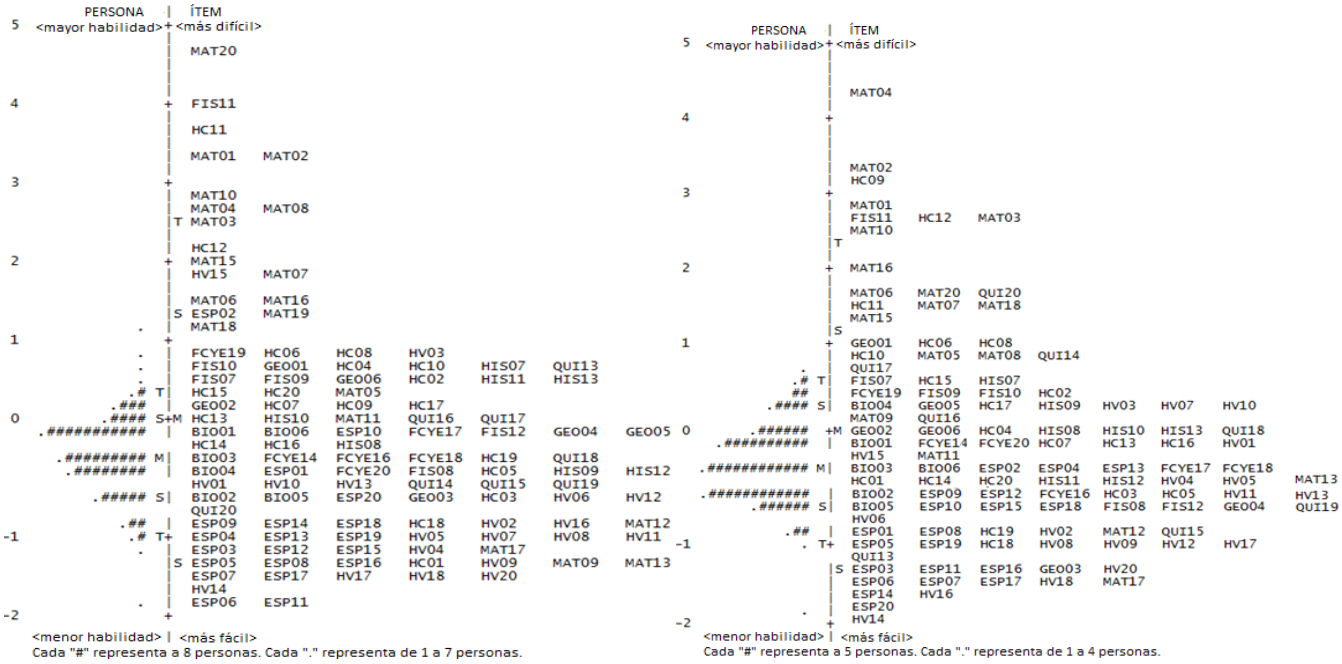

Figura 4. Mapas de Wright de las versiones A (izquierda) y B (derecha) del Excoba/MS

Con base en el análisis Rasch, no se identificaron desajustes serios en ningún reactivo; esto indica una ausencia de problemas de aleatoriedad y de determinismo, tanto cerca como lejos de la zona de medición de cada ítem. Sin embargo, se detectaron algunas deficiencias de correlación o discriminación en los reactivos HV15 y HC07 en ambas versiones; en la versión A se encontraron problemas similares en los reactivos ESP01, QUI13 y QUI16, mientras que en la versión B los reactivos con algunos problemas fueron HC09, ESP02, MAT07, QUI14 y QUI20.

Finalmente, las medidas (habilidades de los estudiantes y dificultades de los ítems) explican el 38.5% y el 37.3% de la varianza para las versiones A y B, respectivamente. El índice promedio de correlación punto-medida fue de 0.29 para la versión A, y 0.28 para versión B. Estos resultados indican valores aceptables y muy parecidos para dos exámenes que representan a un mismo constructo o rasgo latente.

Comparación por área temática del examen

La tabla 4 muestra el comportamiento psicométrico de las seis áreas de las dos versiones del examen. En negritas se marcan los valores fuera del rango deseado. En ambas versiones, las áreas con mayor y menor dificultad son las relacionadas, respectivamente, con Matemáticas y con el lenguaje. Por otro lado, todas las áreas temáticas presentaron muy pocas diferencias en sus niveles de dificultad, siendo la más grande de 0.05, para el caso de Habilidades de lenguaje.

Tabla 4 - *Resultados de los análisis con la TCT del Excoba/MS para las versiones A y B, por áreas temáticas*

Área temática	k	Versión A				Versión B			
		n	*p*	ptbis	α	n	*p*	ptbis	α
H. de lenguaje	19	289	0.68	0.24	0.633	189	0.63	0.21	**0.547**
H. matemáticas	20	396	0.38	0.34	0.784	301	0.38	0.35	0.788
Español	20	290	0.69	0.21	**0.587**	270	0.66	0.30	0.706
Matemáticas	19	396	0.26	0.26	0.655	296	0.24	0.27	0.691
Ciencias naturales	20	289	0.48	0.22	0.612	216	0.47	**0.16**	**0.502**
Ciencias sociales	19	397	0.47	0.48	0.869	298	0.48	0.48	0.877

Nota: k = número de ítems, n = tamaño de la muestra, *p* = dificultad promedio, Ptbis = correlación punto biserial, α = Alpha de Cronbach. En negritas se marcan los valores fuera del rango deseado.

En cuanto al poder de discriminación, medido por el índice de correlación punto-biserial (ptbis) de cada reactivo, se observan puntuaciones promedio similares para la mayoría de las áreas temáticas, con excepción de Español y Ciencias Naturales. En la VB de esta última área temática, no se alcanzó el mínimo solicitado de 0.20. Finalmente, en la confiabilidad de las áreas temáticas, medida por el Alpha de Cronbach, se aprecian variaciones considerables entre ambas versiones en Habilidades del lenguaje (0.09 puntos), Español (0.11 puntos) y Ciencias Naturales (0.11 puntos). Es importante también señalar que las áreas temáticas con mejor confiabilidad fueron Ciencias Sociales y Habilidades matemáticas.

En la tabla 5 se muestran los resultados de los análisis efectuados con el modelo de Rasch. De izquierda a derecha, podremos apreciar en la comparación de ambas versiones que la cantidad de varianza explicada (Imed) varía de un área temática a otra. Las áreas que presentan mayor divergencia en la cantidad de varianza que explican a través de las medidas (habilidades y dificultad), son Matemáticas (diferencia de 18.2 puntos), Español (11.7 puntos) y Habilidades del lenguaje (8.7 puntos); las que presentan menores diferencias son Ciencias Sociales (0.5), y Habilidades matemáticas (0.7). Por otro lado, los promedios de las correlaciones ítem-medida de las distintas áreas temáticas son muy parecidos entre sí y en algunos casos iguales, siendo el área de Español la que mayor diferencia muestra (0.06).

Tabla 5 - *Resultados de los análisis desde el modelo de Rasch del Excoba/MS para las versiones A y B, por áreas*

Área temática	k	Versión A						Versión B					
		n	Var(%)	*Imed*	Problemas			n	Var	*Imed*	Problemas		
					in	out	C/D				in	out	C/D
H. del lenguaje	19	399	38.5	0.35			HV15	298	47.2	0.33			HV15
H. matemáticas	20	401	32.9	0.42	HC07	HC07	HC07	301	32.2	0.42		HC07	HC09
Español	20	398	39.0	0.33				297	27.3	0.39			
Matemáticas*	19	400	75.9	0.35		M9,10,12		300	57.7	0.37			M18
Cs. naturales	20	380	27.6	0.37		FIS11		273	34.5	0.34			
Cs. sociales	19	397	38.6	0.50				298	39.1	0.50	G04	G04	

Nota: (*) en VB solamente pudieron analizarse 18 ítems, debido a que para MAT19 no hubo respuestas correctas.
k = número de ítems, n = tamaño de la muestra, Var = porcentaje de varianza explicada, Imed = promedio de correlación ítem medida, in = infit, out = outfit, C/D = correlación ítem medida y/o índice de discriminación. HC = habilidades matemáticas, HV = Habilidades del lenguaje, M = Matemáticas, FIS = Física, G = Geografía.

Para ambas versiones del Excoba, los ítems que tuvieron algún problema de ajuste (*infit* u *outfit*) fueron ocho, de un total de 117 reactivos. Los reactivos HV15 y HC07 mostraron un comportamiento deficiente en las dos versiones del examen. El resto, HC09, MAT09, MAT10, MAT12, FIS11 y GEO04, ocurrió solamente en una de las pruebas. La mayoría de los problemas se relacionaron con valores altos del indicador *outfit*. Las dos situaciones que se registraron de infit también se dieron por superar el intervalo de ajuste. Estos valores indican demasiada aleatoriedad lejos de la zona de medición del ítem, para el primer caso, y cerca, para el segundo caso.

Para conocer la agrupación de reactivos en cada una de las áreas temáticas, se realizaron análisis factoriales confirmatorios para ambas versiones del Excoba/MS, buscando los modelos que mejor ajustan. La tabla 6 muestra, para las seis áreas temáticas, los indicadores de ajuste y el número de factores identificados en cada caso. En esta tabla se observa que en todos los casos los modelos de agrupación presentan indicadores de ajuste buenos y muy similares entre las versiones A y B. En las áreas de Habilidades matemáticas, Español y Ciencias Sociales se identificó un sólo factor, mientras que en las de Habilidades del lenguaje, Matemáticas y Ciencias Naturales se encontraron dos factores que covarían.

Tabla 6 - *Modelos de agrupación de ítems de las seis áreas temáticas del Excoba/MS, para VA y VB*

Ajuste	H. del lenguaje		H. matemáticas		Matemáticas		Español		Cs. Naturales		Cs. Sociales	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
p	0.79	0.56	0.00	0.03	0.39	0.31	0.08	0.06	0.81	0.09	0.02	0.00
NNFI	1.05	1.02	0.93	0.94	0.98	0.98	0.96	0.91	1.07	0.81	0.97	0.95
CFI	1.00	1.00	0.94	0.95	0.98	0.98	0.97	0.95	1.00	0.84	0.98	0.97
RMSEA	0.00	0.00	0.03	0.03	0.01	0.02	0.02	0.03	0.00	0.03	0.03	0.04
No. factores	2		1		2		1		2		1	
	HV15		HC07				ESP02		BIO01 y QUI19			
Ítems sin carga significativa	HV02 HV03 HV08 HV05 HV16			HC09	MAT16 MAT19	MAT07	ESP01 ESP18		QUI13 QUI16	BIO04 QUI14 QUI18 QUI20		

Los dos factores de Habilidades del lenguaje se relacionan con la lectura y comprensión de textos, así como con la gramática y ortografía. Para el caso del área de Matemáticas se hallaron los factores relacionados, por un lado, con el sentido numérico, pensamiento algebraico y manejo de la información y, por otro lado, con la forma, espacio y medida. Finalmente, en Ciencias Naturales se encontraron factores relacionados, por un lado, con la biología y química y, por el otro, con la física.

Además, en cinco de las seis áreas analizadas, se identificaron reactivos que no presentaron cargas significativas en el modelo respectivo. Para ambas versiones se detectaron los ítems: HV15, HC07, ESP02, BIO01 y QUI19. Las áreas de Habilidades del lenguaje y Ciencias naturales resultaron las menos coincidentes, puesto que se encontraron dos o más ítems con problemas en una sola de las versiones.

A manera de ejemplo de cómo se distribuyen las cargas factoriales de los ítems en un área temática del examen, se presenta el caso de las dos versiones de Habilidades matemáticas. La figura 5 muestra los 20 reactivos de cada versión, donde podemos identificar la semejanza en sus cargas factoriales, así como los reactivos cuyas cargas son insuficientes de acuerdo con nuestros criterios (menores que 0.2). En este caso se encuentran el ítem HC07 (en sus dos versiones) y el ítem HC09 de VB.

Un análisis cualitativo de los modelos de reactivos de esta área permitió identificar que para el caso de HC07 la actividad a desarrollar apela al reconocimiento y a la memoria (reconocer los elementos marcados en una circunferencia); lo cual difiere del resto de las competencias de matemáticas de educación primaria, puesto que la mayoría incluye niveles de comprensión y aplicación. Para el caso de HC09 se compararon los dos ítems y, si bien ambos solicitan el cálculo del área de un triángulo, en un ejercicio se muestra la altura del lado dentro de la figura (VA) y en el otro (VB), fuera de la misma. Suponemos que esta última imagen puede ser confusa, e incluso desorientar al estudiante acerca del cálculo a realizar.

Otra observación importante es que 14 ítems de VA y 16 de VB tienen cargas factoriales mayores a 0.30. Estos resultados nos indican que los dos tests que se generaron automáticamente tienen un comportamiento muy similar, lo que aporta a la validez de los exámenes obtenidos por el GenerEx.

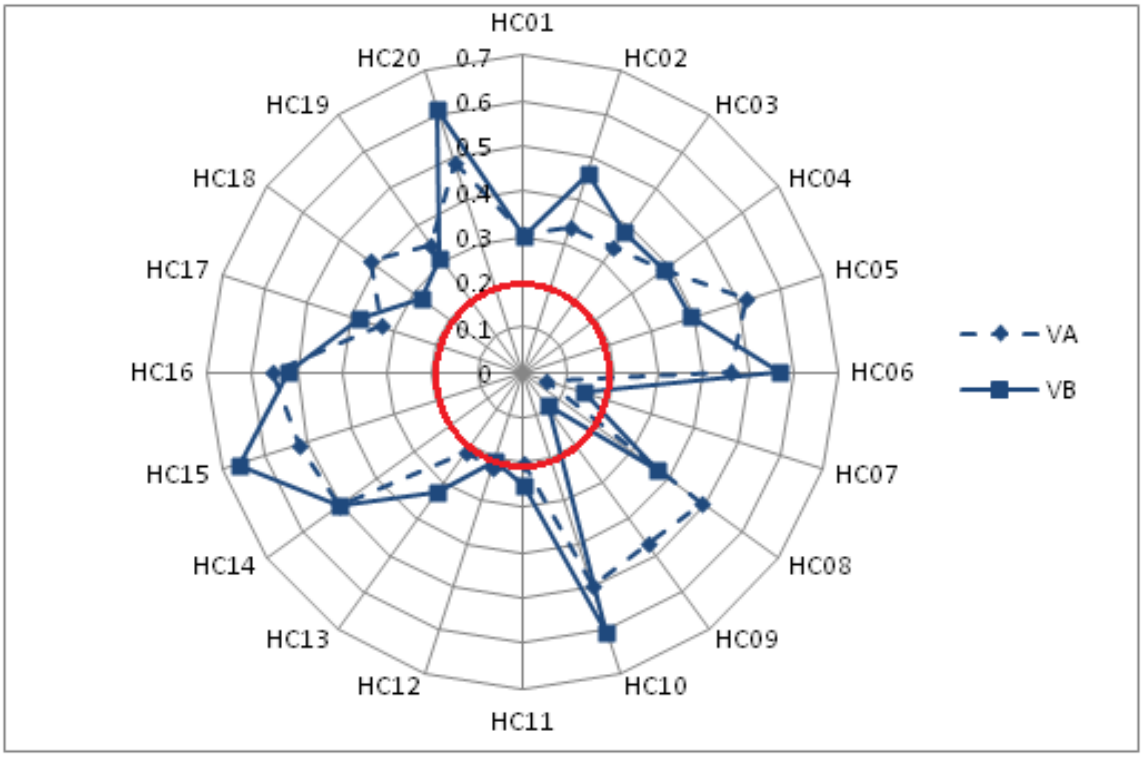

Figura 5. Distribución de las cargas factoriales de Habilidades matemáticas de dos versiones del Excoba/MS

Discusión y conclusiones

Los generadores automáticos de ítems representan un avance considerable en el ámbito de la evaluación psicológica y educativa, ya que permiten diseñar y construir una cantidad importante de reactivos equivalentes, conceptual y psicométricamente (Gierl & Haladyna, 2012). Los GAI resuelven el problema de tener que elaborar pruebas únicas periódicamente, debido al rápido desgaste que sufren cuando se utilizan de manera masiva e intensiva, como es el caso de los exámenes de admisión. El desarrollo de la GAI ha pasado por varias etapas, desde los que se basan en diseño de plantillas de ítems hasta los modelos que se fundamentan en marcos conceptuales cognitivos (Haladyna, 2012). Sin duda, las nuevas herramientas estadísticas y la evaluación asistida por computadora impulsarán fuertemente la consolidación de la GAI en un futuro próximo.

Por lo pronto, el uso de la GAI para propósitos prácticos impone retos inéditos a la psicometría, debido a la necesidad de encontrar formas eficaces y económicas al problema de contar con evidencias de validez de los reactivos y exámenes que se producen de manera automática. No sería factible pensar que el proceso de validación se llevara a cabo con cada uno de los cientos de reactivos generados y miles de exámenes construidos. Por consiguiente, nos propusimos encontrar un método que diera respuesta a esta problemática, tomando como referencia el GenerEx que se fundamenta en una Teoría Débil (Gierl, Zhou & Alves, 2008), ya que no se basa en un modelo de tareas que precise las estructuras cognitivas de las competencias académicas que evalúa, sino que se sostiene en los aprendizajes esperados que se marcan en el currículo mexicano (Ferreyra, 2014; Pérez-Morán, 2014).

Este trabajo descansa en el supuesto de que la evaluación de cada competencia está estructurada por modelos de ítems, que definen familias, de donde se obtienen ítems-hijo que constituyen una versión de examen. En principio, cada modelo precisa y controla tanto la competencia a evaluar como la dificultad de los reactivos que genera, de tal modo que las tareas evaluativas posean parámetros similares para evaluar una misma competencia (Bejar,

2002; Gierl & Lai, 2011; Gierl, Zhou & Alves, 2008).

El método propuesto consideró básicamente tres niveles de análisis: a nivel macro, de los exámenes generados; a nivel meso, de las familias de reactivos, y; a nivel micro, de los ítems-hijo y sus elementos. Para el caso de la validez a nivel de examen (objeto de este trabajo), la idea central de esta propuesta metodológica consistió en comparar la equivalencia psicométrica de dos versiones paralelas del GenerEx, compuestas cada una por 120 reactivos generados al azar, que no tuvieran componentes en común, así como la equivalencia entre las áreas temáticas que conforman cada examen. Estas comparaciones psicométricas se realizaron a través de tres aproximaciones metodológicas complementarias, que se basaron en: la Teoría Clásica de los Test, la Teoría de Respuesta al Ítem (con el modelo de Rasch), y el Análisis Factorial Confirmatorio.

Los resultados del estudio muestran las semejanzas y diferencias de distintos indicadores psicométricos, fundamentales en la TCT, de las dos versiones del examen, como son las medidas de tendencia central, dispersión, normalidad y confiabilidad. Asimismo, se compararon algunos indicadores básicos de la TRI, como son los ajustes interno y externo, la correlación ítem-medida, la discriminación y la varianza explicada, así como la correspondencia entre la dificultad de los reactivos y la habilidad de los estudiantes.

También se analizaron las áreas temáticas (seis en total) de las dos versiones del examen, donde se calcularon las dificultades media y las correlaciones punto-biserial del Excoba/MS. Asimismo, se obtuvieron los indicadores psicométricos básicos de cada área temática con base en la TRI. Finalmente, se realizaron los AFC de las dos versiones de las seis áreas temáticas del Excoba, para comparar los modelos de agrupación de ítems y las cargas factoriales que aporta cada reactivo en los modelos respectivos.

Con los tres tipos de análisis efectuados también se pudieron identificar, para las dos versiones del examen y para las seis áreas temáticas, los ítems que presentan un comportamiento fuera de los rangos aceptables, de acuerdo con los parámetros definidos en este estudio (ver tabla 2).

En síntesis, la metodología desarrollada permitió obtener una buena descripción del funcionamiento del GenerEx y de la validez interna de dos versiones generadas al azar, con herramientas estadísticas básicas. Los resultados se pueden complementar muy bien con un análisis cualitativo de los problemas detectados. Este generador de ítems produce exámenes y reactivos psicométricamente similares, aunque también presenta problemas en algunas áreas temáticas y en ciertos reactivos particulares. Por lo general, las deficiencias detectadas se presentan en ambas versiones del examen, aunque también se identificaron diferentes desajustes entre una versión y otra, lo que nos habla de un problema de definición de contenidos que alimentan al GenerEx, que habrá que estudiar conceptualmente con mayor detalle.

Para terminar, hay que decir que el proceso de validación de cualquier instrumento de medición debe ser permanente, y que los desarrollos en el campo de la GAI son novedosos a nivel mundial. En consecuencia, se divisa un ámbito muy interesante y fecundo de investigaciones en torno a los métodos de validación interna de los generadores automáticos de ítems.

Referencias

Backhoff, E. & Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior*, *21* (3), 95-118. Recuperado de http://metrica.edu.mx/wp-content/uploads/2014/10/1992_Desarrollo_del_EXHCOBA.pdf

Backhoff, E., Ibarra, M. y Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología, 12* (1), 55-62.

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

Bejar, I. I. (2002). Generative testing: From conception to implementation. En S.H. Irvine & P.C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NY: Erlbaum.

Bentler, P. M. (2006). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software, Inc.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64* (4) 407-433. doi: http://dx.doi.org/10.1007/BF02294564

Ferreyra M. F. (2014). *Metodología para analizar la estructura interna de un generador automático de reactivos* (Tesis de doctorado no publicada). Universidad Autónoma de Baja California, Ensenada, México.

Geerlings, H., Glass, C. A. W. & van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76* (2), 337-359. doi: http://dx.doi.org/10.1007/s11336-011-9204-x

Gierl, M. J. & Haladyna, T. M. (2012). Automatic item generation: an introduction. En M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 3-12). Nueva York: Routledge.

Gierl, M. J. & Lai, H. (April, 2011). The Role of Item Models in Automatic Item Generation. Paper Presented at the *Annual Meeting of the National Council on Measurement in Education*. New Orleans, LA.

Gierl, M. J. & Lai, H. (2012). Using weak and theory to create item models for Automatic Item Generation: some practical guidelines with examples. En M. J. Gierl & T. M. Haladyna (Eds.). Automatic Item Generation: Theory and Practice. Nueva York: Routledge.

Gierl, M. J., Zhou, J. & Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *The Journal of Technology, Learning, and Assessment, (7)* 2.

Glas, C. A. W. & van der Linden, W. J. (2003). Computarized adaptive Testing with item cloning. *Applied Psychological Measurement, 27*, 247-261.

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. En M. J. Gierl & T. M. Haladyna (Eds.), *Automatic item generation*: *Theory and practice* (pp. 13-25). Nueva York: Routledge.

Haladyna, T. M. & Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions, 12*, 97-104. doi: http://dx.doi.org/10.1177/016327878901200106

Hively, W., Patterson, H. L. & Page, S. H. (1968). A "universe-defined" system for arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290. doi: http://dx.doi.org/10.1111/j.1745-3984.1968.tb00639.x

Holling, H., Bertling, J. P. & Zeuch, N. (2009). Automatic item Generation for probability word problems. *Studies in Educational Evaluation, 35*, 71-76. doi: http://dx.doi.org/10.1016/j.stueduc.2009.10.004

Hombo, C. & Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Paper presented at the annual meeting of the National Council on Measurement in Education, Seatle, Wa, EE. UU.

Linacre, J.M. (2010). *Winsteps® (Version 3.70.0.2)* [Computer Software]. Beaverton, Oregon: Winsteps.com

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149-174. doi: http://dx.doi.org/10.1007/BF02296272

Pérez-Morán, J. C. (2014). *Análisis del aspecto sustantivo de la validez de constructo de una prueba de habilidades cuantitativas* (Tesis de doctorado no publicada). Universidad Autónoma de Baja California, Ensenada, México.

Rasch, G. (1961). On General Laws and the Meaning of Measurement in Psychology. Proceedings of the *Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributions to Biology and Problems of Medicine*, 321-333. University of California Press: Berkeley, CA. Recuperado de http://projecteuclid.org/euclid.bsmsp/1200512895

Sinharay, S. & Johnson, M. (2012). Statistical modeling of Automatic Item Generation. En M. J Gierl & T. M. Haladyna (Eds.). *Automatic Item Generation: Theory and Practice*. N. Y., New York: Routledge.

SPSS Inc. (2008). *SPSS Statistics for Windows, Version 17.0*. Chicago: SPSS Inc.

NOTAS

[1] Modelo de ítem o plantilla son términos equivalentes, a efectos de este artículo. No debe confundirse con modelo de tarea que es el modelo cognitivo que subyace el rasgo que se evalúa

[2] Por razones históricas, los reactivos de Habilidades del lenguaje se abrevian con HV, haciendo referencia a Habilidades verbales, y los reactivos de Habilidades matemáticas se simplifican con HC, Habilidades cuantitativas. Habilidades verbales y Habilidades cuantitativas son los nombres que reciben las áreas de educación primaria en el EXHCOBA.

[3] Hay que hacer notar que el número de estudiantes que se consideró para esos análisis se redujo considerablemente, debido a que solamente se incluyeron los casos donde el estudiante tuvo una calificación en todos los ítems

AGRADECIMIENTOS

Este trabajo es parte de la investigación doctoral desarrollada por María Fabiana Ferreyra en la Universidad Autónoma de Baja California, gracias a la financiación del Conacyt-México (N° de Registro 247008).

Autores / Authors | **To know more / Saber más**

Ferreyra, Maria Fabiana (fferreyra@metrica.edu.mx).
Investigadora asociada en Métrica Educativa A.C. Ensenada, Baja California (México). Es Profesora de Matemáticas por el Instituto Nacional Superior del Profesorado Joaquín V. González, Buenos Aires (Argentina). Es la autora de contacto para este artículo. Maestra en Ciencias Educativas y doctora en Ciencias Educativas, ambos títulos obtenidos en el del Instituto de Investigación y Desarrollo Educativo de la Universidad de Baja California, México. Su campo de interés es el desarrollo y validación de pruebas de aprendizaje a gran escala, y la enseñanza de las matemáticas. Su dirección postal es: Métrica Educativa, Alvarado 921, Zona Centro. Ensenada, Baja California, C.P. 22800 (México)

Backhoff-Escudero, Eduardo (ebackoff@gmail.com).
Licenciado en psicología por la Universidad Nacional Autónoma de México, Maestro en Educación por la Universidad de Washington y Doctor en Educación por la Universidad Autónoma de Aguascalientes. Consejero de la Junta de Gobierno del Instituto Nacional para la Evaluación de la Educación. Ciudad de México, México. Su campo de interés es el desarrollo y validación de pruebas de aprendizaje de gran escala y la evaluación asistida por computadora. Ha sido Director de Pruebas y Medición del Instituto Nacional para la Evaluación de la Educación (INEE) de México. Actualmente se desempeña como Consejero de la Junta de Gobierno del INEE

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
E-Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]