Validación del área de Ciencias sociales del Examen de Competencias Básicas producido por el Generador Automático de Exámenes (GenerEx)

María Fabiana Ferreyra
Métrica Educativa A.C.
fferreyra@metrica.edu.mx

Yadira Pérez-Garibay
Métrica Educativa A.C.
yperez@metrica.edu.mx

Temática general: Curriculum, conocimientos y prácticas educativas

Tipo de ponencia: Reporte de investigación parcial

Resumen

El Examen de Competencias Básicas (Excoba) es una prueba de admisión a la educación media superior y a la educación superior. Es un examen computarizado creado por un generador automático de reactivos denominado GenerEx que produce familias de ítems isomorfos para evaluar una competencia específica. La Generación Automática de Ítems requiere de una constante validación de los exámenes generados, sobre todo, si es un instrumento que se aplica en gran escala y de alto impacto, como es este caso. El propósito de este estudio fue presentar las propiedades psicométricas y de estructura interna de tres versiones del Excoba producidas por el GenerEx, en el área de Ciencias sociales. La investigación se llevó a cabo con tres tipos de análisis: la Teoría Clásica de los Tests, la Teoría de Respuesta al Ítem y el Análisis Factorial Confirmatorio. Los resultados aportan evidencia de validez de estructura interna con una notable mejoría respecto de un estudio previo; además, reafirman el supuesto de que el GenerEx produce exámenes isomorfos en el área de Ciencias sociales.

Palabras clave: evaluación de estudiantes, evaluación del aprendizaje, validez de las pruebas, análisis estadístico, pruebas de admisión

Introducción

El Examen de Competencias Básicas (Excoba) es un instrumento de evaluación, que se aplica a gran escala, para seleccionar estudiantes que aspiran ingresar a instituciones de educación media superior y de educación superior. Si bien se trata de un test computarizado, su mayor originalidad consiste en que a través de un generador, denominado GenerEx, se pueden producir cientos de ítems que evalúan una misma habilidad y, a su vez, generar miles de versiones diferentes del mismo examen, combinando los distintos ítems (Ferreyra, 2014; Ferreyra y Backhoff-Escudero, 2013).

Para cada competencia definida se elabora un modelo de reactivos. Este modelo contiene las características del contenido a evaluar y una instrucción que funciona como esquema vacío que se completa con diferentes elementos. De modo que, cada combinación de elementos se denomina ítem-hijo de una misma familia de reactivos que evalúan una competencia (ver figuras 1 y 2).

El Excoba/MS, que se administra para el ingreso a la Educación Media Superior, se organiza en seis áreas: dos correspondientes a la educación primaria (Español y Matemáticas) y cuatro de educación secundaria (Español, Matemáticas, Ciencias naturales y Ciencias sociales). Cada área contiene 20 competencias seleccionadas de los planes de estudio vigentes de educación básica de México (SEP, 2011). A su vez, cada competencia se evalúa con un reactivo. De este modo, el examen se conforma de un total de 120 ítems.

Antes de la primera aplicación del Excoba, se inició un proceso de validación de estructura interna del GenerEx. Para ello, se propuso una metodología, se efectuaron pilotajes del

examen, se obtuvieron los primeros resultados y se realizaron algunas modificaciones, en función de las recomendaciones que surgieron de los análisis (Ferreyra, 2014).

Durante 2014, se iniciaron las administraciones del Excoba en algunas de las instituciones usuarias, se generaron bases de datos y se planteó la necesidad de poner a prueba nuevamente el método de validación para confirmar o refutar los resultados obtenidos tras los pilotajes; como así también, constatar qué tan apropiados habían sido los cambios realizados tras el primer proceso de validación. Dada la extensión que implica todo el trabajo, se decidió presentar una de las seis áreas del examen: Ciencias sociales.

De acuerdo con lo expuesto, la pregunta que guió la investigación es: ¿cuáles son las propiedades psicométricas y la estructura interna de las distintas versiones producidas por el GenerEx, en el área de Ciencias sociales? Para responderla se propuso: analizar tres versiones de Ciencias sociales del Excoba, producidas, al azar, por el GenerEx, con fin de aplicar la metodología propuesta en Ferreyra (2014) y así, aportar evidencias de validez de estructura interna del Excoba/MS.

Contenido

El área de Ciencias sociales está compuesta por las asignaturas: Geografía de México y del mundo, Historia, y Formación Cívica y Ética (ver tabla 1).

El tipo de preguntas que se utilizan en esta área es de respuesta semiconstruida y de crédito parcial. Para cada ítem, el estudiante examinado debe contestar más de una pregunta y armar la respuesta entre múltiples opciones (ver ejemplo de ítem en la figura 2). Cada reactivo contiene de tres a cinco preguntas que se califican parcialmente con un valor de hasta 1 punto, según el número de respuestas correctas. Si el ítem contiene 3 preguntas, cada acierto vale 0.33 y si son 5, cada uno es de 0.20. Por lo tanto, el área total tiene un puntaje máximo de 20 puntos.

Método

Para realizar el estudio comparativo del GenerEx, se generaron al azar tres exámenes paralelos, en el área temática de Ciencias sociales. Estos exámenes 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 grupos de aspirantes a ingresar a las escuelas de Educación Media Superior correspondientes a la Universidad Autónoma de Aguascalientes. 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.

Los estudios comparativos de los reactivos del área se realizaron 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 efectuaron con el modelo de Rasch para ítems de crédito parcial (Masters, 1982).

En relación con el objetivo del trabajo, se analizó el isomorfismo (dificultad y pertenencia al constructo) de los ítems-hijo de una misma familia, como parte del área de Ciencias sociales, en las tres versiones. Para ello, se obtuvieron los siguientes indicadores de la TCT: dificultad, correlación punto-biserial y confiabilidad; del modelo de Rasch: medida, nivel de ajuste (interno y externo), correlación punto-medida y discriminación; y además, los índices y las cargas factoriales correspondientes a la agrupación de ítems en esta área.

Los análisis estadísticos se realizaron con la ayuda de los programas Winsteps (Linacre, 2010) y EQS 6.1 (Bentler, 2006). En la tabla 2 se muestran los criterios y límites establecidos por el Comité Técnico de Métrica Educativa, para evaluar la calidad del área y de los ítems en lo individual, de acuerdo con el modelo psicométrico utilizado.

Resultados

La tabla 3 muestra que las tres versiones del área de Ciencias sociales poseen, desde la TCT, distribuciones similares. Las medias y dispersiones de respuestas correctas resultan parecidas, lo mismo que su simetría y curtosis (considerando el error, las curvas son mesocúrticas y con asimetría hacia la izquierda, debido a la presencia de un grupo aproximado de 60 personas con calificación nula). La confiabilidad de las versiones es alta y similar, lo que refuerza estos resultados.

Se estudió también la dificultad por ítem de cada familia en los exámenes. La figura 3 indica las distancias entre los índices de dificultad (p) de ítems-hijo de la misma familia en las tres versiones. Cuanto menor es esa distancia, más similares en dificultad son los reactivos que evalúan la misma competencia. Se observaron diferencias inferiores a 0.15 para casi todas las familias de ítems, las mejores fueron: HIS07, HIS11 y FCYE16; los casos más disímiles fueron: GEO02, GEO03, HIS08, HIS09 e HIS14, donde la distancia entre el valor más alto y el más bajo fue de 0.20.

Desde el análisis Rasch (ver figura 4), se observa que para las tres versiones, la media de las dificultades de los ítems es ligeramente mayor que la media de las habilidades de los examinados; también, en los tres casos se nota un grupo alejado de 60 personas, aproximadamente, con habilidad menor que -2. HIS07 es la familia de reactivos más difícil, alejada del resto, con una dificultad de 2, le sigue GEO01 con una dificultad cercana a 1. El resto de los ítems se agrupa entre -1 y 1. Esta información concuerda con los resultados en dificultad aportados por el modelo de la TCT.

En la figura 5 se muestran problemas de aleatoriedad, tanto cerca como lejos de la zona de medición de las familias de ítems de GEO02 y GEO04; esto se reflejó además, en los índices

de discriminación, que son inferiores a la cota de 0.8. También se encontraron ligeros conflictos de determinismo lejos de la zona de medición en las familias de HIS13, FCYE16 y FCYE18, pero que no afectaron a la discriminación.

El otro aspecto importante a analizar es cómo correlacionan los ítems y cuál es su agrupación en constructos. Se obtuvieron dos índices de correlación: punto biserial (TCT) y punto medida (TRI). En cuanto al primero, los valores promedio fueron 0.64 para V1, 0.63 para V2 y 0.60 para V3. Los índices promedio de correlación punto-medida fueron de 0.62, para V1 y V2 y 0.60 para V3. Los reactivos con correlaciones más bajas se encontraron en GEO01 e HIS07; sin embargo, en ambas familias superó el valor mínimo de 0.2, para las tres versiones. El resto de los ítems presentaron índices superiores a 0.5, incluso en algunos casos llegó a 0.8 (HIS13). Estos resultados indican valores muy buenos y parecidos para tres exámenes que representan a un mismo constructo o rasgo latente (ver figura 6).

Para el AFC se probaron diferentes modelos y el que presentó mejores índices de ajuste (ver tabla 4) fue el de tres factores que covarían: (1) Geografía, (2) Historia y (3) Formación Cívica y Ética (ver figura 7). En los tres exámenes el 90% de los reactivos superaron el valor de 0.60 de carga factorial, lo cual implica una fuerte pertenencia al constructo; las cargas más bajas fueron para GEO01 e HIS07, aunque superaron 0.30 para las tres versiones, excepto HIS07 en V3, que fue de 0.25. Estos resultados son totalmente congruentes con las correlaciones calculadas en la figura 6.

Conclusiones

Los resultados obtenidos se pueden comparar con los del pilotaje realizado durante 2012, con excepción de HIS07, puesto que este reactivo no se pudo analizar en la primera aplicación debido a errores técnicos. En la primera oportunidad fueron dos versiones del examen, con un

total de 400 casos para la primera y 300 para la segunda; en esta ocasión, se trató de tres versiones con más de 500 casos en cada una. La otra diferencia importante es que los resultados de 2012 no contaron como calificación para los estudiantes evaluados, mientras que la de 2014 fue considerada como instrumento de ingreso. Es decir, la primera aplicación fue de bajo impacto y la segunda, de alto impacto.

De la tabla 5, se puede inferir que las tendencias se confirmaron y que mejoró la calidad del área. La estructura interna es excelente, de modo que las versiones son muy similares en cuanto a representación del constructo de Ciencias sociales y de ítems-hermanos muy parecidos (en el sentido de que evalúan la misma competencia). Los promedios de las dificultades difieren en una centésima para versiones del mismo año; pero no ocurre lo mismo entre ítems de una misma familia (diferencias de hasta 0.20 en una misma familia).

Como consecuencia del análisis de estos resultados se propone la implementación de algún método de equiparación de pruebas (e.g.: equipercentil o calibración desde la TRI) para solucionar el conflicto de diferencias de dificultad entre ítems-hermano. Además, se sugiere revisar los modelos de ítems de GEO02 y GEO04, a fin de evitar problemas de aleatoriedad.

Para concluir, es importante destacar que ambos estudios, 2012 y 2014, revelan que el GenerEx funciona como un muy buen generador de exámenes isomorfos, al representar un mismo rasgo latente para el área de Ciencias sociales.

Tablas y figuras

Tabla 1.
Distribución de los contenidos de Ciencias sociales por asignatura, grado y bloque, de acuerdo con los programas de estudio de 2011.

Asignatura	Contenido	Grado	Bloques	Clave
Geografía de México y del mundo	Representación del espacio geográfico	1º	I	GEO01
	Diversidad natural de la Tierra y desarrollo sustentable		II y V	GEO02
	Crecimiento, composición y distribución de la población		III	GEO03
	Causas y consecuencias de la migración		III	GEO04
	La globalización y desigualdad socioeconómica		IV	GEO05
	Diversidad cultural y globalización		III y IV	GEO06
Historia	De principios del siglo XVI a principios del siglo XVIII	2º	I	HIS07
	De mediados del siglo XVIII a mediados del siglo XIX		II	HIS08
	De mediados del siglo XIX a 1920		III	HIS09
	El Mundo entre 1910 y 1960		IV	HIS10
	Décadas recientes		V	HIS11
	Las culturas prehispánicas y la conformación del Virreinato de Nueva España	3º	I	HIS12
	Nueva España, desde su consolidación hasta la Independencia		II	HIS13
	Instituciones revolucionarias y desarrollo económico (1910-1982)		III y IV	HIS14
Formación Cívica y Ética	Autorregulación y ejercicio responsable de la libertad	2º	II	FCYE16
	Responsabilidades en la vida colectiva		III	FCYE17
	El reto de aprender a convivir		III	FCYE20
	La democracia como forma de vida		IV	FCYE18
	Organización del Estado mexicano	3º	IV	FCYE19
	La participación social y política en la vida democrática		IV	FCYE15

Tabla 2.
Criterios asumidos para los análisis estadísticos de los ítems del Excoba/MS

Modelos psicométricos	**Estadísticos**	**Criterio**	
		Aceptable	Bueno
TCT	Correlación punto biserial	≥ 0.2	
	Alfa de Cronbach (α)	≥ 0.6	
TRI	Correlación punto medida	≥ 0.2	
	Infit-Outfit MNSQ	≥ 0.8 y ≤ 1.2	
	Discriminación	≥ 0.8	
AFC	Carga factorial	≥ 0.20	≥ 0.30
	χ^2	≥ 0.01	≥ 0.05
	NNFI	≥ 0.90	≥ 0.95
	CFI	≥ 0.90	≥ 0.95
	RMSEA	< 0.08	< 0.05

Tabla 3.
Indicadores de tendencia central, dispersión, normalidad y confiabilidad del área de Ciencias sociales para V1, V2 y V3.

Indicador	V1	V2	V3
N	560	580	571
Media	8.16 (p = .41)	8.24 (p = .41)	8.33 (p = .42)
Desviación Estándar	4.61	4.33	4.20
Rango	0-16.33	0-15.80	0-16.13
Simetría	-0.562	-0.733	-0.760
Curtosis	-0.938	-0.625	-0.471
Confiabilidad (α)	.939	.938	.929

Tabla 4.
Índices de ajuste para el AFC del área de Ciencias sociales para V1, V2 y V3.

Indices	V1	V2	V3
Chi cuadrado (p)	0.00	0.00	0.00
NNFI	0.96	0.95	0.95
CFI	0.97	0.95	0.95
RMSEA	0.05	0.05	0.05
Cov F1-F2	0.88	0.92	0.90
Cov F1-F3	0.71	0.73	0.73
Cov F2-F3	0.87	0.83	0.83

Tabla 5.
Cuadro comparativo de Ciencias sociales para las aplicaciones de 2012 y 2014.

Aplicación		2012	2014
Confiabilidad (α)		0.86 y 0.87	0.94 y 0.93
Distribución		platocúrtica, asimetría hacia la izquierda	mesocúrtica, asimetría hacia la izquierda
Promedio dificultad		0.47 y 0.48	0.41 y 0.42
Diferencias dificultad por ítem		Mayor diferencia = 0.23	Mayor diferencia = 0.21
Ítem más difícil		GEO01	HIS07, GEO01
Ajuste y discriminación		GEO04: *infit/outfit* > 1.5 con discriminación = 0.5	GEO02: *infit/outfit* > 1.4 con discriminación = 0.6 GEO04: *infit/outfit* entre 1.2 y 1.4.
Correlaciones		Promedio = 0.48 (ptbis), 0.53 (ptmed) Máx = 0.62 Mín = 0.22	Promedio = entre 0.60 y 0.64 (ambos tipos de índices) Máx = 0.81 (ptbis) 0.77 (ptmed) Mín = 0.23
AFC	Modelo	3 factores	3 factores
	Ajuste	Muy bueno	Muy bueno
	Menor carga	GEO01	HIS07 y GEO01
	Cargas	90% superiores a 0.40	90% superiores a 0.60

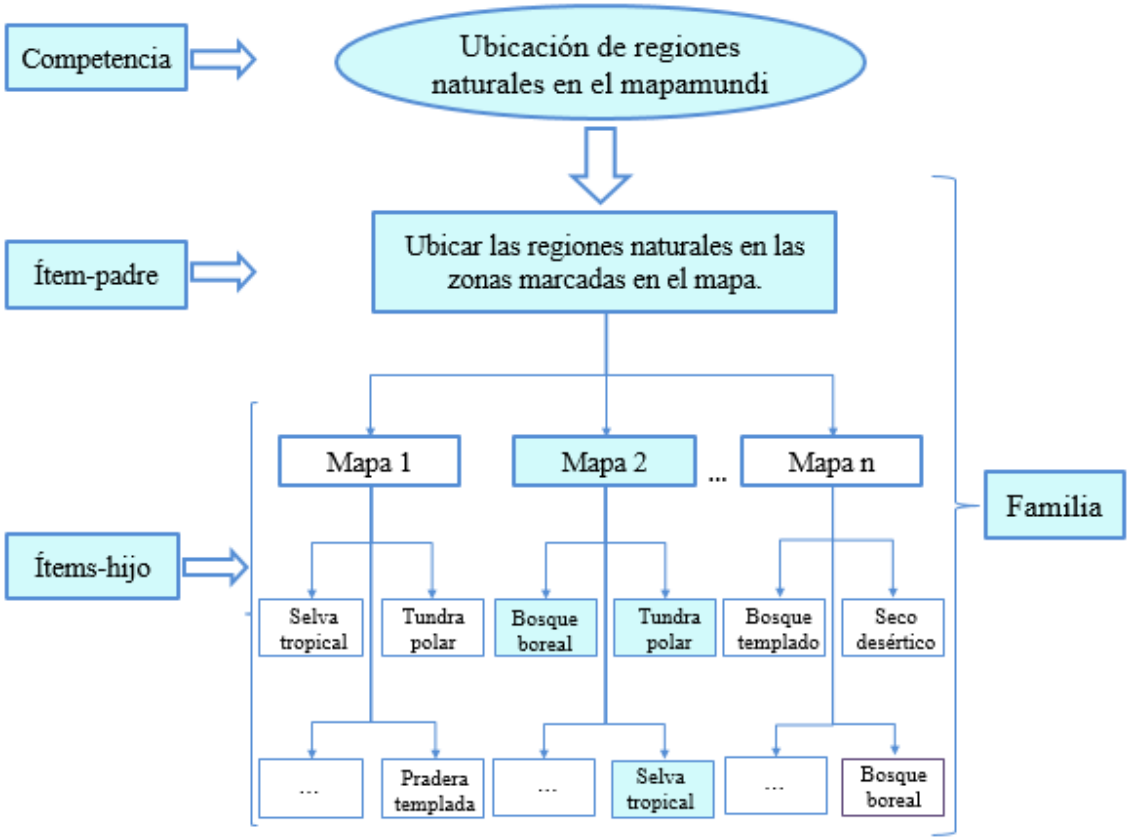

Figura 1. Esquema ejemplo de cómo se estructuran los reactivos de una competencia curricular.

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

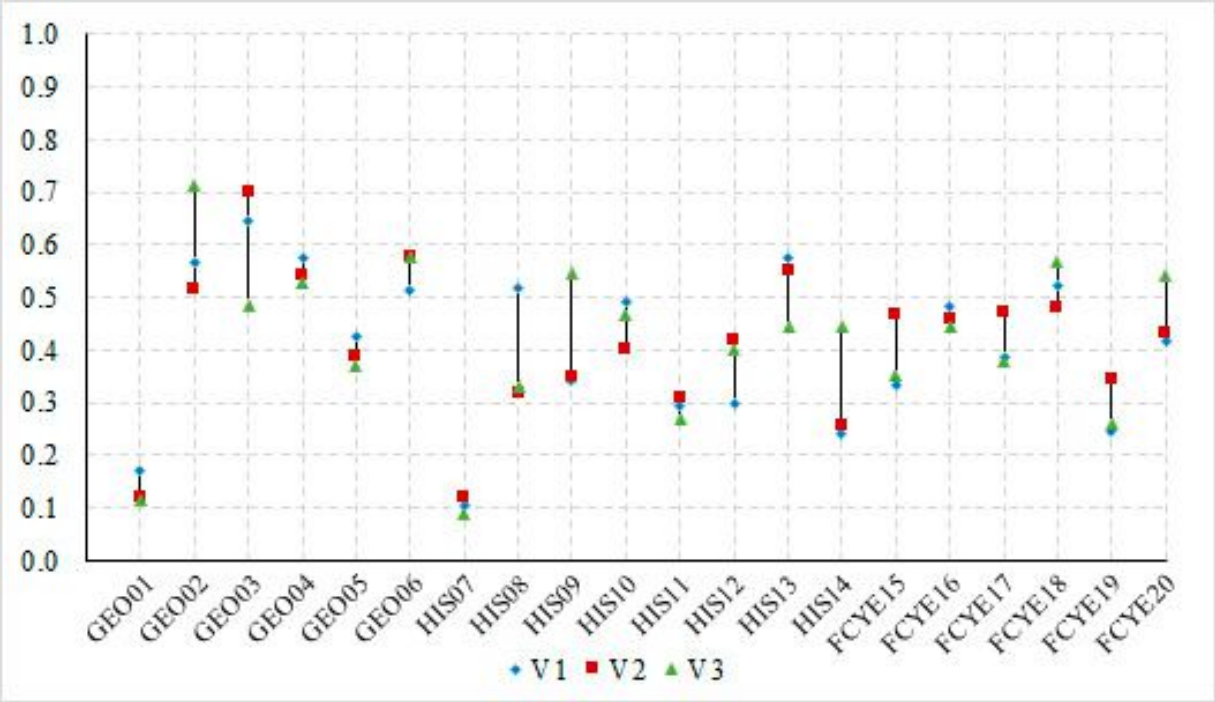

Figura 3. Distribución de dificultades por ítem del área de Ciencias sociales para las versiones V1, V2 y V3.

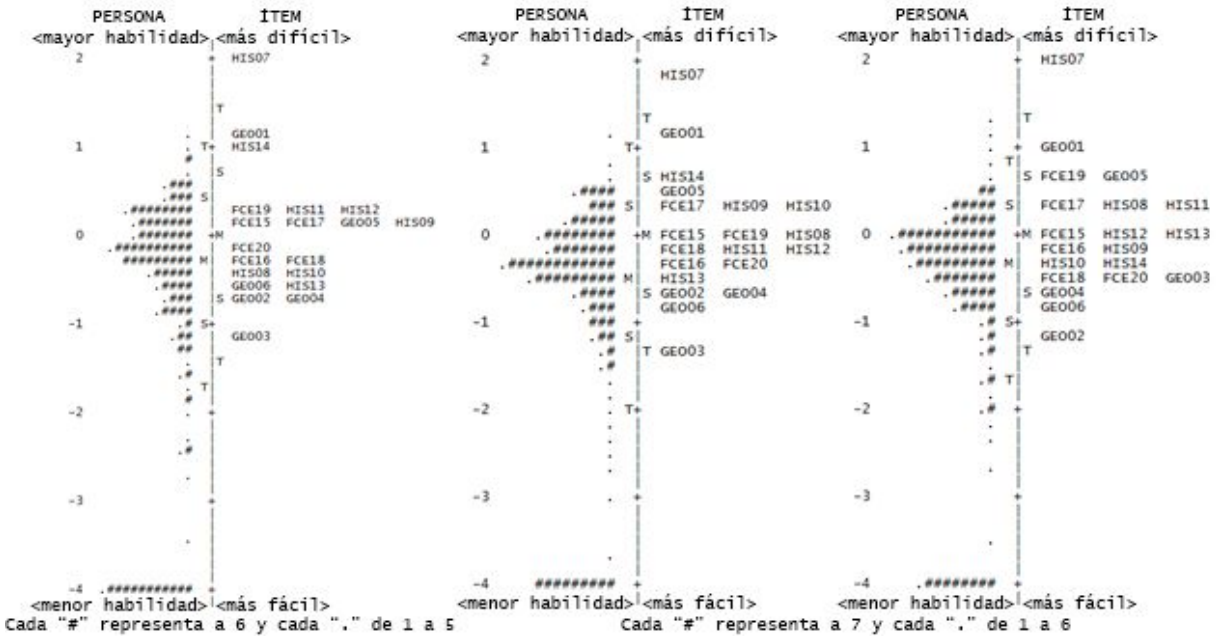

Figura 4. Mapas de Wright del área Ciencias sociales para V1 (izquierda), V2 (centro) y V3 (derecha).

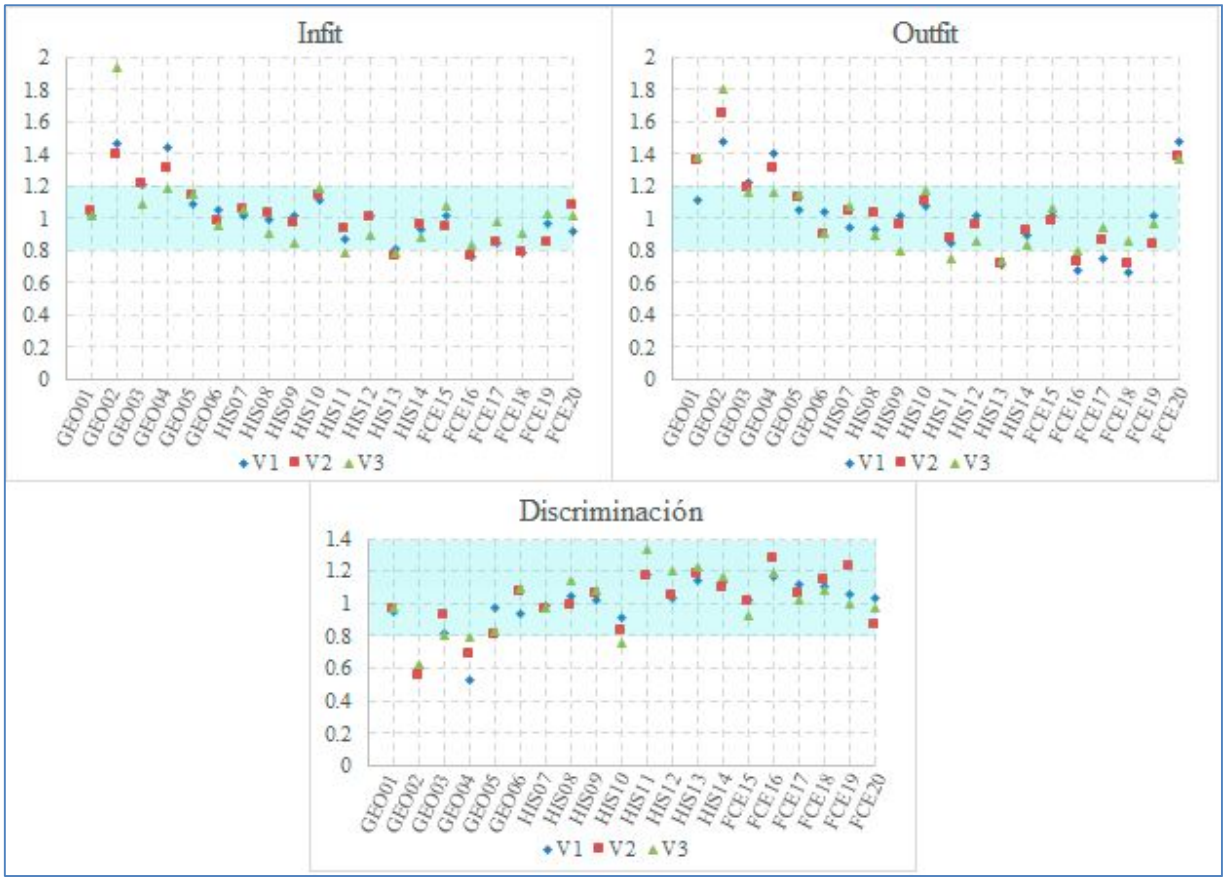

Figura 5. Gráficas de índices de *infit*, *outfit* y discriminación del área Ciencias sociales, para V1, V2 y V3.

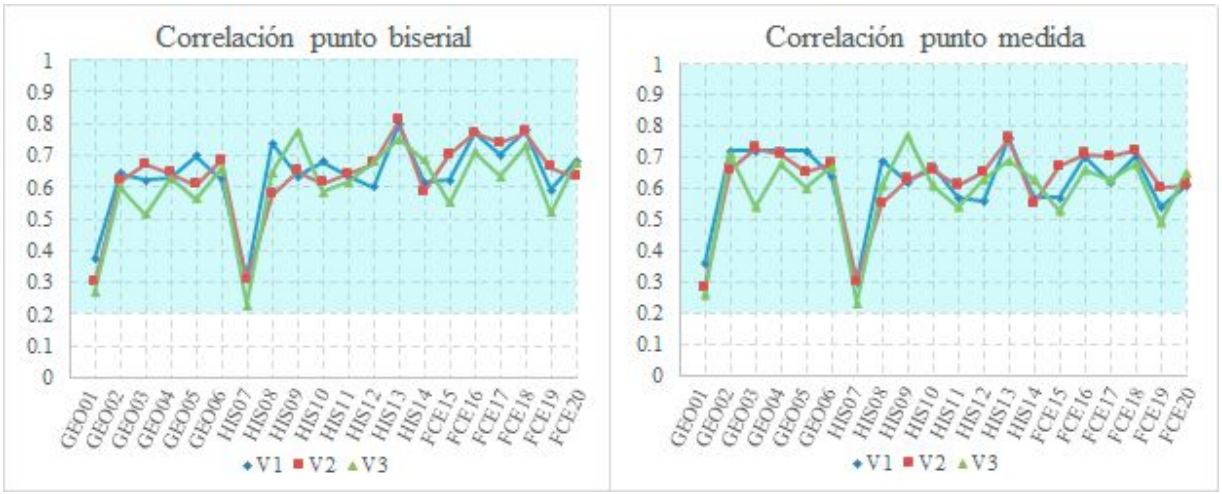

Figura 6. Gráficas de correlaciones del área Ciencias sociales, para V1, V2 y V3

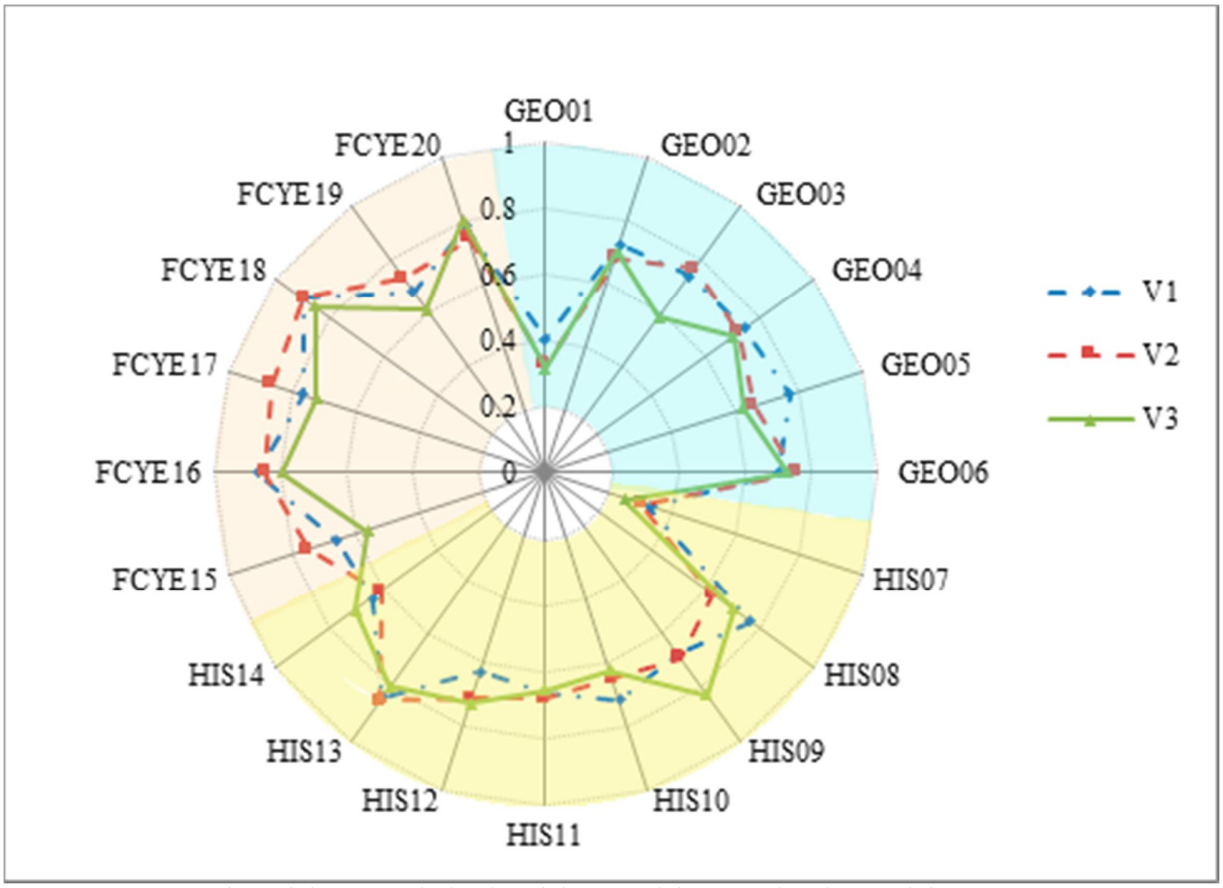

Figura 7. Cargas factoriales estandarizadas del AFC del área Ciencias sociales, para V1, V2 y V3. Modelo de tres factores que covarían.

Referencias

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

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.

Ferreyra M. F. y Backhoff-Escudero, E. (2013, noviembre). Modelo para validar un generador automático de exámenes. Trabajo presentado en el X Congreso Nacional de Investigación Educativa, Guanajuato: México.

Linacre, J. M. (2010b). Winsteps® (Version 3.70.0.2) [Software]. Beaverton, OR: Winsteps.com

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47* (2), 149-174.

Secretaría de Educación Pública. (2011). *Plan de estudios 2011. Educación Básica*. Autor: México.