# MODELO PARA VALIDAR UN GENERADOR AUTOMÁTICO DE EXÁMENES

MARÍA FABIANA FERREYRA / EDUARDO BACKHOFF ESCUDERO
Instituto de investigación y Desarrollo Educativo

**RESUMEN:** El Examen de Habilidades y Conocimientos Básicos (EXHCOBA) es un examen de ingreso a la educación superior y a la educación media superior, desarrollado en 1992 y que se aplica de manera computarizada desde 1993. A partir de 2009 se está en proceso de reestructuración de esta prueba y, en particular, de sus reactivos. Como resultado, se ha obtenido un generador automático de exámenes, el EXHCOBA-R, conformado por ítems de respuesta construida o semiconstruida. Este generador necesita ser validado para poder ser aplicado con certeza y confianza, más aún debido a que se trata de un instrumento a gran escala y de alto impacto.

En este trabajo se presenta un modelo para mostrar evidencias de validez de estructura interna de un generador automático de exámenes: cómo construir las muestras y qué procedimientos estadísticos utilizar; además, se ilustra con tres ejemplos de análisis del EXHCOBA-R.

**PALABRAS CLAVE: evaluación de estudiantes, evaluación del aprendizaje, validez de las pruebas, análisis estadístico, admisión a la universidad.**

## Introducción

El Examen de Habilidades y Conocimientos Básicos (EXHCOBA) es una prueba de admisión a la educación media superior y superior desarrollado por Backhoff y Tirado (1992). Un año después, Backhoff, Ibarra y Rosas (1996) crearon el Sistema Computarizado de Exámenes (SICODEX) con el objeto de administrar dicha prueba. El examen se estableció como requisito de ingreso tanto en la Universidad Autónoma de Baja California (UABC) como en otras instituciones educativas mexicanas. El EXHCOBA computarizado lleva más de dos décadas de aplicaciones, tan solo en 2012 más de 15

instituciones educativas recibieron los servicios del examen y hubo alrededor de 120,000 evaluados.

Esta prueba, al igual que la mayoría de los tests a gran escala, utiliza reactivos de opción múltiple debido a que hacen posible obtener una muestra más amplia de contenidos a evaluar, y a la objetividad y rapidez para ser calificados. Sin embargo, estos prototipos sufren ciertas limitaciones; por ejemplo: presentan una manera un tanto artificial de evaluar el aprendizaje, pueden prestarse a la adivinación, necesitan una constante actualización debido al desgaste de su uso y no aprovechan todas las ventajas de las tecnologías digitales.

Considerando estas limitaciones, se ha realizado un nuevo planteamiento de la evaluación de las competencias escolares, apoyado en las ventajas de los recursos digitales. Como resultado, se ha desarrollado un generador automático de exámenes basado en los planes de estudio vigentes de la educación básica y media superior, el *EXHCOBA-R*, que puede originar miles de exámenes similares con tipos de ítems de respuesta construida o semiconstruida. Estos nuevos ítems se fundamentan en el enfoque conceptual de la teoría cognoscitiva, particularmente en la propuesta de Ausubel, Novak y Hanesian (1983).

El EXHCOBA-R evalúa competencias que se adquieren en tres niveles educativos: educación primaria, secundaria y bachillerato. Para cada nivel se precisan las áreas a evaluar; para cada área se definen 20 competencias básicas y, a su vez, cada competencia es evaluada con un conjunto de reactivos equivalentes. Cada examen generado se conforma de 40 ítems de primaria, 80 de secundaria y 60 de bachillerato (estos últimos correspondientes a tres asignaturas que se eligen según la especialidad que el aspirante desee estudiar en la educación superior); de modo que el examen de ingreso a la educación superior consta de 180 ítems.

Para cada una de las competencias curriculares incluidas en el EXHCOBA-R se elabora un modelo de evaluación. Este modelo contiene dos grandes apartados: (a) la información general de la competencia: eje temático, tema, subtema, nombre y definición del contenido, y justificación de la competencia seleccionada y (b) la plantilla con las reglas, y elementos conceptuales y gráficos de cada competencia, que permiten generar

una gran cantidad de reactivos estructurados en *familias* que contienen *ítems-padre* e *ítems-hijo*.

La figura 1 muestra cómo se estructuran estos reactivos a través de un ejemplo: la competencia de "Representación de fracciones." Para esta competencia se definen dos familias de reactivos: una que le pide al estudiante seleccionar las partes de una figura geométrica que corresponden a una fracción dada, y otra que le solicita escribir la fracción que corresponde a una figura fraccionada. Las dos familias de reactivos se componen de un conjunto de ítems-padre asociados a diferentes figuras geométricas (cuadrado, pentágono, etc.) y cada ítem-padre está conformado por diferentes ítems-hijo que representan las fracciones a identificar o a calcular (ver figura 1).

El generador automático de reactivos selecciona una de las familias, luego un ítem-padre (figura geométrica) y, finalmente, un ítem-hijo (fracción); de esta forma presenta un reactivo específico como el que se muestra en la figura 2.

Es importante aclarar que existen reactivos donde la respuesta es única (figura 2) y solo puede ser calificada como correcta o incorrecta (*ítems dicotómicos*); pero también hay otros ítems donde se solicitan dos o más respuestas al estudiante, que se califican parcialmente en función del número de aciertos (*ítems de respuesta parcial*).

En el caso del EXHCOBA-R, para los ítems de crédito parcial, cada acierto se computa, en partes iguales, para conformar el total de 1 punto. Por ejemplo, si el reactivo solicita ubicar cuatro fracciones en la recta numérica, cada fracción ubicada correctamente se computa como 0.25 puntos.

De lo expuesto se tiene, por un lado, reactivos novedosos tanto por el modo de contestar (respuesta semi-construida) como por la calificación (dicotómica y crédito parcial); por otro lado, un generador automático de reactivos que puede producir, para cada competencia, una gran cantidad de ítems parecidos, según sea el número de ítems-padre e ítems-hijo. Las combinaciones crecen aún más al multiplicarse por los 180 reactivos que contiene el EXHCOBA-R.

Como puede verse, con el generador automático de exámenes, EXHCOBA-R, se pueden crear miles de reactivos que se combinan para crear diferentes pruebas

representativas del currículo y equivalentes conceptualmente. Para poder aplicar este instrumento con la certeza y la confianza deseadas, es indispensable conocer las propiedades psicométricas de los ítems, obtener evidencias de que estas pruebas miden el constructo que pretenden medir, y de que los ítems-hijo son similares, tanto conceptual como psicométricamente, lo cual constituye el tema central de este trabajo.

## Objetivo

Desarrollar y validar un modelo para aportar evidencias empíricas de validez de la estructura interna del EXHCOBA-R, que es producido por un generador automático de reactivos.

## Objetivos específicos

a. Proponer el tipo de muestras necesarias para realizar los análisis estadísticos.
b. Proponer los análisis estadísticos apropiados para obtener las propiedades psicométricas del examen, en general, y de los ítems, en particular.
c. Ilustrar con ejemplos los análisis estadísticos señalados en el objetivo b.

## Contenido

A continuación, se describe un método para la validación empírica del EXHCOBA-R, para el nivel de educación media superior. Aquí se indican: (1) las muestras que dan origen a las distintas bases de datos a analizar, (2) los procedimientos estadísticos a seguir y (3) la descripción de cómo se analiza un examen generado.

(1) Las bases de datos surgidas de las aplicaciones del pilotaje del EXHCOBA-R constituyen los instrumentos de análisis. Para generar dichas bases se consideran dos tipos de muestras (ver figura 3):

- Muestra tipo 1: un examen completo generado para la educación media superior, con un ítem-hijo por cada competencia.
- Muestra tipo 2: prueba parcial de un área del examen, con *6* ítems-hijo diferentes por competencia evaluada.

Las muestras de tipo 1 sirven para indagar sobre las propiedades del examen en general, estudiar su comportamiento y comparar diferentes versiones. Las muestras de tipo 2 se utilizan para analizar reactivos de una misma competencia, comparar sus

propiedades psicométricas y observar si se agrupan en el constructo definido por dicha competencia o rasgo latente.

(2) Para el análisis estadístico de los datos se recurrió a la Teoría Clásica del Test (TCT), a la técnica de Análisis Factorial Confirmatorio (AFC) y la Teoría de Respuesta al Ítem (TRI), a través del modelamiento de Rasch. Los reactivos se consideraron como entes dicotómicos o de crédito parcial, según fuera el caso.

Dentro de la TCT, se utilizaron cuatro parámetros para evaluar el comportamiento de los ítems: el índice de dificultad, la correlación punto biserial (relación entre lo que mide el ítem y la escala), la varianza, y el índice de confiabilidad (Alpha de Cronbach) para estudiar la consistencia interna de la prueba.

En el AFC se utilizaron parámetros de ajuste que evaluaron la calidad del modelo. Se reportaron cuatro índices de ajuste, todos de buena ejecución en muestras pequeñas. Dos fueron de ajuste absoluto, chi cuadrado ($\chi^2$) y Root Mean Square Error Aproximation (RMSEA); y dos de ajuste incremental, el Índice de Ajuste No Normalizado (NNFI) y el índice Comparativo de Ajuste (CFI).

El análisis, según el modelo de Rasch, se efectuó a través de: (a) los estadísticos *infit* y *outfit*, con el objeto de conocer qué tan bien se ajustan las respuestas de los evaluados a cada reactivo; (b) la *correlación ítem medida*, que es la correlación entre las observaciones sobre un ítem y las correspondientes medidas de la persona (este parámetro es mejor que la correlación punto biserial cuando hay datos perdidos) (Linacre, 2009), (c) el *mapa de Wright*, para obtener la distribución de la dificultad de los ítems en relación con la habilidad de los evaluados, y (d) la *discriminación* de los ítems (capacidad de ítem para identificar a los estudiantes según su habilidad).

Para la TCT se utilizaron los programas estadísticos Excel 2007 y SPSS 17. Para los cálculos de AFC se empleó el paquete EQS 6.1 (Bentler, 2010). Los estudios Rasch se realizaron con Winsteps, versión 3.70.0.2 (Linacre, 2009).

El análisis de la información se organizó en cinco niveles, que se describen a continuación.

- Nivel 1: estudio de la prueba completa de nivel básico (muestra tipo 1) con los tres tipos de análisis estadísticos (TCT, AFC y Rasch) para cada versión y la comparación entre versiones diferentes.
- Nivel 2: análisis del examen de nivel básico (muestra tipo 1), por áreas. Se presenta un retrato estadístico de cada área (parámetros de TCT y TRI) y su estructura, de acuerdo con la organización teórica del EXHCOBA-R. Se comparan áreas de distintas versiones.
- Nivel 3: estudio de las muestras realizadas por áreas (muestra tipo 2). Esto, con el objeto de mostrar el comportamiento de un examen completo con ítems exclusivos de un área.
- Nivel 4: análisis de los ítems-hijo de una misma competencia (muestra tipo 2) para conocer su similitud psicométrica y su agrupamiento en el constructo definido por cada competencia.
- Nivel 5: estudio de los componentes de los reactivos de crédito parcial, con el objeto de identificar cuán homogéneos son dichos componentes. Para ello, se utilizaron muestras del tipo 2.

(3) Durante 2012, estudiantes voluntarios, pertenecientes a instituciones con las cuales la UABC tenía convenio, resolvieron el EXHCOBA-R. Los evaluados fueron 1,860 estudiantes, que pertenecían a los siguientes niveles académicos: (a) acababan de cursar los primeros semestres de licenciatura, la mayoría el segundo semestre, o (b) aspiraban a ingresar a l nivel de Educación Media Superior.

Especialistas en sistemas informáticos del EXHCOBA-R administraron, a modo de pilotaje, los distintos exámenes y generaron ocho bases de datos con las respuestas de los estudiantes evaluados, dos bases de las muestras tipo 1 (dos versiones del examen) y seis bases del tipo 2 (una por cada área de las secciones de competencias de primaria y de secundaria). Se depuraron y organizaron dichas bases y con ellas se efectuaron los análisis.

Dada la corta extensión del presente trabajo, se seleccionaron solamente tres ejemplos referentes a los dos últimos niveles de análisis (4 y 5), debido a la originalidad que implica el estudio de ítems-hijo semejantes. El primero pertenece a la TCT, el segundo es un AFC y el tercero es un modelamiento Rasch.

- ✓ *Ejemplo 1*. En el caso de la muestra tipo 2 de cada área, un resultado interesante fue obtener las dificultades media de los ítems de la misma competencia y contrastarla con sus varianzas. Para este ejemplo se utilizó el área de Ciencias sociales con seis ítems-hijo por cada una de las 20 competencias que se evalúan en esta sección del examen. Los resultados mostraron que las cinco familias de Formación Cívica y Ética fueron las que registraron menor varianza, quien más varianza presentó fue la competencia de Geografía G3 (que mide la interpretación de gráficas poblacionales); aunque en ningún caso superó a 0.015. Las dificultades se ubicaron entre 0.40 y 0.87. Las mayores varianzas se concentraron, en general, en los ítems de dificultad media (ver figura 4).
- ✓ *Ejemplo 2*: En este caso se escogió el AFC de una familia de reactivos, la correspondiente a HIS08 (que mide las implicaciones de los cambios políticos y socioeconómicos en Europa y en América durante las revoluciones atlánticas). Se consideraron los seis ítems asociados, se agruparon en un factor y se incorporaron covarianzas de errores entre los reactivos. Posteriormente, se eliminaron aquellas covarianzas que no aportaban carga al modelo, según el test de Wald. Los resultados reportan muy buenos índices de ajuste: p = 0.583 (debe ser superior a 0.01), NNFI = 1.008 (debe ser cercano a 1), CFI = 1 (debe ser superior a 0.95) y RMSEA = 0.00 (debe ser menor que 0.05). Las cargas estandarizadas se distribuyeron de la siguiente manera: ítem 1 con 0.63, ítem 2 con 0.50, ítem 3 con 0.60, ítem 4 con 0.57 e ítem 6 con 0.84; por lo tanto, todos los ítems aportan, en buen grado, al constructo definido por la competencia.
- ✓ *Ejemplo 3*. En este caso, se presenta un análisis Rasch de los elementos de los ítems-hijo de la misma competencia HIS08. Cada ítem contiene cinco preguntas a responder (a efectos del análisis, cada preguntas se denomina elemento), lo que equivale a 0.20 puntos por cada respuesta correcta. Como la muestra considerada consta de seis ítems-hijo para la competencia HIS08, en total se analizaron 30 elementos.

Los resultados mostraron que el elemento 5 del ítem 3 (ítem3.5) presentaba un infit estandarizado fuera del rango (-2,2) y, además, que la correlación fue muy pequeña (ver figura 5). Una revisión de los contenidos permitió verificar que este elemento no correspondía al periodo histórico que se evaluaba en el contenido y, por lo tanto,

provocaba "ruido" entre los examinados. Otro elemento que no presentaba buena correlación fue el ítem 1.3. En ese caso pudo deberse a que la pregunta fue muy difícil, en comparación con el resto.

## Conclusiones

Después de haber realizado los análisis estadísticos en las muestras piloto del EXHCOBA-R, se identificaron los siguientes beneficios del método:

1. se obtuvo una radiografía detallada de dos versiones del examen donde se identificaron sus semejanzas y diferencias;
2. para las dos versiones del examen se constató la agrupación de ítems según el área de conocimiento. De este modo, se localizaron los ítems que aportaban al constructo definido y los que no. Al repetir el ejercicio en dos versiones, los hallazgos son más sólidos;
3. en cada familia de ítems, se identificó cuánto se acercaban en dificultad, diferentes ítems-hijo y se verificó su agrupación en el constructo que los define;
4. se indagó las propiedades de los elementos que conforman los distintos ítems-hijo de una familia. El análisis permitió identificar elementos con buenos parámetros estadísticos o con desajustes;
5. debido a que se realizaron tres tipos de análisis (TCT, Rasch y AFC), los resultados coincidentes dieron mayor certeza en los hallazgos y aquellos que no concurrieron, se consideraron como menos relevantes;
6. se mostraron evidencias de que el uso de ítems de crédito parcial puede ayudar a la calidad estadística de los reactivos, ya que los errores de algún elemento del ítem-hijo pueden compensarse con las bondades de otros elementos, de mejor calidad, en el mismo ítem (como ocurrió con la competencia HIS08).

El modelo también presenta limitaciones, entre ellas se citan las siguientes:

1. como los análisis son complementarios, no todos convergen necesariamente en una misma dirección;
2. debido al gran número de ítems, se requieren muestras grandes, para poder ejecutar los análisis;
3. se necesitan otros análisis que complementen los resultados obtenidos (por ejemplo: modelos TRI de tres parámetros o modelos cognitivos);
4. estos análisis no permiten obtener la información completa de todos los exámenes que se puedan generar; sin embargo, es una muy buena aproximación para

detectar el funcionamiento del generador de reactivos y, sobre todo, para identificar los problemas.

*Generador automático de exámenes* y *generador automático de reactivos* son equivalentes a efectos de este artículo.

## Tablas y figuras

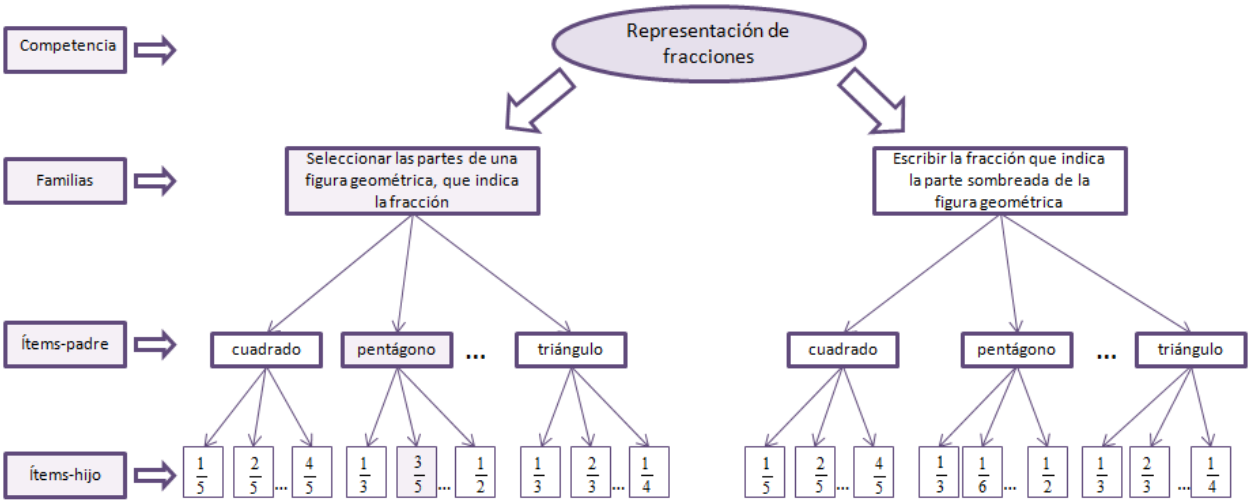



*Figura 1.* Ejemplo de esquema de cómo se estructuran los reactivos de una competencia curricular.

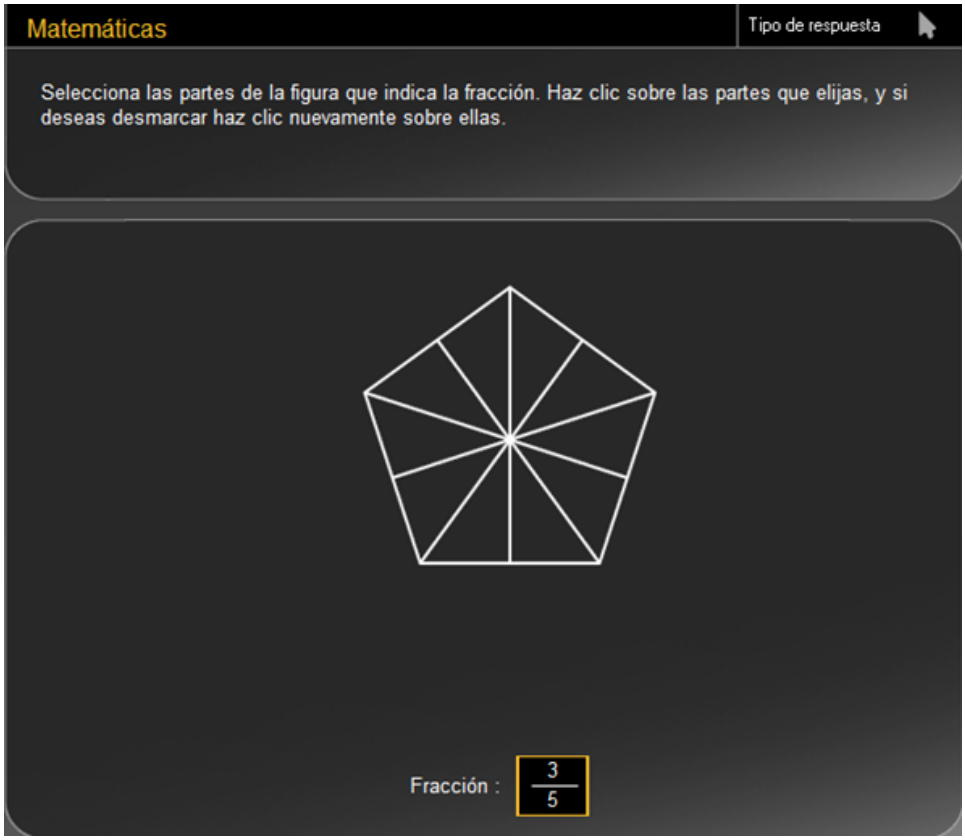
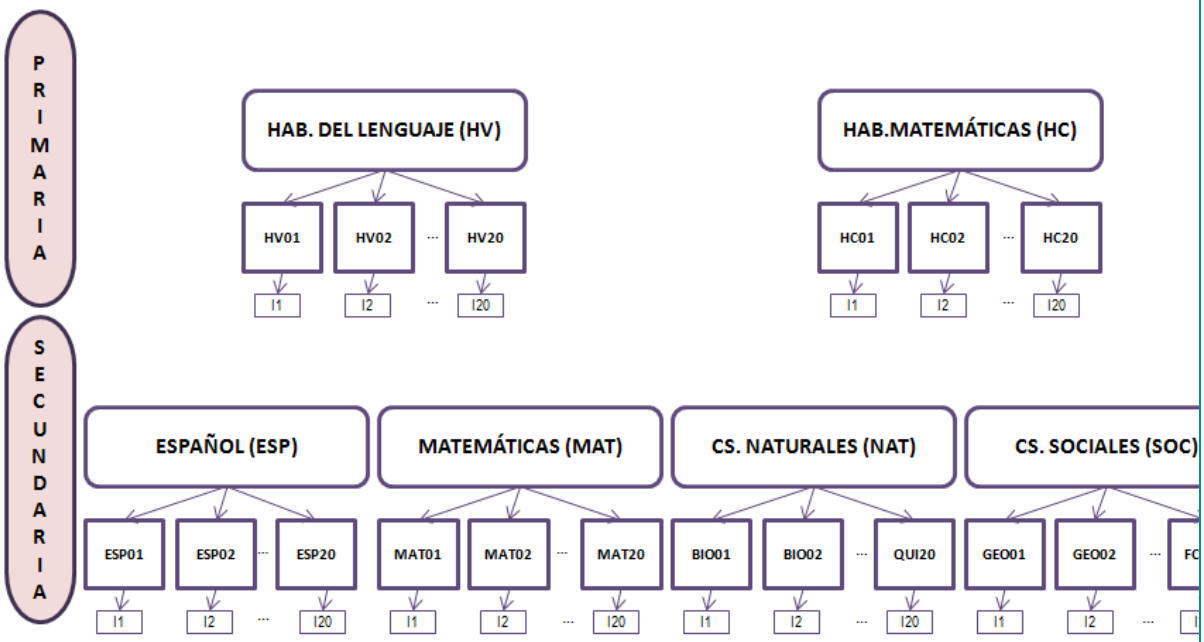



*Figura 2.* Ejemplo de ítem-hijo para la competencia: Representación de fracciones, del área de matemáticas.

**Muestra tipo 1:**




**Muestra tipo 2:**

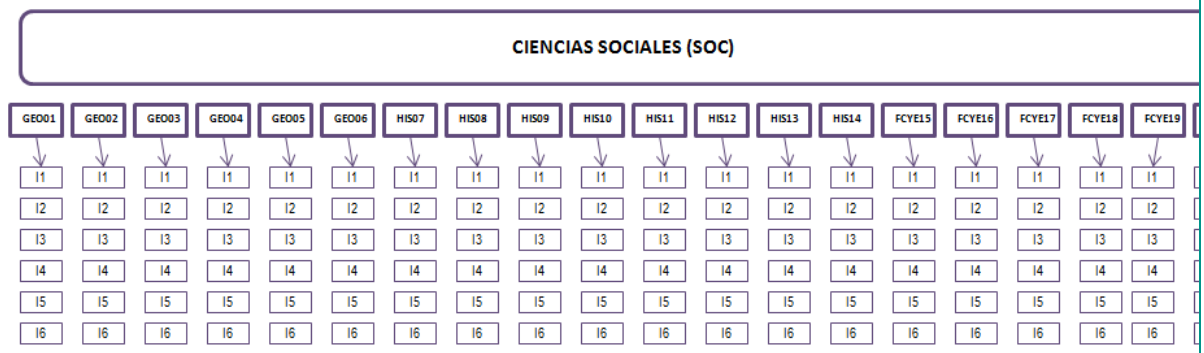



*Figura 3.* Muestra 1: Versión de examen EXHCOBA-R para educación media superior: 6 áreas y 120 ítems. Muestra 2: Ejemplo de una muestra para el área de Ciencias Sociales: SOC. Análisis de los 120 ítems en total, seis ítems-hijo por cada uno de los 20 contenidos. GEO corresponde a competencias de Geografía, HIS a Historia, y FCYE a Formación Cívica y Ética.

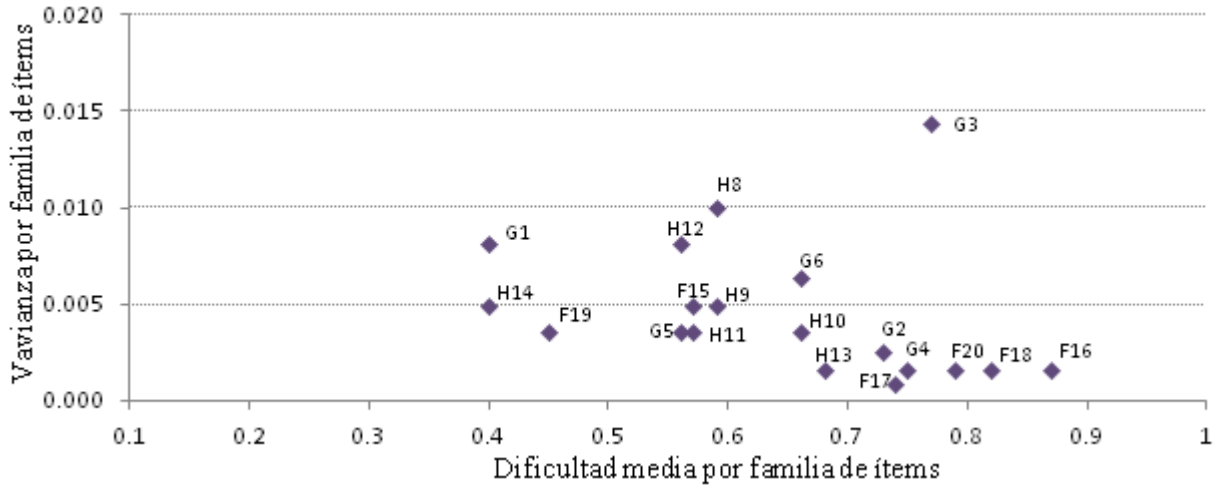



*Figura 4*. Gráfica de la dificultad media por familia de 6 ítems de SOC vs. Varianza. G: Geografía, H: Historia, F: Formación Cívica y Ética. Falta la familia de H7 que no pudo ser evaluada por problemas técnicos durante el pilotaje.

| ELEM | MEDIDA | INFIT ESTANDARIZADO | OUTFIT ESTANDARIZADO | CORR. |
|---|---|---|---|---|
| | - + | -3 -2 -1 0 1 2 3 | -3 -2 -1 0 1 2 3 | PT-MED |
| ITEM1.1 | | | | .58 |
| ITEM1.2 | | | | .50 |
| ITEM1.3 | | | | .15 |
| ITEM1.4 | | | | .44 |
| ITEM1.5 | | | | .46 |
| ITEM2.1 | | | | .34 |
| ITEM2.2 | | | | .49 |
| ITEM2.3 | | | | .26 |
| ITEM2.4 | | | | .44 |
| ITEM2.5 | | | | .48 |
| ITEM3.1 | | | | .57 |
| ITEM3.2 | | | | .43 |
| ITEM3.3 | | | | .58 |
| ITEM3.4 | | | | .40 |
| ITEM3.5 | | | | .09 |
| ITEM4.1 | | | | .42 |
| ITEM4.2 | | | | .30 |
| ITEM4.3 | | | | .39 |
| ITEM4.4 | | | | .43 |
| ITEM4.5 | | | | .58 |
| ITEM5.1 | | | | .20 |
| ITEM5.2 | | | | .56 |
| ITEM5.3 | | | | .25 |
| ITEM5.4 | | | | .30 |
| ITEM5.5 | | | | .38 |
| ITEM6.1 | | | | .62 |
| ITEM6.2 | | | | .64 |
| ITEM6.3 | | | | .57 |
| ITEM6.4 | | | | .56 |
| ITEM6.5 | | | | .63 |

*Figura 5.* Análisis Rasch: Medida, infit, outfit estandarizados y correlación punto biserial de los 30 elementos correspondientes a los 6 ítems de la competencia HIS08 de la muestra tipo 2 de Ciencias sociales. ELEM: elemento, MEDIDA: índice de dificultad medido en

lógitos, INFIT y OUTFIT ESTANDARIZADOS: medidas de ajuste de los ítems, CORR. PT-MED: correlación punto medida.

## Bibliografía

Ausubel, D., Novak, J. y Hanesian H. (1983). *Psicología educativa: Un punto de vista cognoscitivo* (2da. ed.). México: Trillas.

Backhoff, E., Ibarra, M. y Rosas, M. (1996). Desarrollo y validación del Sistema Computarizado de Exámenes SICODEX. *Revista de la Educación Superior, 25* (1). 41-54. Recuperado de: [http://www.exhcoba.mx/pdf/1996_Desarrollo_y_validacion_del_sistema_computarizado_de_examenes_SICODEX.pdf].

Backhoff E. y Tirado F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior*, *21* (3), 95-118. Recuperado de: [http://www.exhcoba.mx/pdf/1992_Desarrollo_del_EXHCOBA.pdf].

Bentler, P. M. (2010). *EQS, A Structural Equation Program*. Multivariate Software, Inc. Version 6.1 (C) 1985 - 2010 (B97).

Linacre, J. M. (2009). Winsteps a Rasch Analysis Computer Program, versión 3.70.0.2 Chicago, EE. UU.: Winsteps Publications.