Evaluación de competencias con reactivos de crédito parcial producidos por generadores automáticos de ítems

Manuel Jorge Gonzalez-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

Temática general: Curriculum, conocimientos y prácticas educativas

Tipo de ponencia: Reporte de investigación parcial

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: Generación Automática, Respuesta Semi-construida, ítems de crédito parcial, propiedades métricas.

Propuesta de Palabras Clave: Teoría de Respuesta al Ítem, Psicometría, Competencias, Educación básica, Exámenes de ingreso o Pruebas de admisión, validez de pruebas (deben ser 5 como máximo)

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.

En este caso los ítems 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 de la escala.

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

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

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.

Con ello se ilustra que el análisis de propiedades métricas de reactivos de crédito parcial producidos por generadores automáticos de ítems es implementable y riguroso.

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 egreso del 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.

Tablas y figuras

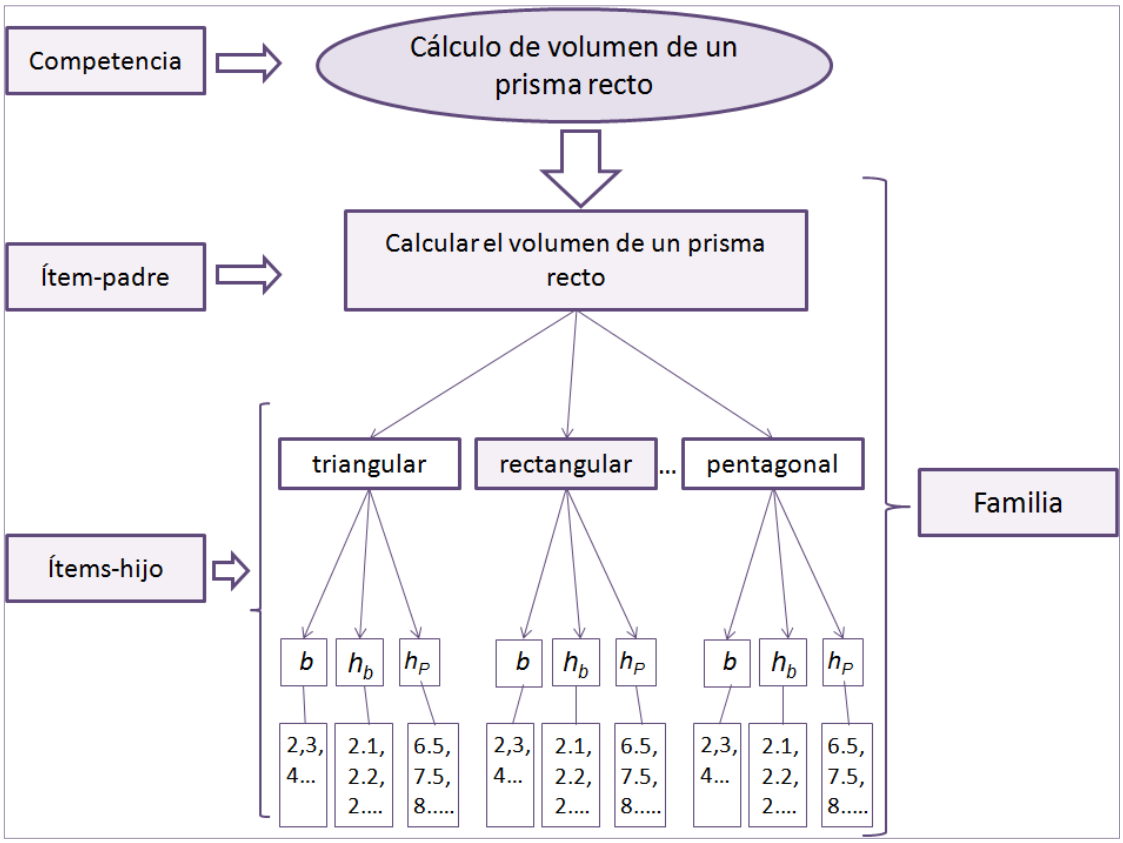

Figura 1. Esquema ejemplo de cómo se estructuran los reactivos de una competencia curricular.

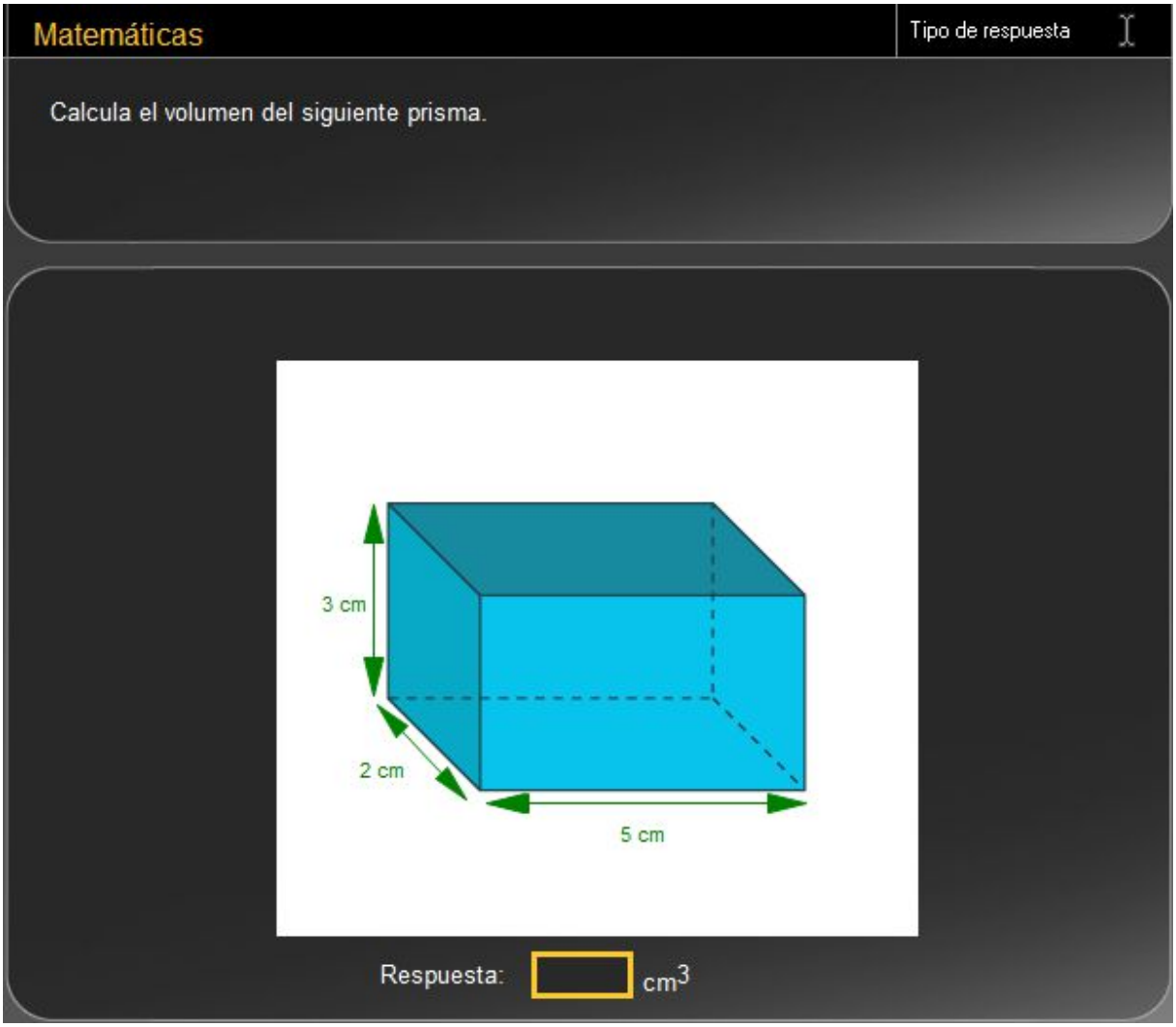

Figura 2. Ejemplo de ítem de Matemáticas con respuesta dicotómica, tomado del Demo del Excoba/MS. Reproducido con la autorización de Métrica Educativa A.C.

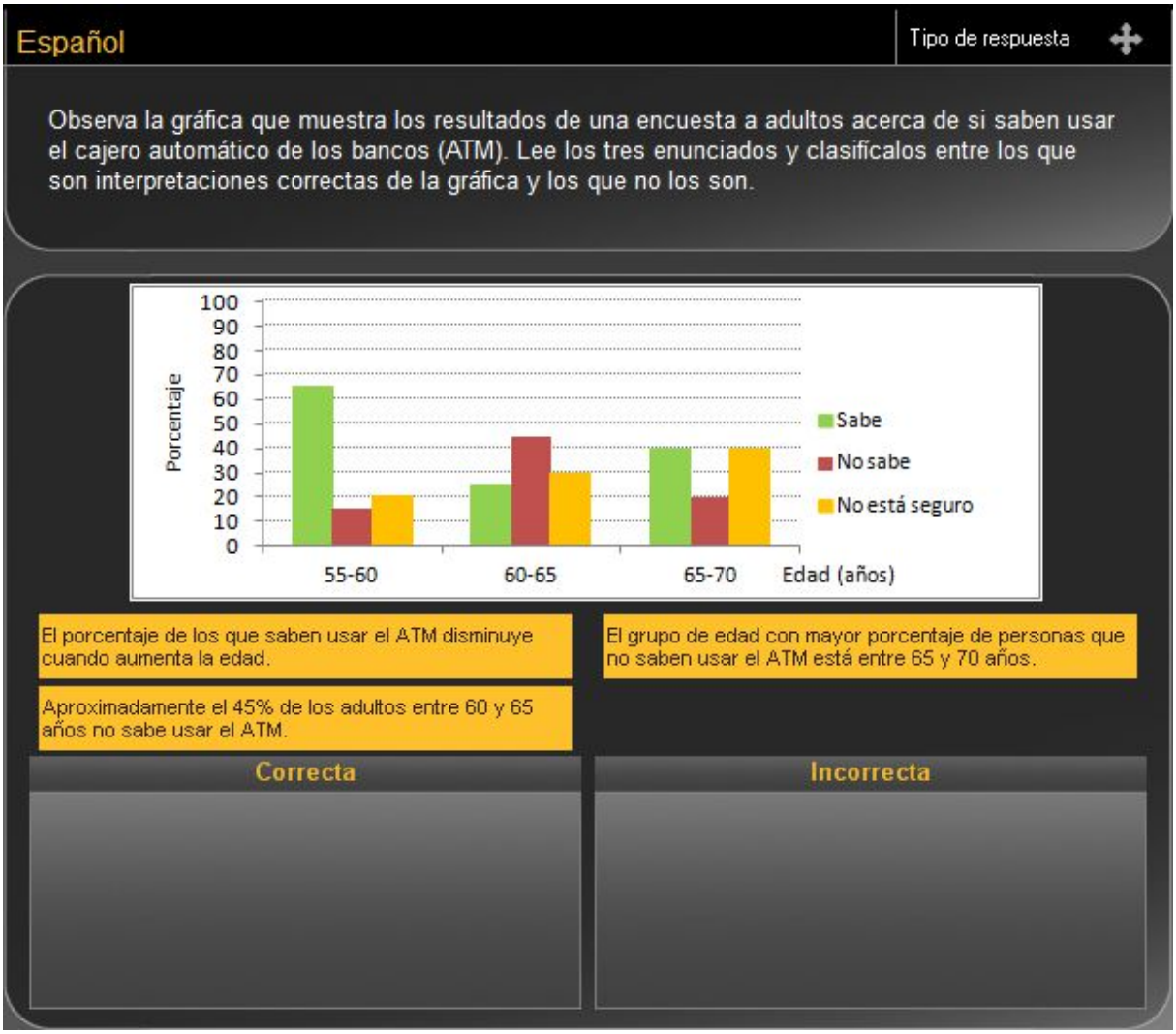

Figura 3. Ejemplo de ítem de Español, con respuesta de crédito parcial, tomado del Demo del Excoba/MS. Reproducido con la autorización de Métrica Educativa A.C.

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZEMP	OUTFIT MNSQ	OUTFIT ZEMP	PTMEASUR-AL CORR.	PTMEASUR-AL EXP.	EXACT MATCH OBS%	EXACT 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

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ESTIM DISCR	ITEM	G
19	4	810	4.54	.50	.99	.2	.57	-.9	.08	.06	99.5	99.5	1.02	M19	0
1	66	810	1.56	.13	1.03	.3	1.25	1.4	.20	.21	91.6	91.6	.98	M01	0
15	87	810	1.23	.12	1.00	.0	.98	-.1	.24	.24	88.9	89.0	1.00	M15	0
20	271	810	.90	.07	1.08	1.4	1.08	1.1	.33	.38	65.3	70.0	.91	M20	0
7	137	810	.88	.08	1.10	1.1	1.14	.8	.28	.31	84.3	84.7	.96	M07	0
8	356	810	.60	.07	1.07	1.4	1.15	2.6	.36	.41	60.4	65.3	.89	M08	0
18	535	810	.19	.05	.92	-1.5	1.03	.4	.52	.49	55.9	53.3	1.05	M18	0
6	277	810	.08	.06	.99	-.2	.86	-.8	.42	.41	71.7	71.0	1.01	M06	0
16	210	810	.02	.09	1.06	1.5	1.03	.5	.29	.33	73.4	75.0	.90	M16	0
14	227	810	-.10	.08	.88	-3.3	.80	-3.4	.44	.34	76.5	73.6	1.26	M14	0
10	286	810	-.50	.08	.83	-5.6	.79	-4.6	.50	.37	76.9	69.4	1.46	M10	0
11	407	810	-1.24	.08	.87	-5.1	.85	-4.1	.51	.41	74.0	66.3	1.48	M11	0
13	454	810	-1.53	.08	.83	-6.4	.79	-5.6	.55	.42	74.4	67.2	1.56	M13	0
9	1379	810	-1.64	.05	1.13	2.7	1.17	3.4	.55	.62	50.0	52.4	.82	M09	0
17	590	810	-2.46	.09	.99	-.1	1.03	.5	.45	.45	78.5	77.3	.99	M17	0
12	2013	810	-2.53	.05	1.29	3.5	2.83	8.4	.56	.66	63.0	67.6	.82	M12	0
MEAN	456.2	810.0	.00	.11	1.00	-.6	1.08	.0			74.0	73.3			
S.D.	507.2	.4	1.69	.11	.12	2.9	.48	3.3			12.8	12.4			

ENTRY NUMBER	TOTAL SCORE	TOTAL COUNT	MEASURE	MODEL S.E.	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD	PT-MEASURE CORR.	PT-MEASURE EXP.	EXACT MATCH OBS%	EXACT MATCH EXP%	ESTIM DISCR	ITEM	G
2	0	810	7.84	1.82		MAXIMUM MEASURE			.00	.00	100.0	100.0		M02	1
3	0	810	7.84	1.82		MAXIMUM MEASURE			.00	.00	100.0	100.0		M03	1
4	0	810	7.84	1.82		MAXIMUM MEASURE			.00	.00	100.0	100.0		M04	1
5	0	810	7.84	1.82		MAXIMUM MEASURE			.00	.00	100.0	100.0		M05	1
19	4	810	3.60	.50	.99	.2	.55	-1.0	.07	.05	99.5	99.5	1.02	M19	3
1	66	810	2.31	.13	.99	.0	1.43	2.3	.18	.19	91.6	91.6	1.00	M01	1
15	87	810	1.99	.12	.96	-.4	.94	-.3	.22	.21	88.7	89.0	1.02	M15	1
7	137	810	1.25	.09	1.25	2.5	1.10	.7	.26	.26	85.4	83.9	1.01	M07	2
16	210	810	.86	.08	.96	-.6	.97	-.3	.28	.31	73.8	74.8	1.00	M16	1
14	227	810	.75	.08	.78	-4.0	.73	-3.7	.41	.32	76.5	73.5	1.22	M14	1
20	271	810	.48	.08	1.11	1.8	1.09	1.3	.31	.34	66.9	70.1	.84	M20	1
6	277	810	.48	.06	1.29	4.3	1.04	.4	.39	.36	70.3	69.3	1.12	M06	2
8	356	810	.18	.06	.80	-3.8	1.23	2.5	.34	.39	60.6	63.5	.87	M08	2
11	407	810	-.18	.07	.61	-7.6	.73	-4.9	.49	.40	73.7	63.3	1.39	M11	1
18	535	810	-.67	.06	1.42	6.2	1.41	6.1	.48	.44	46.4	60.3	.42	M18	1
17	590	810	-.85	.06	.54	-9.4	.62	-7.4	.46	.46	72.4	59.1	1.45	M17	1
10	286	810	-1.39	.08	.83	-5.9	.79	-4.2	.47	.35	76.9	68.9	1.49	M10	3
9	1379	810	-2.37	.05	1.08	1.5	1.13	2.5	.56	.62	51.2	54.8	.79	M09	2
13	454	810	-2.41	.08	.83	-6.4	.79	-5.6	.53	.41	74.4	67.0	1.57	M13	3
12	2013	810	-4.04	.06	1.53	7.2	2.17	8.1	.59	.69	61.4	66.0	.71	M12	1
MEAN	364.9	810.0	1.57	.45	1.00	-.9	1.05	-.2			73.1	72.2			
S.D.	488.9	.0	3.56	.69	.27	4.7	.38	4.1			13.7	12.3			

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