# Validez de familias de ítems producidas por el Generador Automático de Ítems (GenerEx)

María Fabiana Ferreyra[1]
Norma Larrazolo Reyna[2]
Eduardo Backhoff Escudero[3]

## Resumen

Los Generadores Automáticos de Ítems son herramientas informáticas, desarrolladas con el propósito de producir familias de reactivos isomorfos. Con esta idea se creó el GenerEx, generador automático de reactivos del Examen de Competencias Básicas (Excoba). La Generación Automática de Ítems implica un gran avance para la evaluación psicológica y educativa; sin embargo, validar la cantidad de reactivos y exámenes que se generan es un reto metodológico para la psicometría. El propósito de este trabajo fue presentar una propuesta para analizar las propiedades psicométricas y la estructura interna de las familias de ítems producidas por el GenerEx, así como describir los resultados, a través del análisis de las 20 familias de Matemáticas de educación primaria. El estudio se fundamentó en la forma de seleccionar las muestras de reactivos y en tres tipos de análisis con marcos conceptuales complementarios: la Teoría Clásica de los Test, la Teoría de Respuestas al Ítem y el Análisis Factorial Confirmatorio. Los resultados muestran una descripción eficiente del funcionamiento psicométrico del GenerEx en el área de matemáticas. Este generador produce familias de ítems psicométricamente similares, con algunos problemas puntuales. Los análisis se podrían complementar con un estudio cualitativo de las deficiencias detectadas.

**Palabras clave:** Generador Automático de ítems, validez interna, familia de ítems, ítem-hijo, Examen de Competencias Básicas (Excoba)

[1] Métrica Educativa A.C. Para notificaciones dirigirse a fferreyra@metrica.edu.mx
[2] Métrica Educativa A.C.
[3] Instituto Nacional para la Evaluación de la Educación

## Introducción

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

Los generadores automáticos de ítems son las herramientas informáticas encargadas de producir grupos, denominados familias, con decenas o centenas de reactivos conceptualmente equivalentes dentro de cada familia. Los modelos de ítems definen las características y delimitaciones de los diferentes elementos que conforman los reactivos, y proveen de reglas para formación de ítems. Con estas instrucciones, el generador desarrolla una familia de reactivos. Estos ítems similares, llamados ítems-hijo, permiten construir cientos o miles de exámenes paralelos.

El GenerEx (su nombre proviene de Generador de Exámenes) es un ejemplo de generador de ítems. Por medio de él se construyen familias de ítems con sus respectivos ítems-hijo. Cada ítem-hijo se utiliza para conformar las diferentes versiones del Examen de Competencias Básicas (Excoba). En la figura 1 se observa cómo se conforma una familia de reactivos del Excoba, dada una competencia.

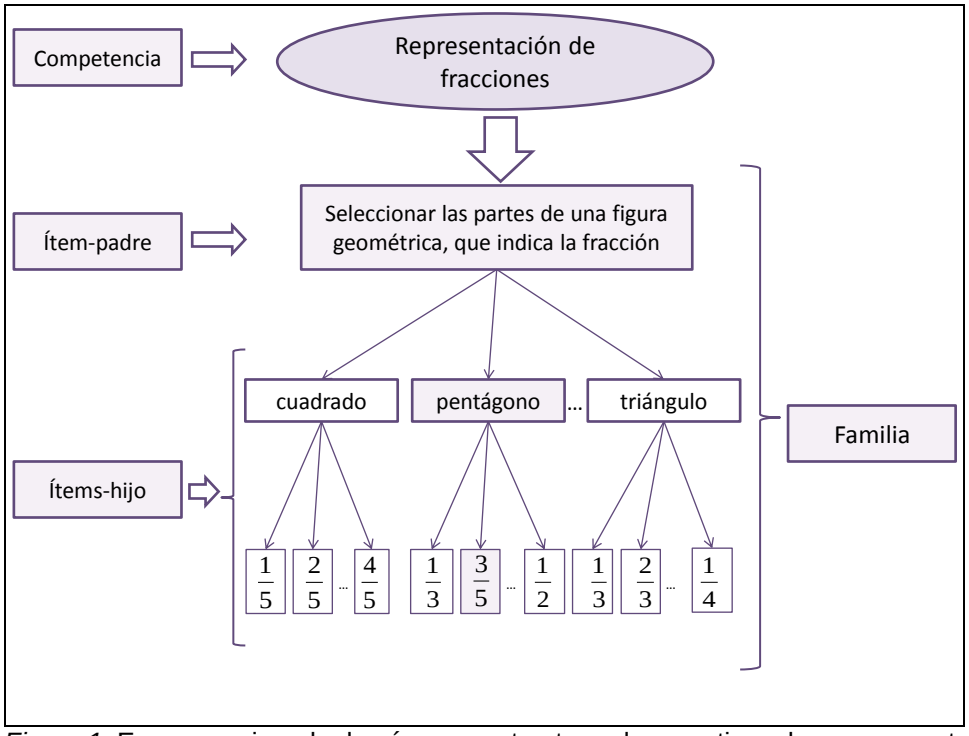


*Figura 1.* Esquema ejemplo de cómo se estructuran los reactivos de una competencia curricular.

El Excoba es un instrumento de evaluación que se utiliza para selección de estudiantes de ingreso a la educación media superior y la educación superior. El examen evalúa 180 competencias: 40 de nivel primario (lenguaje y matemáticas), 80 de nivel secundario (español, matemáticas, ciencias naturales y ciencias sociales) y 60 de bachillerato (orientadas según la carrera universitaria a la cual el estudiante desea ingresar). Para el caso del Excoba/MS (ingreso a la educación media superior) solamente se consideran las 120 primeras competencias.

El formato de los ítems del Excoba/MS no es el tradicional de opción múltiple. Se denominan ítems de respuesta semi-construida, puesto que se utilizan recursos gráficos y de escritura, entre otros, para armar la respuesta. Otra característica importante es la cantidad de respuestas solicitadas por ítem. El 30% de los ítems son dicotómicos (correcto-incorrecto) y el resto son de crédito parcial (de acuerdo con el grado de solución de la respuesta).

El Excoba y el GenerEx son el producto de cinco años de trabajo colegiado con la participación de profesionales expertos en las diferentes áreas de evaluación y en sistemas informáticos. Si bien se han obtenido exámenes cuidadosamente elaborados y revisados, es necesario garantizar, a través de procesos estadísticos, los estándares de calidad de todo instrumento de evaluación. Más aún para la GAI, donde la cantidad de ítems por familia complica el proceso y plantea un desafío a sortear.

**Objetivos**

Por consiguiente, este trabajo tiene los propósitos de:

- proponer una metodología para estudiar la validez de las familia de reactivos producidas por el GenerEx, y
- mostrar ejemplos de los resultados que se obtienen con esta metodología, particularmente, aquellos orientados a aportar evidencias de validez de las familias de reactivos del área de Matemáticas para nivel de educación primaria.

**Fundamentación teórica**

La estrategia más simple para analizar las respuestas de los ítems generados automáticamente es considerar cada ítem como una entidad única, claro que la mayor limitación a este tratamiento es que se necesitan cientos de estudiantes que resuelvan cada uno de los reactivos generados y así obtener los índices estadísticos para cada caso; lo cual se convierte en un trabajo monumental e ineficaz. Por lo tanto, se necesitan modelos alternativos para analizar los miles de reactivos que se producen a través de la GAI.

Además, existen dos tipos de GAI, las que se sustentan en modelos cognitivos, que se denominan de *teoría fuerte*, y las que no, que reciben el nombre de *teoría débil*. Esta categorización también implica diferentes criterios de análisis.

Como respuesta a este problema, Sinharay y Johnson (2012) identificaron tres categorías de modelos de análisis. La primera, con el objeto de predecir los parámetros de los reactivos, en particular: la dificultad, desde las características de los ítems. La segunda considera la dependencia entre parámetros que pertenecen a una misma familia de reactivos. La tercera categoría combina las dos anteriores.

En cuanto a la primera categoría, investigadores como Embretson (1999) y Holling, Bertling y Zeuch (2009), utilizaron el Modelo Logístico Lineal del Rasgo Latente (LLTM, por su nombre en inglés) 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, los análisis se enfocan en la GAI de teoría fuerte.

En la segunda categoría, los procedimientos más desarrollados son dos: el modelo de hermanos idénticos (*Identical Siblings Model*, ISM) y el modelo de los hermanos relacionados (*Related Siblings Model*, RSM). El ISM, de Hombo y Dresher (2001), asume una función de respuesta única para todos los ítems 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, un modelo jerárquico cuya primera componente es un modelo simple de la Teoría de Respuesta al Ítem (TRI) de tres parámetros. Este trabajo se realizó con ítems dicotómicos y el modelo se aplica, fundamentalmente, para tests adaptativos.

La tercera categoría combina los modelos LLTM y RSM en otro denominado Modelo Lineal de Clonación de Ítems, 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 este modelo también debe utilizarse para una GAI de teoría fuerte.

El Excoba es un ejemplo de GAI basado en teoría débil, ya que no se fundamenta en modelos cognitivos; sino que su construcción está sostenida por los planes de estudio de la educación básica y media superior. Por lo tanto, no se cuenta con modelos de tareas donde se especifiquen las estructuras cognitivas que dan soporte a todo el examen. Solamente se plantean los contenidos a evaluar y las habilidades predominantes para cada familia de ítems. En consecuencia, no se tiene un modelo jerárquico que validar, lo cual hubiera permitido utilizar un LLTM u otra variante. Por lo tanto, se tomó la decisión de utilizar la Teoría Clásica de los Tests (TCT) y la TRI, según lo recomiendan los estándares internacionales y nacionales (American Educational Research Association *et al.*, 1999, Martínez-Rizo *et al.*, 2000), para el estudio de las propiedades psicométricas de los ítems agrupados en familias; y además sustentar los análisis de agrupación de los reactivos en constructos, desde un AFC.

Dentro de la TRI, se optó por un modelo de un parámetro por varias razones: la adivinación no se consideró como variable debido a las características de los nuevos ítems, el tamaño de las muestras fue insuficiente para las demandas de un modelo de dos parámetros y la búsqueda del modelo más simple que explique el comportamiento de los datos (de acuerdo con el principio de *parsimonia*). Asimismo, como se tienen exámenes que combinan ítems dicotómicos y politómicos de crédito parcial, fue necesario utilizar un modelo que ajuste a los dos tipos de ítems; por lo tanto, se escogió el modelo dicotómico de Rasch (Rasch, 1961) y su extensión, el modelo de Rasch para crédito parcial, de Masters (1982). Si bien no se incluyó la discriminación como parámetro, se calculó como un índice, con el objeto de obtener información también sobre este comportamiento.

## Metodología

Coherente con esta propuesta de análisis, para estudiar las familias de ítems se fijaron muestras como producto de la administración de una prueba parcial de cada área del examen, con 6 ítems-hijo diferentes por contenido (ver figura 2). En total se aplicaron seis pruebas, una por cada área del Excoba/MS. Las unidades a analizar consistieron en los 6 ítems de cada familia. Por ejemplo, para Matemáticas se estudiaron las 20 familias, cada una con 6 ítems-hijo seleccionados al azar, con la condición de que no se repitieran reactivos de la misma familia.

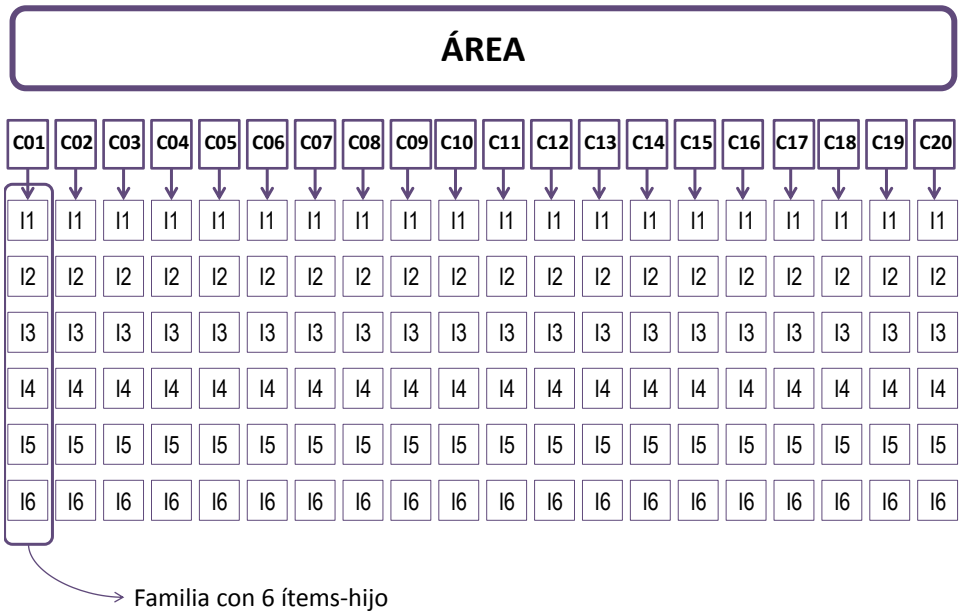


*Figura 2.* Estructura de una prueba parcial por área: 20 contenidos con 6 ítems-hijo por cada uno, dan un total de 120 ítems.

Los participantes, quienes resolvieron las diferentes combinaciones de reactivos, fueron estudiantes de instituciones usuarias del Exhcoba, que aceptaron participar voluntariamente. Las diferentes aplicaciones se efectuaron entre los meses de marzo y junio de 2012, y coincidió con las épocas de administración del Exhcoba (versión original de opción múltiple) como parte de sistema de ingreso a las instituciones participantes. Cada prueba parcial se administró, a modo de pilotaje, a 200 estudiantes, la mayoría de segundo semestre de licenciatura.

Para el análisis de los datos se identificaron dos estrategias: (1) la descripción de las propiedades psicométricas de los ítems y (2) los estudios de agrupación de variables en factores subyacentes, a través del AFC.
(1) Desde la TCT se calcularon las medidas de tendencia central y dispersión por familia. 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 obtuvieron los índices de ajuste, de correlación punto medida y de discriminación.
(2) Para el AFC se calcularon las cargas factoriales correspondientes a la agrupación de ítems para cada familia. Además se obtuvieron: el Índice de Ajuste No Normalizado (NNFI), el Índice Comparativo de Ajuste (CFI) y el Error Medio Cuadrático de Aproximación (RMSEA).
Los programas estadísticos empleados fueron Winsteps, versión 3.70.0.2 (Linacre, 2010) y el paquete EQS 6.1 (Bentler, 2006). En la tabla 1 se establecen los criterios asumidos para evaluar la calidad de los ítems.

Tabla 1.
*Criterios asumidos para los análisis estadísticos de los ítems de las muestras del Excoba/MS*

| | Estadísticos | Criterio | |
|---|---|---|---|
| | | Aceptable | Bueno |
| TCT | | | |
| | Correlación punto biserial | ≥ 0.2 | |
| | Varianza de dificultad por familia | $\rightarrow 0$ | |
| | $\alpha$ de Cronbach (confiabilidad) | ≥ 0.6 | |
| TRI | | | |
| | Correlación punto medida | ≥ 0.3 | |
| | *Infit-Outfit* | ≥ 0.8 y ≤ 1.3 | |
| | Discriminación | ≥ 0.8 | |
| AFC | | | |
| | Carga factorial | ≥ 0.20 | ≥ 0.30 |
| | $\chi^2$ | ≥ 0.01 | ≥ 0.05 |
| | NNFI | ≥ 0.90 | ≥ 0.95 |
| | CFI | ≥ 0.90 | ≥ 0.95 |
| | RMSEA | < 0.08 | < 0.05 |

Especialistas en sistemas informáticos administraron los exámenes y generaron bases de datos con los resultados de las aplicaciones. Se depuraron y organizaron dichas bases y con ellas se efectuaron los análisis previamente descritos. Se concluyó con un informe sobre las propiedades psicométricas de los ítems y sobre el grado de validez basado en evidencias de estructura interna de

las familias de ítems. En los casos pertinentes, se indicaron sugerencias para mejorar el instrumento.

**Resultados**

Dada la corta extensión del presente trabajo, no es posible presentar todos los estudios realizados para las seis áreas temáticas; por lo tanto, se seleccionaron las familias de ítems del área de matemáticas de educación primaria y a continuación se presenta un resumen de los resultados.

A través de la TCT, se calcularon la media de dificultades y la varianza de los 6 ítems de cada una de las 20 familias de Matemáticas (ver figura 3). Las medias de las dificultades por familia se distribuyeron entre 0.05 y 0.75. Las varianzas fueron pequeñas, la mayor se observó en la familia de HC16, que superó ligeramente a 0.02; del resto, la mayoría fue cercana a cero.

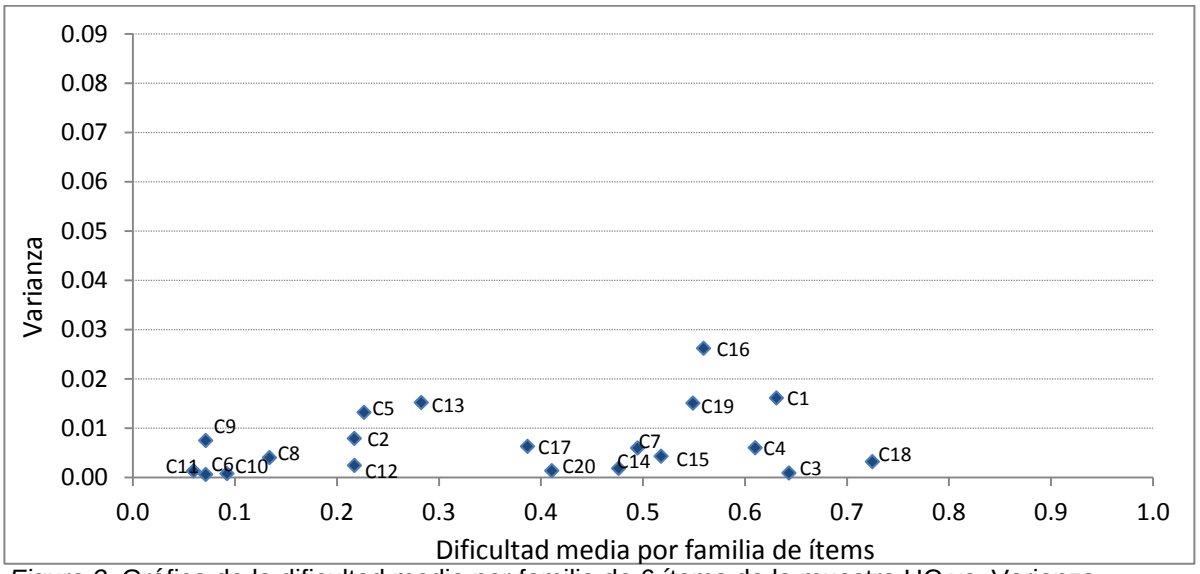


*Figura 3.* Gráfica de la dificultad media por familia de 6 ítems de la muestra HC vs. Varianza.

En la tabla 2 se resumen los demás índices obtenidos por familia. Se encontraron tres casos en que los índices de confiabilidad estaban por debajo de 0.6. El más bajo se refiere a HC09, con una confiabilidad cercana a cero. En el caso de HC05 solo pudieron analizarse 3 de los 6 ítems, debido a fallas técnicas. La familia de HC19 estuvo ligeramente por debajo con 0.54.

Los resultados de la aplicación de la TRI revelaron algunos índices de ajuste fuera de rango, no concentrados particularmente en ninguna familia. En general, el comportamiento de los ítems fue bueno, tanto cerca como lejos de su nivel de dificultad. En cuanto a los índices de discriminación, se encontró un ítem con problemas en la familia HC06, este índice fue negativo. La correlación punto medida fue alta en todas las familias, en la mayoría de los casos, mayores que

0.6; lo cual indica que los reactivos examinados por familia son isomorfos, en el sentido de que evalúan una misma habilidad.

El AFC por familias también arrojó cargas e índices de ajuste muy buenos en general, excepto para un ítem de una familia (HC19). Estos resultados reflejan la calidad en cuanto a la estructura interna de ítems-hijo que evalúan la misma competencia.

Tabla 2.
*Familias de reactivos de Matemáticas de educación primaria y sus deficiencias en los diferentes índices psicométricos*

| | Varianza | Alpha | Infit | Outfit | Disc | Pmed | AFC |
|---|---|---|---|---|---|---|---|
| HC01 | | | | | | | |
| HC02 | | | 2 | 1 | | | |
| HC03 | | | | | | | |
| HC04 | | | 1 | 1 | | | |
| HC05 | | x | | | | | |
| HC06 | | | 2 | 2 | 1 | | |
| HC07 | | | | | | | |
| HC08 | | | | 1 | | | |
| HC09 | | xx | | | | | |
| HC10 | | | | | | | |
| HC11 | | | | | | | |
| HC12 | | | 1 | | | | |
| HC13 | | | 3 | 2 | | | |
| HC14 | | | | | | | |
| HC15 | | | 1 | | | | |
| HC16 | | | 1 | 1 | | | |
| HC17 | | | | | | | |
| HC18 | | | 2 | 2 | | | |
| HC19 | | x | | 1 | | | 1 |
| HC20 | | | 2 | 1 | | | |

*Nota*: Para Alpha de Cronbach, “x” indica Alpha < 0.6. “xx” indica Alpha < 0.2
Para *infit* y *outfit*, discriminación, Pmed (índice de correlación punto medida) y AFC se indica el número de ítems que se encuentran fuera de rango de aceptación.

Los resultados de los análisis psicométricos efectuados a las familias de HC indican que, en general, cada grupo presentó propiedades que avalan el isomorfismo entre los ítems-hermano examinados; es decir, en cada familia los reactivos se asemejaron en dificultad y se asociaron en el constructo que los definió.

También surgieron algunas excepciones que deben considerarse. Desde el AFC, un ítem no cargó en el grupo de HC19; si bien esto no se reflejó notoriamente en el estudio a través del modelo de Rasch. El otro indicio de alerta fue el coeficiente de confiabilidad de HC09; que fue demasiado bajo. Finalmente, el sexto ítem de HC06 reflejó desajustes y discriminación negativa; por lo tanto, sería importante revisar la diferencia de este ítem con respecto a los cinco restantes de su familia.

## Conclusiones y recomendaciones

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 que maneja el Excoba, para el caso del ingreso a la educación media superior.

Este trabajo presentó una propuesta metodológica para aportar evidencias de validez de estructura interna de familias de reactivos desarrollados por la GAI de teoría débil con ítems de respuesta construida o semi-construida. Si bien el estudio no abarcó todas las posibilidades de ítems-hijo, el análisis de la calidad de las familias de reactivos del examen permitió a los desarrolladores del Excoba obtener información de gran utilidad para revisar y perfeccionar el examen.

Se obtuvo un buen diagnóstico del funcionamiento del GenerEx con muestras pequeñas, y con descripciones claras y sencillas. Un aporte interesante es la complementariedad de los análisis, los resultados coincidentes desde las diferentes teorías proporcionan mayor certeza en cuanto al funcionamiento del generador, aquellos que no concurren se pueden considerar menos relevantes.

Deben citarse ciertas limitaciones en la aplicación de la metodología. Esta se ajustó a las necesidades para la validación de una GAI de las características del Excoba; aunque podría adecuarse a otros exámenes de GAI de teoría débil. Las evidencias de validez se obtuvieron a través de muestras, en algunos casos, de menor tamaño a lo requerido, fueron resultado de pilotajes y no de administraciones reales de alto impacto. Además, durante las aplicaciones el editor de reactivos presentó fallas que se reflejaron al recuperar la información y provocaron datos perdidos.

Esta contribución se considera valiosa para el desarrollo de nuevas formas evaluativas amparadas en la tecnología computacional y en desarrollos psicométricos sólidos como son la Teoría Clásica de los Tests y la Teoría de Respuesta al Ítem. Las evidencias de validez posicionan al Excoba a la vanguardia de la evaluación educativa en México con respecto a la aplicación de instrumentos de medición que utilizan reactivos mejor construidos y a los procesos de administración más seguros, al no repetir la prueba en una misma aplicación. De esta forma la calidad del Excoba se exhibe a la comunidad científica y al público en general para ser analizada y valorada con las bondades y limitaciones propias de la metodología utilizada.

## Referencias bibliográficas


American Educational Research Association, American Psychological Association y National Council on Measurement in Education (1999). *Standards for Educational and Psychological Testing,* Washington, American Educational Research Association, 194 p.

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

Embretson, Susan (1999). "Generating items during testing: psychometric issues and models", Psychometrika, volumen 64, número 4, pp. 407-433, doi: 10.1007/BF02294564

Geerlings, Hanneke, Cees Glass y Wim van der Linden (2011). "Modeling rule-based item generation", *Psychometrika*, volumen 76, número 2, pp. 337-359, doi: 10.1007/s11336-011-9204-x

Gierl, Mark y Hollis Lai (2012). "Using weak and theory to create item models for Automatic Item Generation: some practical guidelines with examples", *Automatic Item Generation: Theory and Practice.* N. Y., Routledge, pp. 26-39

Glas, Cees y Wim van der Linden (2003). "Computarized adaptive Testing with item cloning", *Applied Psychological Measurement*, volumen 27, pp. 247-261, doi: 10.1177/0146621603254291

Holling, Heinz, Jonas Bertling y Nina Zeuch (2009). "Automatic item Generation for probability word problems", *Studies in Educational Evaluation*, volumen 35, pp. 71-76, doi:10.1016/j.stueduc.2009.10.004

Hombo, Catherine y Amy Dresher (2001). "A simulation study of the impact of automatic item generation under NAEP-like data conditions*", Annual Meeting of the National Council on Measurement in Education*, Seattle, EE. UU.

Linacre, John (2010). "Winsteps" ® (version 3.70.0.2), *Programa de computación*, Beaverton, Oregon, Winsteps.com

Martínez Rizo, Felipe *et al.* (2000). *Estándares de calidad para instrumentos de evaluación educativa*, México, Ceneval, 51p.

Masters, Geoff (1982). "A Rasch model for partial credit scoring", *Psychometrika*, volumen 47, número 2, pp. 149-174, doi: 10.1007/BF02296272

Rasch, Georg (1961). "On General Laws and the Meaning of Measurement in Psychology". *Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 4: Contributios to Biology and Problems of Medicine*, 321-333, University of californiaPress, Berkeley, CA

Sinharay, Sandip y Matthew Johnson (2012). “Statistical modeling of Automatic Item Generation”, *Automatic Item Generation: Theory and Practice*, New York, Routledge, pp. 183-195