



EVALUACIÓN DE COMPETENCIAS CON REACTIVOS DE CRÉDITO PARCIAL PRODUCIDOS POR GENERADORES AUTOMÁTICOS DE ÍTEMS

MANUEL JORGE GONZÁLEZ-MONTESINOS MARTINEZ

DEPARTAMENTO DE CIENCIAS SOCIALES, UNIVERSIDAD DE SONORA UNIDAD REGIONAL NORTE

mgm@caborca.uson.mx

MARIA FABIANA FERREYRA

MÉTRICA EDUCATIVA A.C.

fferreyra@metrica.edu.mx

RESUMEN

Se presenta una forma de evaluar competencias al egreso del nivel básico y de media superior basada en mecanismos de generación automática de reactivos y que requieren procesos de respuesta semi-construida por parte de los sustentantes, en lugar de respuestas de opción múltiple. Esta alternativa de evaluación es una innovación en proceso de diseño que incorpora además ítems de respuesta binaria e ítems de crédito parcial segmentados en categorías. Se ilustra un procedimiento de control de calidad métrica basado en la extensión PCM-Andrich del modelo de Rasch.

Palabras Clave: Teoría de Respuesta al Ítem, Psicometría, Competencias, Educación básica, Exámenes de ingreso.

INTRODUCCIÓN

La Generación Automática de Ítems (GAI) es el proceso para diseñar y elaborar reactivos 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 (Generador de Exámenes) es un ejemplo de generador de ítems. Esta herramienta produce 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.

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. Para el caso del Excoba/MS (ingreso a la educación media superior) el examen evalúa 120 competencias: 40 de nivel primario (lenguaje y matemáticas) y 80 de nivel secundario (español, matemáticas, ciencias naturales y ciencias sociales).

El formato de los ítems del Excoba/MS no es el tradicional de opción múltiple. Lauren y Daniel Resnick (1992) explicaron que los tests de opción múltiple se ajustan a la psicología conductista; esta psicología educacional plantea dos supuestos básicos: la desagregabilidad (el todo se descompone en sus partes) y la descontextualización (los hechos se estudian aislados de su contexto); estos supuestos no coinciden con la vida real, donde todo aparece entero e inmerso en un mundo mezclado y confuso. Además, Ng y Chan (2009) declararon que este tipo de ítems propician dos clases de respuestas correctas: las que el evaluado conoce y las que acierta por azar. Puede ocurrir que el examinado conozca parte de la solución; sin embargo, este conocimiento parcial no se registra en este tipo de preguntas.

También es cierto que en el Excoba (un examen de alto impacto y a gran escala) es necesario abarcar una cantidad muy vasta de competencias con un gran número de reactivos que las evalúen y el tiempo destinado a cada respuesta no puede ser mucho (1 a 2 minutos por ítem). Además, se requiere una forma de calificación eficaz e inmediata, lo cual complica la variabilidad de respuestas.

De estas necesidades y limitaciones surge un compromiso entre exámenes con respuestas muy creativas versus la precisión y la velocidad de respuesta-corrección. Como solución al conflicto se concibió un nuevo tipo de reactivos que se aproximan a ítems de respuesta semi-construida, puesto que se utilizan recursos gráficos y de escritura, entre otros, para armar la respuesta.

Se pretende que las preguntas reflejen la diversidad de opciones que presenta la vida diaria y que le den la posibilidad al sustentante de crear su respuesta. Asimismo, su desarrollo se sustenta en las posibilidades de la tecnología computacional. Este tipo de evaluación busca acercarse al concepto de evaluación auténtica definido por Wiggins, Grant (1990), en el sentido en que se procura que los





estudiantes construyan explorando diferentes recursos, respondan a preguntas abiertas, combinen destrezas que integran materias o dominios y exploren soluciones múltiples, además de o en lugar de, evaluar solo un producto final o una respuesta única, correcta.

De estas características de los ítems se desprende otra noción importante, la cantidad de respuestas solicitadas en cada reactivo. El 30% de los ítems del Excoba/MS son dicotómicos (correcto-incorrecto) y el resto son de crédito parcial (de acuerdo con el grado de solución de la respuesta) (ver figuras 2 y 3).

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.

Durante 2014 se hicieron las primeras aplicaciones del examen. Entre ellas, la de la Universidad Autónoma de Aguascalientes (UAA), como herramienta de ingreso a su programa de Bachillerato General por competencias. Para ello, se generaron al azar tres versiones diferentes que se denominaron como versión 1 (V1) y versión 2 (V2) y versión 3 (V3). Especialistas en sistemas informáticos del Excoba administraron los distintos exámenes a los aspirantes a ingreso a las escuelas de Educación Media Superior de la UAA. Fue un total de 1711 estudiantes: 560 para V1, 580 para V2 y 571 para V3. Se obtuvo una base de datos con las respuestas de los estudiantes evaluados y, con dicha base depurada y organizada se efectuaron los análisis.

MÉTODO

Para determinar las propiedades métricas de los reactivos generados para integrar las versiones descritas se implementó para las tres versiones un procedimiento de modelación de rasgos latentes (LTM por sus siglas en inglés). Dadas las características de los reactivos y escalas, corresponde la extensión Rasch-Andrich para reactivos de crédito parcial (Embretson, 2000).

Esta variante trata las categorías de crédito parcial como una serie de dicotomías al interior de cada reactivo. Así por ejemplo, si un reactivo consta de cuatro partes, cada una de ellas cuenta .25 y, si un sustentante acierta a las 4 partes, acumula un punto, de lo contrario acumula solo las fracciones a las





que acierte, o bien cero si no acierta a ninguna. En este tipo de LTM los parámetros de interés son por una parte, la competencia del sustentante y las dificultades asociadas a cada categoría de crédito parcial.

La forma general de las probabilidades de respuesta dado un nivel de competencia θ a un reactivo de crédito parcial con 4 segmentos es:

$$p_{i0}(\theta) = 1.0 - p_{i1}^*(\theta)$$

$$p_{i1}(\theta) = p_{i1}^*(\theta) - p_{i2}^*$$

$$p_{i1}(\theta) = p_{i2}^*(\theta) - p_{i3}^*$$

$$p_{i1}(\theta) = p_{i3}^*(\theta) - p_{i4}^*$$

$$p_{i4}(\theta) = p_{i4}^*(\theta) - 0$$

En este caso el ítem que tienen un formato de respuesta uniforme cada ítem se describe por un parámetro de localización de la escala λ_i , que denota la relativa facilidad o dificultad de que el respondente pase de una categoría k a la siguiente $k+1$.

Además, cada una de las categorías $J = k-1$, tiene un umbral que se describe con el parámetro δ_j , y el nivel de rasgo (competencia) de cada respondente es θ .

En otras palabras las categorías de respuesta tienen intersecciones asignadas que se consideran distancias iguales entre sí en todos los ítems en de la escala.

Entonces, bajo esta estructura de ítems las probabilidades de respuesta para cada categoría se calculan por medio de:

$$P_{ix}(\theta) = \frac{\exp\{\sum_{j=0}^x[\theta - (\lambda_i + \delta_i)]\}}{\sum_{x=0}^M\{\sum_{j=0}^x[\theta - (\lambda_i + \delta_i)]\}}$$

Así, en esta expresión, las probabilidades de respuesta correcta a los reactivos de crédito parcial son una función del nivel de competencia θ de cada sustentante, la demanda cognitiva (dificultad o facilidad) de cada categoría de crédito parcial λ_i el umbral de paso de una categoría k a la siguiente $k+1$. Bajo este esquema analítico las categorías de crédito parcial segmentadas en fracciones .20 (5 categorías de CP) o en .25 (4 categorías de CP) o bien cualquier segmentación cuya suma sea igual a 1.00, deben cumplir con criterios de bondad de ajuste empleados como control de calidad métrica en el LTM-Rasch. En este caso se emplean .80 y 1.30 como límites del rango de bondad de ajuste interno y externo. En





esencia, cuando una categoría se mantiene en ese rango en ajuste interno significa que los sustentantes que calibraron en un nivel de competencia cercano a la exigencia cognitiva, aciertan y pasan el umbral hacia la siguiente categoría de crédito parcial. A la inversa, cuando una categoría se mantiene en ese rango en ajuste externo, significa que los sustentantes que calibraron en un nivel de competencia lejano a la exigencia cognitiva, no aciertan y no pasan el umbral hacia la siguiente categoría de crédito parcial.

En la figura 4 se observan 4 ítems que presentan “efecto techo”, 8 que presentan propiedades métricas dentro de los rangos de bondad de ajuste en ambos criterios y 8 que requieren revisión en su proceso de generación y opciones de crédito parcial.

De lo anterior se desprende que 8 de estos serán retenidos para su uso continuado mientras que los 12 restantes se retiran de la puntuación final de los sustentantes con los ajustes correspondientes y pasan a proceso de rediseño. En la figura 5 se observa la distribución uniforme de las categorías de respuesta correcta para un ítem de 4 segmentos de crédito parcial, donde la probabilidad de respuesta correcta a cada categoría $k + 1$ aumenta y disminuye conforme aumenta el nivel de rasgo θ . Finalmente la figura 6 muestra que para el reactivo en conjunto la probabilidad de respuesta correcta aumenta conforme aumenta el nivel de rasgo latente.

Con ello se ilustra que el análisis de propiedades métricas de reactivos de crédito parcial producido por generadores automáticos de ítems está matemáticamente establecido, es implementable con recursos computacionales existentes, y es riguroso en el control de calidad de los ítems sujetos a validación psicométrica.

DISCUSIÓN

La generación automática de reactivos de respuesta semi-construida y construida, con formatos binarios y de crédito parcial es un proceso implementado en México durante 2014 para evaluar competencias académicas al ingreso al nivel medio superior. De esta primera aproximación en esta línea de investigación evaluativa, se desprende que los procedimientos de control de calidad métrica necesarios están implementados y disponibles. En particular, las garantías de control de calidad de ítems e instrumentos administrados mediante sistemas informatizados, pueden y deben obtenerse adaptando recursos técnicos desarrollados en el marco de la psicometría contemporánea y en especial desde la familia de procedimientos de modelación de rasgos latentes (LTM).





TABLAS Y FIGURAS

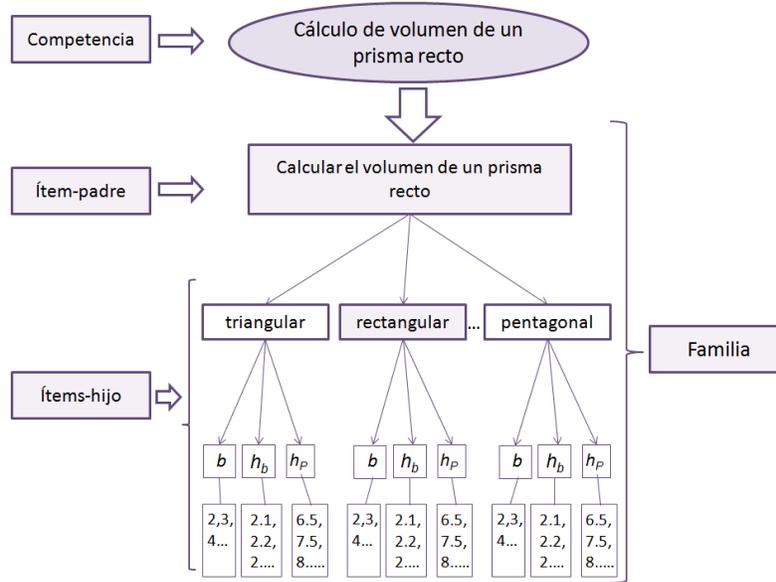


Figura 1. Esquema ejemplo de cómo se estructuran los reactivos de una competencia curricular.

The screenshot shows a digital interface for a mathematics problem. At the top left, the subject is labeled "Matemáticas". On the top right, there is a field for "Tipo de respuesta" (Type of answer). The main instruction is "Calcula el volumen del siguiente prisma." (Calculate the volume of the following prism). Below the text is a 3D diagram of a rectangular prism with dimensions: height is 3 cm, width is 2 cm, and length is 5 cm. At the bottom, there is a response field: "Respuesta: [] cm³".

Figura 2. Ejemplo de ítem de Matemáticas con respuesta dicotómica, tomado del Demo del Excoba/MS. Reproducido con la autorización de Métrica Educativa A.C.





Español Tipo de respuesta

Observa la gráfica que muestra los resultados de una encuesta a adultos acerca de si saben usar el cajero automático de los bancos (ATM). Lee los tres enunciados y clasifícalos entre los que son interpretaciones correctas de la gráfica y los que no los son.

Edad (años)	Sabe (%)	No sabe (%)	No está seguro (%)
55-60	65	15	20
60-65	25	45	30
65-70	40	20	40

El porcentaje de los que saben usar el ATM disminuye cuando aumenta la edad.

El grupo de edad con mayor porcentaje de personas que no saben usar el ATM está entre 65 y 70 años.

Aproximadamente el 45% de los adultos entre 60 y 65 años no sabe usar el ATM.

Correcta **Incorrecta**

Figura 3. Ejemplo de ítem de Español, con respuesta de crédito parcial, tomado del Demo del Excoba/MS. Reproducido con la autorización de Métrica Educativa A.C.





ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	OUTFIT ZEMP	PTMEASUR-CORR.	AL-EXP.	EXACT OBS%	MATCH EXP%	REACT	G		
2	0	756	6.86	1.72	MAXIMUM	MEASURE	.00	.00	100.0	100.0	MAT02	1		
3	0	756	6.86	1.72	MAXIMUM	MEASURE	.00	.00	100.0	100.0	MAT03	1		
4	0	756	6.86	1.72	MAXIMUM	MEASURE	.00	.00	100.0	100.0	MAT04	1		
5	0	756	6.86	1.72	MAXIMUM	MEASURE	.00	.00	100.0	100.0	MAT05	1		
19	30	756	3.53	.30	.97	.0	.42	-.8	.16	.11	98.6	98.6	MAT19	3
20	28	535	2.09	.19	.95	-.1	.71	-.4	.26	.22	94.7	94.6	MAT20	1
1	62	756	1.64	.13	1.02	.1	1.35	.7	.24	.26	91.6	91.6	MAT01	1
18	37	420	1.31	.17	.91	-.3	.61	-.7	.34	.27	91.2	90.8	MAT18	1
8	125	471	1.28	.12	1.17	.5	.75	-.2	.32	.31	92.9	91.2	MAT08	2
7	200	698	1.27	.10	1.37	1.3	1.06	.1	.28	.32	91.5	90.6	MAT07	2
16	173	756	.32	.09	1.15	1.4	1.16	.7	.30	.39	74.8	78.5	MAT16	1
15	189	756	.19	.09	1.05	.6	1.09	.4	.37	.40	75.5	77.1	MAT15	1
6	705	715	.17	.06	1.05	.4	1.32	.7	.50	.50	66.4	67.0	MAT06	2
14	211	756	.01	.09	.94	-.7	.83	-1.0	.47	.42	76.8	75.4	MAT14	1
10	1017	756	-.89	.08	.86	-2.1	.83	-1.5	.57	.48	77.0	70.6	MAT10	3
11	424	756	-1.47	.08	.86	-1.9	.80	-1.7	.59	.51	76.7	71.6	MAT11	1
9	1101	379	-1.82	.08	.89	-.7	2.01	1.8	.61	.66	60.4	65.2	MAT09	2
13	1428	756	-1.84	.08	.93	-.9	.95	-.4	.56	.52	76.3	74.0	MAT13	3
17	562	756	-2.53	.09	.99	.0	1.08	.4	.51	.52	81.1	80.5	MAT17	1
12	503	608	-3.27	.12	1.24	1.4	1.85	1.9	.36	.51	83.7	86.7	MAT12	1
MEAN	339.8	682.7	1.37	.44	1.02	-.1	1.05	.0			81.8	81.5		
P.SD	412.0	123.4	3.18	.64	.14	1.0	.41	1.0			10.4	10.2		

Figura 4. Índices de bondad de ajuste de 20 ítems en la escala de Matemáticas



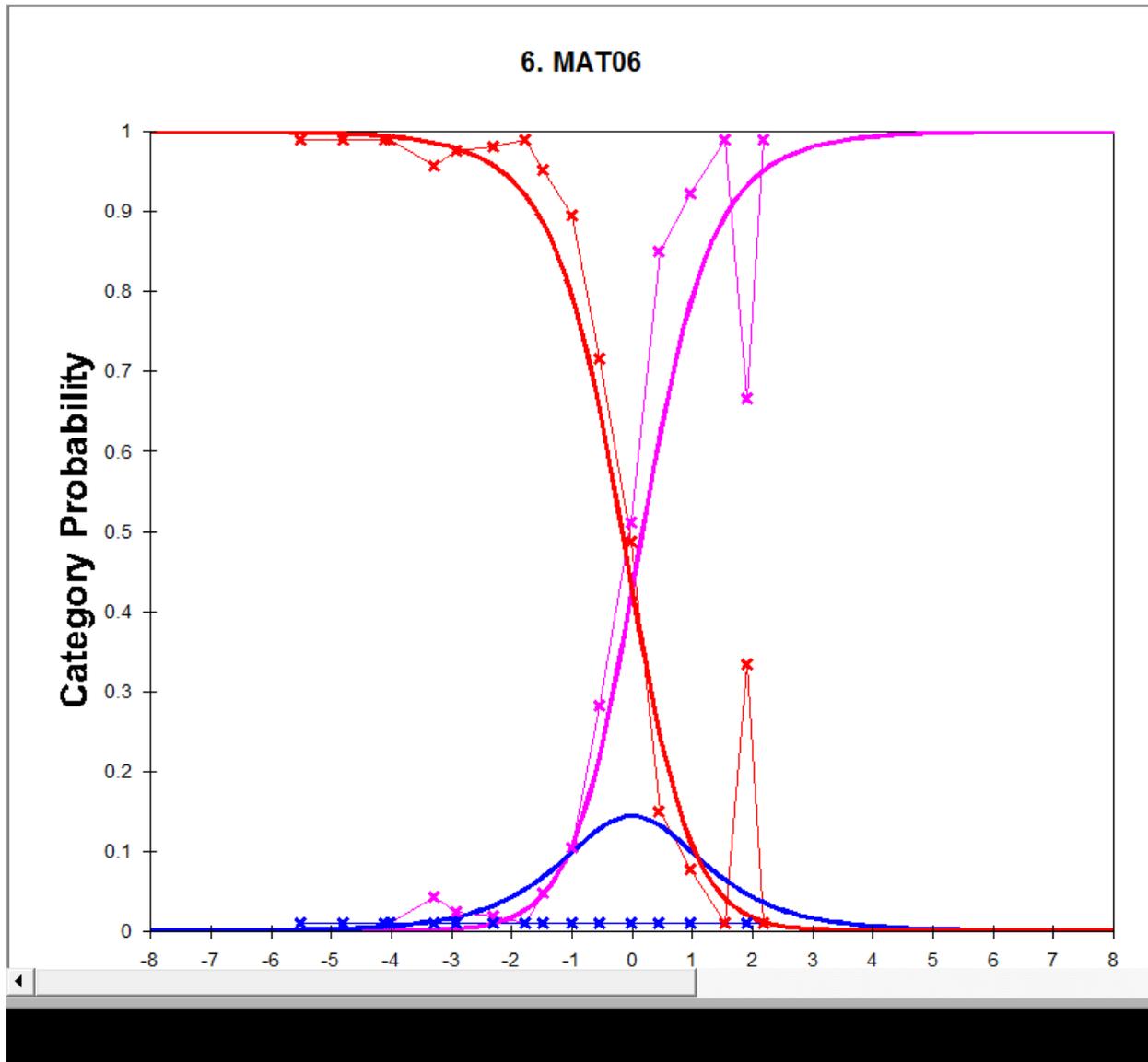


Figura 5. Curvas características de reactivo con 4 categorías CP su trayectoria empírica



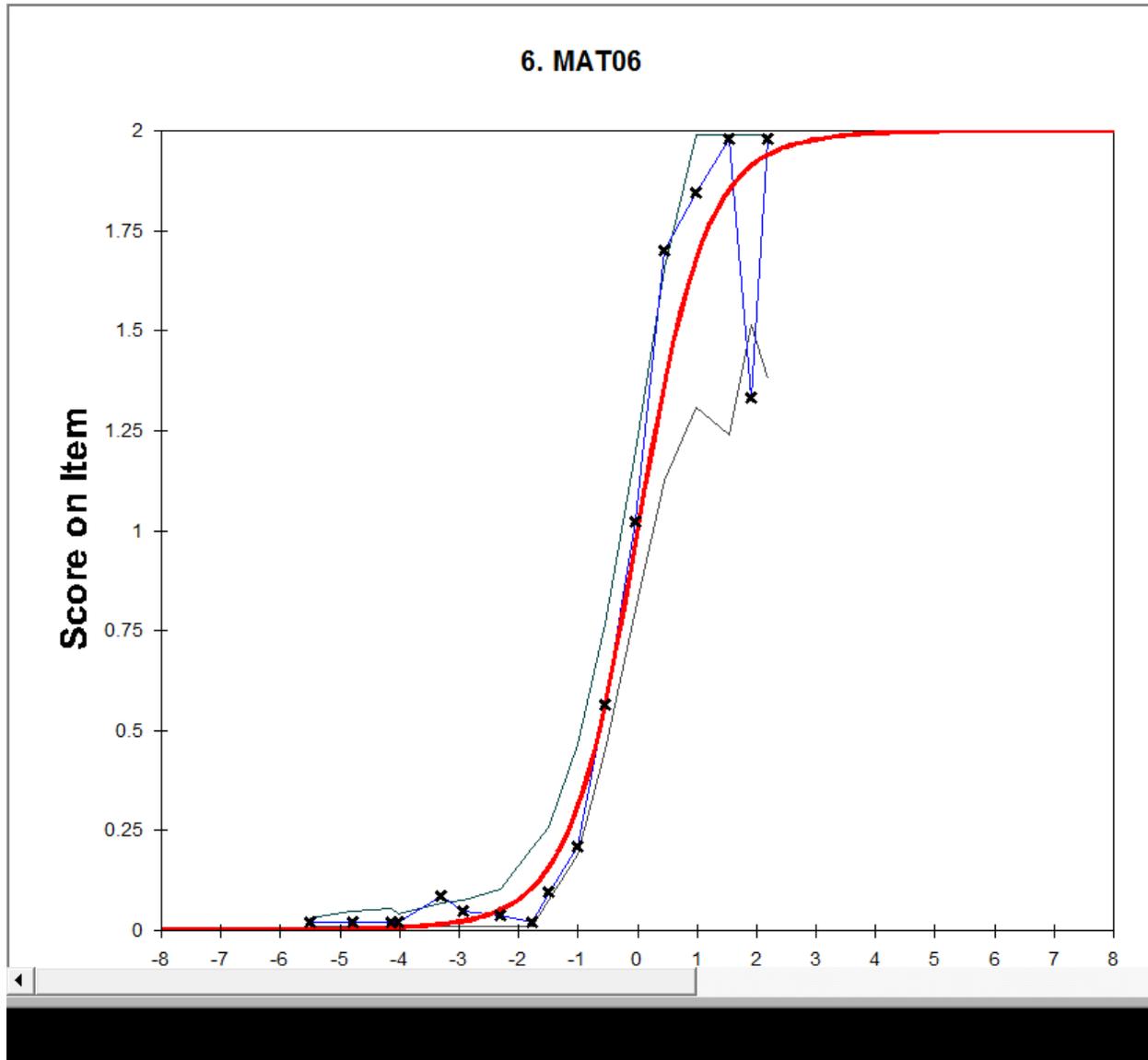


Figura 6. Curvas características del mismo reactivo que ilustra que en conjunto la probabilidad de respuestas correctas aumenta conforme aumenta el nivel de rasgo de -6 hasta +6.





REFERENCIAS:

Embretson, S. y Reise, P. (2000). Item response theory for psychologists.

Lawrence Erlbaum Associates, Mahaw N.J. London.

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

Linacre, J. M. (© 2015). A User's Guide to Winsteps Ministeps: Rasch-Model Computer Programs. Chicago, IL: Electronic Publication. www.winsteps.com

Ng., A. W. Y. y Chan, A. H. S. (marzo, 2009). Different methods of multiple-choice test: Implications and design for further research. En Proceedings of the International MultiConference of Engineers and Computer Scientists 2009: Vol II. Hong Kong: IMECS.

Resnick, L. B. y Resnick, D. P. (1992). Assessing the thinking curriculum: New tools for educational reform. En B. R. Gifford y M. C. O'Connor (Eds.), Changing assessments:

Alternative views of aptitude, achievement, and instruction (pp. 37-75). Boston: Kluwer Academic Publishers.

Wiggins, Grant (1990). The case for authentic assessment. Practical Assessment, Research & Evaluation. Recuperado: 11 de mayo de 2015, de <http://files.eric.ed.gov/fulltext/ED328611.pdf>

