Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Metodología para analizar la estructura interna de un generador automático de reactivos

TESIS

Que para obtener el grado de

DOCTORA EN CIENCIAS EDUCATIVAS

Presenta

María Fabiana Ferreyra

Ensenada, B. C., México, febrero de 2014

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Doctorado en Ciencias Educativas

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

Metodología para analizar la estructura interna de un generador automático de reactivos

TESIS

Que para obtener el grado de

DOCTORA EN CIENCIAS EDUCATIVAS

Presenta

María Fabiana Ferreyra

APROBADA POR:

Dr. Eduardo Backhoff Escudero
Director de tesis

Dr. Manuel Jorge González Montesinos
Sinodal

Dra. Norma Larrazolo Reyna
Sinodal

Dr. Luis Lizasoain Hernández
Sinodal

Dr. Javier Organista Sandoval
Sinodal

Dr. Juan Carlos Rodríguez Macías
Sinodal

DCE
DOCTORADO EN CIENCIAS EDUCATIVAS

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctor en Ciencias Educativas.

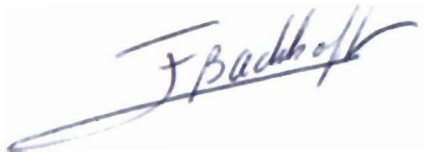
Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente



Dr. Eduardo Backhoff Escudero

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctor en Ciencias Educativas.

Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dra. Norma Larrazolo Reyna

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctor en Ciencias Educativas.

Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Manuel Jorge González Montesinos

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctor en Ciencias Educativas.

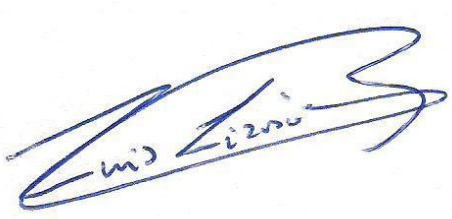
Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente



Dr. Luis Lizasoain Hernández

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo de tesis para el grado de Doctor en Ciencias Educativas.

Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Javier Organista Sandoval

Ensenada, B.C. a 17 de enero de 2014

ASUNTO: Voto aprobatorio al trabajo
de tesis para el grado de Doctor en Ciencias Educativas.

Dr. Lewis Samson McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por la **C. MARÍA FABIANA FERREYRA**, me permito comunicarle que he dado mi **VOTO APROBATORIO** al mencionado trabajo. Con base en lo anterior, dicho documento se considera listo para su defensa en el examen de grado de Doctor en Ciencias Educativas, sobre su trabajo titulado:

Metodología para analizar la estructura interna de un generador automático de reactivos

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Juan Carlos Rodríguez Macías

Dedicatoria

A vos, papá

Agradecimientos

A Dios, por todas las gracias con que me bendijo.

A mi madre, Carlos, Anita y Aramí, las fuentes de amor, paciencia y energía en mi trabajo.

A mi director de tesis, Eduardo Backhoff, a quien admiro, quiero y respeto. Por su apoyo, sus enseñanzas, su entrega al trabajo, que fueron inspiración para mi estudio.

A los miembros del Comité de tesis: Norma Larrazolo, por su calidez y su apoyo incondicional en todo mi trabajo; Luis Lizasoain, por su generosidad y disposición a colaborar en mis incursiones por la psicometría Manuel Montesinos, por compartir conmigo sus amplios conocimientos del modelamiento de Rasch; Javier Organista Sandoval, por estar siempre presente cuando lo necesité para orientar y discutir mi trabajo, y Juan Carlos Rodríguez, por el entusiasmo de sus clases y el fervor con que me transmitió sus conocimientos.

Al Dr. Víctor Corral, por el aporte de su experiencia en el campo de los Modelos de Ecuaciones Estructurales.

A mis compañeros de doctorado, con quienes compartí este intrincado camino de la investigación, en especial a Citlalli y Juan Carlos, compañeros fieles y solidarios.

A Gaby, por su amistad incondicional, su sensibilidad y su calidez. Además, por sus correcciones de estilo, que otorgaron claridad a la presentación de la tesis.

Al personal del EXHCOBA: Martín, José Luis, Eduardo y, especialmente, Yadira, que siempre estuvo atenta a mis comentarios y mis necesidades, dispuesta a ayudar en todo.

A todos los que conforman el IIDE. A los investigadores, de quienes recibí conocimientos que enriquecieron mi vida académica. A los administrativos, en particular a Julio, Estrella y Yésica, por su empeño en que saliera adelante con mis trámites engorrosos.

A la Universidad Autónoma de Baja California que otorgó las facilidades para estudiar en esta institución.

Al Conacyt, por el apoyo económico que hizo posible mi dedicación exclusiva a esta tarea de estudiar un posgrado.

A todos aquellos que, de forma anónima, colaboraron para la culminación de mi doctorado.

Índices

Índice general

Índice de tablas

Índice de figuras

Siglas y acrónimos

ACT	antes, American College Test
AERA	American Educational Research Association
AF	Análisis Factorial
AFC	Análisis Factorial Confirmatorio
APA	American Psychological Association
ATAR	Australian Tertiary Admission Rank
BIO	Asignatura de Biología del área de Ciencias naturales del EXHCOBA-R/MS
CCI	Curva Característica del Ítem
CEER	Comités Elaboradores de Especificaciones y Reactivos
CENEVAL	Centro Nacional de Evaluación de la Educación Superior
CESUES	Centro de Estudios Superiores del Estado de Sonora
CETYS	Centro de Enseñanza Técnica y Superior
CFI	Índice Comparativo de Ajuste
EE. UU.	Estados Unidos de América
EMS	Educación Media Superior
ENEM	Examen Nacional de Enseñanza Media
Enlace	Evaluación Nacional de Logro Académico en Centros Escolares
ES	Instituciones de Educación Superior
ESP	Área de Español del EXHCOBA-R/MS
ETS	Educational Testing Service
EXANI-I	Examen Nacional de Ingreso a la Educación Media Superior
EXANI-II	Examen Nacional de Ingreso a la Educación Superior
Excale	Exámenes de Calidad y Logro Educativos
EXHCOBA	Examen de Habilidades y Conocimientos Básicos
EXHCOBA/MS	Examen de Habilidades y Conocimientos Básicos para el ingreso a la Educación Media Superior
EXHCOBA-R	Examen de Habilidades y Conocimientos Básicos Renovado
EXHCOBA-R/MS	Examen de Habilidades y Conocimientos Básicos Renovado para el ingreso a la Educación Media Superior
FCYE	Asignatura de Formación cívica y ética del área de Ciencias sociales del EXHCOBA-R/MS
FERF	*Family Expected Response Function* (Función de Respuesta Esperada por Familia)
FIS	Asignatura de Física del área de Ciencias naturales del EXHCOBA-R/MS
GAI	Generación Automática de ítems
GEO	Asignatura de Geografía del área de Ciencias sociales del EXHCOBA-R/MS
GLTM	Modelo General del Rasgo Latente
GMAT	Graduate Management Achievement Test
GMT	*gramática de modelo de tarea*
GRE	Graduate Record Examination
UGTO	Universidad de Guanajuato
HC	Área de Habilidades matemáticas del EXHCOBA-R/MS
HIS	Asignatura de Historia del área de Ciencias sociales del EXHCOBA-R/MS
HSTP	High School Placement Test
HV	Área de Habilidades del lenguaje del EXHCOBA-R/MS
IE	Ingeniería de los Tests
INEE	Instituto Nacional para la Evaluación de la Educación

IPN	Instituto Politécnico Nacional
ISEE	Independent School Entrance Examination
ISM	*Identical Siblings Model*
LLTM	Modelo de Test Logístico Lineal
MAT	Área de Matemáticas del EXHCOBA-R/MS
NAT	Área de Ciencias naturales del EXHCOBA-R/MS
NBT	National Benchmark Test
NCEE	National Higher Education Entrance Examination
NCME	National Council on Measurement in Education
NNFI	Índice de Ajuste no Normalizado
PAA	Prueba de Aptitud Académica
PAEG	Pruebas de Acceso a Enseñanzas Universitarias Oficiales de Grado
PFLC	Escuela Preparatoria Federal "Lázaro Cárdenas".
PNFI	Ajuste Normalizado de Parsimonia
PSU	Prueba de selección universitaria
QUI	Asignatura de Química del área de Ciencias naturales del EXHCOBA-R/MS
REESCO	Reactivos Estructurales Constructivos
RMSEA	Error Medio Cuadrático de Aproximación
RSM	*Related Siblings Model*
SAT	Scholastic Aptitud Test
SAT	antes, Scholastic Aptitude Test
SEP	Secretaría de Educación Pública
SICODEX	Sistema Computarizado de Exámenes
SOC	Área de Ciencias sociales del EXHCOBA-R/MS
SON-R 5.5-17	Nombre de un test de inteligencia no verbal
SPSS	Statistical Package for Social Science
STAT	Special Tertiary Admissions Test
TCT	Teoría Clásica de la Medida
TOEFL	Test of English as a Foreign Language
TRI	Teoría de Respuesta al Ítem
UABC	Universidad Autónoma de Baja California
UACJ	Universidad Autónoma de Ciudad Juárez
UAQ	Universidad Autónoma de Querétaro
UGTO	Universidad de Guanajuato
UNAM	Universidad Nacional Autónoma de México
UNISON	Universidad de Sonora
VA	Versión A del EXHCOBA-R/MS para obtención de una muestra
VB	Versión B del EXHCOBA-R/MS para obtención de una muestra

Resumen

Esta tesis presenta una metodología para mostrar evidencias de validez de estructura interna del EXHCOBA-R/MS, examen desarrollado mediante la Generación Automática de Ítems (GAI) de teoría débil. Se incluyen los resultados y el análisis de la aplicación metodológica en diversas muestras de la prueba. Finalmente se lleva a cabo una discusión sobre las bondades y las limitaciones del método propuesto, y se exponen las conclusiones.

Se definieron dos niveles de análisis psicométricos para el desarrollo de la metodología. El nivel examen estudia las propiedades psicométricas de los reactivos de la prueba en su conjunto y de las seis áreas del conocimiento en que se organizó el test (Habilidades del lenguaje, Habilidades matemáticas, Español, Matemáticas, Ciencias naturales y Ciencias sociales). El nivel familia de ítems investiga qué tan isomorfos son los ítems-hijo de una misma familia de reactivos tomando en cuenta el grado de dificultad y la competencia que los define.

Los análisis psicométricos se basaron en la Teoría Clásica de los Tests, la Teoría de Respuesta al ítem (a través del modelo de Masters, derivado del modelo de Rasch) y el Análisis Factorial Confirmatorio (AFC). De este modo, se pudo dar información desde la teoría más básica y tradicional (Teoría Clásica de los Tests) y desde el modelo más parsimonioso de la Teoría de Respuesta al Ítem, como lo es el modelo de Rasch. Además, la inclusión del AFC, permitió corroborar la organización teórica del examen frente a la agrupación empírica de los datos.

Se definieron los tipos de muestras necesarias para dichos análisis. Posteriormente, se administraron, a modo de pilotaje, los diferentes exámenes producidos por la GAI, se obtuvieron las muestras y se analizaron las bases de datos asociadas. Los resultados de las aplicaciones

indican un buen comportamiento, en general, del examen. En el análisis parcial por áreas, dos de ellas presentaron un buen comportamiento estadístico: Ciencias sociales y Habilidades matemáticas (educación primaria). Con respecto a Español (educación secundaria), Matemáticas, Ciencias naturales y Habilidades del lenguaje (educación primaria), la primera de ellas presentó una dificultad demasiado baja con buenos índices de ajuste y correlación, la segunda reflejó extrema dificultad para las muestras, mientras que las dos últimas mostraron algunas deficiencias que deberán atenderse y subsanarse.

El método permitió identificar la calidad psicométrica de los ítems que se obtienen tras la GAI. La metodología implementada se considera una dirección apropiada de validación para exámenes no adaptativos de GAI de teoría débil que abarcan cantidades considerables de contenidos y diferentes áreas de conocimiento, como sucede en el EXHCOBA-R/MS. Este método resulta una alternativa efectiva y sencilla cuando no existe una teoría cognitiva subyacente que avale la GAI y soporte la implementación de modelos componenciales. La metodología podría perfeccionarse si se establece una cantidad apropiada de ítems-hermano para el análisis y si se utilizan otros modelos de la TRI; por ejemplo, el de dos parámetros.

La *validación de la GAI* es un campo novedoso, sobre todo en México, donde el EXHCOBA-R ha dado los primeros pasos en este nivel. Se trata de un área de investigación prometedora, dada la necesidad de generadores de exámenes para la evaluación a gran escala y el escaso desarrollo que presenta, particularmente en este país.

Palabras clave: Generación Automática de Ítems, Validez, Estructura interna, EXHCOBA-R

1

Introducción

El uso de exámenes a gran escala ha aumentado de manera notable y a nivel mundial durante las últimas décadas. Países, estados y distritos escolares utilizan particularmente estas evaluaciones para tomar decisiones de alto impacto con efectos en los estudiantes. Es por esto último que resulta importante y necesario garantizar la calidad de las pruebas utilizadas (Heubert y Hauser, 1999).

En México existen pruebas criteriales a gran escala dirigidas a estudiantes de nivel básico y de educación media superior, como los Exámenes de Calidad y Logro Educativos (Excale) y la Evaluación Nacional de Logro Académico en Centros Escolares (Enlace). Como ejemplos de pruebas normativas aplicadas para el ingreso a la Educación Media Superior (EMS) o a la Educación Superior (ES) pueden citarse el Examen Nacional de Ingreso a la Educación Media Superior (EXANI-I) y el Examen Nacional de Ingreso a la Educación Superior (EXANI-II), ambos desarrollados por el Centro Nacional de Evaluación para la Educación Superior (CENEVAL); el PIENSE II y la Prueba de Aptitud Académica (PAA) elaborados por el College Board (College Board, 2008 y Keller, Deneen y Magallán, 1991) y el Examen de Habilidades y Conocimientos Básicos (EXHCOBA).

El EXHCOBA fue desarrollado en 1992 por Backhoff y Tirado. La administración computarizada de esta prueba surgió un año más tarde (Backhoff, Ibarra y Rosas, 1996). Este examen se estableció como requisito de ingreso en la Universidad Autónoma de Baja California (UABC) y en otras instituciones educativas de educación superior mexicanas. La prueba lleva más de dos décadas de aplicaciones. En el año 2012 quince instituciones recibieron los servicios del examen y fueron evaluados alrededor de 120,000 estudiantes.

El EXHCOBA cuenta hasta el momento con diferentes versiones, todas referentes al desarrollado originalmente en 1992. Sin embargo, los cambios curriculares en la educación básica mexicana han generado la necesidad de redefinir los contenidos y las habilidades. Asimismo, el prototipo de ítems de opción múltiple utilizado en este examen ha reflejado ciertas limitaciones como la manera artificial de evaluar el aprendizaje, la propensión a la adivinación, la necesidad de una constante actualización debido al desgaste de su uso y el poco aprovechamiento de las nuevas ventajas de las tecnologías digitales.

Es por eso que se ha realizado un nuevo planteamiento de la evaluación de las competencias escolares apoyado en las virtudes de los recursos digitales. El resultado ha sido el EXHCOBA-R, un examen basado en los planes de estudio vigentes de la educación básica y de la educación media superior. El examen cuenta con un generador automático de reactivos que puede originar miles de exámenes similares con tipos de ítems de respuesta construida o semiconstruida.

Junto al desarrollo del EXHCOBA-R ha surgido también la necesidad de asegurar su validez y confiabilidad. Por ello, el objetivo principal de esta investigación fue el desarrollo de un método estadístico que aportara evidencias acerca de la validez de estructura interna de este examen aplicado en su versión para el ingreso a la Educación Media Superior (EHCOBA-R/MS). A continuación se presentan una reseña sobre los orígenes, virtudes y limitaciones del EXHCOBA para comprender la introducción del EXHCOBA-R como parte de los exámenes producidos por Generación Automática de ítems (GAI), la exposición del problema, las preguntas que guiaron la investigación, los objetivos y la justificación de este estudio.

1.1. Del EXHCOBA al EXHCOBA-R

El EXHCOBA, que surgió en 1992, se aplica sistemáticamente y de manera computarizada desde hace 22 años. Se utiliza como parte de los diferentes sistemas de ingreso a la ES. También cuenta con una versión simplificada para ingreso a la EMS. Entre las instituciones usuarias se encuentran la Universidad de Guanajuato (UGTO), la Universidad Autónoma de Querétaro (UAQ), la Universidad de Sonora, la Universidad Autónoma de Ciudad Juárez (UACJ), el Colegio Madrid y la Escuela Preparatoria Federal Lázaro Cárdenas (PFLC).

Los contenidos del examen fueron seleccionados por especialistas en contenidos curriculares con el objetivo de evaluar conocimientos y habilidades *básicos* para cursar el nivel próximo de estudios. El propósito del EXHCOBA es identificar en qué grado quedaron en el estudiante los saberes escolares fundamentales después de cursar 12 años de instrucción. Esto incluye contenidos de educación básica y bachillerato.

El examen consta de 190 preguntas de opción múltiple para el caso de la ES y de 130 para la EMS. Cada ítem presenta cuatro opciones de respuesta y una quinta opción denominada "no sé". Esta última tiene el objetivo de minimizar la adivinación (Tirado y Backhoff, 1999). Además, la prueba incluye diferentes versiones de reactivos que se seleccionan al azar para formar el examen que responderá el estudiante.

El sistema computarizado para administrar este instrumento se denomina *Sistema Computarizado de Exámenes* (SICODEX). Se desarrolló un año después del surgimiento del EXHCOBA y se aplica de forma ininterrumpida desde su creación. Este sistema surgió con el propósito de presentar, resolver y calificar completamente el examen por computadora (Backhoff, Ibarra, Rosas, 1995). El SICODEX ha experimentado numerosos perfeccionamientos durante estos años con el fin de aumentar la consistencia y la seguridad en las aplicaciones

(Rosas, Ramírez, Larrazolo, 2009). Aunque su diseño es complejo, su aplicación resulta fácil y ágil, ya que el trabajo se realiza en una red local. Asimismo se cuida que la red local no presente saturación de tráfico, las preguntas aparezcan de manera instantánea y las respuestas se ejecuten con la misma velocidad.

La resolución del examen por computadora es sencilla y amigable, requiere solamente que los estudiantes tengan conocimiento sobre el uso de las funciones básicas del teclado y del ratón. El examinado puede contestar las preguntas en cualquier orden, modificar sus respuestas tantas veces lo desee y llevar un control del tiempo y de las preguntas contestadas. La prueba se califica en forma inmediata. Por lo que el tiempo, los costos y los recursos humanos (ahorra el empleo del lector óptico y la posibilidad de error humano) del examen se optimizan a la vez que este instrumento otorga seguridad y transparencia en los resultados.

El uso del EXHCOBA interviene en la decisión del futuro de aproximadamente 120,000 estudiantes al año. Por su importancia es una necesidad garantizar la validez y confiabilidad de este instrumento. A tal efecto, existen numerosos estudios que aportan evidencias de validez de la prueba (e.g.: Backhoff y Larrazolo, 2001; González-Montesinos, 2004). En la página oficial del examen (http://www.exhcoba.mx/) se exhibe una lista de investigaciones asociadas al tema.

El EXHCOBA es un examen sólido y confiable. Sin embargo, está basado en una estructura conceptual de hace 20 años y con un prototipo de ítems de opción múltiple. De cierto modo está limitado para los nuevos conceptos de evaluación y por el aprovechamiento escaso de las ventajas aparecidas en la tecnología digital en las últimas dos décadas desde el desarrollo del instrumento.

El EXHCOBA-R surgió como una necesidad de superar estas limitaciones. Además de una renovación en los contenidos, el nuevo instrumento incluye dos grandes cambios en la

estructura original del examen: los Reactivos Estructurales Constructivos y la Generación Automática de Ítems. En los siguientes párrafos se presentan estos dos aspectos, los cuales se describen con mayor profundidad en el capítulo dos.

Reactivos Estructurales Constructivos (REESCO). Estos reactivos son un nuevo prototipo de ítems organizados en diferentes tipos que permiten respuestas construidas o semiconstruidas. Los REESCO solicitan al sustentante elaborar una respuesta. Dicha respuesta consiste en escribir la solución numérica o algebraica de un ejercicio, ubicar elementos en categorías, efectuar una selección múltiple u organizar datos, entre otras. Algunos de estos ítems son dicotómicos (correcto-incorrecto) y otros son de crédito parcial (admiten respuestas más o menos correctas).

Generación Automática de Ítems (GAI). La GAI consiste en el proceso que —por medio de modelos de ítems— genera una gran cantidad de ítems calibrados con ayuda de la tecnología computacional. Un sistema de cómputo, diseñado especialmente para el EXHCOBA-R, contiene todos los elementos y reglas necesarias para producir, de manera aleatoria y automática, un conjunto de ítems semejantes (conceptual y psicométricamente) que evalúan una competencia determinada. Por lo tanto, este generador también es capaz de originar una cantidad considerable de versiones de exámenes.

1.2. Planteamiento del problema

En los párrafos anteriores se infiere una nueva generación del examen denominado EXHCOBA-R. Por un lado, este examen presenta reactivos novedosos en cuanto al modo de contestar (respuesta semiconstruida o construida) y a la calificación (dicotómica y crédito parcial). Por otro lado, contiene un generador automático de reactivos que puede producir, para cada competencia, una gran cantidad de ítems equivalentes. Las combinaciones crecen aún más al

multiplicarse por los 120 reactivos que contiene la prueba de ingreso a la EMS, o los 180 para la ES. Estos arreglos de ítems innovadores generan miles de pruebas representativas de un único examen, el EXHCOBA-R.

Los criterios de calidad de los tests estandarizados exigen que estos acrediten diferentes pruebas de validación. Wright y Stone (1999) coincidieron en que no existe un criterio único para decidir los caminos a seguir que determinen la validez de las interpretaciones de un examen; lo importante es delimitar aquellos elementos que son necesarios de los que son optativos y los que son decisivos de los que son aconsejables. El Comité Técnico del EXHCOBA-R ha acordado que tres tipos de evidencias son las fundamentales para evaluar la calidad del instrumento: (a) evidencias de validez de contenido por tratarse de un instrumento basado en los planes de estudios vigentes de la educación básica y media superior de México, (b) evidencias de validez del proceso y de la estructura cognitiva por ser indispensable analizar qué tipo de habilidades o competencias efectúa el evaluado en el momento de resolver la prueba, y (c) evidencias de validez de estructura interna, con un previo análisis de las propiedades psicométricas de los ítems, que den sostén estadístico a la calidad de los reactivos y a la organización teórica del instrumento.

Con respecto a la validación de su estructura interna, el proceso presenta particularidades que lo hacen diferente. No se trata de una única prueba, sino de una gran cantidad de exámenes posibles producidos por la GAI, y debe decidirse cómo analizarlos. Además, no se utilizan los tradicionales reactivos de opción múltiple, sino ítems de respuesta breve o de múltiples selecciones, lo que agrega novedad y al mismo tiempo, complejidad al análisis. Es por eso que resulta necesario considerar diferentes enfoques estadísticos y proponer una metodología que

permita aportar evidencias de validez de estructura interna de los distintos exámenes y reactivos que se producen por la GAI.

1.3. Preguntas de investigación

Conforme a lo expuesto surge la siguiente pregunta de investigación:

- ¿Cómo obtener evidencias de validez, basadas en la estructura interna, de exámenes producidos a través de la Generación Automática de Ítems?

Esta pregunta puede descomponerse en otros cuestionamientos más puntuales:

- ¿Cómo analizar las propiedades psicométricas de los exámenes que se generan a través de la Generación Automática de Ítems?
- ¿Cómo identificar en qué grado la estructura conceptual de los exámenes generados a través de la GAI concuerda con su estructura empírica?
- ¿Cómo estudiar los diferentes ítems que miden una misma competencia académica para decidir si poseen propiedades psicométricas similares y si se agrupan en un mismo constructo?
- ¿Cómo examinar las propiedades psicométricas de los elementos que componen los reactivos de crédito parcial?

1.4. Objetivos

Para responder a estas las preguntas planteadas se proponen el objetivo general y los objetivos específicos:

Objetivo general. Proponer una metodología para aportar evidencias de validez de la estructura interna de los exámenes producidos por el Generador Automático de Ítems, que se utiliza con el EXHCOBA-R.

Objetivos específicos.

- Definir el tipo de muestras necesarias para recabar evidencias de validez.
- Determinar los análisis estadísticos apropiados para obtener las propiedades psicométricas de los exámenes y de las distintas familias de ítems producidos a través de la GAI.
- Determinar los análisis estadísticos para obtener evidencias de validez de estructura interna de los exámenes y de las distintas familias de ítems producidos a través de la GAI.
- Aplicar la metodología propuesta para aportar evidencias de validez de estructura interna del EXHCOBA-R/MS.
- Formular recomendaciones, con base en los resultados obtenidos, para mejorar la estructura del examen.
- Evaluar la eficacia de la metodología propuesta para la validación de exámenes producidos por GAI.

1.5. Justificación y delimitación de la investigación

Todo instrumento de medición que se utiliza a gran escala, y que ejerce alto impacto a quienes se les aplica, exige evidencias de validez que permitan garantizar interpretaciones correctas y el buen uso de los resultados. El EXHCOBA-R tiene un desarrollo de cuatro años de trabajo colegiado con la participación de profesionales expertos en las diferentes áreas de evaluación. Se ha construido un instrumento cuidadosamente elaborado y revisado. Sin embargo, se vuelve necesario garantizar los estándares de calidad de un instrumento de evaluación. El análisis de la

calidad de los ítems y de la estructura interna del examen permitió a los desarrolladores del EXHCOBA-R obtener información de gran utilidad para revisar y perfeccionar el examen.

Las evidencias de validez otorgan confianza y transparencia al proceso de admisión de las instituciones educativas usuarias. Asimismo, permiten garantizar la calidad del instrumento que mide los conocimientos que dice medir, de manera similar, en cada una de las versiones producidas por la GAI. Dichas evidencias de validez posicionan al EXHCOBA-R 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.

Este estudio cumple también con la publicación de los resultados del proceso de obtención de evidencias de estructura interna de un test. Publica y expresa bajo qué condiciones se obtuvieron los datos, de acuerdo con los *Standards*[1] *for Educational and Psychological Testing* (*American Educational Research Association, American Psychological Association and National Council on Measurement in Education*[2], 1999). De esta forma la calidad del EXHCOBA-R/MS se exhibe a la comunidad científica y al público en general para ser juzgada y valorada con las limitaciones propias de la metodología utilizada.

También existe un valor teórico sustancial en la presente investigación. En primer lugar, la documentación acerca del desarrollo del EXHCOBA-R, como una aplicación de la GAI, con la descripción de los nuevos tipos de reactivos y de sus modelos de ítems. En segundo lugar, una propuesta metodológica para aportar evidencias de exámenes desarrollados por la GAI de teoría débil con ítems de respuesta construida o semi-construida. Ambas contribuciones se consideran valiosas para el desarrollo de nuevas formas evaluativas amparadas en la tecnología

[1] En el texto se referirá a ellos como los *Standards*.
[2] AERA, APA y NCME se utilizará para identificar a: American Educational Research Association, American Psychological Association y National Council on Measurement in Education.

computacional y en desarrollos psicométricos sólidos como son la Teoría Clásica de los Tests y la Teoría de Respuesta al Ítem.

Finalmente, es importante aclarar que el trabajo consta de una metodología para validación de una GAI de las características del EXHCOBA-R. Las evidencias de validez se remiten a la parte del examen correspondiente a la educación básica (primaria y secundaria) y se realizaron a través de muestras (en algunos casos, de menor tamaño a lo requerido) obtenidas por pilotajes que no fueron administraciones reales de alto impacto. También, durante las aplicaciones el editor de reactivo presentó fallas que se reflejaron en el momento de recuperar la información y provocó datos perdidos.

1.6. Contenido de la tesis

Esta tesis se organiza en cinco capítulos y 6 anexos. En este primer capítulo se presentó el EXHCOBA-R, el problema de su validación, las preguntas que guían la investigación, el objetivo general y los objetivos específicos que concretan las actividades a seguir. Además se explicaron las razones que respaldan este trabajo y se delimitó la investigación.

En un segundo capítulo se definen los exámenes a gran escala y de alto impacto, así como ejemplos a nivel internacional y nacional. Esto con el objeto de percibir la novedad del EXHCOBA-R en este campo de la evaluación. Luego, se presenta la GAI, cómo surgió, desde qué tipo de teorías se puede producir y sus dos componentes fundamentales: los modelos de tareas y los modelos de ítems. Los modelos de tareas como herramientas indispensables para una teoría fuerte y los modelos de ítems como requisitos básicos para cualquier GAI. Después se realiza una descripción detallada del EXHCOBA-R/MS con sus nuevos tipos de reactivos: los REESCO. Se hace referencia a los estándares internacionales y nacionales para determinar la calidad de las pruebas educativas, se explica acerca de la evolución del concepto de validez y se

establece la definición de este último, que se utilizó para la presente tesis. Finalmente, se dedican tres apartados para presentar los modelos psicométricos que se utilizan, mayormente, para obtener evidencias de validez de estructura interna para tests tradicionales y para tests producidos por GAI. Tras reflexionar acerca de la factibilidad de estos modelos, se opta por los análisis psicométricos adecuados para el caso del EXHCOBA-R, desde la Teoría de Clásica de los Tests, la Teoría de Respuesta al ítem (modelo de Rasch y modelo de Masters) y el Análisis Factorial Confirmatorio (AFC).

En el tercer capítulo se desarrolla el método. Se definen dos niveles de análisis del EXHCOBA-R/MS, así como el tipo y la cantidad de muestras necesarias. Se precisa acerca de la forma de efectuar los análisis estadísticos y los índices a utilizar. También se explica bajo qué circunstancias se administraron las distintas versiones de los exámenes y qué características presentó la población participante.

En el cuarto capítulo se presentan los resultados obtenidos acerca del grado de validez de estructura interna del EXHCOBA-R/MS para cada nivel de análisis. Se detalla la información relacionada con el área de Habilidades matemáticas. Si bien se describen los estadísticos de las áreas restantes, estas gráficas y tablas aparecen en la sección de anexos.

El quinto y último capítulo contiene una síntesis de los resultados de los análisis efectuados al EXHCOBA R/MS. Aquí se reflexiona acerca del porqué del comportamiento estadístico en cada área y familia del examen. También se valora en qué medida el método utilizado respondió a las preguntas planteadas en la introducción de la tesis. Una vez expuesta la descripción de los alcances y las limitaciones del estudio se mencionan nuevas líneas de investigación que complementen y consoliden la metodología para obtener evidencias de validez

de estructura interna de exámenes obtenidos a través de la GAI. Para terminar, son incluidos 6 anexos que complementan el trabajo.

2

Marco de referencia

En este capítulo se presentan los fundamentos teóricos para el estudio de validación del EXHCOBA-R/MS. La definición de los exámenes a gran escala de alto impacto, así como los más conocidos a nivel internacional y nacional, aparecen en el primer apartado. En segundo lugar se introduce el concepto de Generación Automática de Ítems (GAI), el sustento teórico por el cual se desarrollan tests con GAI y ejemplos de exámenes de este tipo.

El EXHCOBA-R como un examen a gran escala, de alto impacto, producido por GAI se describe en el tercer apartado. En cuarto lugar se plantea la necesidad de mostrar evidencias de validez de los exámenes de alto impacto, la definición y evolución del concepto de validez, así como las evidencias que este estudio pretende aportar. Los apartados cinco y seis presentan una descripción de las teorías que se utilizan para obtener las propiedades psicométricas de los ítems y los análisis de validez de estructura interna del EXHCOBA-R/MS. El séptimo punto incluye algunas teorías en desarrollo que se emplean para la validación de exámenes producidos por GAI. Por último se explican las razones de la postura elegida con respecto a los modelos psicométricos utilizados para definir la metodología de validación del EXHCOBA-R/MS.

2.1. Exámenes educativos a gran escala de alto impacto

Los exámenes educativos a gran escala son aquellos donde participan una gran cantidad de evaluados. Por ejemplo: estudiantes de una institución, un estado o un país. Los resultados sirven para describir un sistema educativo, tomar decisiones acerca de los examinados o revisar políticas educativas, entre otros usos.

Los exámenes de alto impacto son utilizados para tomar decisiones que tendrán consecuencias importantes en los evaluados o en el sistema al cual pertenecen. Estas pruebas tienen una línea precisa de división, una calificación de corte, entre los que aprueban y los que desaprueban. Según el resultado obtenido se toman las decisiones. Son ejemplos de este tipo de exámenes las pruebas para obtener la licencia de conducir, los exámenes de ingreso o egreso de un nivel educativo (EMS, ES) y los tests aplicados a estudiantes que evalúan a los maestros o al sistema educativo (con el objeto de otorgar premios o recursos a distintas entidades).

Este tipo de tests ha suscitado críticas severas debido a las consecuencias que imprimen en la sociedad (e.g. Noddings, 2004; Nichols y Berliner, 2007). Como respuesta a estas acusaciones se han realizado numerosas investigaciones, informes y marcos de referencia para promover el buen uso de los resultados de estos exámenes (e.g. Carnoy, Elmore y Siskin, 2003; Heuber y Hauser, 1999; Messick, 1998; Sackett, Borneman y Connelly, 2008).

2.1.1. Exámenes de ingreso a la EMS y a la ES

Un tipo de pruebas educativas a gran escala y de alto impacto son los exámenes de ingreso a la EMS y a la ES. Generalmente, este tipo de tests miden los conocimientos adquiridos en niveles educativos anteriores o habilidades generales con el objeto de elegir a los mejores candidatos para estudiar en el bachillerato o en la universidad.

En este apartado se presentan los ejemplos más relevantes de exámenes de ingreso a la EMS o a la ES, a nivel internacional y nacional. Esto con el propósito de identificar las características básicas de este tipo de pruebas.

2.1.1.1. Experiencia internacional

Existe una gran cantidad de tests y no es propósito de esta tesis citar la lista completa de ellos. Es por eso que se eligieron los países con mayor desarrollo en psicometría a nivel mundial para identificar sus exámenes más relevantes (Hambleton, cit. en Rojas-Tejada, 2001). También se buscaron países con un desarrollo económico alto y una población numéricamente grande, que representaran a cada continente. Después se enlistaron aquellas pruebas más relevantes y aplicadas de manera oficial o extensiva a nivel nacional. Esta lista aparece en la tabla 2.1.

Tabla 2.1
Pruebas más relevantes de selección para ingreso a la EMS o a la ES por países

País	Examen	Ingreso	Áreas temáticas	Tipo de ítems	Administración	Estand[a]
EE.UU.	Independent School Entrance Examination (ISEE)	EMS	razonamiento verbal, razonamiento cuantitativo, comprensión lectora , desempeño matemático, ensayo	• OM • Respuesta abierta (ensayo)	Computadora Lápiz y papel	Sí
	Sitio web oficial (ISEE): http://erblearn.org/schools/admission/isee					
	High School Placement Test (HSPT)	EMS (evalúa cada año escolar)	Habilidades verbales, Habilidades cuantitativas, Lectura, Matemáticas, Lenguaje, Opcionales: ciencias, aptitudes mecánicas, religión católica	• OM	Lápiz y papel	Sí
	Sitio web oficial (HSPT): http://www.ststesting.com/hsptpg9.html					
	SAT (antes, Scholastic Aptitude Test)	ES	Lectura, Matemáticas, Escritura	• OM • RA (ensayo)	Computadora Lápiz y papel (ensayo)	Sí
	SAT subject test	ES	Batería de exámenes de 20 materias.	• OM	Computadora	Sí
	Sitio web oficial (SAT): http://sat.collegeboard.org/home					
	ACT (antes, American College Test)	ES	Inglés Matemáticas Lectura Ciencia Ensayo (opcional)	• OM • RA (ensayo)	Computadora	Sí
	Sitio web oficial (ACT): http://www.actstudent.org/					

continúa tabla

Tabla 2.1 (continuación)
Pruebas más relevantes de selección para ingreso a la EMS o a la ES por países

País	Examen	Ingreso	Áreas temáticas	Tipo de ítems	Administración	Estand[a]
España	Pruebas de Acceso a Enseñanzas Universitarias Oficiales de Grado (PAEG)	ES	Fase general (obligatoria): Castellano y literatura, Lengua extranjera, Historia (España o Filosofía), Lengua Cooficial Fase específica (voluntaria): máximo de 4 asignaturas.	• RA, de 4 a 6 ejercicios por asignatura	Lápiz y papel	No
Países Bajos	No hay examen general	EMS/ ES	Diploma y examen de egreso			
Australia	Special Tertiary Admissions Test (STAT)[b]	ES	Competencias verbales: Humanidades y ciencias sociales Competencias cuantitativas: Matemáticas y Ciencias	• OM • RA (para una versión de escritura en inglés)	Lápiz y papel	Sí
	Sitio web oficial (STAT): http://www.uac.edu.au/stat/					
Francia	Baccalauréat	ES	Tres versiones: Científico, Económico y social, y Literario	• mayoría de RA • OM con justificación	Lápiz y papel Oral	Semi
	Sitio web oficial (Baccalaureat): http://www.education.gouv.fr/cid143/le-baccalaureat.html					
Japón	Daigaku Nyūshi Sentā Shiken[c]	ES	Cívica, geografía e historia, literatura japonesa, lengua extranjera, ciencias, matemáticas	• OM	Lápiz y papel	Sí
	Sitio web oficial (National Center): http://www.dnc.ac.jp/					
China	NCEE (National Higher Education Entrance Examination)	ES	Obligatoria: chino, matemáticas y lengua extranjera. Dos áreas a seleccionar: Ciencias (física, química y biología) Humanidades (historia, geografía y política)	• S/D	Lápiz y papel	Semi
Sudáfrica	National Benchmark Test (algunas universidades)	ES	Alfabetización académica Alfabetización cuantitativa Matemáticas	• OM	Lápiz y papel	Sí
	Sitio web oficial (NBT): http://www.nbt.ac.za/					

Continúa tabla

Tabla 2.1 (continuación)
Pruebas más relevantes de selección para ingreso a la EMS o a la ES por países

País	Examen	Ingreso	Áreas temáticas	Tipo de ítems	Administración	Estand[a]
Brasil	Vestibular [d]	ES	Portugués (lengua y literatura), Matemáticas, Historia, geografía, Biología, Física, Química y lengua extranjera	• OM • Ensayo (si aprueba OM)		
	Examen Nacional de Enseñanza Media-ENEM	ES	Ciencias naturales, ciencias humanas, matemáticas, portugués y lengua extranjera	• OM • Ensayo		Sí
	Sitio web oficial (ENEM): http://www.brasil.gov.br/para/estudiar/acceso-a-la-universidad/enem/br_model1?set_language=es					
Chile	Prueba de selección universitaria (PSU)		Matemáticas, lengua, ciencia e historia	• OM	Lápiz y papel	Sí
	Sitio web oficial (PSU): http://www.demre.cl/psu.htm					

Nota: OM = opción múltiple. RA = respuesta abierta. S/D = Sin datos. [a] test estandarizado. [b] Es para estudiantes que no participan en el sistema de ingreso Australian Tertiary Admission Rank (ATAR). [c] Test de admisión a la universidad del centro nacional. [d] Vestibular es el nombre de los exámenes de ingreso a la universidad, en Brasil. Cada universidad tiene su vestibular, aquí se consideró el de la universidad de San Pablo (que utilizan también otras instituciones de ES).

Según la tabla 2.1, los exámenes de ingreso son mayormente estandarizados, de opción múltiple (con la salvedad de escritura de ensayos y el caso de Francia) y administrados a lápiz y papel, con la excepción de EE. UU. que implementa exámenes computarizados. Todas las pruebas incluyen matemáticas y el idioma oficial, como materias básicas. Posteriormente se agregan materias relacionadas con los estudios a seguir.

2.1.1.2. La experiencia nacional

Las pruebas estandarizadas más importantes aplicadas para ingreso a la EMS o a la ES en México son cinco. Estas se destacan por dos criterios: (a) son exámenes estandarizados y confiables, y (b) se aplican a poblaciones de gran tamaño (superan los 10,000 evaluados) y en más de una institución educativa del país. Las evaluaciones son el Examen Nacional de Ingreso a la Educación Media Superior (EXANI-I), el Examen Nacional de Ingreso a la Educación Superior (EXANI-II), la Prueba de Aptitud Académica (PAA), el PIENSE II y el EXHCOBA.

El EXANI-I y el EXANI-II son pruebas de conocimientos y habilidades básicos elaboradas y administradas por el CENEVAL. El EXANI-I está dirigido a egresados de la escuela secundaria que desean ingresar a las instituciones educativas de la EMS en México. Se administra en las instituciones públicas miembros de la Comisión Metropolitana de Instituciones Públicas de Educación Media Superior (COMIPEMS) de la Zona Metropolitana de la Ciudad de México (entre ellas, el Colegio de Bachilleres, la Escuela Nacional Preparatoria de la Universidad Nacional Autónoma de México (UNAM) y el Centro de Estudios Tecnológicos del Instituto Politécnico Nacional (IPN). El EXANI-II está dirigido a quienes desean ingresar a la ES y las instituciones usuarias son aproximadamente 100, las cuales representan a 30 entidades federativas del país.

La Prueba de Aptitud Académica (PAA) y el PIENSE II pertenecen al College Board, entidad privada de EE.UU. que elabora y administra exámenes desde inicios del siglo pasado. La primera institución mexicana en utilizar la PAA, como parte de sus requisitos de ingreso, fue el Tecnológico de Monterrey, quien continúa desde 1963. También aplican pruebas del College Board para el ingreso a sus bachilleratos o a la ES la Universidad Autónoma de Coahuila, la Benemérita Universidad Autónoma de Puebla, el Centro de Enseñanza Técnica y Superior, la Universidad Anáhuac y la Universidad de Guadalajara. La PAA está dirigida a estudiantes que

han finalizado la EMS y desean ingresar a la universidad. Sin embargo, algunos establecimientos educativos lo utilizan como parte de la admisión a la EMS, con un puntaje menor al exigido para el nivel superior. Por ejemplo, PIENSE II es una prueba orientada al ingreso a la EMS.

El EXHCOBA (Backhoff y Tirado, 1992) es el examen computarizado estandarizado más antiguo desarrollado en México. Este examen es aplicado para la selección de estudiantes a la ES y cuenta con una versión simplificada para ingreso a la EMS. Actualmente ofrece servicios a 15 instituciones educativas aproximadamente. Entre ellas se pueden mencionar la Escuela Preparatoria Federal Lázaro Cárdenas (PFLC), el Colegio Madrid de México, la Universidad Autónoma de Guanajuato (UGTO), la Universidad de Sonora (UNISON), la Universidad Autónoma de Querétaro (UAQ) y la Universidad Autónoma de Ciudad Juárez (UACJ).

Si se comparan los tipos de exámenes de ingreso que se utilizan a nivel mundial con los de México se observa que en México se refleja lo ocurrido a nivel internacional. En este país, los tests incluyen aspectos básicos de matemáticas, español y agregan conocimientos específicos. Son generalmente pruebas de opción múltiple, de lápiz y papel. Mientras que el EXANI admite dos tipos de administraciones, lápiz y papel, y por computadora, el EXHCOBA sólo cuenta con la aplicación computarizada (ver tabla 2.2).

Tabla 2.2
Pruebas más relevantes de selección para ingreso a la EMS o a la ES en México

Examen	Contenidos	Reactivos	Administración
EXANI-I	Selección: razonamiento verbal, razonamiento lógico-matemático, español, matemáticas Diagnóstico: Ciencias naturales (biología, física y química), Ciencias sociales (Historia, geografía, y formación cívica y ética), Inglés.	OM	Lápiz y papel computadora
EXANI-II	Selección: razonamiento verbal, razonamiento lógico-matemático, español, matemáticas, tecnologías de la información y comunicación. Diagnóstico: (módulos, según la especialidad, cinco áreas por módulo). Áreas: cs. administrativas, cs. agropecuarias, cs. de la salud, cs. naturales y exactas, cs. sociales, humanidades, ingenierías y tecnologías, psicología, pedagogía y bases de la educación, y docencia.	OM	Lápiz y papel computadora
Sitio web oficial (CENEVAL): http://www.ceneval.edu.mx/ceneval-web/content.do?page=0			
PAA	Razonamiento verbal, razonamiento matemático, redacción de forma indirecta	OM/ Escritura de números	Lápiz y papel
PIENSE II	Habilidad cognoscitiva, conocimiento de español, conocimiento de matemáticas, conocimiento de inglés	OM	Lápiz y papel
Sitio web oficial (College Board América Latina y El Caribe): http://www.collegeboard.com/ptorico/latinam/lamain.html			
EXHCOBA	Versión para ingreso a la EMS: Habilidades cuantitativas, Habilidades verbales, Español, Matemáticas, Ciencias naturales (biología, física, química) y Ciencias sociales (historia, geografía, formación cívica) Versión para ingreso a la ES: Agrega 3 áreas, según la especialidad: Matemáticas, Estadísticas, Ciencias sociales, Humanidades, Económico administrativas, Lenguaje, Biología, Física, Química	OM	Computarizada
Sitio web oficial (EXHCOBA): http://www.exhcoba.mx/			

Nota: OM = opción múltiple.

Las pruebas de ingreso a instituciones de la EMS y de la ES expuestas en estos apartados son producto de desarrollos únicos y puntuales para cada administración, es decir, ninguna de ellas, menos aún las de lápiz y papel, se producen por métodos de generación de ítems de manera automática, con recursos computacionales. En el próximo apartado se describe el desarrollo de la

GAI, se citan algunas de sus aplicaciones y se presenta al EXHCOBA-R como un examen de ingreso elaborado por medio de este proceso.

2.2. La Generación Automática de Ítems

Según Gierl y Haladyna (2012), los mayores estados de cambio en el campo de la evaluación se reflejan en tres áreas: los tests computarizados, el diseño de exámenes y los tests de diagnóstico cognitivo. Los tests computarizados se originaron con la Batería de aptitud vocacional para las fuerzas armadas en los años 1960's (Sands, Waters y McBride, 1997). Después, el uso de la computadora se extendió hacia los exámenes adaptativos (Luecht, 1998; Wainer y Kiely, 1987). Test muy conocidos que se administraban con lápiz y papel han cambiado al formato computarizado en la actualidad (e.g. el Graduate Record Examination-GRE-, el Graduate Management Achievement Test-GMAT- o el Test of English as a Foreign Language-TOEFL-). Gierl y Lai (2011) reportaron que durante 2009, fueron administrados exámenes educativos por computadora en 27 estados de EE. UU.

El advenimiento de la evaluación computarizada trajo nuevos desafíos, sobre todo en el área de diseño de tests, porque los ítems se aplican con más frecuencia, están más expuestos y por lo tanto, necesitan un banco muy amplio y costoso de reactivos. Así, han surgido diferentes estrategias como la Ingeniería de los Test (IT) promovida por Gierl y Haladyna (2012), que pertenece a Luecht (2012). Este nuevo enfoque destinado a las prácticas de medición con principios básicos de ingeniería y procesos tecnológicos se utiliza tanto para el desarrollo de los tests como para el análisis, calificación y reporte de resultados.

La IT se fundamenta en teorías cognitivas para generar familias de ítems psicométricamente parecidos y que miden un mismo contenido. Estos modelos incluyen ítems similares que se desarrollan y analizan en conjunto. Desde el punto de vista tradicional, el

desarrollo de ítems incluye un proceso donde cada ítem se trata de manera individual y aislada. Por lo tanto, la unidad de análisis es el ítem individual de cada test. Esta comparación entre la IT y el diseño tradicional permite resaltar las fortalezas del nuevo enfoque de la generación de ítems.

Sin embargo, dentro de la IT existen preguntas que no han sido respondidas totalmente: cómo asegurar que para cada constructo las demandas cognitivas estén correctamente definidas, operacionalizadas por los modelos y medidas por los ítems generados; y cómo asegurar que el contenido y las propiedades psicométricas de los ítems son similares dentro de una misma clase de modelo de tareas (Gierl y Haladyna, 2012).

2.2.1. Perspectiva histórica

Gierl y Lai (2012) definieron a la Generación Automática de Ítems (GAI) como el proceso para generar *ítems estadísticamente equilibrados* a través de *modelos de ítems* y con la ayuda de la *tecnología computacional*. Es por eso que en este procedimiento se utilizan desarrolladores de modelos de ítems, tecnología computacional aplicada y métodos estadísticos para confirmar la calidad de los ítems. De acuerdo con esta definición y la de IT se puede inferir que la GAI está incluida en la IT, ya que la GAI es un método para producir reactivos, que es parte del proceso de la IT.

Drasgow, Luecht y Bennett (2006) marcaron tres eventos importantes en el desarrollo de la GAI. Según estos autores, la GAI tuvo sus orígenes en el movimiento de los exámenes basados en criterios de la década de los sesentas. El hecho fundamental fue la introducción de la *forma de ítem* de Hively (Hively, Patterson y Page, 1968). Los ítems se generaban por formatos que contenían elementos fijos y variables. Esto dio como resultado reactivos similares en dificultad, pero no necesariamente homogéneos, es decir, involucraran las mismas habilidades

cognitivas. La GAI progresó con el surgimiento de métodos cognitivos para la instrucción y el diagnóstico. Sin embargo, estos métodos se concentraron en la enseñanza y no en los tests, por lo que no se exploraron las implicaciones psicométricas. En un tercer paso se dio la integración de la psicometría y las perspectivas cognitivas desde dos metodologías, la teoría débil (Bejar, 1993) y la teoría fuerte (Embretson, 1999).

También Haladyna (2012) presentó una reseña histórica sobre la evolución de la GAI. En ella destacó que lo preponderante había sido su diversidad, así como la falta de cohesión entre sus contribuyentes. Este autor definió cinco hechos clave del desarrollo de la GAI; estos son: (1) GAI basada en la prosa (1970); (2) la teoría de faceta (1953, 1959); (3) las formas de ítems (1974); (4) la formación de conceptos (1970); y (5) el seminario por invitación de la *Educational Testing Service* (ETS) de 1998 que dio origen al libro *Item Generation for Test Development* de Irvine y Kyllonen (2002).

1. La GAI basada en la prosa pertenece a Bormuth (1970) y surgió con la necesidad de mejorar la redacción de los ítems que en ese momento era subjetiva e ineficiente. El investigador introdujo el supuesto con respecto a que la redacción de los reactivos no debería ser subjetiva, así como que dos elaboradores de ítems, con el mismo contenido y las mismas especificaciones, deberían producir ítems similares y de calidad. Esta teoría fue refutada, ya que se pudieron presentar ítems muy diferentes bajo las condiciones que Bormuth establecía (Roid y Haladyna, 1978). Sin embargo, Haladyna (2012) aseveró que la GAI necesitaba una teoría para la redacción de los ítems, la cual sería un gran avance para esta ciencia.

2. El diseño de faceta fue implementado por Guttman en 1959 con el objetivo de operacionalizar un constructo y validarlo a través de modelos estadísticos. El recurso utilizado es un mapeo de oraciones que se desarrollan y van generando los diferentes ítems. Este diseño

también tiene sus limitaciones porque implica diferentes mapeos según cada desarrollador de ítems. No obstante, Haladyna (2012) señala que este método tampoco registra resultados exitosos.

3. Las formas de ítems fueron propuestas por Osburn (1968). Este investigador ideó un formato que generara ítems con una estructura sintáctica fija; por ejemplo: una oración con espacios en blanco que se reemplazan por números o conceptos para formar un ítem. Estas formas deben crearse con cuidado para no producir reactivos fuera de la realidad, es decir, deben incluirse restricciones. Estos formatos son útiles para cuestiones numéricas. Su desventaja es que deben crearse muchos formatos diferentes para cubrir todo el dominio de conocimiento a evaluar. Al parecer, las formas de ítems se abandonaron hacia finales del siglo pasado; sin embargo, la GAI las ha recuperado para su desarrollo.

4. Haladyna (2012) aseguró que el aprendizaje de conceptos y su evaluación es una de las razones más sencillas para generar un examen. La explicación de la evaluación de los aprendizajes proviene actualmente de la psicología cognitiva; sin embargo, el aprendizaje de conceptos podría provenir de la investigación de los psicólogos de 1960’s y 1970’s.

5. El libro *Item Generation for Test Development* de 2002 es otro referente de los avances en la GAI. Este material incluye los productos del seminario de la ETS. En esa conferencia se hizo referencia a la generación de exámenes con énfasis en el concepto clave *isomorfo*. Allí también se definió el concepto *modelo de tareas*. En este documento se introdujeron algunos principios, tales como que el análisis de constructo es esencial para obtener una GAI efectiva y que un método obligatorio para el desarrollo de ítems es el uso de los modelos de reactivos (también conocidos como *item shells*).

Un paso fundamental en el desarrollo de pruebas es definir el constructo a medir. Esta medición se ejerce por medio de ítems que invocan habilidades, las cuales se categorizan en demandas cognitivas. La GAI ha buscado conjugar estas relaciones; si logra mejorar la validez de las pruebas o si, al menos, garantiza un desarrollo más eficiente de los ítems y mantiene el grado de validez se le augura un futuro exitoso (Haladyna, 2012).

2.2.2. Teoría débil y teoría fuerte que sustentan la GAI

Un tema importante es con respecto a qué teoría sustenta la creación de los modelos de ítems. Según Gierl y Lai (2012) no existían, hasta el momento de elaborar su capítulo, estudios publicados que describieran los principios o las prácticas necesarias para desarrollar modelos de ítems. Su aporte consistió en describir e ilustrar cómo estos modelos se pueden crear a través de dos aproximaciones teóricas para la GAI, una *débil* y otra *fuerte*. Para ello, los autores retomaron las recomendaciones de Drasgow, Luecht y Bennett (2006).

Desde la teoría débil se utilizan guías de diseño con el objetivo de crear modelos de ítems para generar reactivos isomorfos. Esta teoría recomienda considerar un ítem-padre que muestre la estructura subyacente en el ítem, del cual se conozcan sus propiedades psicométricas; propone un punto de referencia para nuevos reactivos denominados ítems-hijo. En este caso, los especialistas en contenido deben manipular los elementos que constituyen diferentes hijos y decidir sobre la dificultad de los ítems con los recursos disponibles (su experiencia profesional y sus expectativas).

Esta propuesta tiene sus limitaciones. Una de ellas es que los elaboradores de los modelos deben predecir las propiedades estadísticas de los reactivos, lo cual no siempre ocurre de manera precisa. Otro inconveniente es que para asegurar ítems equivalentes, en muchas ocasiones, la

variedad de propuestas es limitada y se obtienen ítems demasiado "iguales" (llamados, peyorativamente, *clones*).

La teoría fuerte se basa en modelos cognitivos. Estos prototipos especifican y manipulan, a través de un registro teórico, los elementos que afectan el nivel de dificultad de los ítems generados en los exámenes de desempeño. La teoría cognitiva ayuda no solamente a revelar el conocimiento de los evaluados y las habilidades requeridas para resolver un ítem, sino también las características de contenido de los ítems que afectan a su dificultad. Según Gierl y Lai (2012), si se modelan las interacciones entre el examinado y el contenido es posible predecir y controlar las propiedades psicométricas de los ítems.

Estos autores señalaron que la teoría fuerte posee poco desarrollo debido a que se ha focalizado en aquellos procesos psicológicos donde ya existían modelos cognitivos de tareas de ejecución. Lamentablemente es escaso el desarrollo de teorías cognitivas comparativas que guían el desarrollo de reactivos para la cantidad de contenidos que se necesitan en los exámenes educativos.

De acuerdo con ambas teorías, la teoría débil se ajusta más para aquellos exámenes que abarcan muchos contenidos, aunque con la desventaja que genera el compromiso de asegurar dificultades similares para los ítems-hijo y a su vez, una cantidad de opciones que permitan generar mayor diversidad de reactivos que no se conviertan en *clones*. Además, como no existe información disponible para guiar el desarrollo de modelo de ítems con teoría débil, los especialistas deben apoyarse en sus propios juicios y en guías un tanto ambiguas. El desafío que plantearon Gierl y Lai (2012) como solución al problema de los modelos de teoría débil sería convertirlos en modelos de estructuras cognitivas. De este modo no serían necesarios pilotajes extensos, ya que las propiedades cognitivas y psicométricas estarían modeladas y controladas.

Los autores declararon que se necesita aún mucha investigación acerca de cómo diseñar, desarrollar y evaluar los métodos para crear modelos de ítems, tanto desde la teoría débil como de la fuerte. También se debe estudiar cómo evaluar las propiedades de estos modelos desde su capacidad generativa. Esta rama de la investigación sería clave para la aplicabilidad de la GAI.

2.2.3. Los modelos de tareas y los modelos de ítems

Un elemento clave para la GAI son los *modelos de tareas* (Luecht, 2012). Estos representan una alternativa para reemplazar las tradicionales especificaciones de los tests. De acuerdo con el autor, estos modelos son especificaciones cognitivamente orientadas para una clase o familia de ítems. Estas especificaciones integran los componentes del conocimiento declarativo, las relaciones entre dichos componentes, las habilidades cognitivas, así como el contenido relevante, el contexto y las características auxiliares que afectan la complejidad cognitiva de una tarea.

Cada modelo define una combinación única de habilidades cognitivas aplicadas a un contenido declarativo en una región específica de una escala de medida basada en un constructo. Los modelos de tareas constituyen un diseño detallado que describe la complejidad cognitiva y que ayuda a mantener la dificultad estadística y otras propiedades psicométricas (e.g. discriminación del ítem, confiabilidad) de una familia de ítems asociados a ellos.

El modelo se construye a través de una *gramática de modelo de tarea (GMT)*. Esta gramática ofrece una descripción formal de: (a) las combinaciones del conocimiento declarativo y las habilidades necesarias para resolver una tarea, (b) la información acerca de la complejidad de los componentes de la tarea, (c) la información auxiliar o las herramientas que facilitan o complican la tarea y (d) otros atributos relevantes asociados a los componentes que también pueden afectar a la dificultad del ítem. Las habilidades se especifican mediante verbos. Ejemplos de un formato de GMT son: $f(x_1)$ (*f* indica una habilidad aplicada a un objeto de conocimiento

x_1), $f(x_1, x_2)$ (una habilidad *f* donde se manipulan dos objetos de conocimiento, x_1 y x_2), $f[g(x_1, x_2), x_3]$ (una habilidad *f* aplicada al resultado de una relación entre x_1 y x_2, con x_3). Así se pueden conseguir modelos de tareas desde sencillos a complejos.

Cada modelo de tarea constituye la base para crear múltiples modelos de ítems[3] que, a su vez, genera múltiples ítems. Solano-Flores, Shavelson y Schneider, (2001, p.2) utilizan esta definición: "un *template* es un conjunto de instrucciones para desarrollar ejercicios". Haladyna y Shindoll (1989) lo llamaron *item shell* y lo describieron como una estructura externa vacía "llena de huecos" que, al completarse, permite generar conjuntos de reactivos semejantes. Sus ventajas son que posibilitan el desarrollo de pruebas de respuesta construida, formalizan las propiedades estructurales de los ítems, estandarizan los formatos de respuesta y calificación, y regulan el proceso de desarrollo de los exámenes (Solano-Flores *et al*., 2001).

En el contexto evaluativo el concepto como *forma del ítem* surgió probablemente con Osburn (1968) y fue desarrollado por Hively, Patterson y Page (1968), quienes lo aplicaron de manera específica para reactivos de aritmética. El concepto de *item shell* surgió para formalizar los procedimientos de generación de ítems de opción múltiple; ejemplo de ello es el trabajo de Haladyna y Shindoll (1989). Estos moldes también se han utilizado para generar ítems de respuesta construida en exámenes de ciencias naturales (Solano-Flores, Jovanovic, Shavelson y Bachman, 1999, Solano-Flores y Shavelson, 1997). A partir del siglo XXI se ha extendido el uso de los modelos de ítems que funcionan como estructuras para generar, con la ayuda de la computadora, gran cantidad de ítems referentes a un mismo constructo, y similares en demandas cognitivas y en dificultad (e.g. Bejar, 2002; Embretson, 2002; Gierl, Zhou y Alves, 2008).

[3] *Modelo de ítem, forma de ítem, plantilla, molde o* sus nombres en inglés: *template* e *ítem shell* son términos equivalentes, a efectos de esta tesis. No debe confundirse con *modelo de tarea* que es el modelo cognitivo que subyace al test.

Según Liecht (2012), una plantilla consta de tres componentes: un modelo de ejecución según el tipo de reactivo, un modelo de datos y un calificador de ítems. De este modo quedan determinados el formato de presentación, el contenido que se manipulará y las reglas de calificación para cada plantilla. El modelo de ejecución es reutilizable a través de los diferentes ítems, ya que sólo se cambia el contenido para generar los múltiples reactivos.

Gierl y Lai (2011) también definieron el concepto de modelo de ítem, quienes consideran que debe incluir una base o raíz de reactivo, las opciones e información auxiliar. La base del reactivo contiene el contexto, el contenido y la pregunta que el examinado debe responder. Las opciones deben incluir la respuesta correcta y uno o más distractores. En el caso de ítems de respuesta construida no se requieren distractores. La información auxiliar incluye cualquier material adicional necesario en la generación de los ítems (textos, imágenes, tablas, diagramas, sonido o video). Tanto la base del reactivo como las opciones de respuesta pueden subdividirse en *elementos* (frases, palabras, letras, símbolos y números).

Si los reactivos generados con el modelo de ítem pretenden medir un contenido con niveles de dificultad similares son llamados *isomorfos*. En ese caso, los desarrolladores de ítems manipulan aquellos *elementos incidentales*, que son características superficiales del reactivo y que no alteran su dificultad, para producir los ítems isomorfos. En necesario aclarar que la mayoría de los métodos de la GAI generan ítems de opción múltiple (Mortimer, Stroulia, Vosoughpour y Yazdchi, 2012). La figura 2.1 presenta un modelo de ítems para el caso de respuesta construida.

Base del reactivo:
María ha pagado \$**A** por la colocación de cerámicos en el piso de una habitación de su casa. El costo de instalación es de \$**B**/m^2. Si el piso donde se instalaron los cerámicos es de forma cuadrada, ¿cuánto mide el lado de dicha habitación?
Elementos:
A $\rightarrow$ rango de valores: 1525-1675 (de 25 en 25). **B** $\rightarrow$ rango de valores: 30-45 (de 5 en 5).
Respuesta correcta:
$\sqrt{\frac{A}{B}}$
Ejemplo de reactivo generado con el modelo:
María ha pagado \$1600 por la colocación de cerámicos en el piso de una habitación de su casa. El costo de instalación es de \$30/m^2. Si el piso donde se instalaron los cerámicos es de forma cuadrada, ¿cuánto mide el lado de dicha habitación?

Figura 2.1. Modelo de ítem con dos elementos enteros.

2.2.4. Aplicaciones de la GAI

La mayor cantidad de aplicaciones de la GAI se encuentra en los exámenes adaptativos informatizados (Rojas-Tejada, 2001), cuya principal característica es que los ítems administrados se van adaptando al nivel de competencia del examinado en las respuestas de los reactivos, previos al mismo test. Para ello es necesario un banco de ítems muy extenso con distintos niveles de dificultad. Estas pruebas, en muchos casos, pertenecen a la GAI de teoría fuerte porque el sustento para la producción de gran cantidad de ítems similares está dado por las teorías cognitivas (análisis de procesos, estrategias y estructuras del conocimiento). En la tabla 2.3 se presentan ejemplos de tests que utilizan GAI, en diferentes países; uno de ellos es el GRE, un examen a gran escala de alto impacto.

Tabla 2.3
Tests que se generan mediante la GAI, teoría que los sustenta, país donde se desarrollaron y tipo de reactivos que utilizan

País	Adaptativo	Test	Tipos de ítems	Teoría de GAI
EE. UU.	Sí	GRE (ETS)	OM y desarrollo	Débil (Drasgow, Luecht y Bennet, 2006)
España	Sí	Test Adaptativo Informatizado de Análisis Lógico	OM	Fuerte (Revuelta y Ponsoda, 1998)
Países Bajos	Sí	Test de inteligencia no verbal: SON-R 5.5-17	OM	Fuerte (Geerlings, Glass y Van der Linden, 2011)
Austria	Ambas posibilidades	intelligence structure battery: INSBAT	OM - Ítems de rta. abierta numérica	Fuerte (Arendasy, Sommer, Gittler, Hergovich, 2006)
Austria	No	Bilingual word fluency test (inglés y alemán)	Ordenamiento de letras en una palabra	Fuerte (Arendasy, Sommer y Mayr, 2012)

Nota: OM = opción múltiple.

2.3. Descripción del EXHCOBA-R

Durante los 22 años de aplicaciones del EXHCOBA se han desarrollado nuevas versiones del instrumento, siempre bajo su estructura original; también se han efectuado ajustes a los reactivos a fin de calibrar el examen. Aunque los conocimientos y las habilidades evaluados son básicos y se mantienen estables independientemente de los cambios de planes de estudio, se consideró necesario poner en marcha una reestructuración del examen, a la luz de las nuevas corrientes del conocimiento y de los cambios curriculares desarrollados por la Secretaría de Educación Pública (SEP) a partir de 2006 en el caso de la educación secundaria, de 2009 y 2011 en el caso de la educación primaria y de 2011 con el Plan General del Bachillerato.

El formato de opción múltiple se vio limitado para las exigencias surgidas con esta nueva concepción del examen. La propuesta fue desarrollar un nuevo tipo de ítems que se aproximen a una situación real y cotidiana del salón de clases. Esta necesidad de cambio se vio favorecida por

la evolución que experimentó la tecnología informática, lo cual abrió posibilidades para el desarrollo de un examen con una interfaz superior a la actual.

2.3.1. Concepción del aprendizaje, subyacente al EXHCOBA-R

El constructivismo es la orientación dominante en psicología de la educación. Este movimiento establece que el conocimiento y el aprendizaje no son resultado de la experiencia directa, sino el fruto de la actividad mental constructiva y donde las personas interpretan esa experiencia (Carretero, 2004). Esta es una posición compartida por distintas tendencias de la investigación psicológica y educativa. Entre ellas pueden identificarse claramente tres: el constructivismo cognitivo con raíces en la psicología y epistemología genéticas, el constructivismo de orientación sociocultural inspirado en las ideas y planteamientos de Vygotsky y el constructivismo asociado al construccionismo social con la irrupción de enfoques postmodernos que sitúan el conocimiento en el uso del lenguaje y en las prácticas lingüísticas (Coll, 2001).

El EXHCOBA-R está basado, particularmente, en algunos aspectos del constructivismo cognitivo, el cual sustenta que el conocimiento tiene lugar en la mente de las personas. El EXHCOBA-R está diseñado para evaluar conceptos y habilidades básicos que existen en los evaluados; estas informaciones son requisito indispensable para que los estudiantes puedan relacionarlas después con las experiencias de aprendizaje que tendrán lugar en el nivel próximo de estudios.

Ausubel *et al.* (1983) formularon dos dimensiones ortogonales en los procesos de los aprendizajes escolares. La primera va desde el aprendizaje *por recepción* hacia el aprendizaje *por descubrimiento*, mientras que la segunda va desde el aprendizaje *por repetición* y hacia el *significativo*. En el aprendizaje *por recepción* el contenido total a aprender se presenta al alumno en su forma final y el estudiante debe internalizarlo o incorporarlo. En el aprendizaje *por*

descubrimiento el alumno debe reorganizar la información, integrarla a su estructura cognoscitiva existente o transformarla para obtener el producto final deseado. En la segunda dimensión, el aprendizaje *significativo* implica dos condiciones: (a) el objeto de aprendizaje se relaciona, de modo sustancial y no arbitrario, con lo que el alumno ya sabe y (b) el estudiante adopta una actitud a favor del aprendizaje significativo. Si faltara alguna de estas condiciones se daría un aprendizaje *por repetición*, es decir, el estudiante incorpora los conceptos o procedimientos, simplemente, de modo arbitrario.

Según Novak (1982), la mayor eficacia del aprendizaje significativo reside en tres grandes ventajas: producir una retención más duradera de la información, facilitar nuevos aprendizajes relacionados y producir cambios profundos que perduran más allá del olvido de detalles concretos. El aprendizaje repetitivo será superior sólo en el caso de que la evaluación del aprendizaje solicite un recuerdo literal del original. Pozo (1996) señaló que los dos tipos de aprendizaje no son totalmente dicotómicos, sino que conforman un continuo y deben coexistir. Sin embargo, Ausubel (2002) priorizó el aprendizaje significativo, por lo que la repetición o memorización quedó relegada.

De acuerdo con lo expuesto en este apartado, el EXHCOBA-R pretende recuperar los aprendizajes significativos que tuvieron lugar tanto en la educación básica como en el bachillerato, ocurridos por recepción o por descubrimiento. El examen evalúa aquellos conceptos y procedimientos que el estudiante logró relacionar sustancialmente, los cuales permanecieron de manera organizada y no arbitraria.

Para Pellegrino, Chudowsky y Glaser (2001), el modelo de aprendizaje subyacente tiene que servir como elemento unificador que brinde cohesión entre el currículo, la instrucción y la evaluación. Esta relación es crucial porque un examen de ingreso a instituciones educativas no es

un instrumento aislado, sino que debe estar alineado al currículo y a la instrucción con el propósito de apoyar nuevos aprendizajes. Por lo tanto, el EXHCOBA está organizado según las estructuras curriculares de la educación básica y de la educación media superior vigentes. Estos currículos remarcan que los alumnos son capaces de adquirir el conocimiento bajo procesos constructivos y que los maestros junto con los estudiantes deben construir los aprendizajes en colectivo (Plan de estudios de educación básica, SEP, 2011).

2.3.2. Su modelo y estructura

Para diseñar instrumentos de selección de calidad se requiere el trabajo colegiado de expertos, fundamentado en un método sólido y de acuerdo con los propósitos que se persiguen. Por ello, el modelo de desarrollo de exámenes propuesto por el Instituto Nacional para la Evaluación de la Educación (INEE, 2005), con algunas modificaciones por el Comité Técnico del EXHCOBA para adaptarlo a las necesidades del examen, se eligió para la estructuración del EXHCOBA-R.

En el desarrollo de la evaluación intervienen diversos especialistas, los cuales se agrupan en distintos órganos colegiados: Comité Técnico, Consejo Consultivo, Comités Académicos, Comités Elaboradores de Especificaciones y Reactivos (CEER), y Comités de Validación y Sesgo. Cada uno de estos órganos cumple una función específica y complementaria en el proceso de construcción, su trabajo se programa en forma escalonada y sus productos se convierten en insumos de las siguientes etapas, por lo que en el proceso de generación de este tipo de pruebas se considera, en parte, el de su validación (Contreras, 2000; Contreras, Backhoff y Larrazolo, 2003).

La tabla 2.4 presenta las siete fases y 13 etapas del desarrollo del EXHCOBA-R. En cada etapa se utilizan diversos procedimientos, entre los cuales destacan: (1) la documentación de procesos de construcción de pruebas normativas de gran escala, realizados por instituciones de

reconocida calidad internacional; (2) la capacitación dirigida a los comités de especialistas y docentes que participan en el proceso; (3) la elaboración de materiales para el trabajo de los comités; y (4) el trabajo colegiado, donde se toman las decisiones de mayor importancia.

Tabla 2.4.
Proceso de diseño, construcción, aplicación y validación del EXHCOBA-R

Fases	Etapas	Participantes	Procedimientos	Productos
1. Planeación general	1. Diseño del plan general de evaluación	• Consejo Consultivo • Comité Técnico • Personal técnico	1. Documentación 2. Seminarios 3. Trabajo colegiado 4. Reuniones periódicas de trabajo 5. Pruebas del funcionamiento del sistema	a. Manual Técnico para el Diseño del EXHCOBA-R
	2. Diseño y desarrollo del sistema informático	• Especialistas en bases de datos • Especialistas en diseño gráfico • Especialistas en sistemas de información • Personal técnico		b. Marco de Referencia de los Cuestionarios de Contexto c. Cuestionarios de contexto del alumno (versión lápiz y papel)
	3. Diseño y elaboración de cuestionarios de contexto	• Especialistas en diseño de cuestionarios de contexto • Especialistas en evaluación del aprendizaje • Personal técnico		d. Plataforma informática y sistema computarizado del examen e. Plataforma informática y sistema computarizado de los cuestionario de contexto f. Sistema para calificar exámenes y generar reportes estadísticos g. Sistema de información general de usuarios del EXHCOBA-R
2. Estructuración del EXHCOBA-R	4. Diseño del EXHCOBA-R	• Comité Técnico • Comités Académicos • Personal técnico	1. Análisis curricular 2. Preparación de materiales (retículas y formatos) 3. Capacitación 4. Trabajo colegiado 5. Documentación del proceso	h. Manual Técnico para el Diseño del EXHCOBA-R i. Tabla de Contenidos de cada área del examen con justificaciones
3. Construcción del EXHCOBA-R	5. Especificaciones con plantillas de reactivos	• Comité Técnico • Comités Elaboradores de Especificaciones y Reactivos • Personal técnico	1. Preparación de materiales (formatos y Tablas de contenidos) 2. Capacitación 3. Trabajo colegiado 4. Revisión y corrección de las especificaciones 5. Documentación del proceso	j. Manual técnico para la Elaboración de Especificaciones con plantillas y generadores de reactivos k. Especificaciones de reactivos de cada área del examen l. Versiones de especificaciones con sus plantillas y generadores de reactivos

Continúa tabla

Tabla 2.4 (continuación)

Proceso de diseño, construcción, aplicación y validación del EXHCOBA-R (elaborada por el Comité Técnico del EXHCOBA)

Fases	Etapas	Participantes	Procedimientos	Productos
4. Construcción de la interfaz	6. Programación del editor de reactivos	• Comité Técnico • Especialistas en bases de datos • Personal técnico	1. Especificación de requerimientos del editor 2. Programación del editor para diferentes tipos de reactivos 3. Pruebas y ajustes al editor 4. Documentación del editor	m. Editor de reactivos n. Manual Técnico del Editor
	7. Diseño gráfico de la interfaz	• Comité Técnico • Especialistas en diseño gráfico • Personal técnico	1. Especificación de requerimientos de la interfaz 2. Diseño y elaboración de la interfaz 3. Documentación de la interfaz	o. Interfaz gráfica del examen. p. Manual Técnico de la Interfaz.
	8. Montaje en red del examen	• Comité Técnico • Especialistas en sistemas de información • Personal técnico	1. Alimentación de la información para la generación de reactivos 2. Programación y montaje de los reactivos en el sistema computarizado 3. Pruebas de funcionamiento del examen	q. Sistema computarizado del EXHCOBA-R r. Manual Técnico del Sistema
5. Administración del EXHCOBA-R	9. Piloteo	• Comité Técnico • Especialistas en sistemas de información • Estudiantes • Personal técnico	1. Revisión de contenidos del examen en línea 2. Administración del examen a muestra de estudiantes 3. Generación de bases de datos	s. Manual Técnico para la revisión de contenidos del examen en línea t. Manual Técnico para la Administración del examen y de Cuestionarios de Contexto u. Base de datos con resultados del piloteo v. Resultados de los Cuestionarios de Contexto
	10. Administración real del examen	• Investigadores asociados especialistas en tecnología y procesos de evaluación • Personal de soporte técnico. • Personal de instituciones usuarias.	1. Capacitación de personal técnico. 2. Instalación del examen 3. Administración del examen a estudiantes 4. Bases de datos	w. Manual de capacitación para personal técnico x. Base de datos de resultados y. Informes de resultados a instituciones usuarias
6. Análisis e interpretación de resultados del EXHCOBA-R	11. Análisis psicométricos para Comité Técnico	• Comité Técnico • Asesores en medición	1. Análisis estadísticos de resultados 2. Reuniones informativas	z. Informe Técnico sobre el Comportamiento Psicométrico del examen
	12. Análisis de resultados para el Consejo Consultivo	• Comité Técnico • Consejo consultivo		aa. Informe Técnico sobre los resultados de la evaluación asociados con las variables de contexto
7. Recopilación de evidencias de validez del EXHCOBA-R	13. Estudios de validación	• Comité Técnico • Asesores en validación • Personal técnico	1. Documentación 2. Investigación	bb. Informes Técnicos de Estudios de Validación

Nota: Tabla elaborada por el Comité Técnico del EXHCOBA. La tabla aún necesita revisión por parte de dicho comité.

El EXHCOBA-R conserva a grandes rasgos la estructura del EXHCOBA original. El nuevo instrumento evalúa 120 contenidos para el caso del ingreso a la EMS: 40 pertenecientes a la educación primaria y los 80 restantes a la educación secundaria. Para el ingreso a la ES, se agregan 60 contenidos relacionados con la especialidad a estudiar. Cada contenido abarca un campo del conocimiento que es explorado por un reactivo. De este modo, el examen está organizado en tres secciones: la primera sección está compuesta de las habilidades adquiridas durante la educación primaria relacionadas con el uso del lenguaje y las matemáticas; la segunda sección corresponde al nivel de secundaria dividida en cuatro áreas (español, matemáticas, ciencias naturales y ciencias sociales), donde a su vez el área de ciencias naturales se subdivide en biología, física y química, y ciencias sociales se reparte en historia de México e historia universal, geografía, y formación cívica y ética; la tercera sección incluye los conocimientos adquiridos en la EMS y está dividida en nueve áreas de especialidad. En la figura 2.2 se muestra la nueva estructura con la correspondiente distribución de reactivos, acordada por el Comité Técnico del EXHCOBA.

Nivel		Areas								
Primaria	Habilidades básicas	del lenguaje 20 ítems				matemáticas 20 ítems				
Secundaria	Conocimientos básicos	Español 20 ítems		Matemáticas 20 ítems		Ciencias naturales 20 ítems		Ciencias sociales 20 ítems		
Bachillerato	Conocimientos de la especialidad	Lenguaje 20 ítems	Ciencias sociales 20 ítems	Económ. adminis trativa 20 ítems	Comuni-cación 20 ítems	Humani-dades 20 ítems	Matemá-ticas 20 ítems	Física 20 ítems	Química 20 ítems	Biología 20 ítems

Figura 2.2. Estructura conceptual del EXHCOBA-R.

Aunque la estructura del examen permanece similar, el EXHCOBA-R ya no es una prueba con varias versiones, pues para cada contenido se desarrolla un modelo de ítem que

permite generar tantos reactivos diferentes según sean los elementos y las condiciones impuestos en su creación. Por lo tanto, el EXHCOBA-R es un ejemplo de GAI y su desarrollo se ha efectuado desde la teoría débil, de acuerdo con Gierl y Lai (2012) y Drasgow et al. (2006). A continuación se definen los tipos de ítems que conforman el examen y se detalla cómo fueron elaboradas las especificaciones con los modelos de reactivos necesarios para la generación automática de reactivos.

2.3.3. Reactivos Estructurales Constructivos (REESCO)

El EXHCOBA ha utilizado, hasta el momento, el prototipo de opción múltiple para el desarrollo de sus ítems. Los exámenes de opción múltiple ofrecen grandes ventajas para las pruebas a gran escala: la posibilidad de obtener una muestra amplia de contenidos a evaluar, la objetividad en su calificación, la agilidad en su aplicación y la rapidez para reportar resultados. Pero, también presentan problemas como a continuación se explica.

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.

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.

Asimismo, los ítems de opción múltiple presentan entre tres y cinco posibilidades donde solamente una es la correcta (o la mejor). De este modo se restringe al evaluado a elegir entre las

limitadas opciones disponibles y se lo inhibe para que desarrolle y escriba su respuesta. Otra desventaja es el desgaste de los reactivos, que deben renovarse continuamente en sucesivas aplicaciones. Por último, el EXHCOBA ya utiliza los recursos computacionales para su aplicación, dados los avances de la tecnología digital, pero podrían explotarse estas capacidades en mayor grado.

También es cierto que en exámenes a gran escala es más sencillo poseer una plantilla de respuesta correcta única por ítem y compararla con las de los sustentantes en lugar de obtener un número muy alto de variabilidad de respuestas que compliquen la calificación. Otra característica de este tipo de exámenes es que pueden abarcar una cantidad muy vasta de contenidos; por ello, necesitan de un gran número de reactivos que los evalúen y el tiempo destinado a cada respuesta no puede ser mucho. En el caso del EXHCOBA se estima de 1 a 2 minutos por ítem.

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, los cuales reflejan la diversidad de opciones que presenta la vida diaria o que le dan la posibilidad al sustentante de crear su respuesta. Asimismo, su desarrollo se sustenta en las posibilidades de la tecnología computacional.

Estos prototipos se denominan Reactivos Estructurales Constructivos (REESCO). El concepto de *reactivo estructural* es porque cada ítem apela a un campo del conocimiento y no a un contenido puntual. Por lo tanto, el evaluado debe mostrar la capacidad para responder a un sistema de conceptos coherentes enlazados y no a un concepto único y desarticulado. El nombre de *reactivo constructivo* refiere al constructivismo cognitivo, ya que pretende evaluar aprendizajes significativos (ocurridos por recepción o por descubrimiento) de la educación básica o media superior, posición sustentada en el cognitivismo de Ausubel (2002).

Los REESCO descartan el formato de las cuatro opciones de los ítems tradicionales y buscan adaptarse a las exigencias de los diferentes contenidos y a las competencias a evaluar. Por tanto, no pueden reducirse a un único formato, sino que se distribuyen en diferentes tipos. En muchos casos de aritmética, física, química, geometría e incluso álgebra, los ítems solicitan al evaluado escribir la respuesta; el sustentante debe anotar la solución numérica o literal, según se requiera.

En geografía se presentan mapas y los estudiantes deben ubicar elementos en esas figuras. Un recurso similar se utiliza para trabajar con líneas del tiempo, ejes coordenados, formularios, entre otros. Otro procedimiento es clasificar la información, es decir, deben ubicarse en categorías (e.g.: proteínas, vitaminas, carbohidratos, lípidos) según sean los datos (e.g.: diferentes alimentos). Estos últimos ítems son útiles en ciencias naturales, en ciencias sociales y en español.

Otro aspecto interesante fue resolver el problema de la escritura para los ítems del área de lenguaje. Una solución es presentar textos con frases marcadas. Esas frases contienen opciones para ser reemplazados (e.g.: hablando/ablando, has/haz, vaya/valla). Lo esencial en este tipo de reactivos es que los elementos a modificar no aparecen aislados, sino contextualizados. Otra posibilidad es mostrar textos donde se solicita marcar fragmentos (e.g.: idea principal, frases que resuman, frases incorrectas). Para poder construir párrafos se proponen oraciones sobre un mismo tema y el estudiante debe reorganizarlas de tal manera que construya un texto claro y coherente.

Otra particularidad de los REESCO es la cantidad de respuestas solicitadas por reactivo. Existen ítems donde el resultado es único y solo puede ser calificado como correcto o incorrecto (*ítems dicotómicos*); pero, hay otros ítems donde se requieren dos o más respuestas que se

califican en función de lo que el estudiante pudo contestar correctamente. Estos últimos se conocen como *ítems de crédito parcial*.

2.3.3.1. La especificación de reactivos con el modelo de ítems

Para poder desarrollar ítems es necesario elaborar documentos que precisen las competencias escolares a evaluar, que indiquen los procesos intelectuales a medir y los detalles técnicos necesarios para generar los ítems. Este conjunto de características que definen a un ítem se denomina, para efectos de esta tesis, *especificación de reactivos* (INEE, 2005). La meta principal de la especificación de reactivos es asegurar que cada ítem evalúe, de manera precisa, el contenido seleccionado que, a su vez, representa el currículo del cual se extrajo. La información debe ser completa y clara con el fin de obtener reactivos válidos y confiables.

En el caso del EXHCOBA-R cada especificación de reactivos define las pautas para evaluar una competencia escolar a través de un conjunto de reactivos similares. Además, los REESCO exigen, dentro de las especificaciones, un formato especial: una plantilla que permita incorporar toda la información para la GAI.

En resumen, las especificaciones de reactivos diseñadas para el EXHCOBA-R deben proporcionar la siguiente información: (1) los *datos del elaborador y de los revisores*, (2) la *descripción del contenido a evaluar* y (3) la *plantilla* con las reglas y elementos conceptuales para elaborar un conjunto de reactivos similares. A continuación se describen los tres componentes.

1. En el espacio de *datos del elaborador y de los revisores* se plasman los datos de quién o quiénes redactan la especificación y de los revisores. También se incluyen los nombres y los apellidos completos con la fecha de elaboración, revisión o corrección (según corresponda). Una

especificación contiene tantas revisiones como sean necesarias, con sus respectivas correcciones, antes de ser aceptada por el Comité Técnico.

2. En el apartado de *descripción del contenido a evaluar* se define el contenido y su ubicación en el programa de estudios correspondiente. La información incluye: (a) el nombre del área; (b) el nivel educativo; (c) el eje temático (para Matemáticas), ámbito (para Español) o asignatura (para Ciencias Naturales y Ciencias Sociales); (d) el tema; (e) el subtema; y (f) el nombre del contenido a evaluar con su definición. Luego se incluyen las características del contenido: importancia, conocimientos necesarios, habilidades involucradas y delimitación, en toras palabras, se precisa con cuánta profundidad se evaluará el contenido. Un ejemplo de delimitación en aritmética consiste en aclarar las características de los números utilizados (como los números naturales, enteros, decimales o fraccionarios), las operaciones a realizar, el rango o la cantidad de números que se utilizan, si se incluyen problemas y de qué tipo, el estilo de la redacción, las fórmulas necesarias, las magnitudes, las unidades, entre otras características. Todo esto con el fin de fijar la extensión del contenido y por tanto, el grado de dificultad del reactivo.

3. Una vez obtenida la información que contextualice el contenido a evaluar se incluye la *plantilla*. El especificador debe decidir qué tipo de reactivo se ajusta mejor para evaluar dicho contenido. De acuerdo con esa elección, se procede a dar reglas para elaborar un conjunto de ítems organizados en una *familia de reactivos* con sus *ítems-padre* e *ítems-hijo*.

La figura 2.3 expone cómo se estructuran estos reactivos a través de un ejemplo: la competencia de *Representación de fracciones*. Para esta habilidad se definen una familia de reactivos, la cual contiene un ítem-padre que a su vez especifica la acción a realizar y en qué contexto. Para dicha instrucción se pueden seleccionar distintas figuras geométricas y fracciones diferentes por cada figura. Las figuras geométricas junto con las fracciones se denominan

elementos emergentes, ya que no deberían alterar la dificultad del ítem. Así, se selecciona una figura con una fracción y se generan los *ítems-hijo*, también denominados *ítems-hermano*. Cabe aclarar que según sea la figura geométrica serán las divisiones que contenga y las posibles fracciones a marcar; estas son restricciones para la elaboración de reactivos (con el objeto de producir ítems coherentes y enmarcados en situaciones posibles).

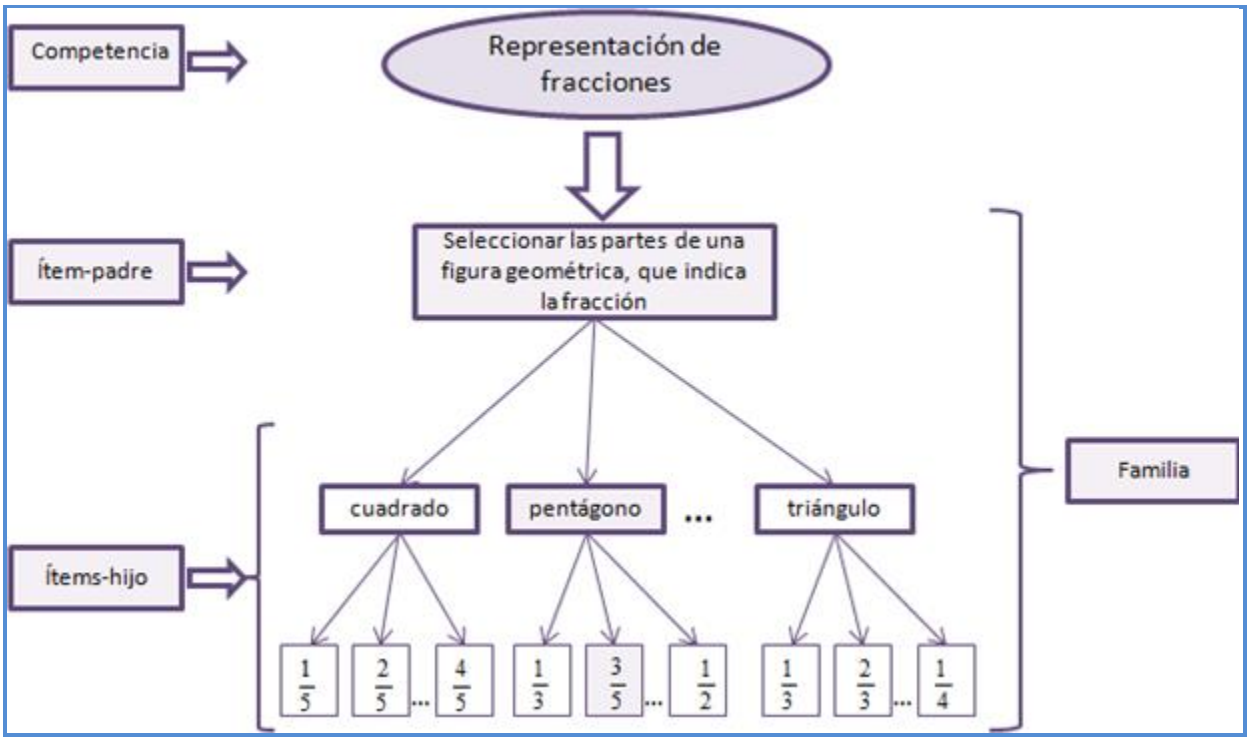


Figura 2.3. Ejemplo de esquema de cómo se estructuran los reactivos de una competencia curricular.

En la figura 2.4 se presenta un ejemplo del modelo de ítem para la competencia "Representación de fracciones" (para una especificación completa, véase Anexo A). La plantilla contiene diferentes secciones, las cuales son:

- estrategia de evaluación, donde se plantea de qué forma va a evaluarse el contenido, es decir, el tipo de actividad a realizar por el alumno;

- base del reactivo o directriz del ítem, la cual debe ser clara y lo más breve posible; contiene tres elementos: (1) un enunciado donde se presente el ejercicio (podría omitirse si el ítem fuera muy directo y sencillo), (2) la indicación de lo que se debe responder y (3) cómo contestarlo. Debajo de esta instrucción se presenta el esquema vacío del problema en sí;
- reactivo ejemplo con su respuesta correcta, o sea, un ejemplo como ilustración del reactivo, lo que permite precisar las características visuales del ítem (mapas, textos, cuadros o esquemas); también se exhibe la respuesta correcta del ejemplo; y,
- datos para el programa de cómputo que proporcionan la información necesaria para generar los diferentes reactivos con las respuestas correctas para cada caso. Se divide en dos apartados: (1) elementos y reglas para la construcción de reactivos, donde se incluyen los elementos que permiten generar los diferentes reactivos, cómo seleccionar dichos elementos y cómo presentar los ítems. Según sea el caso, se incluyen las particularidades de las figuras, las fórmulas necesarias, las características de los mapas, de las gráficas, de los esquemas, de los textos, etcétera. (2) Respuesta, que informa cuál es la respuesta correcta para cada caso; se aclara también del formato de la respuesta y los criterios que se deben observar para calificar.

Contenido: Representación de fracciones.

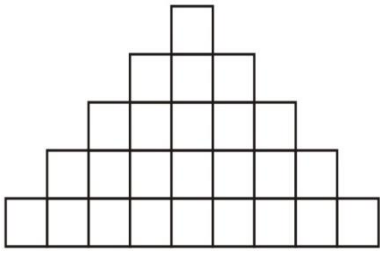
Estrategia de evaluación:
Se presenta una fracción propia y se pide colorear su equivalente en una cuadrícula.

Base del reactivo:
Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar, haz clic nuevamente sobre ellas.

Fracción	Figura

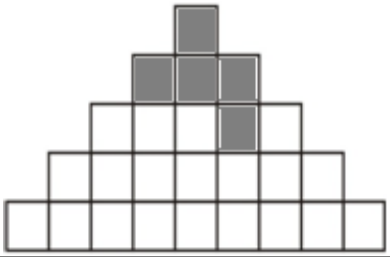
Reactivo ejemplo:
Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar, haz clic nuevamente sobre ellas.

$\frac{3}{15}$



Respuesta del reactivo ejemplo:
Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar, haz clic nuevamente sobre ellas.

$\frac{3}{15}$



Datos para el programa de cómputos:

- Elementos y reglas para la construcción de reactivos:
  - Se presenta al sustentante como base del reactivo el siguiente enunciado:
    a. Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar, haz clic nuevamente sobre ellas.
  - De manera aleatoria, el sistema debe elegir uno de las cinco **Figura**s propuestas en la **Tabla para la elaboración de reactivos.**
    Para la figura elegida, seleccionar una **Fracción** de la misma fila de la tabla.
    A continuación, presentar la fracción con la figura.
- Respuesta:
  - El sustentante, mediante un clic, podrá pintar (o sombrear) cada porción de la figura que se le presente.
    El sistema debe permitir al sustentante corregir; es decir, borrar alguna porción sombreada, mediante un segundo clic en la misma.
    Las partes pintadas podrán estar en cualquier lugar de la figura.
    Para cotejar la respuesta del sustentante, revisar en la columna "**R**", si el número de porciones

pintadas coincide con la de la respuesta.
Se otorgará un punto cuando el número de partes pintadas, en cualquier posición, coincida con **R**

Tabla para la elaboración de reactivos

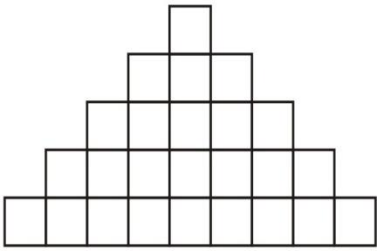
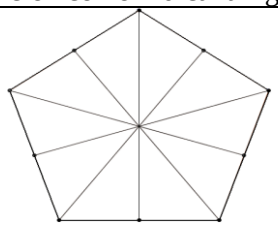
Figura	**Fracción =** $\frac{a}{b}$	**R** (n° de partes a pintar)
 **Observación:** cada una de las partes debe ser un cuadrado. Todos los cuadrados deben tener el mismo tamaño. Debe haber un total de 25 cuadrados colocados en la posición como indica la figura.	$\frac{1}{5},\frac{2}{5},\frac{3}{5},\frac{4}{5}$ $\frac{2}{10},\frac{4}{10},\frac{6}{10},\frac{8}{10}$ $\frac{3}{15},\frac{6}{15},\frac{9}{15},\frac{12}{15}$	$\boldsymbol{Fracción \times 25}$
 **Observación:** La figura debe ser un pentágono regular y los triángulos deben ser congruentes. Deben ser 10 triángulos.	$\frac{1}{5},\frac{2}{5},\frac{3}{5},\frac{4}{5}$ $\frac{3}{15},\frac{6}{15},\frac{9}{15},\frac{12}{15}$	$\boldsymbol{Fracción \times 10}$
Figura 3 **Observación:**	Fracciones 3	Respuesta 3
Figura 4 **Observación:**	Fracciones 4	Respuesta 4
Figura 5 **Observación:**	Fracciones 5	Respuesta 5

Figura 2. 4. Ejemplo de plantilla de reactivos del área de Habilidades matemáticas, del contenido: "Representación de fracciones".

Es importante mencionar que para el caso particular del EXHCOBA-R/MS se evaluaron tres contenidos, cada uno, a través de dos familias de reactivos; uno de los contenidos pertenece a Habilidades matemáticas y dos a Matemáticas. Para el resto (117 competencias) se utilizaron plantillas de una sola familia de reactivos con uno o varios ítems-padre.

2.3.3.2. Tipos de REESCO

Ya se mencionó que los nuevos reactivos no se limitan a un único formato. Existen 20 tipos de REESCO desarrollados para el EXHCOBA-R/MS, que se distribuyen en los 120 ítems de la prueba (ver tabla 2.5). Según esta tabla, los tipos de reactivos de mayor uso son: *elemento categoría* con el 38% del total (46 de 120 competencias), *RN fórmulas* y *Selección elementos* con el 14% y el 12.5%, respectivamente. Hay siete tipos de ítems que fueron creados para evaluar una única competencia del EXHCOBA-R/MS. En el caso del área de Ciencias sociales, el 85% de los ítems pertenecen a *elemento categoría* (17 de 20). En Ciencias naturales y español, poco más de la mitad también pertenece a este último tipo. En cambio, para Habilidades matemáticas, lo más utilizado es el tipo *RN fórmulas*[4] (60%). En Habilidades del lenguaje la mayoría pertenece a *Selección elemento*, aunque no llega al 50% (8 de 20), en Matemáticas, la mayor frecuencia es de *R algebraica*, sin embargo solamente representa al 25% del total.

[4] *RN fórmulas* refiere al tipo de reactivos de *Respuesta numérica con fórmulas*.

Tabla 2.5.
Distribución de tipos de reactivos en el EXHCOBA-R/MS, según el área de conocimiento y el tipo de respuesta

N°	Tipo	Área del conocimiento							Tipo de respuesta		
		HV	HC	ESP	MAT	NAT	SOC	TOTAL	A	E	S
3	Elemento imagen	1	2	-	-	-	3	6	✓		
4	Selección elementos	8	-	6	-	1	-	15			✓
5	Elemento categoría	4	-	11	2	12	17	46	✓		
6	Orden oraciones	1	-	-	-	-	-	1	✓		
7	RN fórmulas	-	12	-	2	3	-	17		✓	
8	RN ecuaciones	-	-	-	2	-	-	2		✓	
9	RN triángulos	-		-	1	-	-	1		✓	
10	RN pendiente	-	-	-	1	-	-	1		✓	
11	RN rangos	-	1	-	-	1	-	2		✓	
12	RN y selección	1	-	-	-	-	-	1		✓	✓
13	RN sucesiones	-	1	-	1	-	-	2		✓	
14	Orden números	-	1	-	-	-	-	1	✓		
15	RN etiquetas	-	2	-	3	-	-	5		✓	
16	Selección frase	3	-	1	-	1	-	5			✓
17	Orden elem. múltiple	-	-	-	-	2	-	2	✓		
18	frase imagen	2	-	2	-	-	-	4	✓		
19	RN iluminación	-	1	-	-	-	-	1		✓	✓
20	RN gráficas	-	-	-	1	-	-	1		✓	
21	RN R algebraica	-	-	-	2	-	-	2		✓	
22	R algebraica	-	-	-	5	-	-	5		✓	
	TOTAL	20	20	20	20	20	20	120	60	40	22

Nota: HV = Habilidades del lenguaje, HC = Habilidades matemáticas, ESP = Español, MAT = Matemáticas, NAT = Ciencias naturales, SOC = Ciencias sociales, A = Arrastre, E = Escritura, S = Selección. R = respuesta. RN = respuesta numérica.

Conforme a la actividad motriz involucrada en la respuesta solicitada, el 50% de los ítems son de arrastre, el 33% son de escritura y el 17% restante son de selección. En la tabla 2.6 se ordenan los tipos de reactivos y se describen según las diferencias de programación.

Tabla 2.6
Tipos de REESCO y sus características, según la respuesta motriz requerida

Tipo	Descripción según la actividad motriz
	Arrastre
Elemento categoría	Se presentan elementos (conceptos, enunciados o imágenes) y nombres de categorías. Se deben clasificar y ubicar los elementos en las categorías correspondientes.
Elemento imagen	Se presentan elementos y una figura (e.g. imagen, diagrama, recta, línea de tiempo). Se deben ubicar los elementos en los sectores correspondientes de la figura. El número de sectores disponibles es mayor que el de elementos.
Frase imagen	Se presentan elementos (palabras o expresiones marcadas en un texto) que se deben ubicar dentro de una figura con espacios en blanco.
Orden elemento múltiple	Es similar a *frase imagen*, la diferencia radica en que los elementos son imágenes.
Orden oraciones	Se presentan oraciones que deben ordenarse en un párrafo coherente.
Orden números	Se parece a *frase imagen*, pero los elementos son números a ordenar.
	Escritura
RN fórmula	Se presenta un ejercicio en lenguaje coloquial. Se solicita responder escribiendo un número. En algunos casos, junto al espacio de respuesta, se agregó una ventana donde se despliega una lista de unidades de medida y se debe elegir una. El problema puede estar acompañado de una tabla, una figura o un conjunto de datos.
R algebraica	Ejercicio donde se solicita escribir una expresión algebraica mediante el uso del teclado (signos: +, -, *, /, ^, números y letras).
RN R algebraica	Este tipo admite las posibilidades de los reactivos *RN fórmula* y *R algebraica*. Es capaz de generar aleatoriamente ejemplos de cualquiera de los dos tipos.
RN gráficas	Se asemeja a *RN fórmula*. La diferencia consiste en que el ejercicio siempre está apoyado en una gráfica de frecuencias generada aleatoriamente.
RN ecuaciones	Se parece a *RN fórmula*, pero el enunciado va acompañado de ecuaciones.
RN/pendiente	Se asemeja a *RN fórmula*; se presenta con la gráfica de función generada, aleatoriamente, en el plano cartesiano.
RN sucesiones	Se parece a *RN fórmula*; se presenta una lista de números que se debe continuar.
RN/rangos	Se asemeja a *RN fórmula.* La diferencia radica en que para cada respuesta puede determinarse un rango diferente de respuestas correctas.
RN triángulos	Se asemeja a *RN fórmula*. La diferencia consiste en que el enunciado va acompañado de una figura con datos generados aleatoriamente.
RN etiquetas	Es similar a *RN triángulos*. Los datos de la figura no se generan al azar, se capturan en el editor y luego se seleccionan al azar.
	Selección
Selección elemento	Se presentan textos o fórmulas con espacios vacíos o expresiones marcadas. Al hacer clic sobre lo marcado, se despliega una ventana con opciones, se debe elegir una que complete correctamente el texto o la fórmula.
Selección frase	Se presenta un texto con expresiones marcadas. La tarea es seleccionar una o más de esas expresiones, según se solicite.
	Escritura y selección
RN y selección	Similar a *RN fórmula*, la diferencia radica que en la ventana de opciones aparece una lista más extensa (e.g. países del mundo).
RN iluminación	Se presentan una figura dividida en partes iguales. Se solicita iluminar cierta cantidad de partes de dicha figura. También puede aparecer la figura con partes iluminadas y se solicita escribir un número como respuesta. (e.g. figura 2.6).

Nota: RN refiere a respuesta numérica.

Otra característica importante es la cantidad de respuestas solicitadas al examinado. De ahí, surgen ítems dicotómicos e ítems de respuesta parcial, con reactivos desde 2 hasta 21 respuestas. En la tabla 2.7 se presenta un resumen de las distribuciones del número de respuestas por área en el EXHCOBA-R/MS. De ella se puede inferir que aproximadamente el 30% de los ítems son dicotómicos, casi todos de Habilidades matemáticas o de Matemáticas; por lo cual, la gran mayoría de reactivos admite crédito parcial. De estos últimos, la cantidad de respuestas más frecuente representa casi el 34% con cinco respuestas y la siguiente más utilizada representa el 22%, con tres.

Tabla 2.7
Distribución del número de respuestas solicitadas por ítem según el área de aprendizaje considerada en el EXHCOBA-R/MS

N° de rtas. solicitadas	Áreas						Total
	HV	HC	ESP	MAT	NAT	SOC	
1 [a]	1	15	1	14	3	0	34
2	1	1	1	1	3	0	7
3	8	2	9	3	1 [b]	3	26
4	2	1 [c]	2	0	0	0	5
5	7	0	5	0	12	16	40
6	0	0	0	0	1	0	1
13 o 14	0	0	1 [d]	0	0	0	1
19, 20 o 21	0	0	1 [d]	0	0	0	1
1 o 2 [e]	0	1	0	2	0	0	3
Total	19 [f]	20	20	20	20	19 [f]	118

Nota: [a] Son los ítems dicotómicos.
[b] Se califica como de dos respuestas, porque la última está condicionada a las respuestas anteriores.
[c] Se califica como de tres respuestas porque la última está condicionada a las respuestas anteriores.
[d] Se agrupó para calificar como cuatro respuestas.
[e] Estos contenidos son evaluados por dos familias, una con una respuesta y la otra con dos.
[f] Al momento del pilotaje, faltaron dos plantillas, una para HV y otra para SOC. Posteriormente, en HV se agregó un modelo con ítems de cuatro respuestas y en SOC uno con reactivos de tres respuestas.

2.3.4. El generador automático de ítems

Toda la información de la especificación de reactivos se captura en un sistema de cómputo creado especialmente para el EXHCOBA-R. A su vez, lo capturado se almacena en una base de datos en línea. El sistema de cómputo contiene un módulo que recibe todos los insumos de la plantilla (bases de reactivos, reglas y elementos, imágenes) y con base en esta información se generan reactivos de manera aleatoria y automática. Este módulo se denomina *generador automático de ítems*.

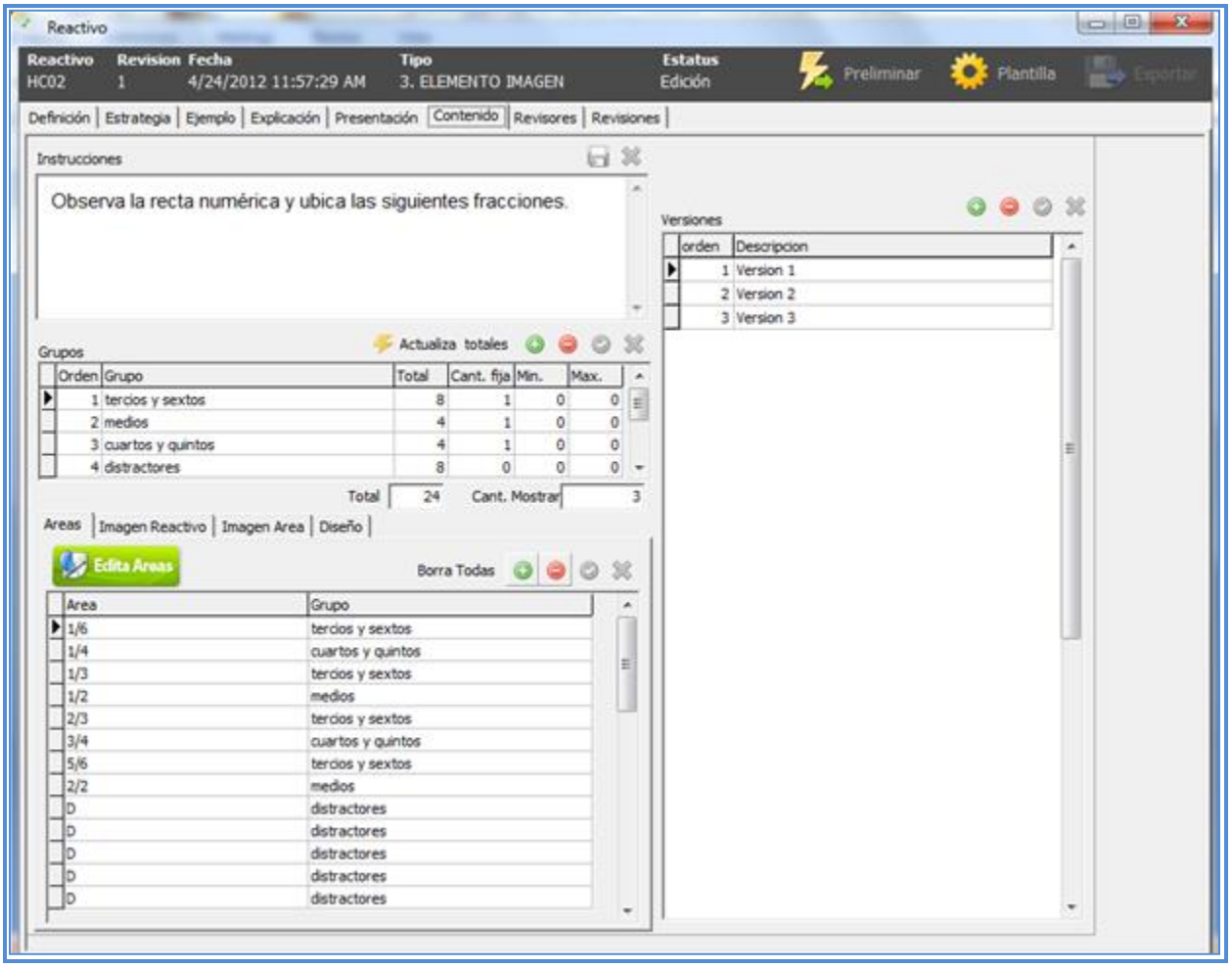


Figura 2.5. Imagen de la captura de los elementos de una familia de reactivos (HC02).

En la figura 2.5 se ejemplifica cómo son capturados los datos de la plantilla de una familia de reactivos. Esta familia contiene tres ítems-padre (versión 1, versión 2 y versión 3). Las *Áreas* representan los elementos que darán origen a los diferentes ítems-hijo. La zona de *Grupos* es donde se especifican las reglas de selección de los elementos. La base del reactivo es general para todos los ítems y se encuentra en la zona de *Instrucciones*.

Para entender el funcionamiento del generador, por ejemplo, en el caso de la plantilla de la figura 2.4, el generador automático de reactivos selecciona una figura (pentágono) y luego, una fracción ($\frac{3}{5}$); de esta forma presenta un reactivo denominado *ítem-hijo* como aparece en la figura 2.6.

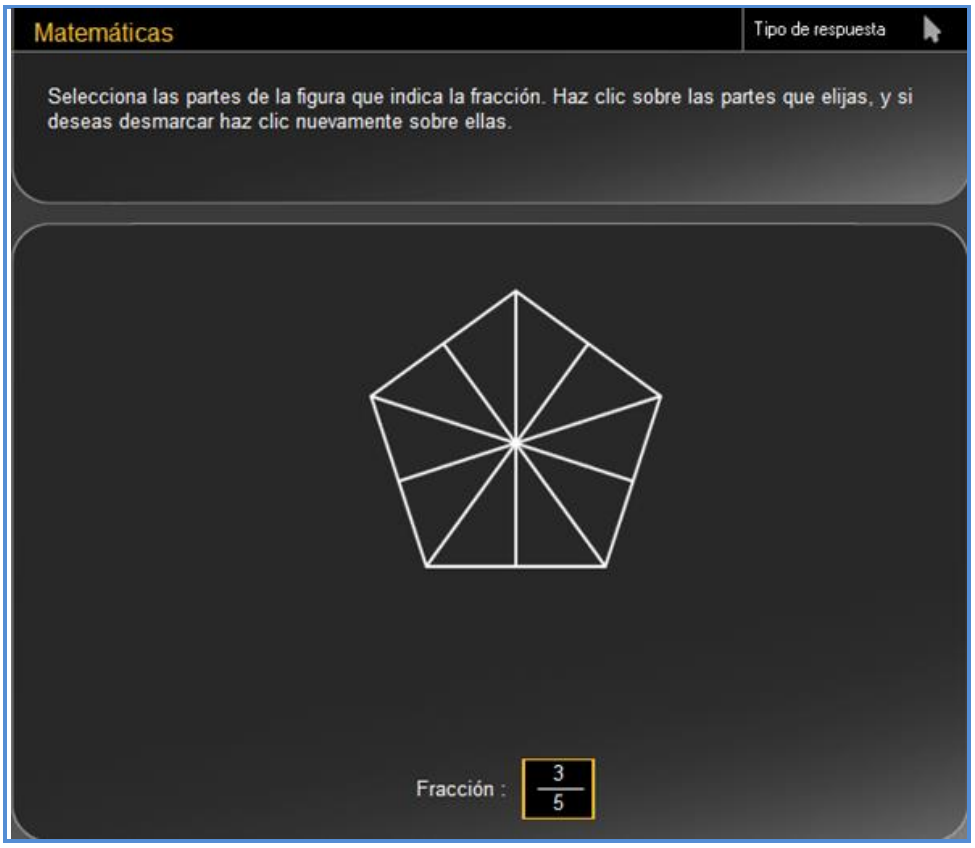


Figura 2.6. Ejemplo de ítem-hijo para la competencia "Representación de fracciones", del área de Habilidades matemáticas.

2.4. La validez como un criterio de calidad de las pruebas educativas de alto impacto

Todo instrumento utilizado en el ámbito educativo para evaluar a gran escala y particularmente, para tomar decisiones que afectan la vida académica de los evaluados, requiere de ciertas referencias que sirvan de modelo para garantizar exámenes eficientes que proporcionen los datos necesarios para establecer juicios justos y acertados. Es por esta razón que los exámenes educativos a gran escala y de alto impacto deben ser pruebas estandarizadas. A efectos de esta tesis, test estandarizado refiere a un examen que se desarrolla según patrones de calidad en su elaboración, aplicación y calificación (Martínez-Rizo et al., 2000). Estas pruebas tienen el objetivo de ofrecer, dentro de lo razonablemente posible, oportunidades justas e iguales para todos.

En el caso del EXHCOBA-R, esos marcos de referencias son proveídos por los *Standards* (AERA, APA y NCME; 1999) a nivel internacional y a nivel nacional, *Los estándares de calidad para instrumentos de evaluación educativa* (Martínez-Rizo *et al.*, 2000).

2.4.1. Estándares de calidad de las pruebas educativas

Los *Standards* (AERA, APA y NCME; 1999) plantean 15 tipos de patrones a seguir. El primero de todos es la validez y en él se citan 24 estándares. Entre otras especificaciones, en este patrón se indica la necesidad de proveer evidencias de estructura interna de un test, de publicarlas y de expresar bajo qué condiciones se obtuvieron los datos. El patrón número 2, con 20 estándares, se dedica a la confiabilidad y los errores de medición de los instrumentos. En él se declara la necesidad de reportar índices de confiabilidad para escalas, subescalas o toda combinación de puntuaciones que vayan a ser interpretadas. El patrón número 3 se dedica al desarrollo de los tests y a su revisión. En el estándar 3.9 se precisa la necesidad de obtener las propiedades psicométricas de los ítems, ya sea a través de la Teoría Clásica del Test (TCT) o de la Teoría de

Respuesta al Ítem (TRI). Para el caso de la TCT especifica que deben reportarse índices de dificultad y discriminación, mientras que para la TRI deben agregarse parámetros de ajuste.

El documento de Martínez-Rizo et al. (2000) incluye doce tipos de patrones de la calidad de los instrumentos. El estándar 4.4 establece que los reactivos deben pilotearse y someterse a análisis de dificultad y discriminación antes de ser utilizados para aplicaciones oficiales. Los estándares número 5 se refieren a la confiabilidad de los tests; indican que es necesario analizar constantemente la confiabilidad, precisar el método de estimación de la misma, reportar los resultados y las condiciones en que se estimaron estos valores. Los estándares número 6 están dedicados a la validez; el 6.1 determina que para todo instrumento de evaluación debe ejecutarse, primeramente, un análisis de validez en una etapa de prueba, y análisis posteriores de manera regular; estos resultados deben publicarse.

> Los instrumentos de evaluación deberán caracterizarse por su elevado nivel técnico en todas las dimensiones que deben atender los instrumentos psicométricos, en particular las diferentes variantes de la validez y la confiabilidad, de modo que se asegure la comparabilidad y objetividad de los resultados (Martínez-Rizo et al. , 2000, p. 18).

De lo expuesto se infiere que la validez de un examen es esencial para garantizar su calidad. A continuación se desarrollan cuestiones con respecto a la validación de instrumentos de evaluación educativa.

2.4.2. Evolución del concepto de validez

Messick (1993, p. 13) definió la validez como “un juicio evaluativo integrado del grado con el cual la evidencia empírica y los fundamentos teóricos sustentan la adecuación y la precisión de las inferencias y acciones basadas en las puntuaciones de una prueba o de otro modelo

evaluativo". Por lo tanto, la validez involucra una comparación de evidencias (práctica) y fundamentos (teoría) con las interpretaciones de las calificaciones de un test, y el uso de dichas interpretaciones. Por un lado, este juicio no es cuestión de todo o nada, sino de un nivel en una escala. Por otro lado, como toda prueba está inmersa en una condición social que cambia continuamente, su validez no es un hecho consumado, sino una propiedad en constante evolución y la validación es un proceso inacabable.

Históricamente, el primer tipo de validación fue criterial. La *validez de criterio* demuestra en qué medida los puntajes en el test de un examinado permiten inferir la ejecución que tendrá en una variable externa llamada criterio. Esta validez apunta a la relación de las puntuaciones en la prueba (generalmente, en términos de correlaciones o regresiones) con una medida criterio para un propósito determinado (ingreso a un nivel educativo superior, predicciones de aprovechamiento escolar, entre otros). Esta categoría se divide en dos sub categorías: predictiva y concurrente. En el primer caso, el criterio se fija en el futuro; por lo tanto, el estudio demandará un tiempo considerable. Si el criterio se fija en el presente, se trata de validez concurrente (los resultados del instrumento se correlacionan con el criterio en el mismo momento o punto del tiempo).

Messick (1993) advirtió que una prueba que se basa sólo en evidencias de validez referidas a un criterio podría ser vulnerable a la falta de evidencias en cuanto a relevancia de contenido o a consecuencias sociales adversas. Este modelo también enfrentó un problema de fundamento (Kane, 2006): ¿cómo validar el criterio? Aun cuando un segundo criterio puede validar el primero, esto solamente empuja el problema a un siguiente paso. El modelo criterial es muy útil en una validación secundaria de un atributo, cuando se asume que esta medida está disponible para ser asumida como criterio; pero, en todo caso, en algún punto, el criterio debe ser

validado de otro modo.

Posteriormente, surgió la *validación de contenido*. Esta validez señala el grado con que una muestra de ítems, tareas o preguntas de un test son representativas de un universo (dominio) de contenido definido. Las evidencias se basan en juicios de expertos, quienes sustentan la relevancia y representatividad de contenido de una prueba, y no en evidencias referidas a las puntuaciones de la pruebas. Según esta afirmación, la validez de contenido no se incluiría como validez; sin embargo, las consideraciones de relevancia y representatividad de contenidos influye notoriamente en la naturaleza de las inferencias de las puntuaciones en otro tipo de evidencias de validez (Messick, 1993).

Este modelo juega un rol importante en la validación; no obstante, es problemático cuando se asume para argumentar la validez de afirmaciones de procesos cognitivos u otros atributos teóricos; cuando es necesario ir más allá de estas interpretaciones se requiere otro tipo de evidencias (Kane, 2006).

Cronbach y Meehl (1955) desarrollaron el concepto de *validez de constructo* como una alternativa a los otros tipos de validez. Esta validez se enfoca en los puntajes del test como una medida de la característica psicológica de interés, por ejemplo, la inteligencia. Es la medida de un atributo, constructo, que no está definido operacionalmente, ya que los constructos son construcciones teóricas acerca de la naturaleza de la conducta humana.

En 1971, la validez de constructo continuaba como uno de las posibles aproximaciones a la validación; aunque Cronbach (1971) enfatizaba la necesidad de integrar los distintos tipos de evidencias de validez en la evaluación de las interpretaciones y usos de las puntuaciones de los tests. Los *Standards* de 1985 (AERA, APA y NCME) consideraron a la validez como un concepto unificado y reconocieron distintos tipos de evidencias. De todos modos, el documento

continuó con la división: las evidencias de validez relativa al criterio se debían utilizar para exámenes de ingreso, las evidencias de validez relativas al contenido serían para exámenes de aprovechamiento, y las evidencias relativas al constructo, para explicaciones basadas en la teoría.

Más tarde, Messick afirmó que todas las interpretaciones de las puntuaciones de los tests debían ser sustentadas por la validez de constructo y que el buen uso de las puntuaciones, con un propósito en particular, debía ser evaluada en términos de la relevancia que poseía el constructo con referencia al propósito dado (Messick, 1993).

Los *Standards* de 1999 aclararon dos aspectos fundamentales: (1) la validez es única y lo que se obtienen son distintas *evidencias* basadas en el enfoque y en las interpretaciones que se deseen hacer de las puntuaciones de los tests; la acumulación de evidencias aporta al grado de validez del instrumento; y (2) como todos los instrumentos son medidas de algún constructo, *validez de constructo* y *validez del instrumento* son términos equivalentes.

Con respecto al último punto del párrafo anterior, es necesario aclarar que una de las dificultades que planteaba la validez de constructo era que solamente podría aplicarse a tests psicológicos, y no a pruebas educativas. Como respuesta a este problema, los *Standards* (AERA, APA y NCME, 1999, p. 5 y p. 172) definieron al constructo, en una extensión amplia, como "el concepto o la característica para la cual se diseña un test". Todo test tiene el objetivo de medir el constructo por el cual fue desarrollado. Por eso, *todo* instrumento evaluativo debe contar con un constructo a medir. Ningún examen escapa al alcance de los estándares.

Los *Standards* (AERA, APA y NCME, 1999) destacaron cinco tipos de evidencias, según en qué se basen: en el contenido de los tests, en los procesos de respuesta, en la estructura interna, en la relación con otras variables o en las consecuencias de evaluar. De ellos, se definen a continuación las evidencias basadas en la estructura interna, concepto esencial para esta tesis.

> El análisis de la estructura interna de un test puede indicar el grado en que las relaciones entre los ítems del test y los componentes (teóricos)[5] del test se alinean al constructo sobre el cual se basan las interpretaciones propuestas de las puntuaciones del test. (AERA, APA y NCME, 1999, p. 13).

Para evitar confusión entre validez y validez de constructo, de aquí en adelante se consideran como sinónimos. Ambas se distinguen de las evidencias de validez basadas en la estructura interna del test.

2.4.3. Debate actual sobre el concepto de validez

Es necesario aclarar que existe una corriente que se opone al concepto de validez (específicamente, a la validez de constructo) defendido por Cronbach (1971), Messick (1993) y Kane (2006), entre otros. Entre los opositores se encuentran Borsboom, Mellenbergh y Heeerden (2004), quienes recuperaron la definición propuesta por Kelley en 1927 cuando afirmó que un test es válido si mide lo que pretende medir. Es más, un test es válido para medir un atributo si y solo si: (a) el atributo existe y (b) variaciones en el atributo producen variaciones en los resultados del procedimiento de medición. La idea general se basa en la teoría causal de la medición. Los autores aseveran que la teoría de la validez se funda en ontología, referencia y causalidad y no, en epistemología, concepto y correlación. Si bien las cuestiones epistemológicas son esenciales a la validación y las consecuencias son claves para el uso del test, ambas son irrelevantes al concepto y definición de la validez en sí misma.

Borsboom, Cramer, Kievit, Scholten y Franic (2009) afirmaron que la teoría de validez de constructo sostiene que la validez es una propiedad de las interpretaciones de las puntuaciones en los tests en términos del constructo que refleja la fortaleza de las evidencias para estas

[5] *teóricos* fue agregado para una mejor comprensión del tipo de componentes al que se refiere la definición.

interpretaciones. Ellos propusieron otro punto de vista, el cual sostiene que la validez es una propiedad de los instrumentos que codifica si estos instrumentos son sensibles a variaciones en un atributo objetivo.

Esta propuesta no es trivial y genera otra perspectiva dentro de la investigación evaluativa. Esta tesis se sustenta en el concepto de validez tal como lo definen los clásicos psicómetras estadounidenses y bajo la concepción que proponen los *Standards* (AERA, APA y NCME, 1999), por considerar esta teoría sustentada en un gran número de investigaciones y que ha proporcionado importantes aportes a la evaluación educativa.

2.5. Las propiedades psicométricas de los ítems desde la TCT y la TRI

Las evidencias basadas en la estructura interna de un test indican si las relaciones entre los ítems, y las dimensiones definidas teóricamente, permiten confirmar la existencia de los constructos teóricos que el test pretende medir. Para poder ejecutar los análisis correspondientes que permitan obtener dichas evidencias, previamente, es necesario averiguar las propiedades psicométricas de los ítems. Según lo plantean los *Standards* (AERA, APA y NCME, 1999), se puede realizar desde la Teoría Clásica del Test (TCT) o desde la Teoría de Respuesta al Ítem (TRI). En este apartado se presentan ambas teorías y se incluye el Análisis Factorial Confirmatorio (AFC) como un método para establecer evidencias de estructura interna de instrumentos de medición.

Los avances en métodos de evaluación educativa incluyen el desarrollo de modelos formales (psicometría) que representan una forma particular de razonamiento, a partir de evidencias. Estos modelos proveen de reglas para integrar los diferentes trozos de información conformados por las distintas tareas de una evaluación. Particularmente, los exámenes a gran escala requieren de las posibilidades de las estructuras estadísticas formales que permitan

analizar modelos complejos de aprendizaje y grandes volúmenes de datos.

Con el fin de aportar evidencias de validez del EXHCOBA-R/MS, se deben interpretar y utilizar correctamente las propiedades psicométricas y sus fórmulas asociadas. Para ello es necesario conocer los fundamentos en los que se basan dichas fórmulas, es decir, las dos grandes teorías que guían la construcción y el análisis de la mayoría de los tests: la Teoría Clásica de los Tests (TCT) y la Teoría de Respuesta al Ítem (TRI). Por lo tanto, en primer lugar se expone la TCT como primera gran estructura teórica de la psicometría, sus alcances y limitaciones. En segundo lugar se introduce la TRI como una solución a algunos problemas surgidos del enfoque clásico.

2.5.1. La Teoría Clásica de los Tests

La TCT ha sido el enfoque predominante en la construcción y análisis de los tests psicológicos, tiene origen en los trabajos de Spearman de 1904, 1907 y 1913, cuenta con 100 años de historia, con los cuales se ha ganado el adjetivo de clásica. La TCT debe asegurar que el instrumento que se utiliza para medir una característica, lo hace con precisión. Para ello se debe asumir que la puntuación que una persona obtiene en un test (X, puntuación empírica) está formada por dos componentes: la puntuación verdadera (V) de esa persona y un error (e) debido a muchas causas que no se pueden controlar. Traducido en términos matemáticos, sería:

$$X = V + e$$

Spearman basó su teoría en tres supuestos: (1) la puntuación V se define como la puntuación media que obtendría una persona si se le aplicara infinitas veces el test (es una definición teórica, ya que este procedimiento es imposible en la práctica); (2) no existe relación entre el tamaño de la puntuación con el error que se comete (puede haber puntuaciones altas con errores pequeños o grandes, indistintamente); y (3) los errores de medida de las personas en un

test no están relacionados con los errores de medida en otro test distinto. En otras palabras, los errores cometidos en una ocasión no varían sistemáticamente con los cometidos en otra ocasión. Además, se definen *tests paralelos*, como aquellos que miden exactamente lo mismo, pero con ítems diferentes. De este modo, las puntuaciones verdaderas y las varianzas de los errores de medida son las mismas en ambos tests (Muñiz, 2010).

Si bien el enfoque de la TCT ha sido muy útil y eficaz para el desarrollo de tests, presenta ciertas limitaciones. Una de ellas es que las mediciones no son invariantes respecto al instrumento utilizado. Los puntajes de dos tests que miden el mismo rasgo, aunque se estandaricen, suelen ser distintos debido a que cada test tiene su propio conjunto de ítems y cada ítem tiene propiedades distintas; los errores de cada ítem son distintos y no se ajustan a la estandarización del test. Otra restricción es que no pueden separarse las características de los examinados de las características del test, solamente cabe examinar una en el contexto de la otra (si un test es "fácil", el evaluado se reportará como con mucha aptitud, mientras que si el test es "difícil", el examinado parecerá tener poca aptitud) (Cortada de Kohan, 2004).

2.5.2. La Teoría de Respuesta al Ítem

La TRI, también conocida como Teoría de los Rasgos Latentes, es una familia de modelos estadísticos que se utilizan para analizar los datos de un ítem de un test. Esta teoría provee un proceso estadístico unificado para estimar características estables de los ítems y de los examinados, y define cómo estas características interactúan en la descripción del ítem y el desempeño en el test.

Los supuestos de la TRI están relacionados con la probabilidad de que un examinado, con un nivel particular de habilidad, produzca una respuesta específica a un ítem determinado. Los modelos de la TRI relacionan las puntuaciones de los ítems con los niveles de habilidad del

evaluado (β) y los parámetros del ítem, a través de funciones no lineales. La TRI produce un rango amplio de predicciones incondicionales (para los grupos examinados) y condicionales (para cada individuo, según su nivel de habilidad); estas predicciones son más flexibles que las de la TCT y se expresan en unidades de puntuación distintas a las unidades que apelan a la cantidad de respuestas correctas observadas (Yen y Fitzpatrick, 2006). Es decir, la TRI se concentra en los ítems individuales, en sus características estadísticas y en sus contribuciones independientes a los tests.

Muñiz y Hambleton (1992) recopilaron información acerca de los primeros cincuenta años de la TRI; según estos autores, Binet y Simon hicieron un primer acercamiento a principios del siglo pasado, Thurstone y Ackerson realizaron los primeros bosquejos durante la década de los 20's (aunque en ese momento no se presumía llegar a una teoría); Ferguson en los 40's con el planteo de curvas características; en la misma época surgieron los aportes de Lawley, Tucker y Lazarsfeld, este último fue quien acuñó el término de "latente". Sin embargo, el nacimiento formal se podría señalar con Lord, a través de su tesis doctoral de mediados de siglo pasado (1952). Luego, en 1960, Rasch publicó su libro donde desarrolló el modelo logístico de un parámetro, que se ha convertido en el más popular y más utilizado. En 1968 otro impulso lo generaron Lord y Novick en su libro donde dedicaron cinco capítulos a la nueva teoría.

Si bien en ese momento surgió un nuevo camino para la psicometría, treinta años después se impuso en ese campo. Esto probablemente se debió a la dificultad del modelo y a la carencia de programas computacionales que pudieran manejar y analizar el gran número de datos que implicaba la teoría. En la década de los setenta aparecieron los primeros programas, pero fue en los 80's y 90's cuando se expandió y se afianzó su aplicación, con el gran apoyo del libro de

Lord "*Application of Item Response Theory to Practical Testing Problems*" en 1980 (Muñiz y Hambleton, 1992).

Una de las principales aportaciones de la TRI a la psicología y la educación es la multiplicidad de modelos, según las necesidades técnicas de los instrumentos. Dentro de estos modelos, los más utilizados y estudiados son los logísticos de 1, 2 y 3 parámetros (Yen y Fitzpatrick, 2006), cuyos supuestos asumen a la función logística como la función matemática que asigna a los valores de la variable medida, una probabilidad de acertar el ítem; con lo cual se genera una Curva Característica del Ítem (CCI). Además, la suma de las CCI origina la Curva Característica del Test. Si para la función se considera solamente el parámetro de dificultad (es decir, se asume que acertar o fallar un ítem depende solamente de la dificultad), se denomina modelo logístico de un parámetro; si se consideran dificultad y discriminación, el modelo será de dos parámetros; y el de tres parámetros será el que agregue la posibilidad de aciertos al azar.

La relación entre el contenido del ítem y el orden de dificultad empírica de los ítems, producto de la manera en que los individuos responden a ellos, verifica, mejora o contradice el significado de la variable cuyos ítems pretenden ejecutar. Es decir, cuando se utilizan ítems con personas se puede comparar el orden conceptual esperado con el orden empírico real que proveen los datos para analizar en qué medida se confirman las expectativas. Esta validez de orden del ítem operacionaliza dos tipos de evidencias de validez que proponen los *Standards* (AERA, APA y NCME, 1999): la de contenido y la de estructura interna. El orden de dificultad de los ítems define el significado de la variable y por lo tanto, ambos tipos de validez (Wright y Stone, 1999).

Otro aporte fundamental de la TRI es la unificación de la métrica para distintos tests que se utilicen y la determinación del instrumento más adecuado para obtener la mejor información

acerca del punto de corte entre los sujetos que conocen un dominio y aquellos que no; esto último es de gran importancia práctica, ya que muchas instituciones deben tomar decisiones en torno a un examen.

A modo de resumen, la Tabla 2.8 describe las diferencias entre ambas teorías. De ella se desprende que la TRI es más sólida que la TCT; aunque la simplicidad de la última provee de una descripción más sencilla de las propiedades psicométricas de los ítems.

Tabla 2.8
Diferencias entre la TCT y la TRI

Características	TCT	TRI
Tipo de modelo	Lineal	No lineal
Complejidad del modelo	Menor	Mayor
Supuestos	Débiles (fáciles de cumplir con los datos)	Fuertes (difíciles de cumplir con los datos)
Invarianza de las mediciones	No	Sí
Invarianza de las propiedades del test	No	Sí
Escala de puntuaciones	De 0 a la puntuación máxima del test	Desde -∞ hasta +∞ (usualmente de -3 a +3)
Énfasis	Test	Ítem
Relación ítem-test	Sin especificar	CCI [a]
Descripción de los ítems	Índices de dificultad y de discriminación	Parámetros de dificultad, discriminación y azar

Fuente: basado en Muñiz (2010). [a] CCI = Curva Característica del Ítem.

2.5.2.1. El modelo de Rasch

El modelo de Rasch o Modelo Logístico de un Parámetro es un modelo de la TRI, basado únicamente en un parámetro; calcula la probabilidad de una respuesta específica (bivalente: respuesta correcta o incorrecta) a través de una función logística de la persona (conocida como habilidad o rasgo del sujeto) y del parámetro del ítem (dificultad del ítem) (Yen y Fitzpatrick, 2006). Es un modelo estocástico (en otras palabras: probabilístico, no determinista), donde la medida del rasgo y la del ítem aplicado se ubican en una misma escala lineal con un origen común (Tristán-López, 2001).

La escala que utiliza el modelo de Rasch sirve tanto para medir la habilidad en las personas como para calibrar la dificultad de los reactivos; permite comparar el dominio de un individuo con los de otros individuos o con respecto a un criterio. Esta escala es lo suficientemente extensa, de modo que puede medir a todos y lo suficientemente precisa, con divisiones igualmente espaciadas, lo que posibilita mayor exactitud. La unidad de medida se denomina "*lógito*" (traducción libre de '*log odd ratio unit'*) y se define como el logaritmo natural del *momio*[6]; por ser un logaritmo, los lógitos pueden tomar cualquier valor real.

El modelo de Rasch permite calcular la probabilidad de lo que sucede cuando una persona utiliza su habilidad al enfrentarse con la dificultad de un ítem. De acuerdo con Wright y Stone (1998), la fórmula que vincula rasgo-dificultad del ítem se puede pensar de la siguiente manera. Primeramente, se combinan los parámetros β_n (habilidad de la persona n) con δ_i (dificultad del ítem i) a través de su diferencia ($\beta_n - \delta_i$). Esta resta puede tomar cualquier valor real; si se le aplica una operación exponencial de base *e* ($e^{(\beta_n - \delta_i)}$), se acotan los resultados desde cero a más infinito[7]. Como lo que se pretende es buscar una probabilidad, los valores deben restringirse al intervalo [0; 1]; por lo tanto, se aplica la siguiente razón y se obtiene:

$$P_{1ni} = \frac{e^{(\beta_n - \delta_i)}}{1 + e^{(\beta_n - \delta_i)}} \qquad (1)$$

La fórmula (1) se entiende como la probabilidad de acierto de una persona n con habilidad β_n al ítem i con dificultad δ_i. Debido a que la respuesta puede tomar únicamente dos

[6] Momio es un término utilizado en las apuestas, donde se arriesgan, con puntos a favor o en contra, acerca de que un jugador, un caballo, un boxeador, etc. pueda ganar.
[7] e = número de Neper = 2.718…

valores: 1 (correcta) y 0 (incorrecta), las expresión que se desprende de obtener la *probabilidad de respuesta incorrecta* es:

$$P\{x_{vi} = 0 \mid \beta_v, \delta_i\} = 1 - \frac{e^{(\beta_n - \delta_i)}}{1+e^{(\beta_n - \delta_i)}} = \frac{1}{1+e^{(\beta_n - \delta_i)}} \quad (2)$$

Cabe aclarar que no cualquier ecuación permite estimar β_n y δ_i, de modo que las estimaciones de β_n no dependan de los efectos de δ_i, y viceversa; Rasch y Andersen, entre otros (cit. en Wright y Stone, 1998), probaron que no existe otra fórmula matemática que exprese esta relación de manera correcta, y que permita la estimación de las medidas de las personas y la calibración de los ítems, de forma independiente.

Resulta oportuno entonces, recordar que la relación (X = V + e) en la TCT es de tipo lineal, mientras que en el modelo de Rasch, se trata de funciones exponenciales. Es interesante remarcar la parsimonia de este modelo, sólo necesita un parámetro para describir una gran cantidad de datos.

Para el modelo de Rasch, la Curva Característica del Ítem (CCI) es una gráfica que describe la probabilidad de respuesta correcta a un ítem determinado, de acuerdo con la habilidad del evaluado con que se enfrenta. En general, en la TRI, existen dos tipos de modelos de curvas, las que se ajustan al reactivo y las de contraste. Estas últimas definen las características y propiedades que se buscan en un ítem y luego, se contrastan con los puntos que define el reactivo en la práctica, es decir, se establece previamente un modelo de cómo se deben comportar los reactivos para decidir si el ítem se ajusta o no a ese modelo. Rasch es un ejemplo de modelo de contraste y la función que define esta CCI es:

$$P_i(\beta) = P_i(x_i = 1|\beta) = \frac{1}{1+e^{[-(\beta-\delta_i)]}} \qquad i = 1, 2, 3, \ldots, k \qquad (3)$$

Donde:

x_i = es la puntuación para el ítem i (en este caso es acierto)

β = es el valor de la habilidad para cada individuo evaluado.

P_i = es la probabilidad de que un examinado elegido al azar con aptitud β conteste correctamente el ítem i

δ_i = parámetro de la dificultad del ítem i

k = número de ítems del test

Para una interpretación gráfica de la función (4), en la figura 2.7 se muestra la CCI para tres ítems; allí se puede apreciar que dado un ítem, la probabilidad de respuesta correcta se aproxima a 0 para niveles muy bajos de habilidad y a 1, para niveles muy altos de habilidad. Las líneas punteadas señalan que el ítem 1 tiene un valor de δ_1 = -1, mientras que en el caso del ítem 2, δ_2 = 0 y el ítem 3, δ_3 = 1. De la gráfica se infiere, por ejemplo, que el ítem 3 es más difícil que los ítems 1 y 2 (un examinado necesita mayor habilidad para tener el 50% de probabilidad de contestar correctamente el ítem 3 que para los ítems 1 y 2). A su vez, el ítem 2 es más complejo que el ítem 1. Se debe destacar que, de acuerdo con la idea de modelo estocástico, una medida de conocimiento de un sujeto no es un valor preciso y exacto de los conceptos y habilidades que maneja esa persona; sino que es una probabilidad de respuesta. Aquel que tenga una medida alta de conocimiento, tendrá una mayor probabilidad de dar una respuesta correcta a un ítem.

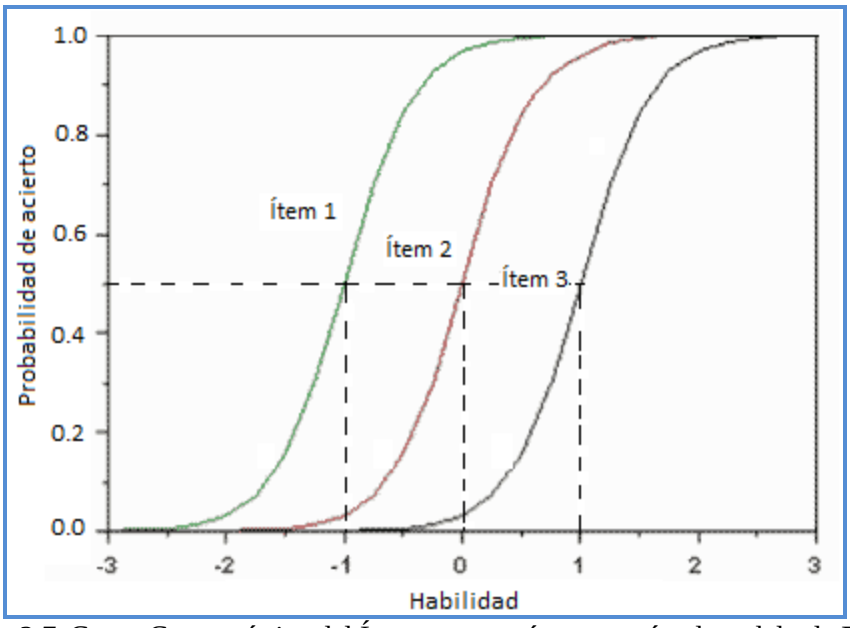


Figura 2.7. Curva Característica del Ítem para tres ítems, según el modelo de Rasch.

Embretson y Reise (2000) indicaron las propiedades de la CCI para el modelo de Rasch: (1) las probabilidades de responder correctamente aumentan a medida que aumenta el nivel de habilidad, (2) las pendientes son iguales para todas las curvas (ya que los ítems solo difieren en dificultad), las curvas convergen, pero no se cortan y 3) el punto de inflexión de cada curva (donde la tasa de cambio pasa de crecer aceleradamente a decrecer desaceleradamente) ocurre cuando la probabilidad de acertar un ítem es de 0.5

En términos numéricos, la CCI queda determinada por tres valores: la dificultad (δ), la discriminación (a) y la adivinación sistemática (c). En el caso de Rasch, el valor δ se encuentra en el punto de inflexión de la curva (donde la probabilidad es 0.5), la discriminación es $a = \frac{1}{1.7}$ para todos los ítems de cualquier test, y $c = 0$, también para todos los reactivos (Tristán-López, 2001).

Una gran consecuencia de estos modelos unidimensionales es que implica independencia local, es decir, un ítem no depende del resto de los ítems (es una curva en un plano); si no se

caería en una contradicción, pues la respuesta no dependería únicamente de la variable medida (β).

Si bien, originalmente, el modelo de Rasch se aplica para datos dicotómicos, existen diferentes extensiones que conforman una familia de modelos de Rasch. Masters (1982) en su artículo: *Un modelo de Rasch para calificación de crédito parcial*[8], amplió el modelo básico a ítems que admiten calificación parcial (respuestas parcialmente correctas). De este modo, se permite calificar en dos o más categorías ordenadas, y las alternativas son libres de variar en número y estructura, de ítem a ítem. Las calificaciones se ordenan en 0, 1, 2, 3, etc. indicando un orden. El nivel más alto de ejecución lo otorga el número mayor posible en cada ítem. Bajo este formato, el número de pasos en cual se divide cada reactivo varían según el ítem.

La fórmula que se utiliza para el modelo de Masters es similar a la (1), solamente que ahora la probabilidad se reparte en pasos y con el requisito de que la persona *n* debe responder alguna de las categorías disponibles, se obtiene la expresión general para calcular la probabilidad de una persona *n* de obtener una puntuación *x* en el ítem *i*:

$$P_{xni} = \frac{e^{\sum_{j=1}^{x}(\beta_n - \delta_{ij})}}{\sum_{r=0}^{m_i} e^{\sum_{j=0}^{r}(\beta_n - \delta_{ij})}} \qquad x = 0, 1, \ldots m_i \qquad (4)$$

La ecuación (4) establece que la probabilidad de dar una respuesta de una calidad determinada (*x*) para un sujeto, dado su nivel de habilidad (β) y el conjunto de dificultades asociadas a cada paso del ítem *i* (δ_{i1}, δ_{i2}, δ_{i3}) corresponde a la diferencia entre la habilidad del sujeto y la dificultad del paso asociado a cada categoría, considerando todos los pasos del ítem

[8] Título original: *A Rasch model for Partial Credit Scoring*.

en conjunto. Un ítem de *m* categorías tendrá un número de pasos *m*-1. Esto implica que la probabilidad de alcanzar una determinada categoría está condicionada al grupo de sujetos que alcanzan esa categoría o la categoría anterior. Los parámetros de localización de cada categoría no pueden ser interpretados como la dificultad de esa categoría, porque también incluyen información acerca de la categoría anterior.

De este modo se obtiene un modelo unidimensional de Rasch para categorías de respuesta ordenadas. Este modelo de crédito parcial permite parametrizar las dificultades de una serie de pasos en cada ítem.

2.6. Otros modelos estadísticos para aportar evidencias de validez basadas en la estructura interna

Existen diferentes modelos estadísticos que permiten obtener evidencias de validez basadas en la estructura interna. Entre ellos, el análisis factorial exploratorio, uno de los procedimientos más antiguos, es una técnica estadística perteneciente al Modelo Lineal General (Martínez-Arias, Hernández-Lloreda, M. V. y Hernández-Lloreda, M. J., 2006). En este tratamiento se pretende reducir la información a través de la búsqueda de dimensiones o constructos latentes a partir de las correlaciones entre las variables observadas, con el menor número de factores y la mayor varianza condensada. El procedimiento es gradual e inducido por los datos, no presupone un número determinado de factores.

Otro recurso son los modelos componenciales, que se aplican en las teorías cognitivas para identificar los componentes cognitivos que sirven para comprender los constructos que miden los reactivos. Entre ellos, destacan el Modelo de Test Logístico Lineal (LLTM) de Fischer y el Modelo General del Rasgo Latente (GLTM) de Embretson. Los supuestos, los requisitos y

cómo evaluar el ajuste entre datos y estructura son cuestiones importantes a considerar en la evaluación de la utilidad de estos modelos (Yen y Fitzpatrick, 2006).

Otro procedimiento muy utilizado es el modelamiento de ecuaciones estructurales. Este incluye una gran variedad de técnicas que se distinguen por dos características: (1) la estimación de múltiples e interconectadas relaciones de dependencia y (2) la habilidad para representar conceptos no observables en dichas relaciones. Entre las diferentes técnicas, Hair, Anderson, Tatham y Black (1992) citaron el análisis de estructura de covarianza, el análisis de la variable latente y el Análisis Factorial Confirmatorio (AFC). Dada la importancia, para la presente tesis, de la técnica de AFC, a continuación, se dedica un apartado para su definición y la descripción de sus características.

2.6.1. El Análisis Factorial Confirmatorio

El Análisis Factorial (AF) es un conjunto de técnicas estadísticas que estudian la superposición de varianza para agrupar, a través de las correlaciones entre un número grande de variables observadas, en una cantidad menor de variables hipotéticas subyacentes, denominadas *factores*. El fundamento del AF radica en que la existencia de correlaciones altas entre un número de variables observadas se debe a la existencia de variables latentes o constructos, que explican estas relaciones (Zamora-Muñoz, Monroy-Casorla y Chávez-Álvarez, 2010). La interpretación de lo que cada factor significa queda en manos de la evaluación subjetiva del investigador (Williams y Monge, 2001). Estos procedimientos matemáticos se pueden utilizar para obtener evidencias de validez de estructura interna de un test (Martínez-Arias, *et al.*, 2006).

El AFC, como parte de los modelos estructurales, involucra variables latentes, (e.g.: habilidad verbal, ansiedad, calidad de vida) que se proponen desde una teoría. Estos constructos definen variables observables y medibles (e.g.: la respuesta a un ítem en un examen, el número

de aciertos en un test, el número de computadoras en una vivienda). El análisis consiste en *poner a prueba un modelo teórico*, donde se establecen las relaciones entre las variables observadas que se asocian a los constructos definidos para constatar si refleja el comportamiento de los datos (variables observadas).

El modelo de medida que precisa cómo las variables observadas expresan a las variables latentes se expresa a través de la matriz de varianzas-covarianzas:

$$\sum = \Lambda_x \Phi \Lambda_x' + \Theta_\delta$$

$\sum =$ matriz de varianzas-covarianzas del modelo teórico, simétrica, con varianzas de las variables en la diagonal principal y las covarianzas como elementos externos a dicha diagonal.

Λ_x = matriz de saturaciones de las variables en los factores.

Φ = matriz de covarianzas entre los factores comunes.

Θ_δ = matriz de covarianzas entre los factores únicos o residuos, que será diagonal cuando no se permitan covariaciones entre estos factores.

El análisis consiste en encontrar los valores de $\widehat{\Lambda_x}$, $\widehat{\Phi}$ y $\widehat{\Theta}_\delta$ para que la matriz de covarianza poblacional ($\widehat{\Sigma}$) que presenta el modelo sea lo más próxima posible a la matriz de covarianzas muestral (S) que aportan los datos. Para ello, se define una función de ajuste, $F(S, \widehat{\Sigma})$, que mide la proximidad entre $\widehat{\Sigma}$ y S. Para cada valor de F, se obtiene un escalar que representa la distancia entre ambas matrices. Esta distancia recorre un continuo no negativo, y es cero cuando ambas matrices son iguales.

Al respecto, Corral-Verdugo y Obregón-Salido (1998) determinaron que el AFC, como parte de los modelos de ecuaciones estructurales, requiere de tres fases de análisis: la especificación, la prueba de bondad de ajuste y la estimación de los parámetros. En la primera

fase se definen las variables latentes (constructos) y sus indicadores (qué ítems evalúan a dichos constructos), se fijan las varianzas y covarianzas entre variables y los errores de las mismas. En un segundo paso se debe contrastar el modelo inclusivo obtenido a partir de los datos y un modelo teórico restringido (propuesto por los desarrolladores de los tests). Finalmente, en la tercera fase se estiman los parámetros para cada relación específica.

Si la estructura de correlaciones se explica por un único factor, entonces se trata de un modelo unifactorial o unidimensional; en el caso de que el modelo se defina con más variables latentes, se denomina multifactorial o multidimensional. En este último tipo de modelos se espera que las variables que se agrupan en un factor se correlacionen altamente entre sí y pobremente con las que componen otro factor.

Nunnally y Bernstein (1995) plantearon algunas condiciones básicas que deben cumplir las variables para poder aplicar un AFC. Las variables deben ser continuas y de distribución normal, cada variable observable debe correlacionarse altamente con, al menos, una de las variables restantes y se deben tomar muestras grandes (para impedir que los agrupamientos sean solo efectos de errores muestrales).

¿Qué sucede si se desea efectuar un análisis de datos donde cada variable es un ítem dicotómico? Cuando se realizan estudios de AFC con ítems dicotómicos pueden surgir serias dificultades ya que estos ítems, al presentar solo dos posibilidades, no se distribuyen normalmente, lo que provoca correlaciones pequeñas y dificultad para conformar factores. En estos casos es común la conformación de factores por razones meramente estadísticas (lo que se denomina *factores de dificultad*), es decir, los ítems solamente tienen en común sus índices de dificultad. Esto puede resultar en presentar como multidimensionales datos que en realidad son unidimensionales (Martínez-Arias et al., 2006; Nunnally y Bernstein, 1995).

Surge, entonces, la necesidad de sortear este inconveniente. En general, existen dos grandes caminos: uno de ellos es tratar los datos dentro de un modelo de la TRI ("es importante señalar la equivalencia formal de la TRR[9] y el análisis factorial clásico como modelos"; según Nunnally y Bernstein (1995, p. 641), el otro es utilizar algún enfoque analítico factorial. Dentro de este último, una solución en realizar un AF clásico con la matriz de correlaciones tetracóricas, con el inconveniente de que a veces la matriz no está definida positiva (propiedad necesaria para el AF).

2.7. Modelos estadísticos para obtener las propiedades psicométricas de un GAI

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.

Como respuesta a este problema, Sinharay y Johnson (2012) propusieron tres categorías de modelos. 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 combina las dos anteriores.

En cuanto a la primera categoría, investigadores como Embretson (1999) y Holling, Bertling y Zeus (2009) utilizaron el Modelo de Test Logístico Lineal o Modelo Logístico Lineal del Rasgo Latente (LLTM, por su nombre en inglés, *Linear Logistic Test Model*, propuesto por

[9] TRR (Teoría de Respuesta al Reactivo) es la notación equivalente a la TRI, que se utiliza en el presente texto.

Fischer en 1973), que es una extensión del modelo de Rasch. El procedimiento emplea las respuestas de un conjunto de ítems para estimar los efectos de las *covariables* (*covariates,* en inglés) de los ítems y así, utilizar estos estimativos para predecir la dificultad de los ítems. Este modelo recurre a una fundamentación cognitiva, ya que las covariables serían habilidades subyacentes a cada ítem. Para ello se requiere un modelo cognitivo que dé soporte a cada contenido y por lo tanto, a los ítems generados. Freund, Hofer y Holling (2008) comunicaron que la correlación entre las dificultades estimadas por Rasch y por el LLTM era de 0.714 y que, además, la diferencia entre las habilidades estimadas por ambos métodos no era significativa. Janssen, Schepers y Peres (2004) y Holling et al (2009) obtuvieron resultados semejantes.

En la segunda categoría se utilizan modelos donde los ítems generados se pueden agrupar en familias con el objetivo de estimar los parámetros del modelo a nivel familia. Una vez estimados los parámetros por familia, se emplean estos en vez de los parámetros por ítem. 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, si bien es sencillo, contiene ciertas limitaciones porque no considera las variaciones dentro de una misma familia. Algunos investigadores argumentaron que el uso de ISM para estimar parámetros puede conducir a pérdida de precisión para estimar la habilidad de los evaluados (Glass y Van der Linden, 2003). Estos autores propusieron el RSM con el objeto de resolver el problema, por medio de la incorporación de una estructura asociada entre los ítems de una misma familia. Este es un modelo jerárquico cuyo primer componente es un modelo simple de la TRI de tres parámetros. El RSM se aplica fundamentalmente para tests adaptativos, donde la habilidad de los evaluados juega un rol primordial.

La tercera categoría combina los modelos LLTM y RSM en otro denominado Modelo Lineal de Clonación de Ítems (LICM), desarrollado por Geerlings, Glas y Van der Linden (2011). Estos investigadores utilizan un modelo de ojiva normal de tres parámetros para especificar la probabilidad de responder correctamente a un ítem j.

2.8. Modelos a seguir para el caso del EXHCOBA-R/MS

El objetivo de esta tesis es la de proponer una metodología para obtener evidencias de validez de estructura interna de exámenes originados por GAI, como lo es el EXHCOBA-R/MS. Por lo cual, expuestos los diferentes requisitos de validación, las características del instrumento a validar y los métodos conocidos, fue necesario tomar una decisión acerca de qué metodología emplear en este caso concreto.

De acuerdo con los *Standards* (AERA, APA y NCME, 1999), se utilizaron las dos teorías recomendadas, TCT y TRI, con el objeto de conocer las propiedades de los ítems. La pregunta que se desprende es qué modelo de la TRI es el apropiado para los análisis. Probablemente, sería conveniente utilizar modelos de 1 o 2 parámetros (contemplan variación en *dificultad*, en el primer caso y agregan *discriminación*, en el segundo) y no de 3 parámetros, ya que los nuevos tipos de reactivos no se prestan a la *adivinación* y no sería necesario agregarla como variable (Sinharay, Johnson y Williamson, 2003).

El tamaño de la muestra también obliga a tomar decisiones. Según Wright y Stone (1998), para 20 ítems se necesitan 200 evaluados si se aplican modelos de un parámetro y aún más para modelos de más parámetros. Cuando se trata de ítems de crédito parcial, la muestra debe ser aún más numerosa, por ejemplo: 25 ítems con 500 examinados, según Samejima (1969). El problema radica en que debe haber un suficiente número de casos en los extremos, esto es necesario cuando los ítems son muy fáciles o muy difíciles. De lo anterior se desprende que los

tamaños de las muestras obtenidas en los diferentes pilotajes fueron insuficientes para las demandas de un modelo de dos parámetros (ver capítulo 3, tabla 3.2). Por lo que se optó por el modelo de Rasch, así como agregar el cálculo de la discriminación, considerada como un índice y no como parámetro. De este modo, se podría obtener una información más fidedigna de los análisis (Yen y Fitzpatrick, 2006, pp. 125 y 126). Otro argumento a favor de un modelo de un parámetro es la parsimonia, puesto que se debe buscar el modelo más simple que explique el comportamiento de los datos. Asimismo, cuando se tienen exámenes que combinan ítems dicotómicos y politómicos de crédito parcial se debe utilizar un modelo que ajuste a los dos tipos de ítems. Por lo tanto, resulta pertinente emplear el modelo dicotómico de Rasch y el modelo de crédito parcial de Masters, conceptualmente compatible con él (Yen y Fitzpatrick, 2006, p. 118).

El EXHCOBA-R es un ejemplo de GAI basado en teoría débil. Este no se cuenta con un modelo de tareas donde se especifique 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. Lo que sí existe, desde su estructura orgánica, es un constructo unificador para cada familia de reactivos, sin las habilidades cognitivas que atraviesan a todos los reactivos de una misma área. Por lo tanto, se creyó conveniente efectuar un estudio del examen, desde ambas teorías (TCT y TRI), a través de la comparación de diferentes muestras y así identificar qué tanto se asemejan.

Como el EXHCOBA-R está estructurado, teóricamente, de acuerdo con los planes de estudio, se consideró apropiado proponer evidencias de su organización a través de agrupación de los ítems en constructos, desde un AFC. Si bien los datos no son continuos, la mayoría provienen de ítems de crédito parcial (estos reactivos generan una escala ordenada de 3 a 6

valores), lo que indica que no se presentaría el problema de los ítems dicotómicos. Esta decisión, de aplicar un AFC, se aúna a la de muchos investigadores en ciencias sociales y de la conducta que tratan datos intervalares con modelos estructurales, puesto que lo que se desea estudiar son tendencias más que precisiones (e.g.: Herrero, 2010; Oliveira, Almeida, Ferrándiz, Ferrando y Prieto, 2009; Phakiti, 2008; Diseth, 2007; Li y Tompkins, 2004; García y Castañeda, 2006; Santos Melgoza y Castañeda Figueiras, 2006 y Sánchez Hernández, Bazán Ramírez y Corral Verdugo, 2009).

Otro aspecto importante es que los desarrolladores de las especificaciones de ítems con las plantillas generadoras de reactivos fueron elaborados con la intención de que dichos ítems fueran isomorfos en sus propiedades psicométricas (en particular, su dificultad). En consecuencia, se consideró necesario analizar estadísticamente, cada familia, desde la TCT y desde la TRI y apoyar su pertenencia al constructo, a través de un AFC.

3

Método

Con el objeto de responder a la pregunta ¿cómo obtener evidencias de validez, basadas en la estructura interna, de exámenes producidos a través de la Generación Automática de Ítems? se implementó una metodología para el caso concreto del EXHCOBA-R/MS. Este apartado presenta la organización estructural de dicho examen (niveles, áreas y número de ítems). También se definen los tipos de muestras, los exámenes generados y los participantes que resolvieron las pruebas. Finalmente se describen los análisis estadísticos de las bases de datos obtenidas mediante las muestras.

3.1. Conformación del EXHCOBA-R/MS

El EXHCOBA-R/MS consta de 120 ítems distribuidos en dos niveles académicos (educación primaria y educación secundaria) y en seis áreas de conocimiento (dos para primaria y cuatro para secundaria). La tabla 3.1 presenta la distribución del número de reactivos por nivel, área, y ámbito, eje temático o asignatura, según el nombre que recibe en cada área.

Tabla 3.1
Estructura del EXHCOBA-R/MS y distribución de los ítems por área del conocimiento

Nivel	Educación primaria	k	Educación secundaria	k
Área	Habilidades del lenguaje		Español	
Ámbito	Estudio	16	Estudio	9
	Literatura	2	Literatura	4
	Participación comunitaria y familiar	2	Participación ciudadana	7
k		20		20
Área	Habilidades matemáticas		Matemáticas	
Eje temático	Sentido numérico y pensamiento algebraico	6	Sentido numérico y pensamiento algebraico	8
	Forma, espacio y medida	7	Forma, espacio y medida	6
	Manejo de la información	7	Manejo de la información	6
k		20		20
Área			Ciencias naturales	
Asignatura			Biología	6
			Física	6
			Química	8
k				20
Área			Ciencias Sociales	
Asignatura			Geografía	6
			Historia I y II	8
			Formación Cívica y Ética	6
k ítems				20

Nota: k = Número de ítems.

3.2. Construcción de las muestras

Con base en esta organización y en la capacidad de producir familias de reactivos similares a través de la GAI se definieron dos niveles de análisis: (a) *nivel examen* para indagar sobre las propiedades de una prueba completa, estudiar el comportamiento de cada área y comparar diferentes versiones del test, y (b) *nivel familia* 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. Estos dos niveles de análisis permitieron fijar dos tipos de muestras (ver figura 3.1).

Muestra tipo 1. Como resultado de la aplicación de un examen completo generado para la educación media superior, con un ítem-hijo por cada uno de los 120 contenidos del EXHCOBA-R/MS. Estas muestras obedecieron al primer nivel de análisis, el examen completo.

Muestra tipo 2. Como producto de la administración de una prueba parcial de un área del examen, con 6 ítems-hijo diferentes por contenido. Las muestras de tipo dos se utilizaron para el segundo nivel de análisis, por familias de reactivos.

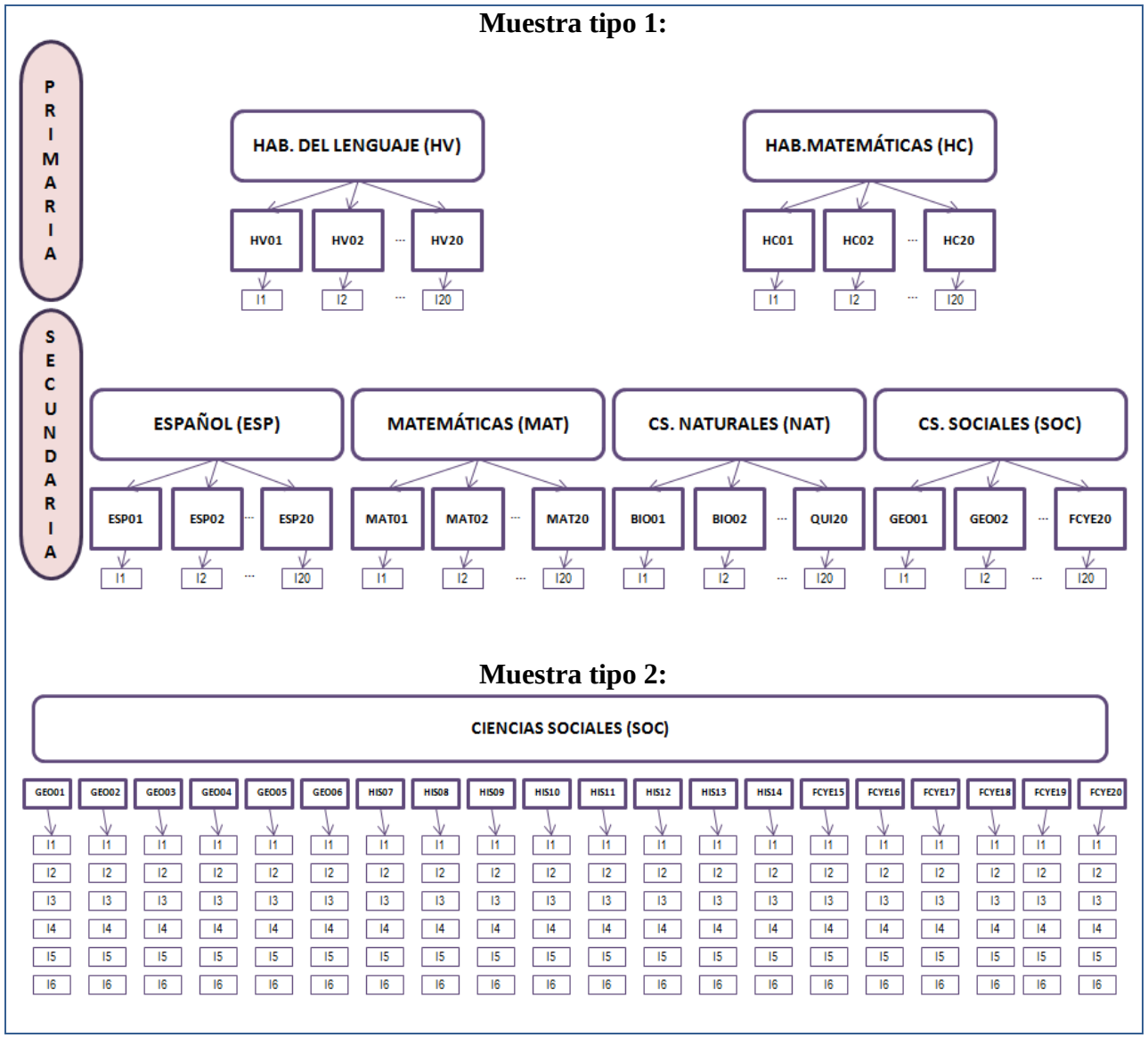


Figura 3.1. Tipos de muestra del EXHCOBA-R/MS. Muestra 1, versión de examen EXHCOBA-R/MS: 6 áreas y 120 ítems. Muestra 2, ejemplo de una muestra de un área (Ciencias Sociales –SOC-): 120 ítems (seis ítems-hijo por cada uno de los 20 contenidos). GEO corresponde a competencias de Geografía, HIS a Historia, y FCYE a Formación Cívica y Ética.

Para la obtención de las muestras del tipo uno se aplicaron dos versiones fijas: **VA** y **VB**, constituidos por 120 ítems, uno por cada contenido del EXHCOBA-R/MS. Cada reactivo se eligió al azar a través del Generador Automático de Ítems del EXHCOBA, con la condición de que los ítems de VB fueran diferentes a los de VA (que no fueran evaluados los mismos ítems-hermano). Para el caso de las muestras del tipo dos se administraron seis exámenes parciales correspondientes a cada una de las seis áreas del EXHCOBA-R/MS. Estas pruebas se denominaron ***HV***, ***HC***, ***ESP***, ***MAT***, ***NAT*** y ***SOC*** para Habilidades de lenguaje, Habilidades matemáticas, Español, Matemáticas, Ciencias Naturales y Ciencias Sociales, respectivamente. Para estos exámenes se seleccionaron seis reactivos distintos (elegidos al azar) de cada uno de los 20 contenidos del área evaluada. Así, se conformaron seis tests de 120 ítems cada uno.

Cada contenido del EXHCOBA-R/MS contó con un modelo de ítem, a excepción de tres casos en donde se elaboraron dos modelos de ítem para un mismo contenido. Como cada modelo tiene el objetivo de generar una familia de reactivos isomorfos, se puede afirmar que en el EXHCOBA-R/MS 117 contenidos fueron evaluados cada uno por una familia y los tres contenidos restantes fueron evaluados, cada uno, por dos familias. Estas situaciones particulares se dieron para HC05, MAT08 y MAT20. En VA y VB se eligieron ítems de familias diferentes para MAT08 y MAT20. HC05 fue evaluado desde una misma familia tanto en VA como en VB. En las muestras HC y MAT se emplearon tres ítems representantes de cada familia para evaluar estos tres contenidos.

3.2.1. Participantes

Los participantes, quienes resolvieron las diferentes combinaciones de reactivos del EXHCOBA-R/MS, fueron alumnos que pertenecían a instituciones educativas con las cuales la UABC tenía

convenio durante 2012, y que participaron voluntariamente. Los estudiantes evaluados pertenecían a los niveles académicos siguientes:

- para las muestras tipo 1: alumnos que aspiraban a ingresar a la Escuela Preparatoria Federal "Lázaro Cárdenas" (PFLC);
- para las muestras tipo 2: alumnos que acababan de cursar los primeros semestres de licenciatura, la mayoría de segundo semestre.

Los tests se resolvieron a modo de pilotaje. Se obtuvo información de 1,849 participantes, cuya distribución aparece en la tabla 3.2. 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 del sistema de ingreso a las instituciones participantes.

Con respecto a los estudiantes universitarios, no se obtuvo información sobre la edad, el sexo, las calificaciones en el pregrado y las carreras a las que pertenecían los alumnos. En cuanto a los estudiantes examinados de la PFLC que pretendían ingresar a la EMS, se estima que el rango de edades estaría entre 14 y 16 años de edad. La población se repartió entre el 59% y 41%, de mujeres y hombres, respectivamente. El promedio de las calificaciones de la educación secundaria de dichos alumnos fue de 9.14 y el rango, de 8 a 10.

La última de las administraciones fue en la PFLC. Se decidió allí efectuar el estudio del examen completo, ya que de este modo, desde la TCT, se podrían comparar ambas versiones (VA y VB) porque pertenecían a la misma población. Además, como se trató de la última aplicación, ya se habían subsanado errores del editor y se podía tener información de casi todos los ítems (117 de los 120).

Tabla 3.2.
Distribución de estudiantes evaluados por examen, según la institución educativa de pertenencia

M	Examen	Institución	N
HV	Habilidades del lenguaje	UGTO	167
HC	Habilidades matemáticas	UACJ (Chihuahua)	100
		CESUES (Hermosillo)	56
ESP	Español	UAQ	220
MAT	Matemáticas	CESUES (Hermosillo)	239
NAT	Ciencias naturales	UACJ (Chihuahua)	60
		CESUES (San Luis Río Colorado)	101
SOC	Ciencias sociales	UAQ	206
A	EXHCOBA-R/MS	PFLC	401
B	EXHCOBA-R/MS	PFLC	299
Total			1,849

Nota: M = Nombre de la muestra. UGTO: Universidad de Guanajuato. UACJ: Universidad Autónoma de Ciudad Juárez. CESUES: Centro de Estudios Superiores del Estado de Sonora. UAQ: Universidad Autónoma de Querétaro. PFLC: Escuela Preparatoria Federal "Lázaro Cárdenas".

Si bien se trató de una participación voluntaria en todos los casos, personal del EXHCOBA ofreció en cada institución sortear una laptop entre los 100 primeros resultados del examen piloto que se aplicara. Además, se exhortó a participar en el pilotaje y resolver los ejercicios con interés antes de llevarlo a cabo, ya que los resultados aportarían prestigio a la institución participante y a los estudiantes de estas. En el caso de la UAQ, esta institución obsequió una unidad de memoria USB de 8 Gb a cada examinado.

Una consideración especial merece el problema del tamaño de la muestra. Según lo plantearon Martínez-Arias et al. (2006), existen reglas prácticas en función de la razón del número de sujetos sobre el número de variables utilizadas en el análisis; los valores van de 10 a 30. Gorsuch (1983) propuso un mínimo de 5 sujetos por variable y no menos de 100 casos. De acuerdo con la exigencia mínima propuesta por Gorsuch se estimaron 100 casos para el análisis de 6, 18 y 20 reactivos, 150 personas para el de 30 ítems y 600 casos para el examen completo de

120 ítems. La última exigencia (600 datos) no se pudo cumplir, mientras que las otras dos presentaron algunas limitaciones específicas en cada situación (véase el capítulo de Análisis de resultados).

3.3. Análisis estadísticos

Para el análisis de los datos se identificaron dos estrategias: (1) la descripción de las propiedades psicométricas de los ítems y la exploración acerca de la unidimensionalidad por medio de la TCT y del modelo de Rasch, y (2) los estudios de agrupación de variables en factores subyacentes a través del AFC.

1. Las propiedades psicométricas de los ítems se obtuvieron desde dos teorías: la TCT y la TRI. Dentro de los numerosos modelos que incluye la TRI se utilizó el de Rasch para ítems dicotómicos y para ítems de respuesta parcial (este último también conocido como modelo de Masters[10]).

Desde la TCT se emplearon cuatro índices para evaluar el comportamiento de los ítems: el índice de dificultad, la correlación punto biserial, la varianza de las dificultades y el coeficiente de consistencia interna (Alpha de Cronbach). Asimismo se obtuvieron las gráficas de distribución de las calificaciones.

Índice de dificultad de un ítem. Se define como la razón entre el número de aciertos y el número total de respuestas a un reactivo. Este número se encuentra en un rango de 0 a 1. Una dificultad igual a 0 indica que ningún alumno pudo contestar correctamente el ítem; por el contrario, una dificultad igual a 1 señala que todas las respuestas fueron correctas. En otras palabras, índices mayores implican dificultades menores. Una dificultad media estaría en 0.5,

[10] A modo de simplificación, se utilizará el modelo de Rasch para aludir a éste y a su asociado, el modelo de Masters.

donde la mitad de los examinados pudo contestar acertadamente. A esta razón se la suele denotar con una *p* y su fórmula para ítems dicotómicos es la siguiente (Abad, Garrido, Olea y Ponsoda, 2006):

$$p_i = \frac{A_i}{N_i}$$

p_i = índice de dificultad del ítem *i*

A_i = número de aciertos en el ítem *i*

N_i = número de aciertos más número de errores en el ítem *i*

Coeficiente de correlación punto biserial o correlación ítem total. Es la correlación del producto de momento de Pearson entre las respuestas calificadas (dicotómicas o politómicas) y el total de las calificaciones (los casos perdidos se omiten por pares). Esta correlación se calcula para cada ítem con el fin de controlar si éste es consistente con el comportamiento promedio de los demás. Una correlación pequeña muestra que el ítem no mide el mismo constructo que los reactivos incluidos en el test. Un coeficiente menor que 0.2 o 0.3 indica que el reactivo no funciona muy bien con respecto a la escala completa; por lo tanto, debe ser revisado, modificado o eliminado (Field, 2005). La notación utilizada para este coeficiente es generalmente r_{pb} y su fórmula para reactivos dicotómicos (Magnusson, 1967):

$$r_{pb_i} = \frac{M_p - M_q}{S_n}\sqrt{pq}$$

Donde:

r_{pb_i} = correlación punto biserial del ítem *i*

M_p = media de la variable continua para los casos con valor 1 en la variable dicotómica.

M_q = media de la variable continua para los casos con valor 0 en la variable dicotómica.

s_n = desviación Típica obtenida con n casos.

n_1 = número de casos con valor 1 en la variable dicotómica.

p = proporción de casos con valor 1.

q = proporción de casos con valor 0 (q = 1 – p).

Varianza de las dificultades. La varianza es una medida de la dispersión de los datos. En el caso concreto del EXHCOBA-R/MS se calculó la varianza de las dificultades de los seis ítems seleccionados por contenido. El objeto fue explorar cuánto se alejaban los ítems en dificultad con respecto a su media. La varianza es siempre no negativa y, en estas circunstancias, se pretende que los ítems-hermano sean lo más similares posibles. La aspiración es que las varianzas tiendan a 0. El símbolo usualmente utilizado para denotar esta medida es s^2 y su fórmula tomada de Magnusson (1967) y adaptada a los datos es:

$$s^2 = \frac{\sum_{i=1}^{k}(p_i - m)^2}{k}$$

Donde:

s^2 = varianza de las dificultades

k = número de ítems

p_i = índice de dificultad del ítem i

m = dificultad media de los ítems considerados

Coeficiente de consistencia interna (Alpha de Cronbach). La confiabilidad es la capacidad de una prueba para demostrar estabilidad y consistencia en sus resultados. Una manera de realizar este análisis es administrar formas diferentes de una prueba en un intervalo de tiempo a un mismo grupo. Otro camino es estudiar su consistencia interna a través de coeficientes que

den evidencias del grado con el cual los ítems de una prueba funcionan de una misma manera; depende de la consistencia del desempeño de un individuo de un reactivo a otro y se basa en la desviación estándar de los reactivos por separado (Williams y Monge, 2001). El nivel de confiabilidad de una escala es fundamental. No se pueden establecer predicciones ni inferencias a partir de puntuaciones de baja confiabilidad, ya que esta reduce las correlaciones entre medidas y, por lo tanto, afecta a la calidad del test (Martínez Arias et al., 2006).

Un cálculo muy utilizado es el del Alpha de Cronbach. Este coeficiente puede variar en un rango de -1 a +1, aunque los valores negativos no son significativos para la confiabilidad. Cuanto más se aproxime este número a su valor máximo, 1, mejor es el comportamiento de la escala. Según el NCME (1999), valores entre 0.90 y 1 indican una alta calidad de una prueba. Valores superiores a 0.50 implican una confiabilidad adecuada para evaluar individuos sólo si se promedia con varias evaluaciones cuyas confiabilidades sean similares, como sucedió con las 20 familias analizadas por área del EXHCOBA-R/MS.

La notación característica del Alpha de Cronbach es α, y su fórmula (Cronbach, 1951):

$$\alpha = \frac{k}{k-1}\left[1 - \frac{\sum_{i=1}^{k} s_i^2}{s^2}\right]$$

Donde:

α = coeficiente Alpha de Cronbach

k = número de ítems de la prueba

s_i^2 = varianza del ítem *i*

s^2 = varianza de la prueba

En el caso de la TRI se aplicó el modelo de Rasch en su variante para ítems de crédito parcial, pues permite que cada reactivo tenga su propio patrón de respuesta. Para ello, se implementó la instrucción ISGROUP = 0, la cual especifica que cada ítem tiene su propia estructura de respuesta, ya sea dicotómica o politómica (con la cantidad de categorías que especifique cada ítem).

Cabe recordar que los ítems de crédito parcial admiten niveles ordenados de ejecución. Así se otorga parte de la calificación total para cada ítem en función de los aciertos parciales. El motivo de este tipo de reactivos es la expectativa de obtener estimaciones más precisas que la simplificación acierto/error de las preguntas dicotómicas. Bajo este formato, el número de pasos en que se divide cada ítem puede variar; no necesita ser el mismo para todos los reactivos del test.

En el EXHCOBA-R no existe una jerarquía cognitiva con respecto a los pasos de la resolución de un ítem, ya que cada crédito se otorga por acumulación de respuestas correctas. A mayor cantidad de aciertos se tiene un mayor dominio de la habilidad o competencia evaluada. Por lo tanto, el análisis estadístico se centró en el comportamiento del ítem en su totalidad y no en cada categoría o paso. Evaluar cada paso, implicaría estudiar qué tan provechoso es el número de créditos parciales y no qué tanto se ha adquirido la habilidad evaluada en cada categoría. Para el análisis psicométrico de los elementos que conforman los ítems politómicos se aplicó el modelamiento de Rasch de datos dicotómicos. Esto último con el objeto de estudiar la calidad de cada componente.

La calibración de los ítems, según el modelo de Rasch, se efectuó a través de la información proporcionada por: la medida del ítem, los índices de ajuste *infit* y *outfit*, el índice de correlación punto medida, el índice de discriminación del ítem y el mapa de Wright.

Medida del ítem. Para calibrar un reactivo (decidir acerca de su grado de dificultad) se calcula la medida del ítem. La unidad utilizada es el *lógito*. La medida se calcula como el logaritmo natural del cociente de fallas entre éxitos, así: $D = ln\frac{q}{p}$, (p es el índice de dificultad de la TCT, conocido como el cociente del número de respuestas correctas que tuvo el ítem entre el número total de individuos que lo respondieron; además, $q = 1 - p$). Estos valores pueden ser cualquier número real, pero lo común es considerar el intervalo [-3; 3]. Un reactivo con una dificultad de 0.5 tiene una calibración de 0 lógitos, un ítem más fácil se calibrará con valores menores que 0 y uno más difícil con números mayores que 0. Cuanto más difícil es un reactivo más positivo es el número de lógitos que lo calibre; cuanto más fácil es un reactivo más pequeño será el número de lógitos negativos. Cada división corresponde a una dosificación de dificultad en los reactivos. Por lo tanto, se requiere de un conjunto de reactivos que expresen una extensa gama de dificultad para poder medir con precisión y estimar qué tanto sabe o ignora una persona (Tristán-López, 2001).

Índices de ajuste infit y outfit. Estos indican con cuánta precisión o predicción los datos se ajustan al modelo. Tienen también el propósito de identificar aquellos datos que responden o no a las especificaciones del modelo de Rasch. *Infit* significa ajuste sensible a datos erróneos dentro de la distribución estadística; detecta desajustes en las desviaciones cerca de la zona de medición del ítem; estos problemas suelen ser difíciles de detectar y de corregir. *Infit* es el promedio *ponderado* de las desviaciones (o las diferencias) cuadráticas estandarizadas entre el desempeño observado y el esperado. *Outfit* es un estadístico de ajuste sensible a los datos erróneos en las colas de una distribución estadística de un conjunto de datos; identifica incoherencias en las desviaciones lejos de la zona de medición del ítem. *Outfit* es el promedio de las desviaciones (o

diferencias) cuadráticas estandarizadas entre el desempeño observado y el esperado (Tristán-López, glosario).

Existen dos maneras de reportar estos índices: las medias cuadráticas y las estandarizadas. Las medias cuadráticas son siempre positivas, deben acercarse a 1 para indicar buen ajuste. En el caso de los ajustes estandarizados, los valores deben ser cercanos a 0. Si bien no existen reglas fijas acerca de la interpretación de los diferentes estadísticos, Linacre (2002) estableció los criterios que se muestran en la tabla 3.3. Los límites de ajuste que el Comité Técnico del EXHCOBA propone como aceptables son más estrictos: [0.8 – 1.3] para las medias cuadráticas y [-2, 2] para los valores estandarizados.

Estos índices de ajuste son de gran importancia dentro del modelo. Por una parte, las medias cuadráticas menores a .80 indican determinismo o predictibilidad, no captan suficiente *señal* de la variable latente (constructo). Por otra parte, los ítems con valores superiores a 1.30 muestran demasiada aleatoriedad, que se interpreta como ruido o interferencia fuera de la variable latente (constructo). Estos criterios se aplican tanto para el *infit* como para el *outfit*. Si los valores *infit* o *outfit* están dentro del intervalo de tolerancia se dice que el ítem ajusta al modelo (González-Montesinos, 2008).

Linacre (2010) sugirió que se utilicen las medidas de ajuste estandarizadas cuando los medias cuadráticas escapen del rango de aceptabilidad y cuando las muestras sean pequeñas o los tests sean cortos. Si los valores estandarizados se encuentran en el rango [-2; 2] el ítem es métricamente productivo; si ocurre lo contrario, entonces, el reactivo desajusta con respecto al constructo definido.

Tabla 3.3.
Rangos de valores para los índices de ajuste infit y outfit y sus implicaciones para el examen.

Infit – outfit	Implicaciones para el examen
Medias cuadradas	
> 2	Distorsiona o degrada el sistema de evaluación. Quizás provocado por una o dos observaciones.
(1.5; 2)	Construcción improductiva de la medición, pero que no la degrada.
[0.5; 1.5]	Productiva para la medición
< 0.5	Poco productiva para la medición, pero no la degrada. Puede conducir altas confiabilidades de manera errónea.
Estandarizada	
≥ 3	Datos muy inesperados. Con muestras de gran tamaño, el desajuste sustantivo debe ser pequeño.
(2; 3)	Datos notablemente impredecibles.
[-2; 2]	Los datos son razonablemente predecibles.
< -2	Los datos son demasiado predecibles. Otras dimensiones deben de estar restringiendo los patrones de respuesta.

Nota: Información recuperada de Linacre (2002)

Índice de correlación punto medida. Es una modificación de la correlación punto biserial. Este coeficiente mide el grado de asociación entre el puntaje particular observado para un reactivo en particular y el puntaje total observado en el examen (González-Montesinos, 2008). Según Linacre (2010), este estadístico resulta mejor que la correlación punto biserial cuando hay datos perdidos.

El coeficiente de correlación punto medida se encuentra en un rango [-1; 1]. Un valor negativo indica que el ítem contradice la variable latente que lo define; puede ser un indicativo de que el reactivo fue mal calificado porque se le otorgaron claves de respuesta erróneas. Linacre (2010) sugirió investigar las correlaciones negativas antes que los estadísticos de ajuste de Rasch. Por el contrario, correlaciones positivas indican un mayor vínculo con la escala a la cual pertenece el ítem. Si bien no existe un criterio único acerca de cuál es el valor mínimo aceptable para identificar los reactivos que funcionan en sintonía con el constructo que representan, se pueden considerar los mismos criterios que para la correlación punto biserial.

Índice de discriminación del ítem. Indica la capacidad de un ítem para identificar a los estudiantes según su habilidad. En el modelo de Rasch, durante la fase de estimación, todas las discriminaciones son equivalentes a 1 a fin de ajustarse al modelo. Sin embargo, empíricamente, nunca son exactamente iguales. Por lo tanto, se puede reportar una estimación de estas discriminaciones post-hoc (como un tipo de estadístico de ajuste). En consecuencia, se trata de un estadístico descriptivo y es una indicación del grado en que un ítem ajusta al modelo de Rasch, no es un parámetro. Según Linacre (2010), el índice de discriminación, que se calcula a través del programa Winsteps, es una aproximación del parámetro de discriminación que se obtiene con modelos TRI de 2 parámetros en programas como BILOG (para ítems dicotómicos) o PARSCALE (para ítems de crédito parcial).

El índice de discriminación esperado es 1. Un valor mayor que 1 se interpreta como una discriminación mayor de la esperada; lo contrario ocurre con un valor menor que 1. Generalmente, valores altos de *infit* y *outfit* se corresponden con discriminaciones bajas y viceversa, *infit* y *outfit* pequeños se asocian con altas discriminaciones. Masters (1988) propuso un rango de discriminación apropiado, que va de 0.5 a 2, o incluso podría ir hasta infinito, aunque las medias cuadradas (*infit-outfit*) se vuelven muy pequeñas, lo cual significa que el ítem funcionaría mal de algún modo.

Desde una perspectiva Rasch, los ítems que sobre discriminan tienden a comportarse como llaves (puntos de corte) no como elementos de medición; mientras que los reactivos que sub discriminan no miden ni estratifican. El primer tipo de ítems no se considera negativo en el caso de exámenes normativos (como el EXHCOBA), la desventaja es que estos reactivos no proveen información sobre quienes demuestran pobre desempeño o, incluso, de los desempeños relativos de quienes mostraron buenos resultados.

Mapa de Wright. Esta gráfica provee una imagen con respecto a qué tan bien mide a sus evaluados un examen. En el mapa se ubican, en una misma escala de lógitos, los examinados según su habilidad y los ítems según su dificultad. De este modo se pueden comparar personas con ítems. Un ejemplo de un mapa se presenta en la figura 3.2. Este se organiza en dos histogramas verticales; la distribución de las personas se muestra a la izquierda y la distribución de los ítems a la derecha. Los examinados más hábiles se ubican en la parte superior y los menos hábiles en el extremo inferior. Los ítems se sitúan de más difíciles a más fáciles, de arriba hacia abajo. La M indica la media (de los examinados o de los ítems), S es una desviación estándar y T, dos desviaciones. Para los análisis del EXHCOBA la media de las distribuciones de los ítems se centró en 0 (cero) y la desviación estándar en 1. Cada "#" representa un número fijo de evaluados (18 en este caso) y cada "." es una cantidad menor de la prefijada (de 1 a 17 en este caso). Cada R indica un reactivo acompañado del número correspondiente. Cuando los evaluados y los ítems quedan enfrentados en el gráfico se interpreta como que ambos son comparables, que existen ítems que apuntan a medir las habilidades de la muestra evaluada.

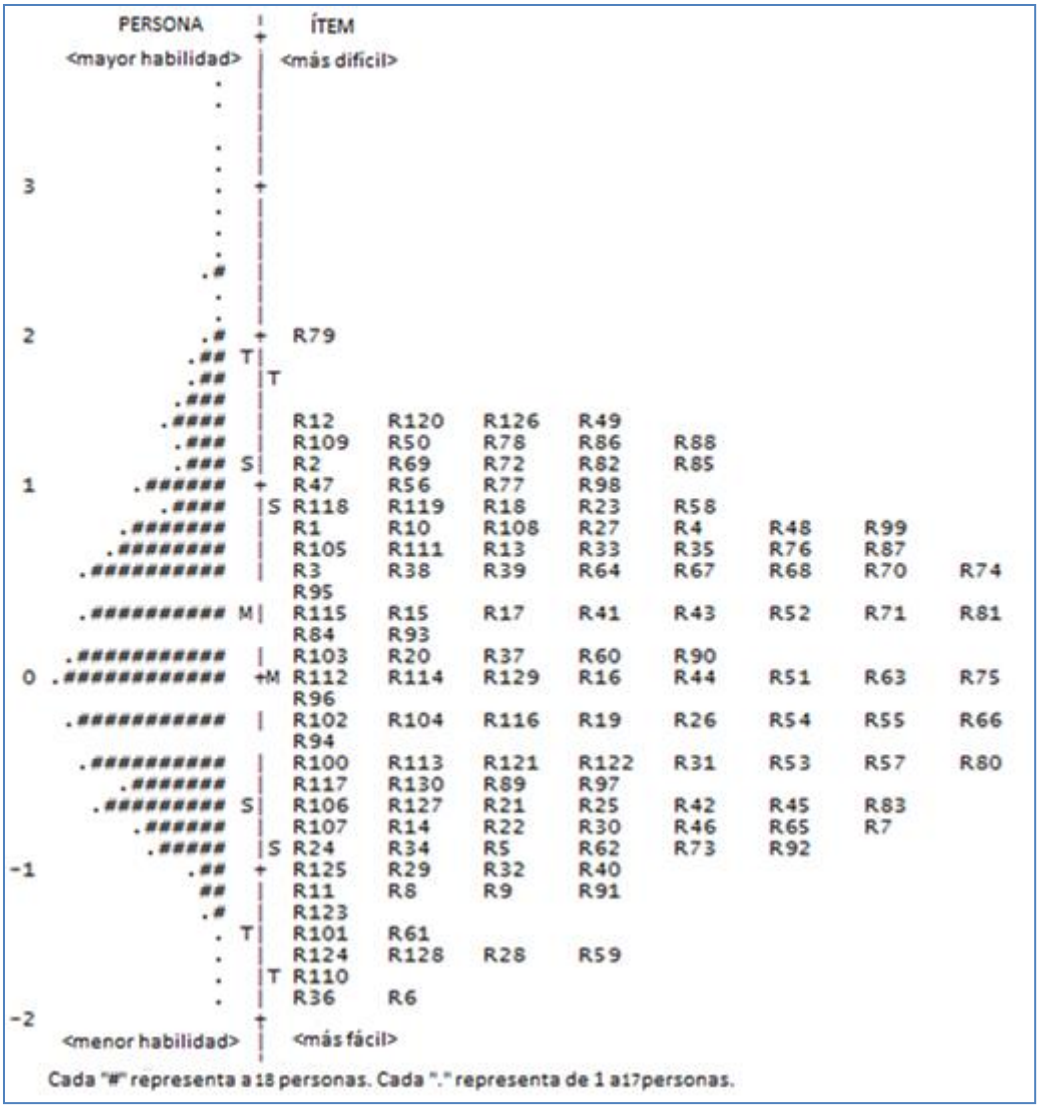


Figura 3.2. Mapa de Wright para una versión del EXHCOBA/MS aplicado a 6000 aspirantes a ingresar a la ES

2. El AFC se aplicó para evaluar la unidimensionalidad por área del examen y por familia de ítems. Se establecieron covarianzas entre los errores de las variables. Posteriormente se eliminaron a través del test de Wald aquellas covarianzas que no mejoraban los ajustes y se volvieron a ejecutar los análisis.

A cada variable se le asocia un número o carga factorial en el AFC. Este valor indica la relación cuantitativa entre la variable y el factor (también denominado como habilidad o rasgo

latente). Cuanto más lejano es el valor de cero se puede afirmar que la habilidad está definida por esa variable. En el caso de cargas estandarizadas, estas no superan a 1 (Gorsuch, 1983).

Hair et al. (1992) propusieron una relación entre las muestras y las cargas factoriales para decidir sobre la significancia de dichas cargas relacionadas con el tamaño de la muestra (ver tabla 3.4). Lógicamente, este juicio también está influenciado por el número de variables a analizar; cuando el número de variables aumenta, el nivel de aceptabilidad para que las cargas sean significativas, disminuye.

Tabla 3.4
Identificación de cargas factoriales significativas, según el tamaño de la muestra

Tamaño de la muestra	Carga factorial	
	de 0.05*	de 0.01*
100	0.19	0.26
200	0.14	0.18
300	0.11	0.15

Nota: basado en Hair et al, 1992. (*) Nivel de significancia.

Una vez puesto a prueba el modelo, y obtenidas las cargas factoriales, se evaluó la bondad de ajuste del modelo con respecto a los resultados empíricos. Para ello, se consideraron diferentes índices. Martínez-Arias *et al.* (2006) reconocieron que no existe acuerdo sobre cuáles son los mejores índices para determinar un buen ajuste. Una de las pruebas más utilizadas es χ^2, que sirve para estimar la discrepancia entre las dos matrices comparadas (teórica y de los datos). Se espera que χ^2 resulte lo más pequeño posible, de modo que la diferencia no sea estadísticamente significativa; se suele pedir que p sea mayor que 0.05 o que 0.01. Cabe aclarar que χ^2 es muy sensible al tamaño de la muestra, ya que cuando estas son grandes prácticamente todos los modelos se vuelven inaceptables (Bentler y Bonnett, 1980).

Un índice muy robusto es el Error Medio Cuadrático de Aproximación (RMSEA), propuesto por Browne y Cudeck (1989). Este índice es absoluto, no comparativo, ya que mide

las diferencias reales entre los elementos de la matriz de hipótesis y la matriz derivada de la muestra empírica (Steiger, 1990). Para un buen ajuste, el RMSEA debe ser menor que 0.05 y para un ajuste aceptable debe estar entre 0.05 y 0.08.

Otros elementos que se reportan son los índices de ajuste comparativo. Estos índices contrastan los datos con un modelo hipotético de base, que es el más simple que se puede aplicar a los datos (sin factores comunes y sin correlaciones entre los factores). Son valores descriptivos que se interpretan de acuerdo a las reglas prácticas sugeridas por sus autores. Entre ellos destacan el Índice de Ajuste Normalizado (NFI) con su modificación NNFI y el Índice Comparativo de Ajuste (CFI). Los índices normalizados se encuentran dentro de un margen de 0 a 1; en caso contrario, pueden superar a 1. Un buen ajuste debe ser mayor que 0.95, mientras que un ajuste aceptable debe sobrepasar 0.90 (Bentler, 1990; Bentler y Bonnet, 1980).

Otro recurso importante es la parsimonia. El objetivo es penalizar aquellos modelos que tienen demasiadas variables. Los índices que se utilizan se denominan *Parsimonian Fit Index* (entre ellos, el Índice de Ajuste Normalizado de Parsimonia –PNFI–) y son valores que relacionan los comparativos con los grados de libertad del modelo.

Para esta tesis se reportaron cuatro índices de ajuste que evaluaron la calidad del modelo, todos de buena ejecución en muestras pequeñas: χ^2, RMSEA (de ajuste absoluto), NNFI y el CFI (estos dos últimos, de ajuste incremental). Para el caso de la parsimonia se reportaron únicamente los grados de libertad, ya que los modelos propuestos son, en su mayoría, unidimensionales o, a lo sumo, de dos factores.

Una situación particular se presentó para el área de Habilidades matemáticas, con 16 ítems dicotómicos, tres politómicos de tres categorías y uno de dos categorías. Por lo tanto, se convirtieron los cuatro reactivos de crédito parcial a dicotómicos, se calculó la matriz de

correlaciones tetracóricas y luego se efectuó un AFE para constatar si los resultados obtenidos en el AFC con los datos originales coincidía con los del AFE.

A modo de resumen, y con base en lo que la literatura propone y en las normas utilizadas en el EXHCOBA tradicional, en la tabla 3.5 se establecen los criterios asumidos para evaluar la calidad de los ítems y del examen, en general.

Es necesario también especificar que para la TCT se utilizaron los programas estadísticos Microsoft Excel 2007 y SPSS 17.0 (SPSS, 2008). Los estudios Rasch se realizaron con Winsteps, versión 3.70.0.2 (Linacre, 2009). Finalmente, para los cálculos de AFC se empleó el paquete EQS 6.1 (Bentler, 2006).

Tabla 3.5.
Criterios asumidos para los análisis estadísticos de los ítems de las muestras del EXHCOBA-R/MS

	Estadísticos	Número de variables	Criterio	
			Aceptable	Bueno
TCT				
	Correlación punto biserial		> 0.2	
	varianza dificultad por familia		$\rightarrow 0$	
	α	6	> 0.5	
		20	> 0.7	
		120	> 0.9	
TRI				
	Correlación punto medida	6	> 0.3	
		20	> 0.2	
		120	> 0.2	
	Infit-Outfit		> 0.8 y < 1.3	
	Discriminación		> 0.8	
AFC				
	Carga factorial		> 0.20	> 0.30
	χ^2		> 0.05	> 0.01
	NNFI		> 0.90	> 0.95
	CFI		> 0.90	> 0.95
	RMSEA		< 0.08	< 0.05

3.3.1. Niveles de análisis: de examen y de familia de ítems

Como se mencionó al inicio de este capítulo, el análisis estadístico se estructuró en dos niveles: nivel de examen y nivel de familia de ítems. La descripción de estos dos niveles de análisis aparece en los párrafos siguientes.

Nivel de examen. Este nivel se organizó en dos categorías, examen completo (R/MS) y cada área del examen.

Examen completo (R/MS). Estudio de la prueba completa de nivel básico (ver figura 3.3). Se efectuaron tres tipos de análisis estadísticos (TCT, Rasch y AFC) para cada versión, VA y VB con la comparación entre ambas versiones. Esto con el objeto de analizar las propiedades psicométricas de dos exámenes que se generan por la aleatorización del GAI.

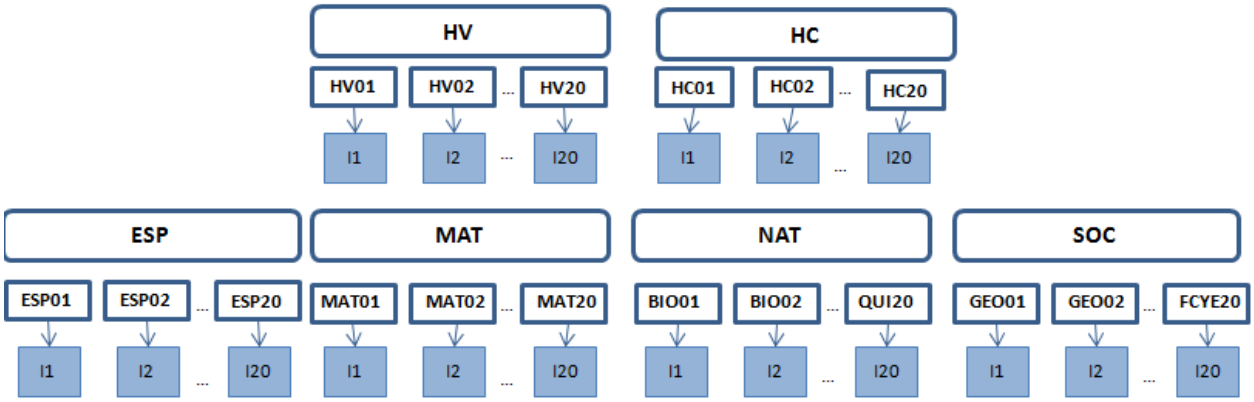


Figura 3.3. Estructura del EXHCOBA-R/MS, 120 ítems en total, con un ítem por cada contenido del examen

Cada área del examen. Análisis de las versiones VA y VB por áreas (ver figura 3.4). Se trata de un retrato estadístico de cada área (estadísticos de TCT y TRI) y de su estructura, de acuerdo con la organización teórica del EXHCOBA-R. Se compararon las seis áreas correspondientes de las dos versiones. Aquí se identificó en qué grado la estructura conceptual del examen concuerda con su estructura empírica.

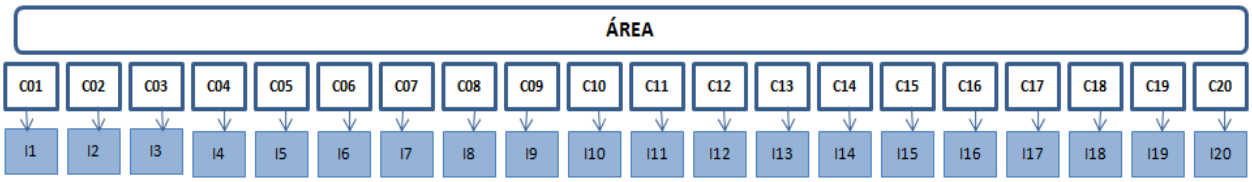


Figura 3.4. Estructura de un área del EXHCOBA-R/MS, 20 contenidos y un ítem por cada contenido

Nivel de familia de ítems. Este nivel se estructuró en tres categorías.

Muestra completa del área. Estudio de las muestras completas realizadas por áreas (HV, HC, ESP, MAT, NAT y SOC) como se muestra en la figura 3.5. Esto con el objeto de mostrar el comportamiento de un examen con ítems exclusivos de un área.

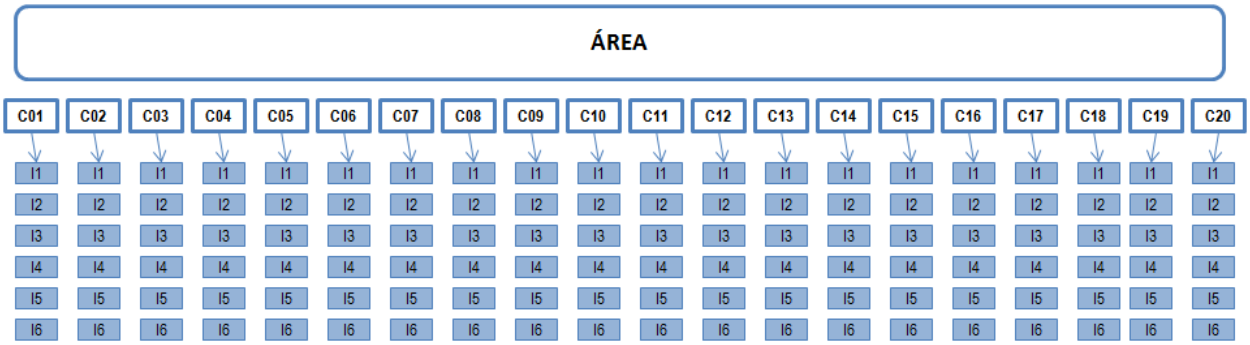


Figura 3.5. Estructura de una muestra por área, 20 contenidos con 6 ítems por cada uno.

Ítems de cada competencia o contenido. Análisis de los ítems-hijo de una misma competencia (para cada una de las 20 competencias de HV, HC, ESP, MAT, NAT y SOC) (véase la figura 3.6). El objeto fue estudiar los diferentes ítems que miden un contenido determinado para decidir si poseen propiedades psicométricas similares y si se agrupan en el constructo que los define.

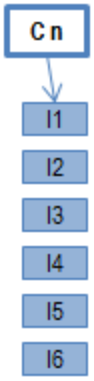


Figura 3.6. Familia con 6 ítems que evalúan el contenido *n*.

Elementos que conforman los ítems. Estudio de los componentes de los reactivos de crédito parcial. Para ello se utilizaron los elementos de una familia de cada muestra de HV, HC, ESP, MAT, NAT y SOC (ver figura 3.7). Este análisis examinó las propiedades psicométricas de los componentes de los reactivos de crédito parcial de un mismo contenido para identificar qué tan homogéneos son dichos componentes.

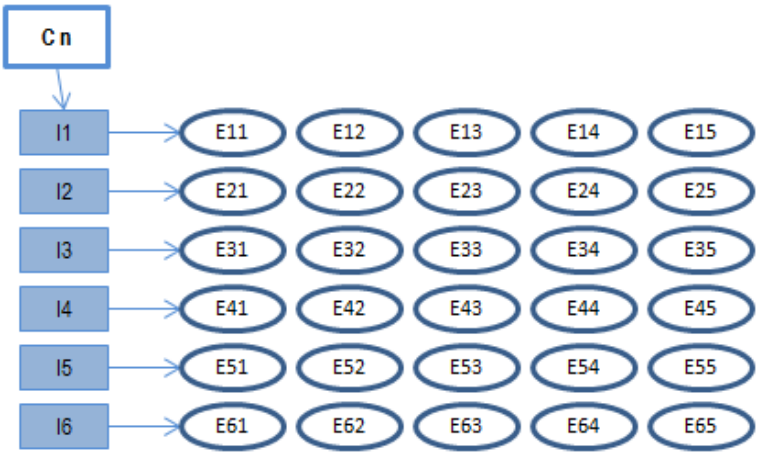


Figura 3.7. Esquema de los elementos de una familia de 6 ítems de crédito parcial.

Especialistas en sistemas informáticos fueron los encargados de administrar los exámenes y generaron bases de datos con los resultados de las aplicaciones. Estas bases originales venían en archivos diferentes, según el tipo de reactivos al que pertenecían y el lugar donde fueron

aplicados los exámenes. La mayoría contenía las respuestas de los estudiantes, en texto y en códigos. Algunas bases incluían los códigos; pero existía una forma de decodificar la información y así, conocer la respuesta dada por cada evaluado. En total se recibieron 53 bases de datos aproximadamente.

Algunos tipos de reactivos ya venían calificados por el sistema de cómputo. En todos los casos se efectuó una revisión exhaustiva de respuestas correctas e incorrectas cotejando con los exámenes (los cuales se conservaron en el sistema) y con las respuestas indicadas en las especificaciones de reactivos.

La calificación de ítems dicotómicos se estableció como "0": incorrecto y "1": correcto. Para los ítems de crédito parcial se computó cada acierto en partes iguales y así, conformar el total de 1 punto. Por ejemplo, si el reactivo solicitaba ubicar cinco fracciones en la recta numérica, por cada fracción ubicada correctamente se otorgó 0.20 puntos. Para los reactivos de siete o más respuestas se fraccionó la calificación en cinco puntajes parciales (0, 0.25, 0.5, 0.75 y 1 punto). En el anexo B se detalla la calificación de cada uno de los 120 ítems.

Finalmente se depuraron y organizaron dichas bases y con ellas, se efectuaron los análisis previamente descritos. Los problemas que se suscitaron en los procesos de recuperación de la información para las aplicaciones pertinentes están detallados en el capítulo de Análisis de resultados.

Después de comparar las dos versiones del instrumento y los ítems de las familias de reactivos de cada uno de los 120 contenidos del examen se concluyó con un informe donde se mencionan las propiedades psicométricas de los ítems y el grado de validez basado en evidencias de estructura interna del EXHCOBA-R/MS, así como sugerencias para mejorar el instrumento en los casos pertinentes. Esta información también aparece en el capítulo de resultados.

4

Resultados

En este capítulo se presentan los resultados de los análisis psicométricos efectuados con las bases de datos de los distintos pilotajes del EXHCOBA-R/MS. Cabe recordar que en el método propuesto se establecieron dos niveles de análisis: nivel examen y nivel familia de reactivos. En el primer nivel se analizó el test tal cual se aplica para el ingreso a la EMS y en el segundo nivel se estudiaron las propiedades psicométricas de los ítems-hijo de una misma familia. De acuerdo con esta organización, a continuación se describen los resultados de ambos niveles con las categorías descritas en el capítulo tres del Método.

4.1. Nivel de examen: análisis estadísticos de las muestras VA y VB

En primer lugar, se muestran los resultados de la prueba completa, de acuerdo con la estructura para el ingreso a la EMS, en las dos versiones piloteadas (VA y VB). Esta información se utilizó para comparar las dos versiones del EXHCOBA-R/MS, producidas aleatoriamente por el generador de reactivos e inferir qué tan parecidos pueden ser los tests producidos por la GAI. En segundo lugar, se insertan los resultados de los análisis estadísticos de VA y VB, por área. Estos sirvieron para cotejar los comportamientos de cada área en dos exámenes diferentes y para analizar la calidad de los reactivos.

Debido a que se utilizan los mismos procedimientos para cada área evaluada, se decidió incluir toda la información relacionada con Habilidades matemáticas (HC), como representativa de los análisis. De las áreas restantes se incluye un resumen de resultados, así como algunos comentarios pertinentes. Para avalar las interpretaciones, todas las figuras y las tablas se incluyeron en el Anexo D.

4.1.1. De los exámenes completos

A modo de recordatorio, la Figura 4.1 presenta la organización de las muestras VA y VB, ambas evalúan todos los contenidos seleccionados en el EXHCOBA-R/MS correspondientes a la educación básica (primaria y secundaria). En consecuencia, son 120 ítems-hijo donde cada uno representa a un contenido diferente del examen.

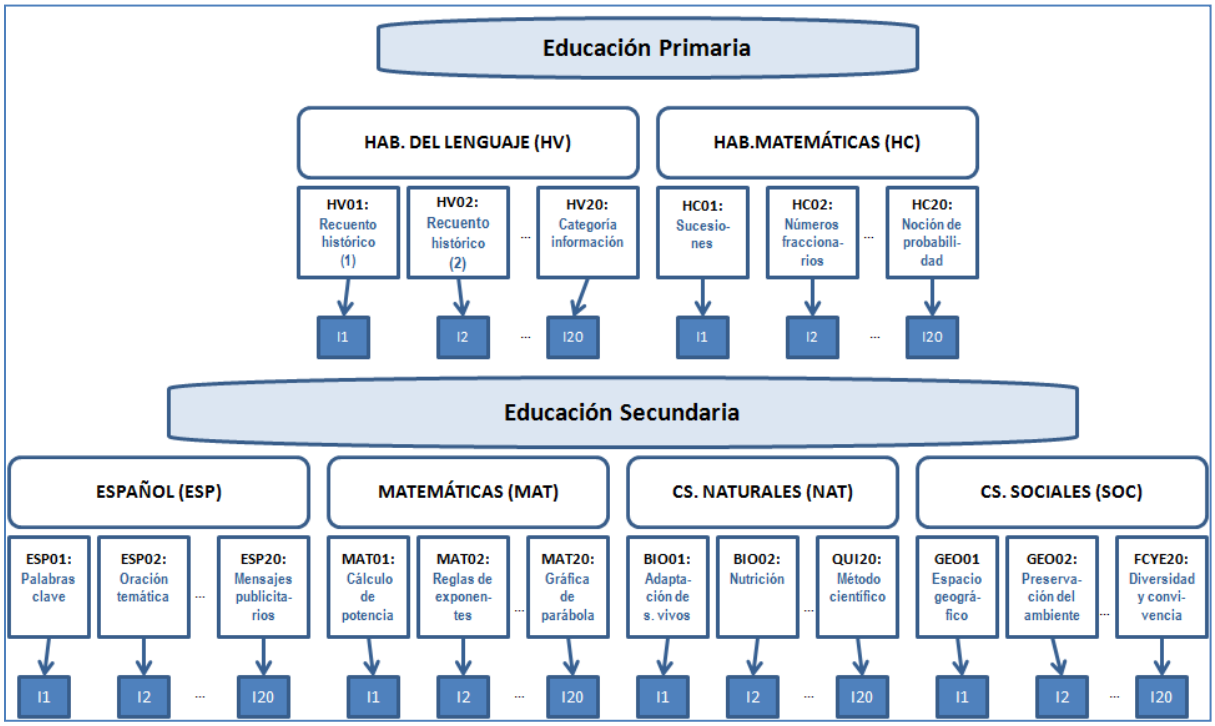


Figura 4.1. Esquema de distribución de reactivos del EXHCOBA-R/MS, versiones A y B.

En la práctica, se obtuvo información de 117 reactivos debido a que, por razones técnicas, hubo que descartar tres, estos son: HV19, MAT14 e HIS06. Los reactivos HV19 e HIS06 se eliminaron de la muestra por serios problemas de diseño que, al momento de los pilotajes, no habían sido subsanados. En cuanto a los ítems de MAT14, si bien fueron aplicados, no se pudieron decodificar los datos; por lo tanto, tampoco se incluyó este contenido en los análisis.

Los examinados fueron aspirantes a ingresar a la escuela Preparatoria Federal "Lázaro Cárdenas" (PFLC), 401 personas resolvieron VA y 301, VB. Si bien las versiones se aplicaron a muestras diferentes, ambas pertenecían a la misma población y la participación en una o en otra de las administraciones fue al azar. Por lo expuesto, se consideró adecuado comparar los resultados de ambas pruebas.

Para el análisis desde la TCT, se incluyeron solamente los datos donde no había casos perdidos. Las medias de las dificultades fueron $\bar{x}_A$ = 60.92 (p = 0.52) y $\bar{x}_B$ = 58.12 (p = 0.50). Los índices de dificultad (p) de cada ítem se encuentran en el Anexo C. Las medias de las dificultades por área fueron parecidas en los dos exámenes; la más distante fue Habilidades del lenguaje (HV) con una diferencia de 5 centésimas. Los índices de confiabilidad fueron aceptables y similares en ambos tests (aproximadamente 0.90). Se registraron confiabilidades aceptables y semejantes para SOC, HC y MAT; más bajas para HV y NAT (especialmente en VB). Un comportamiento extraño se observó en ESP, mientras que en VB fue buena, en VA no llegó a 0.60 (ver tabla 4.1).

Tabla 4.1
Dificultad media, índice de correlación punto biserial y confiabilidad del EXHCOBA-R/MS para VA y VB, examen completo y por áreas

Área	k ítems	VA			VB		
		Dificultad	Pbis	α	Dificultad	Pbis	α
HV	19	0.68	0.24	0.633	0.63	0.21	**0.547**
HC	20	0.38	0.34	0.784	0.38	0.35	0.788
ESP	20	0.71	0.21	**0.587**	0.69	0.30	0.706
MAT	19	0.26	0.26	0.655	0.24	0.27	0.691
NAT	20	0.48	0.22	0.612	0.47	**0.16**	**0.502**
SOC	19	0.47	0.48	0.869	0.48	0.49	0.877
EXHCOBA-R/MS	117	0.52	0.26	0.902	0.50	0.25	0.897

Nota: Pbis = índice de correlación punto biserial, α = Alpha de Cronbach, HV = Habilidades del lenguaje, HC = Habilidades matemáticas, ESP = Español, MAT = Matemáticas, NAT = Ciencias naturales, SOC = Ciencias sociales

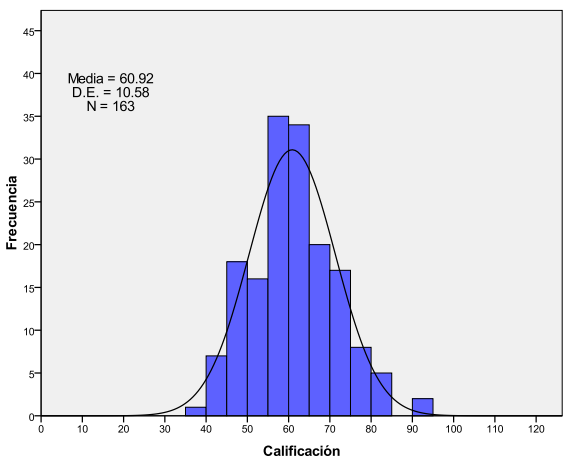
Las correlaciones punto biserial de cada ítem se encuentran también en el Anexo C. En la tabla 4.2 se muestra un resumen. Los reactivos que reflejaron peor pertenencia al constructo, según este índice, fueron: HV15, HC07 y ESP16, para ambas pruebas; exclusivos de VA se encontraron: HV13 y QUI16, y de VB: HC09, ESP02, BIO01 y QUI14. De acuerdo con estos resultados, se infiere un ligero mejor comportamiento de VA con respecto a VB.

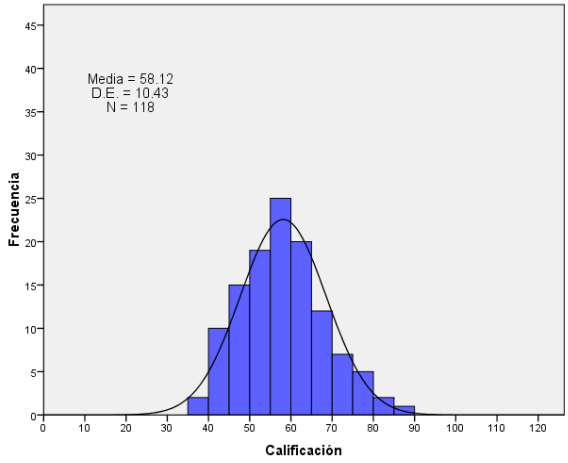
Tabla 4.2
Cantidad de ítems según el rango de la correlación punto biserial para VA y VB

Rango de la correlación	VA	VB
Pbis < 0.1	12	15
0.1 ≤ Pbis < 0.2	22	20
0.2 ≤ Pbis < 0.3	39	43
Pbis ≥ 0.3	44	38
k total	117	116 [a]

Nota: Pbis = índice de correlación punto biserial. [a] Un ítem no pudo ser categorizado porque no se obtuvieron respuestas correctas.

En la figura 4.2 se muestran ambas distribuciones cuyos rangos van de 35 a 95 en el primer caso y de 35 a 90 en el segundo (la calificación mínima posible es 0 y la máxima, 117). Los coeficientes de simetría son similares. Estos números se encuentran dentro del rango [-0.5; 0.5] de aceptabilidad para curvas normales. En cuanto a la curtosis, la distribución de VA es ligeramente leptocúrtica (0.18), mientras que VB tiende a ser platocúrtica (-0.25). Sin embargo, los errores son grandes, lo cual indica que los valores no son significativos. Por lo tanto, estos datos inducen a afirmar que se trata de distribuciones normales.





Simetría	.312	Simetría	.338
Error estándar de simetría	.190	Error estándar de simetría	.223
Curtosis	.185	Curtosis	-.254
Error estándar de Curtosis	.378	Error estándar de Curtosis	.442

Figura 4.2. Distribución de las calificaciones del EXHCOBA-R/MS, VA y VB.

En el mapa de Wright de la figura 4.3 se muestran los niveles de dificultad de los ítems (medidos en lógitos), sin incluir a las personas. De este modo, se aprecian y se comparan las distribuciones de las dificultades de los 117 ítems de los dos exámenes. En ambas pruebas el área más compleja fue Matemáticas (MAT). El área más fácil fue Español (ESP); le siguió Habilidades del lenguaje (HV). Habilidades matemáticas recorrió la gama de dificultades que va de -2 a 3 lógitos; mientras que Ciencias sociales (SOC) y Ciencias naturales se ubicaron en el rango (-1, 1). Existen ítems aislados que, solamente, en una de sus versiones superan la medida de 3 lógitos de dificultad. Estos son: MAT20, FIS11 y HC11 de VA, y MAT04 y HC09 de VB. No se detectaron reactivos con dificultad menor que -2 lógitos.

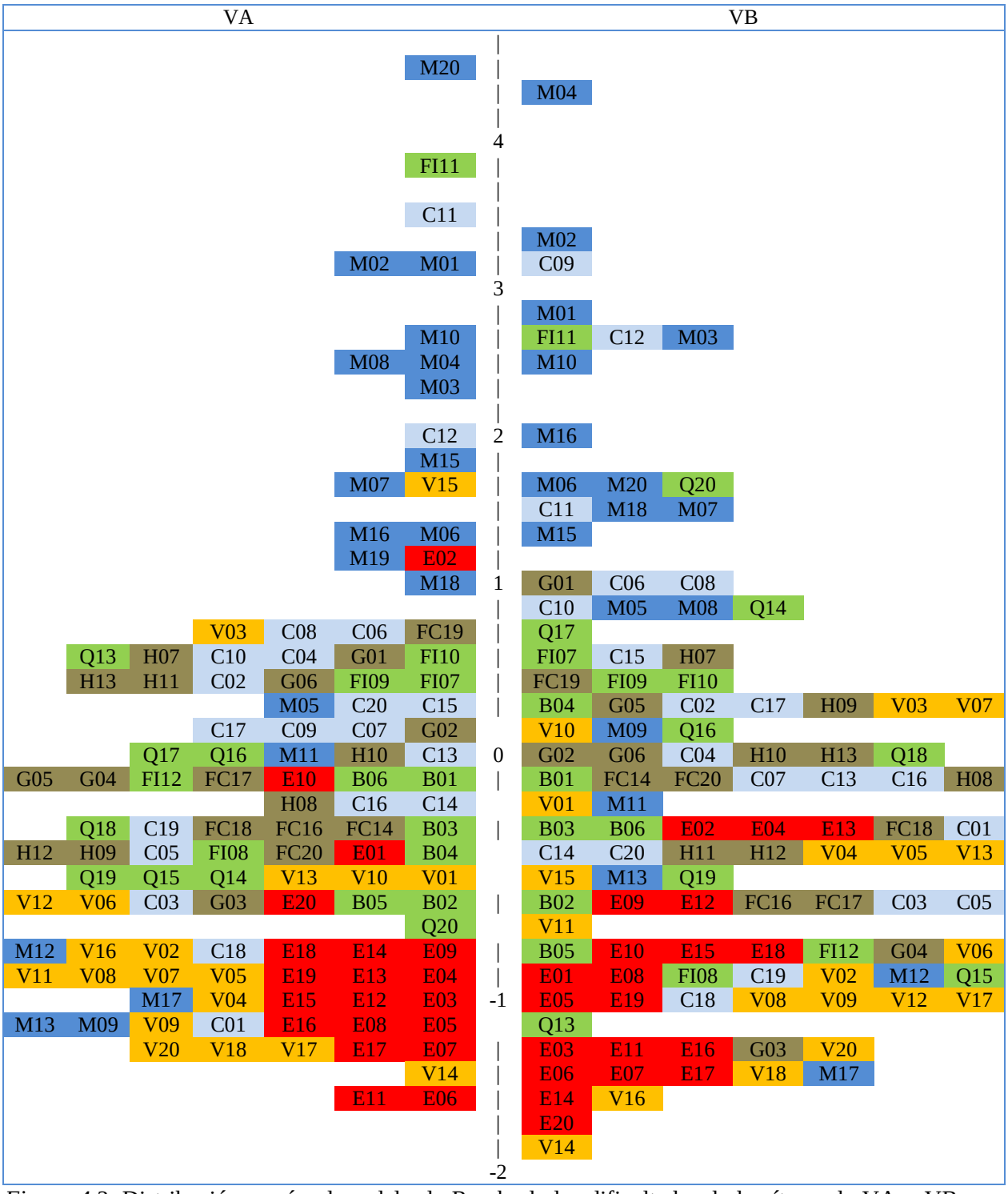


Figura 4.3. Distribución, según el modelo de Rasch, de las dificultades de los ítems de VA y VB, por área. V = HV, C = HC, E = ESP, M = MAT, B = BIO, FI = FIS, Q = QUI, G = GEO, H = HIS, FC = FCYE.

En la tabla 4.3 se indican los ítems con mayores deficiencias, tras la aplicación del modelo de Rasch (para toda la información, consultar el Anexo C). Se observa que tres ítems, HV15, HC07 y QUI19, comparten sus problemas en los dos exámenes. HV15 y HC07 presentaron conflictos de correlaciones asociados a discriminaciones también bajas, mientras que QUI19 reveló desajustes *infit-outfit* en los dos ítems-hermano. ESP02, HC09, MAT07, QUI13, QUI14 y QUI20 mostraron solamente fallas en uno de los tests, lo cual, a su vez, refleja que estas seis familias contienen ítems-hijo no isomorfos.

Tabla 4.3
Infit, outfit, correlación punto medida y discriminación según el modelo de Rasch para VA y VB

VA					VB				
Item	*Infit*	*Outfit*	Pmed	Discr.	*Infit*	*Outfit*	Pmed	Discr.	Item
HV15	1.01	1.07	**0.04**	0.99	1.06	1.06	**0.03**	**-0.01**	**HV15**
HC07	1.15	1.16	**0.09**	**0.70**	1.13	1.19	**0.18**	**0.71**	**HC07**
--					1.00	1.04	**0.04**	1.00	**HC09**
ESP01	1.07	1.06	**0.07**	0.95					--
--					1.08	1.08	**-0.01**	**-0.25**	**ESP02**
ESP18	1.13	1.24	**0.08**	0.71					--
--					1.02	1.04	**0.05**	0.98	**MAT07**
QUI13	1.07	1.08	**0.06**	0.87					--
--					1.14	1.15	**0.08**	**0.76**	**QUI14**
QUI16	1.29	**1.34**	**0.13**	**0.67**					--
QUI19	**1.34**	**1.66**	0.24	**0.68**	**1.33**	**1.59**	0.28	0.81	**QUI19**
--					1.02	1.01	**0.07**	0.98	**QUI20**
Promedio	1.00	1.00	0.28	1.05	1.00	1.00	0.29	1.05	Promedio

Nota: Pmed = índice de correlación punto medida, Discr = índice de discriminación, según el modelo de Rasch

A través del índice de correlación punto medida (ver tablas 4.3 y 4.4), calculado desde el modelo de Rasch, se infiere una ligera inclinación hacia una mayor calidad de ítems de VB, con respecto a VA (medias de 0.29 y 0.28, respectivamente). Estos resultados parecieran contradecirse con el otro tipo de correlación, obtenida desde la TCT; sin embargo, la diferencia es pequeña y podría deberse a la gran cantidad de información que se eliminó en la teoría clásica, por causa de datos perdidos.

Tabla 4.4

Cantidad de ítems según el rango de la correlación punto medida para VA y VB

Rango de la correlación	VA	VB
Pmed < 0.1	5	6
0.1 ≤ Pmed < 0.2	22	14
0.2 ≤ Pmed < 0.3	42	41
Pmed ≥ 0.3	48	55
k total	117	116 [a]

Nota: Pmed = índice de correlación punto medida. [a] Un ítem no pudo ser categorizado porque no se obtuvieron respuestas correctas.

Otro dato interesante es cuánta varianza se puede explicar a través del modelo unidimensional que propone el EXHCOBA-R/MS. VA explica el 38.5% y VB el 37.3%, a través de sus medidas, lo cual indica valores aceptables y parecidos para dos exámenes que representan a un mismo modelo.

Debido a que se contó con la información de los resultados del EXHCOBA tradicional aplicado en esa fecha, a los mismos estudiantes, se cruzó la información para comparar con los puntajes de VA y VB. En las tablas 4.5 y 4.6 se observa una alta correlación de ambas versiones con el EXHCOBA tradicional. Si bien, al cotejar con las calificaciones de la educación primaria y secundaria, los valores no son tan altos, su significatividad se mantiene al nivel de 0.01. Por lo tanto, se infiere que el nuevo examen (en sus dos versiones, VA y VB) se comporta de manera similar que la antigua prueba y, a su vez, refleja los aprendizajes escolares de la educación básica de dichos estudiantes.

Tabla 4.5
Correlaciones de Pearson de la calificación total de EXHCOBA-R para VA con las calificaciones de los estudiantes en su educación básica y en el EXHCOBA tradicional

VA N = 163	EXHCOBA -R/MS	Primaria [a]	Secundaria Std	EXHCOBA Std	Final (E + S)
EXHCOBA-R/MS	1	.493**	.352**	.801**	.783**
Primaria		1	.319**	.424**	.455**
Secundaria Std			1	.374**	.611**
EXHCOBA Std				1	.963**
Final (E + S)					1

Nota: Std = estandarizado. E + S = calificación ponderada del EXHCOBA y de la educación secundaria.
[a] El promedio del nivel primario fue consultado a los estudiantes, no es oficial.
** Correlación significativa al nivel 0.01 (dos colas).

Tabla 4.6
Correlaciones de Pearson de la calificación total de EXHCOBA-R para VB con las calificaciones de los estudiantes en su educación básica y en el EXHCOBA tradicional

VB N = 118	EXHCOBA-R/MS	Primaria [a]	Secundaria Std	EXHCOBA Std	Final (E + S)
EXHCOBA-R/MS	1	.459**	.245**	.721**	.715**
Primaria		1	.297**	.431**	.467**
Secundaria Std			1	.267**	.524**
EXHCOBA Std				1	.961**
Final (E + S)					1

Nota: Std = estandarizado. E + S = calificación ponderada del EXHCOBA y de la educación secundaria.
[a] El promedio del nivel primario fue consultado a los estudiantes, no es oficial.
** Correlación significativa al nivel 0.01 (dos colas).

Después de la exposición de resultados a través de las teorías psicométricas TCT y TRI, se infiere que los dos exámenes dan muestras de similitud (dificultad media, confiabilidad, varianza explicada, correlación con el EXHCOBA original y con los promedios de educación básica). Sin embargo, las áreas de Habilidades del lenguaje, Español y Ciencias naturales difieren en la calidad de ciertos ítems, lo cual también se manifiesta en la escasa confiabilidad de estas áreas, en alguna de sus dos versiones.

Otro hallazgo es que tanto la TCT como la TRI coinciden en bajas correlaciones para HV15 y HC07, en ambos exámenes. Esto permite inferir la escasa pertenencia de estas familias, al constructo de habilidades y conocimientos básicos del EXHCOBA-R/MS. Casos extraños se

presentaron para HV13, QUI16 (los dos de VA), HC09, ESP02, BIO01 y QUI14 (estos cuatro de VB). Estos resultados revelan la falta de homogeneidad de los ítems-hijo de estas familias.

Si bien la cantidad de datos analizados es escasa, en los próximos apartados se observa la continuidad de algunos de los problemas detectados en este análisis, lo cual implica una coherencia en los resultados.

4.1.2. De cada área

En este apartado se muestran los resultados de VA y VB de manera detallada para el área de Habilidades matemáticas y se comparan los resultados. De las áreas restantes, se incluye un resumen y el resto de la información se adjunta en el anexo D. Finalmente, se resumen los hallazgos y se sugieren recomendaciones para mejorar la GAI del EXHCOBA-R/MS.

4.1.2.1. Habilidades matemáticas

A modo de recordatorio, la figura 4.4 presenta un esquema del área de HC. Son 20 contenidos básicos representativos del plan de estudios de la educación primaria (SEP, 2011), particularmente centrados en quinto y sexto grados.

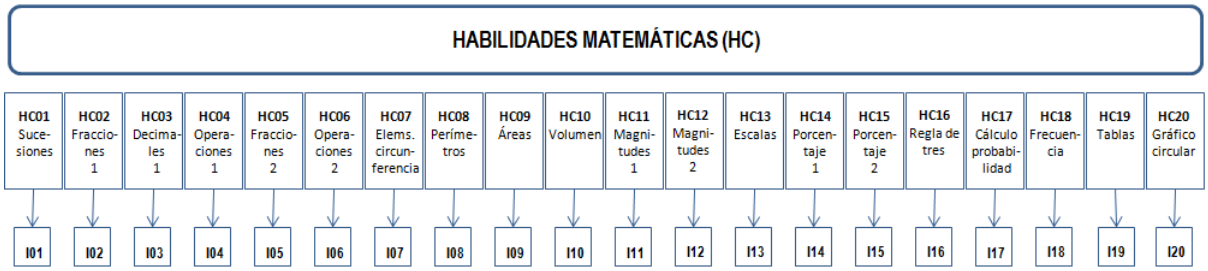
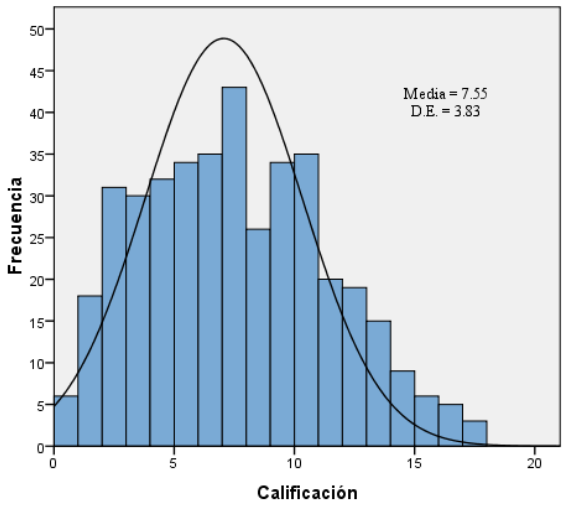


Figura 4.4. Esquema de contenidos del área de Habilidades del lenguaje (VA y VB)

Se analizaron 396 datos de VA y 301 de VB. Según la TCT, los índices indicaron un promedio de dificultad, del total de 20 ítems, de 0.377 para VA y de 0.383 para VB, con una

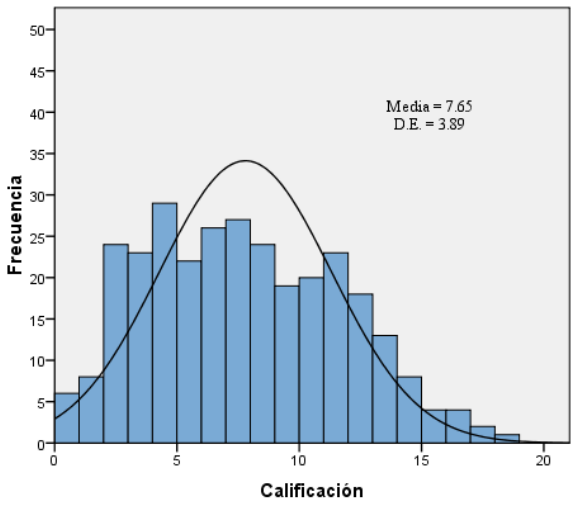
desviación estándar de 0.180 y 0.183, respectivamente. En la figura 4.5 se observa que ambas distribuciones resultaron normales con una curtosis ligeramente negativa (la curva tiende a ser platocúrtica).

VA



Simetría	.295
Error estándar de simetría	.122
Curtosis	-.654
Error estándar de Curtosis	.243
Mínimo	.66
Máximo	17.66

VB



Simetría	.289
Error estándar de simetría	.140
Curtosis	-.684
Error estándar de Curtosis	.280
Mínimo	.00
Máximo	18.00

Figura 4.5. Distribución de las calificaciones del área de Habilidades matemáticas de VA y VB.

En la figura 4.6 se percibe que todos los reactivos tuvieron dificultad inferior a 0.80 y algunos fueron extremadamente difíciles, con *p* cercano a cero. En general, las dificultades de ítems-hermano resultaron similares en ambas versiones. Las mayores diferencias se encontraron en HC09 y HC01, con 0.37 y 0.25, respectivamente; el resto no superó la distancia de 0.15.

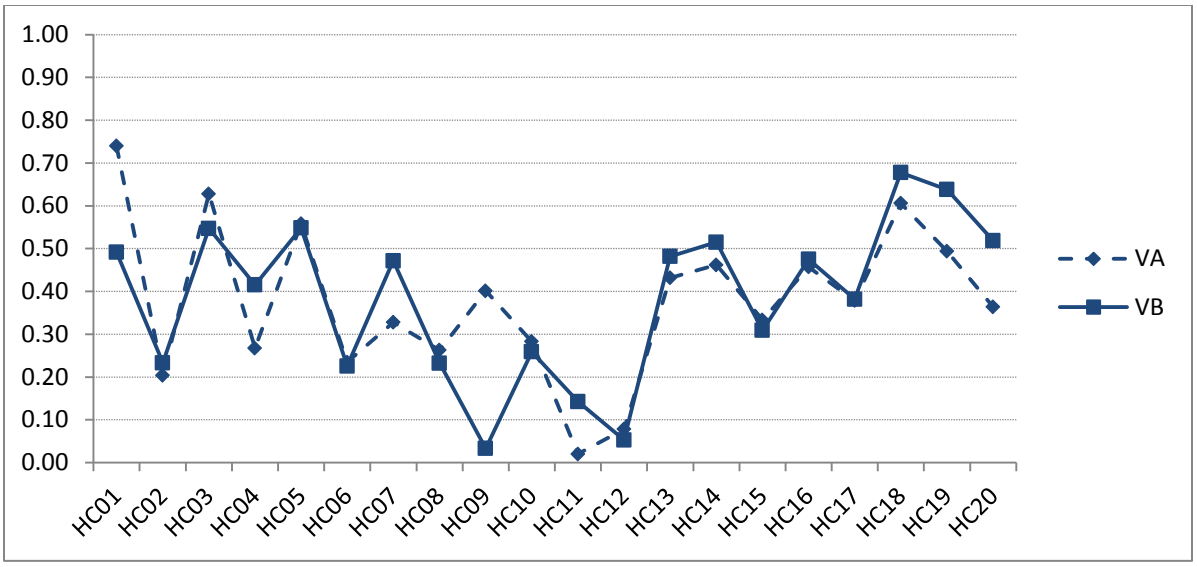


Figura 4.6. Índices de dificultad para el área de Habilidades matemáticas en VA y VB.

Tal como se determinó en el capítulo tres del Método, se consideraron como ítems aceptables aquellos que presentaran una correlación punto biserial igual o superior a 0.2. Según este criterio, se detectaron problemas de correlación en HC07 para ambas versiones y para HC09 de VB (ver figura 4.7), tal como se había registrado en los análisis psicométricos del examen en su totalidad. Para el resto de los ítems, la mayoría de los índices se encontraron en el intervalo [0.30; 0.50]. Estos valores se vieron reflejados en la confiabilidad aceptable de ambas pruebas en esta área ($\alpha \approx 0.78$).

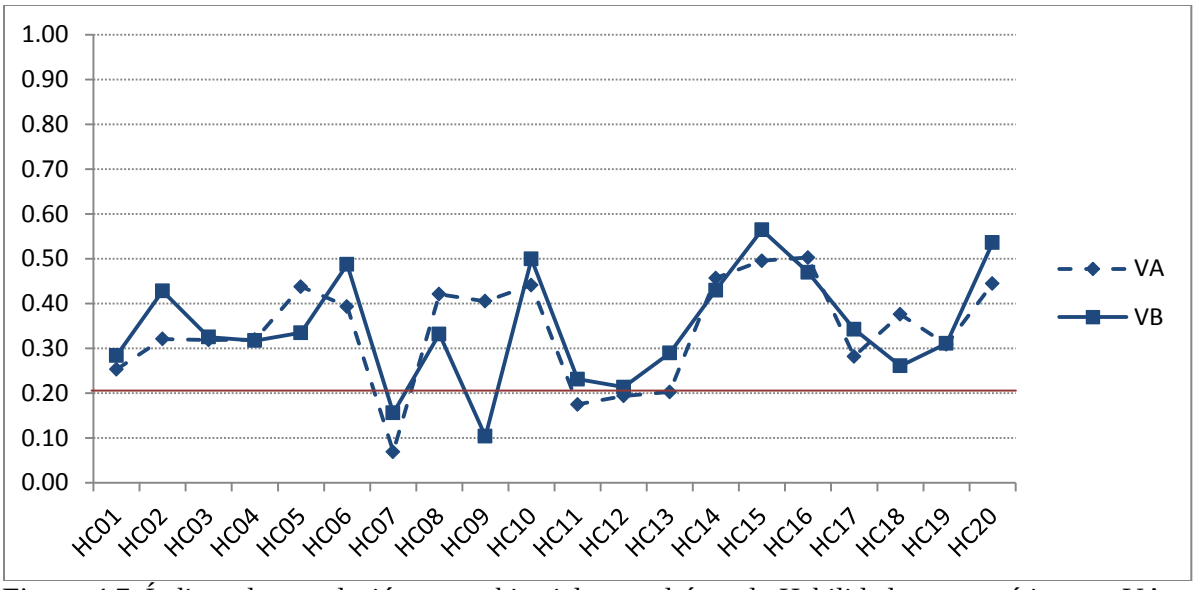
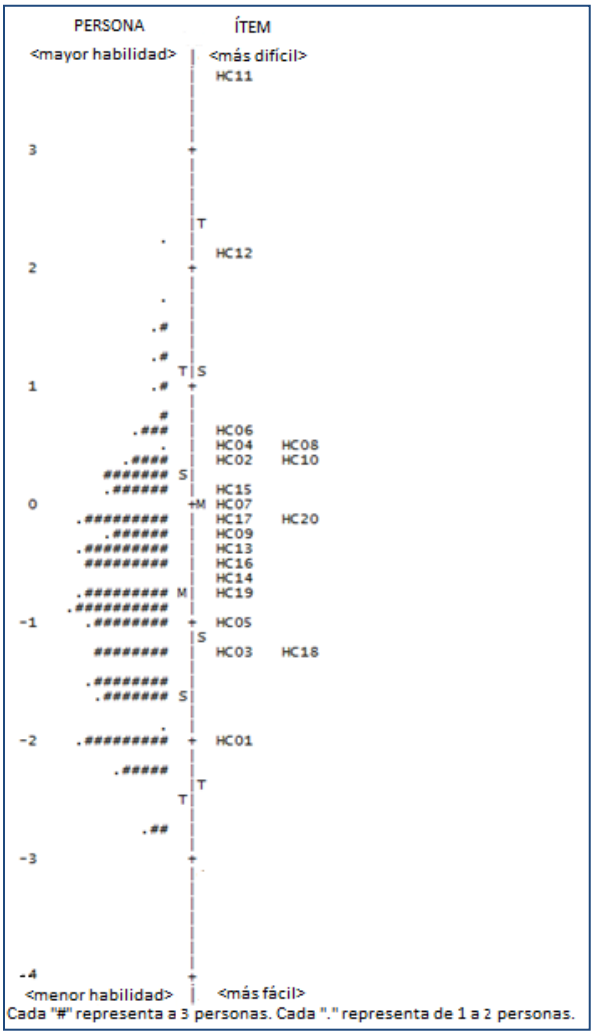


Figura 4.7. Índices de correlación punto biserial para el área de Habilidades matemáticas en VA y VB. Índices de confiabilidad. Alpha de Cronbach: VA = 0.784, VB = 0.788.

En los mapas de Wright (ver figura 4.8), construidos por el modelamiento de Rasch, se identifica en las dos muestras, que la media de dificultad de los ítems fue mayor (casi en una desviación estándar) que la media de habilidad de los examinados. Los reactivos se distribuyeron en dificultades diferentes, de modo que apuntaron a las distintas habilidades de los evaluados; sin embargo, quedaron estudiantes sin reactivos lo suficientemente fáciles para su nivel, asimismo, ítems demasiado complicados para dicha población. En general, las distribuciones fueron similares en ambas pruebas, con la excepción de HC09 que en VA tuvo una dificultad inferior a 0 lógitos y en VB, fue de 3 lógitos.

VA



VB

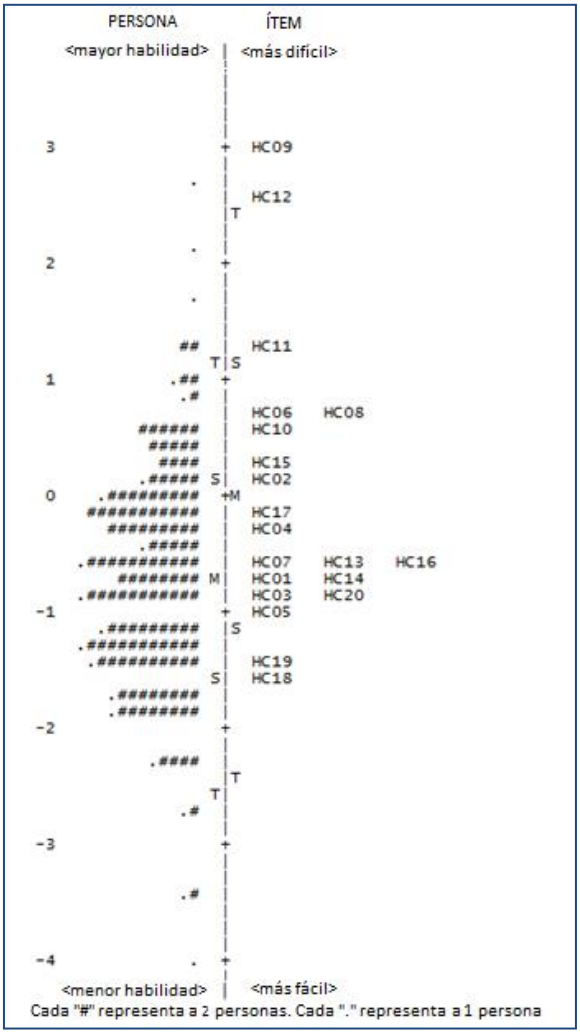


Figura 4.8. Mapas de Wright para Habilidades matemáticas, de VA y VB.

Según lo especificado en el Método, se determinó como ítems con buen ajuste o *productivos* a aquellos cuyos índices expresados en medias cuadráticas estuvieran entre 0.8 y 1.3. Para el caso que quedaran fuera de rango, se calcularon los índices estandarizados, si estos pertenecían al intervalo [-2; 2], se aceptaban los ítems dentro de la escala. Esta última decisión se tomó como una recomendación de Linacre (2010), cuando los datos son pocos. Según este criterio, para ambas versiones, HC07 presentó desajustes tanto de *infit* como de *outfit*, y en menor grado HC03 (ver figuras 4.9 y 4.10).

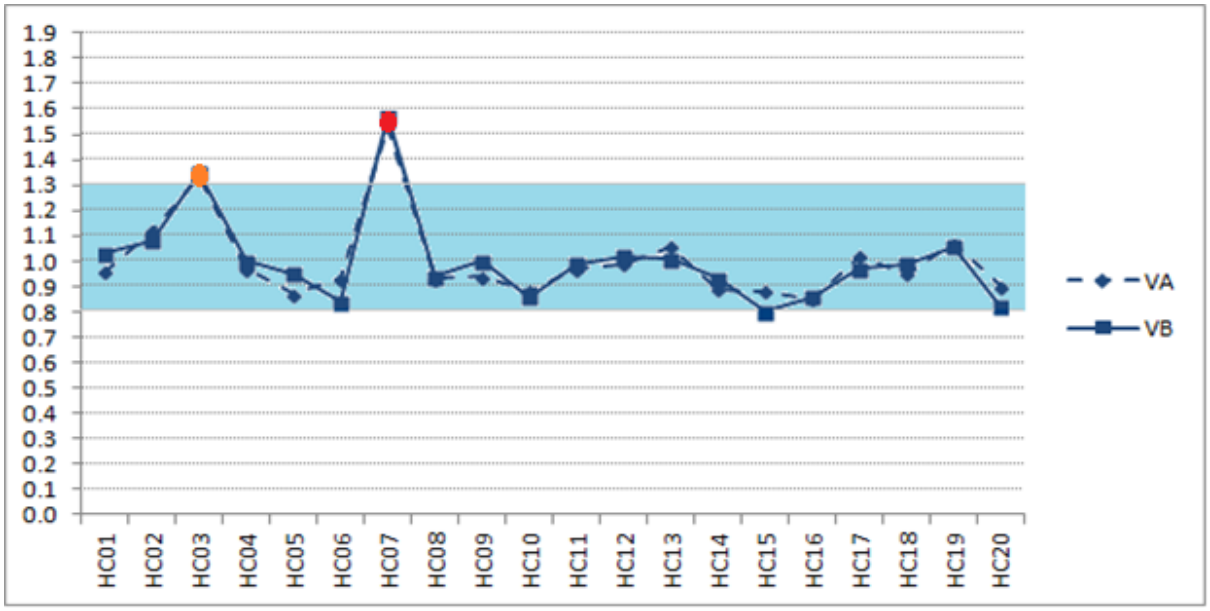


Figura 4.9. Valores de *infit* de cada ítem del área de Habilidades matemáticas de VA y VB.

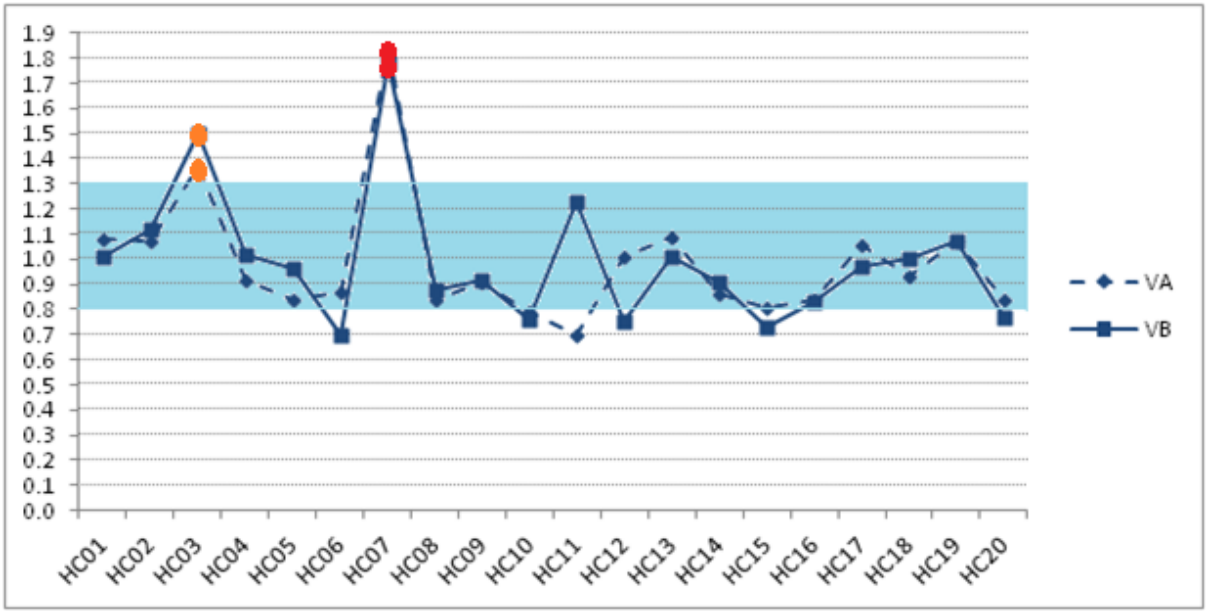


Figura 4.10. Valores de *outfit* de cada ítem del área de Habilidades matemáticas de VA y VB.

La correlación punto medida que arrojó el análisis de Rasch indica valores inferiores a 0.20 para HC09 de VB y HC11 de VA, estos ítems son los más difíciles de ambas versiones (ver figura 4.11). En cuanto a la discriminación (ver figura 4.12), en general, los índices son buenos (superiores a 0.80), excepto nuevamente para HC07, ya que en ambas versiones los valores son muy pequeños.

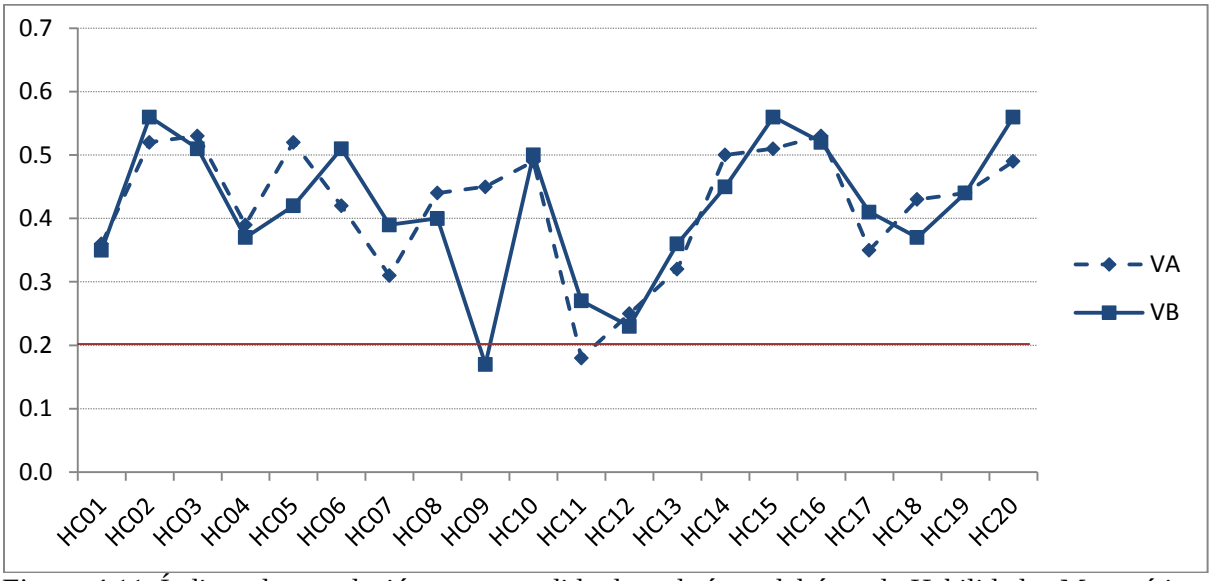


Figura 4.11. Índices de correlación punto medida de cada ítem del área de Habilidades Matemáticas de VA y VB.

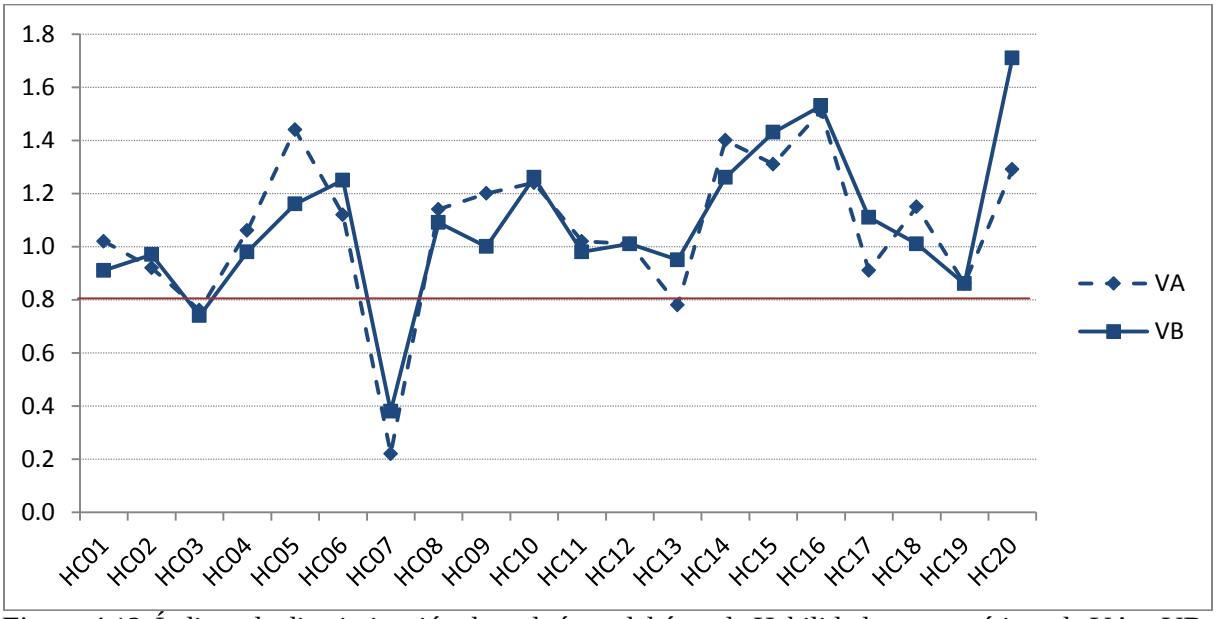


Figura 4.12. Índices de discriminación de cada ítem del área de Habilidades matemáticas de VA y VB.

Para el AFC, se armó un modelo de un factor denominado *Habilidades matemáticas de la educación primaria*. Además, se incorporaron covarianzas de errores entre los reactivos del mismo eje temático. Posteriormente, se eliminaron aquellas covarianzas que no aportaban carga al modelo, según el test de Wald. Los análisis señalaron resultados similares para ambas

versiones. En los dos casos, se reportaron índices de ajuste aceptables (ver tabla 4.7), salvo para *p* de VA. Se probó con modelos de dos factores, sin embargo, el valor de *p* no mejoró; por lo tanto, se optó por el modelo unidimensional, por ser más parsimonioso. Las cargas factoriales resultaron apropiadas para la mayoría de los ítems, salvo para HC07 (en ambas versiones) y para HC09, de VB (ver figura 4.13).

Tabla 4.7.
Área de Habilidades matemáticas. Índices de ajuste de los AFC para VA y VB, por modelo propuesto

Índices	Versión A		Versión B	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Chi cuadrado	220.502	228.803	199.763	195.164
Grados libertad	162	156	165	163
P	**.001**	**.000**	.033	.043
NNFI	0.93	0.91	0.94	0.95
CFI	.94	.92	.95	.95
RMSEA	.030	.034	.027	.026
Covarianza F1-F2		0.92		1.00

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*.

	VA	VB
HC01	.306	.299
HC02	.335	.462
HC03	.339	.381
HC04	.383	.386
HC05	.520	.391
HC06	.458	.567
HC07	**.056**	**.143**
HC08	.492	.368
HC09	.469	**.092**
HC10	.495	.603
HC11	.203	.250
HC12	.223	.206
HC13	.217	.326
HC14	.500	.505
HC15	.520	.663
HC16	.554	.522
HC17	.331	.382
HC18	.420	.275
HC19	.347	.312
HC20	.486	.609

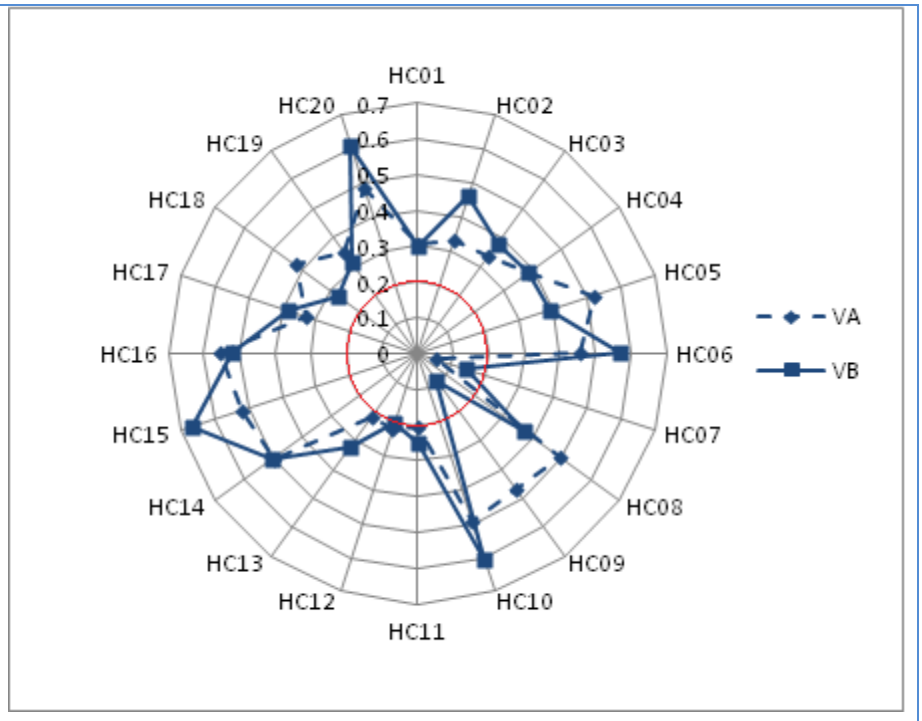


Figura 4.13. Cargas factoriales estandarizadas del AFC para Habilidades matemáticas VA y VB. Modelo de un factor con errores que covarían.
Estadísticos significativos al nivel de 0.05; excepto para HC07 de VA y HC09 de VB.

Debido a que los ítems de HC, en su mayoría, son dicotómicos, se creyó conveniente constatar que la agrupación de los reactivos se debió a las habilidades latentes que invocan los ejercicios y no a aspectos ajenos, por ejemplo, la dificultad. Por lo tanto, se convirtieron los cuatro ítems de crédito parcial al formato correcto-incorrecto para construir la matriz tetracórica de cada versión. Con esta matriz, se efectuó un AFE de un factor para identificar el autovalor o *eigenvalue* (ver tabla 4.8 y figura 4.14) y calcular las cargas factoriales en ambas versiones (ver figura 4.15). En el caso de VB, se eliminó HC09 porque su inclusión originaba una matriz definida no positiva.

Se correlacionaron las cargas resultantes de ambos tipos de análisis (AFE y AFC) para las dos versiones, y dieron significativas al nivel de 0.01. Lo anterior indica que el AFC no se desvirtuó por la existencia de ítems dicotómicos; aunque sí las cargas fueron menores que con el AFE.

Tabla 4.8
Autovalores y porcentaje de varianza explicada para el AFE de HC de VA y VB

	Eigenvalues iniciales		Extracción de sumas de cargas cuadradas	
	Total	% de varianza	Total	% de varianza
VA	6.436	32.180	5.789	28.946
VB	6.253	32.913	5.646	29.718

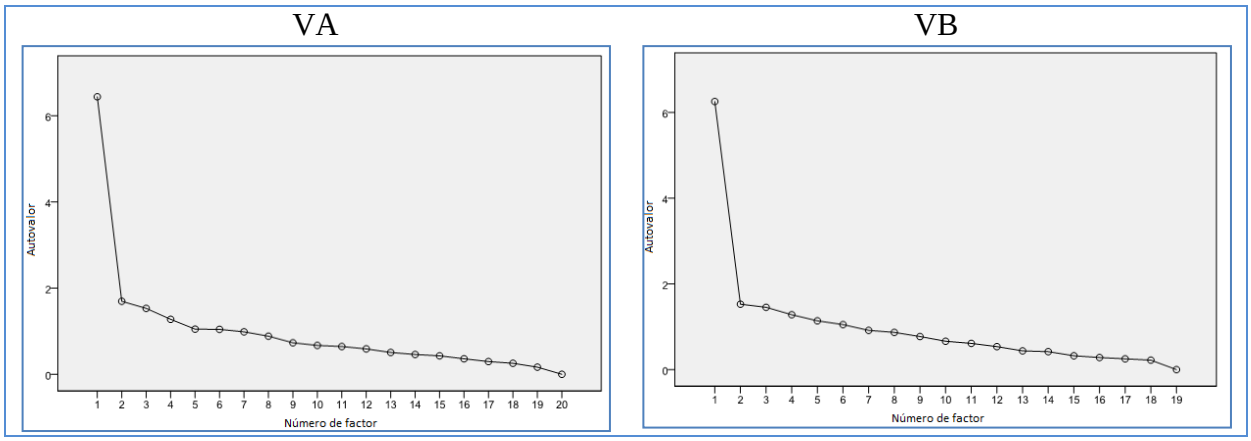


Figura 4.14. Gráficos de sedimentación de los AFE efectuados al área de HC de VA y VB.

	VA	VB
HC01	.393	.411
HC02	.558	.593
HC03	.445	.528
HC04	.501	.463
HC05	.630	.496
HC06	.577	.714
HC07	**.130**	**.150**
HC08	.625	.499
HC09	.585	--
HC10	.650	.734
HC11	.535	.381
HC12	.381	.448
HC13	.246	.420
HC14	.665	.634
HC15	.723	.810
HC16	.682	.651
HC17	.386	.467
HC18	.542	.396
HC19	.419	.300
HC20	.657	.774

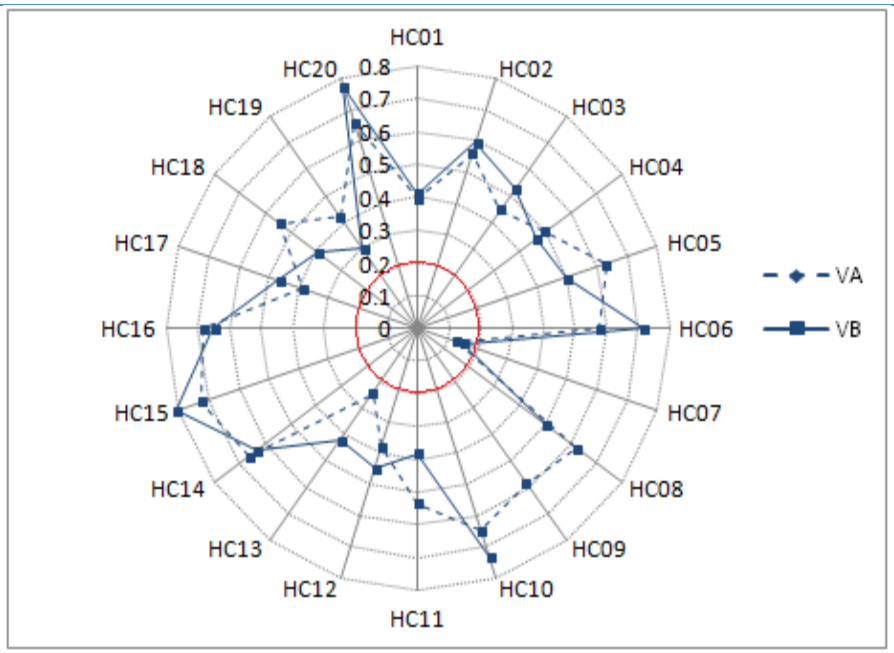


Figura 4.15. Cargas factoriales estandarizadas del AFE de las matrices tetracóricas de Habilidades matemáticas, VA y VB. Modelo unidimensional.

Después de presentar los resultados de los diferentes análisis estadísticos, se infiere que el área se comporta de manera similar en ambas versiones (dificultad, confiabilidad, correlaciones, índices Rasch y AF). Además, los ítems muestran buenas propiedades psicométricas, en general, con algunas excepciones. Los reactivos pertenecientes a la competencia HC07 revelan serios

problemas de ajuste al modelo, según Rasch, y poca representación del constructo, de acuerdo con al TCT y el AF (AFE y AFC). Además, el modelo de Rasch identificó un problema de ajuste en los ítems de HC03, por lo que habría que analizar sus causas. El reactivo HC09 presenta un comportamiento extraño solamente en VB, por lo cual sería apropiado comparar los ítems de ambas versiones y detectar el origen de esa diferencia.

2.1.1.3. Habilidades del lenguaje

Para el área de Habilidades del lenguaje (ver figura 4.16) solamente se obtuvo información de 19 ítems. El HV19 no se pudo administrar por problemas de diseño, como ya se había comentado. Se examinaron 289 datos de VA y 189 de VB, debido a la eliminación de los casos perdidos.

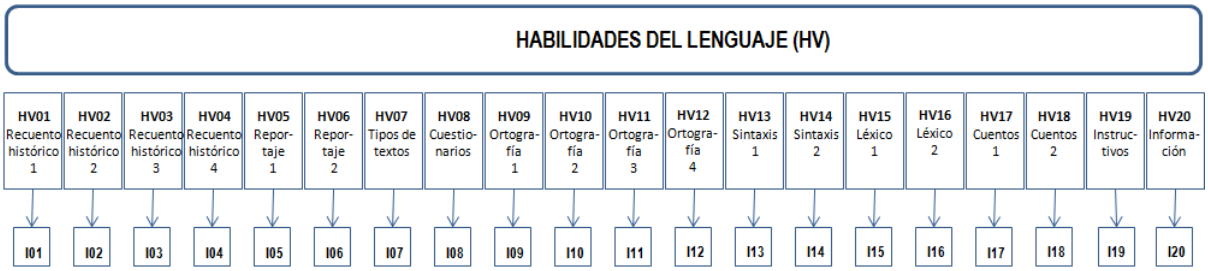


Figura 4.16. Esquema de contenidos del área de Habilidades del lenguaje (VA y VB)

En la tabla 4.9 se muestra un resumen de los estadísticos calculados en la TCT y la TRI, ahí se señalan con negritas las deficiencias psicométricas[11] y se enmarcaron las más sobresalientes. Para revisar toda la información en detalle, dirigirse al Anexo D, al apartado de HV.

[11] En el caso de la *p* (TCT) y la medida (TRI), se marcaron los valores con amplia diferencia de dificultades entre ítems-hijo de VA y VB de una misma familia. Por ejemplo, para HV07, la diferencia de *p* es de 0.42, y la distancia entre las dos medidas es de 1.03 lógitos. Por lo tanto, los dos ítems-hijo de HV07 no son isomorfos en cuanto a su dificultad.

Tabla 4.9
Índices calculados a través de la TCT y de la TRI (modelo de Rasch) para los ítems de HV de VA y VB.

Item	TCT				TRI									
	Dificultad (p)		Pbis		Medida		Infit		Outfit		Pmed		Discriminación	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
HV01	**0.56**	**0.33**	**0.17**	0.23	0.23	0.59	1.15	0.97	1.11	0.94	0.49	0.54	0.84	1.01
HV02	0.67	0.65	**0.13**	0.25	-0.13	-0.13	1.14	0.97	1.20	0.99	0.38	0.43	0.81	1.02
HV03	0.28	0.33	0.22	**0.09**	**1.68**	**0.89**	0.96	1.00	0.93	1.05	0.36	0.24	1.08	0.96
HV04	**0.81**	**0.59**	0.41	0.26	**-0.39**	**0.32**	0.92	0.96	0.86	0.95	0.42	0.33	1.07	1.02
HV05	**0.71**	**0.48**	0.36	**0.17**	**-0.34**	**0.35**	0.91	1.02	0.91	1.02	0.44	0.29	1.13	0.95
HV06	0.62	0.63	0.21	**0.13**	0.13	0.01	1.02	1.00	0.99	1.00	0.27	0.28	0.98	0.99
HV07	**0.72**	**0.30**	0.30	**0.16**	**-0.20**	**0.83**	0.94	0.95	0.90	0.92	0.43	0.41	1.10	1.05
HV08	0.86	0.80	**0.17**	0.22	-0.29	-0.29	1.10	1.04	1.34	1.08	0.33	0.35	0.96	0.97
HV09	0.89	0.84	0.30	0.30	-0.63	-0.41	0.98	1.06	0.92	1.16	0.41	0.29	1.08	0.92
HV10	0.61	0.41	0.33	**0.11**	0.29	0.79	0.90	1.11	0.90	1.12	0.50	0.34	1.10	0.87
HV11	0.75	0.59	0.28	0.24	-0.29	0.20	1.00	1.00	1.03	0.99	0.40	0.38	1.01	1.01
HV12	0.66	0.77	0.31	0.32	0.09	-0.36	1.02	0.95	1.00	0.97	0.41	0.42	0.97	1.05
HV13	0.55	0.49	**0.12**	0.20	0.30	0.41	1.12	0.97	1.12	0.97	0.23	0.35	0.73	1.07
HV14	0.91	0.86	0.35	0.32	-1.02	-1.14	0.93	0.95	0.86	0.96	0.33	0.33	1.04	1.03
HV15	**0.09**	**0.54**	**-0.05**	**-0.07**	**2.81**	**0.41**	1.05	1.07	1.23	1.08	**0.00**	**0.01**	0.95	**-0.10**
HV16	**0.53**	**0.91**	**0.13**	0.24	**0.05**	**-0.80**	1.05	0.96	1.05	0.81	0.20	0.33	0.94	1.02
HV17	0.88	0.73	0.37	0.25	-0.72	-0.44	0.91	0.96	0.90	0.93	0.36	0.36	1.05	1.05
HV18	0.93	0.95	0.28	0.21	-0.74	-0.67	0.95	1.00	1.13	0.96	0.34	0.33	1.01	1.00
HV19														
HV20	0.91	0.85	0.24	0.29	-0.81	-0.58	1.03	1.03	1.02	1.04	0.33	0.33	1.01	0.97
Prom	0.68	0.63	0.24	0.21	0.00	0.00	1.00	1.00	1.02	1.00	0.35	0.33	0.99	0.94

Nota: pbis = índice de correlación punto biserial. pmed = índice de correlación punto medida. Prom = promedio.

Ambas distribuciones estuvieron dentro del rango de normalidad, aunque la gráfica de VA tendió a ser leptocúrtica, mientras que la de VB, fue mesocúrtica. El promedio de dificultad fue de 0.681 para VA y de 0.634 para VB, con una desviación estándar de 0.22 y 0.21, respectivamente. Ambas versiones exhibieron dificultades parecidas para los ítems-hijo de algunas familias (por ejemplo: HV02, HV03, HV06, HV08). Los reactivos con mayores diferencias fueron: HV15, HV07, HV16, HV05, HV01, HV04 y HV10, en ese orden. Estas diferencias resultaron mayores o iguales que 0.20, y la máxima, de 0.45. El rango de dificultad se

encontró entre 0.30 y 0.90, excepto HV15 de VA, que fue resuelto correctamente sólo por el 9% de los estudiantes.

De acuerdo con el modelamiento de Rasch, la media de dificultad de los ítems fue una desviación menor que la media de habilidad de los examinados. Sin embargo, existieron casos aislados, en VA se observa que HV03 y HV15 se separó del resto de los reactivos (resultaron más difíciles en dos desviaciones que la habilidad media de las personas). También se identificaron ítems demasiado fáciles para ambas versiones, estos fueron: HV14, HV20 y HV18.

Seis ítems de VA y seis de VB no superaron la barrera de 0.20 de correlación punto biserial. Lo curioso es que solamente HV15, con índices negativos, coincidió en las dos versiones. Los cinco ítems restantes, con correlaciones bajas, no se repitieron en los dos exámenes (lo cual indica falta de isomorfismo entre estos ítems-hermano). Los problemas de consistencia interna se reflejaron en la confiabilidad del área, ya que los Alpha de Cronbach resultaron relativamente bajos, 0.633 para VA y 0.547 para VB.

En el análisis Rasch, la correlación punto medida, que es más sensible al registrarse casos perdidos, también identificó a HV15 con un índice cero en ambos tests. Tanto la TCT como la TRI señalaron serias fallas de correlación para este reactivo dentro del área y en el examen en general. En el resto de los reactivos no aparecieron índices menores a 0.20, desde el modelo de Rasch.

No se presentaron problemas de ajuste al modelo, los valores de *infit* y *outfit* reflejaron un buen comportamiento de ambas versiones, con valores dentro de lo aceptado. Los índices de discriminación también estuvieron cercanos a 1, excepto para HV15 de VB, cuyo número fue negativo.

Para el AFC, según la organización temática del examen, se armaron varios modelos. En cada caso, se establecieron covarianzas entre los errores y mediante el test de Wald se eliminaron aquellas que no aportaron cargas importantes. Finalmente, se documentaron dos modelos (ver Anexo D, en el apartado de HV).

El primero, el más sencillo, fue de un factor: *Habilidades básicas del lenguaje*, que representa aquellas competencias elementales que debe aprender un niño tras cursar la educación primaria. Si bien este modelo se ajustó bastante bien para VA, no ocurrió lo mismo para VB. El segundo modelo fue de dos factores y reportó mejores índices de ajuste, sobre todo para VB. Si bien la organización en dos factores permitió una mejor distribución de cargas, permanecieron ítems con deficiencias aún en el mejor modelo, estos reactivos son: HV02, HV08, HV13 y HV16, para VA, y HV03 y HV05 para VB. Estos ítems también presentaron índices de correlación punto biserial bajos. Cabe aclarar que en todos los casos debió eliminarse el ítem HV15, debido a grandes problemas de correlación detectados por el test de Wald.

Tras los análisis estadísticos efectuados a HV, en sus dos versiones, se infiere que en cuanto a la dificultad, los reactivos parecen ser más sencillos de lo esperado. Los ítems-hijo de HV15, HV16 y HV07 presentan dificultades muy disímiles en sus dos versiones, lo cual no debería ocurrir si se pretenden ítems-hermano isomorfos. El resto de los reactivos son más similares, en relación con este parámetro.

Si bien, no se identificaron desajustes de *infit* ni de *outfit* en ninguno de los reactivos, se percibe una ligera superioridad de VA sobre VB (confiabilidad y correlación). Se detectaron algunos problemas, que no se comparten en VA y VB, esto refleja falta de isomorfismo entre ítems de una misma familia. En general, el área necesita una mayor cohesión en su constructo. En particular, la discriminación negativa de HV15 aunada a serias deficiencias de correlación,

sugieren que esta familia de reactivos se elimine o cambie sustancialmente.

4.1.2.3. Español

En el área de Español (correspondiente a la educación secundaria) pudieron administrarse y recuperar la información de los 20 ítems. Un esquema del área se observa en la figura 4.17.

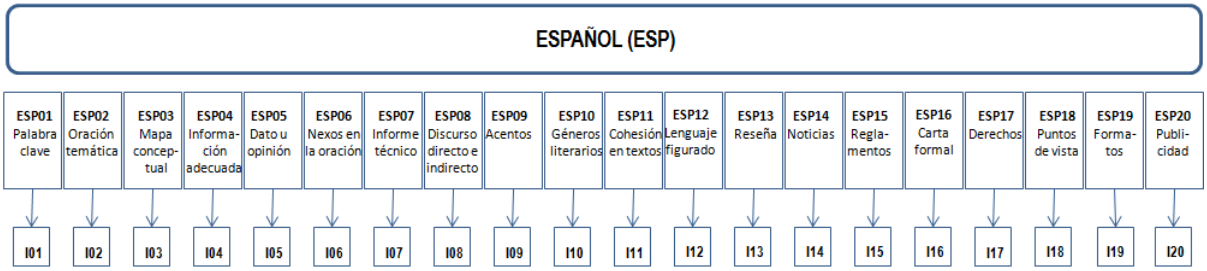


Figura 4.17. Esquema de contenidos del área de Español (VA y VB)

De los análisis efectuados a través de la TCT y del modelo de Rasch, en la tabla 4.10 se listan los valores calculados. Para revisar todos los resultados en detalle, dirigirse al Anexo D, al apartado de Español.

Tabla 4.10
Índices calculados a través de la TCT y de la TRI (modelo de Rasch) para los ítems de ESP de VA y VB

Item	TCT				TRI									
	dificultad		pbis		medida		Infit		outfit		pmed		Discriminación	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
ESP01	**0.46**	**0.75**	**0.11**	0.23	0.52	0.04	1.10	1.20	1.09	1.26	**0.17**	0.36	0.92	0.84
ESP02	**0.18**	**0.53**	**0.09**	**0.06**	**2.45**	**0.60**	0.98	1.04	1.03	1.04	0.21	**0.15**	1.01	**0.58**
ESP03	0.73	0.75	0.26	0.28	-0.12	-0.31	0.94	1.01	0.89	0.99	0.45	0.37	1.07	0.99
ESP04	0.81	0.63	0.25	0.36	**-0.03**	**0.66**	1.02	1.02	1.06	0.99	0.41	0.43	0.99	1.01
ESP05	0.91	0.74	0.21	0.36	-0.46	-0.05	1.01	0.99	1.00	1.03	0.29	0.43	0.99	1.01
ESP06	0.84	0.87	0.28	0.33	-0.92	-0.47	0.94	0.97	0.91	1.01	0.36	0.40	1.06	1.02
ESP07	0.86	0.73	**0.16**	0.32	-0.61	-0.43	0.98	0.99	0.92	0.98	0.29	0.36	1.01	1.01
ESP08	0.84	0.75	0.31	0.34	-0.38	0.03	0.98	1.00	0.99	1.00	0.39	0.43	1.02	1.01
ESP09	0.75	0.60	**0.19**	0.25	0.10	0.39	1.04	1.09	1.10	1.08	0.29	0.39	0.96	0.88
ESP10	**0.45**	**0.65**	0.21	0.25	**0.97**	**0.27**	1.06	1.03	1.05	1.04	0.42	0.43	0.92	0.96
ESP11	0.78	0.70	0.21	0.37	**-0.80**	**-0.23**	1.02	0.95	1.02	0.93	0.27	0.43	0.97	1.10
ESP12	0.70	0.53	0.24	**0.18**	**-0.11**	**0.55**	0.98	1.03	0.98	1.02	0.33	0.27	1.05	0.92
ESP13	**0.71**	**0.45**	0.26	0.20	**0.05**	**0.59**	0.95	1.08	0.93	1.15	0.35	0.28	1.03	0.94
ESP14	0.56	0.58	0.27	0.28	**0.18**	**-0.70**	0.94	1.07	0.93	1.06	0.41	0.34	1.08	0.99
ESP15	0.93	0.78	0.22	0.41	**-0.37**	**0.17**	0.92	0.92	0.90	0.91	0.34	0.49	1.02	1.03
ESP16	0.80	0.77	**0.14**	0.31	-0.34	-0.24	1.03	0.96	1.06	0.98	0.31	0.38	0.97	1.02
ESP17	0.89	0.79	**0.19**	0.39	-0.51	-0.38	1.00	0.90	0.94	0.84	0.29	0.48	1.00	1.09
ESP18	0.67	0.60	**0.10**	0.29	0.13	0.37	1.12	0.96	1.16	0.95	0.27	0.43	**0.78**	1.12
ESP19	0.74	0.76	**0.18**	0.35	-0.02	-0.06	0.99	0.96	1.06	0.96	0.38	0.45	1.02	1.03
ESP20	**0.59**	**0.92**	0.23	0.43	**0.30**	**-0.82**	0.97	0.91	0.97	0.78	0.32	0.46	1.03	1.03
Prom	0.71	0.69	0.21	0.30	0.00	0.00	1.00	1.00	1.00	1.00	0.33	0.39	1.00	0.98

Nota: pbis = índice de correlación punto biserial. pmed = índice de correlación punto medida. Prom = promedio.

Los análisis estadísticos desde la TCT se hicieron sobre 294 datos de VA y 270 de VB, después de eliminar los casos perdidos. Ambas distribuciones se aproximaron a la de una normal (simétricas y mesocúrticas). El promedio de dificultad fue de 0.710 para VA y de 0.694 para VB, con una desviación estándar de 0.19 y 0.12, respectivamente. En algunos casos las dificultades resultaron parecidas para ítems-hijo de igual familia (e.g.: ESP03, ESP14, ESP19). Los ítems con mayores diferencias entre ítems-hermano se observaron en: ESP02, ESP20, ESP01, ESP13 y ESP10 (listados de mayor a menor), todas mayores que 0.20 y menores que 0.36. El ítem más

complejo fue el ESP02 de VA. También hubo muchos reactivos elementales, que superaron el 80% de respuestas correctas, particularmente en VA (siete reactivos).

La dificultad, analizada a través del mapa de Wright, mostró en ambos casos, que la media de dificultad de los ítems fue una desviación menor que la media de la habilidad de los examinados, además el rango de dificultad se encontró entre -1 y 1, lo que significa que los reactivos resultaron relativamente fáciles. Una excepción fue ESP02 de VA, cuya dificultad se acercó a 2.5 (esto no ocurrió para el mismo ítem, de VB). En las dos pruebas quedaron estudiantes sin ítems que evaluaran su habilidad. Las mayores diferencias entre reactivos que pertenecían a la misma familia se identificaron en ESP02, ESP20 y ESP14.

Al igual que HV, el área de Español no mostró problemas de ajuste. Tanto el *infit* como el *outfit* de todos los ítems estuvieron dentro del rango de aceptación. Del mismo modo, los índices de discriminación por ítem fueron buenos, salvo dos valores: ESP02 y ESP20, de VB, que estuvieron ligeramente por debajo de 0.8 (0.58 y 0.78, respectivamente).

En cuanto a la correlación biserial, ocho reactivos de VA y dos de VB no alcanzaron el mínimo de 0.20. Los índices menores de encontraron en ESP02, cuyos valores fueron inferiores a 0.10. Las bajas correlaciones se vieron reflejadas en la confiabilidad, particularmente en VA (Alpha de Cronbach: VA = 0.587, VB = 0.706). El cálculo de la correlación punto medida fue menos riguroso, solamente ESP01 no superó el mínimo en VA, y ESP02 quedó en el límite para VA y por debajo en VB.

Para el AFC, se armó un modelo de un factor denominado *Conocimientos de español de la educación secundaria*. Los análisis arrojaron índices de ajuste similares y aceptables para ambas versiones. Las cargas factoriales fueron apropiadas para la mayoría de los ítems, salvo para ESP02 (tanto para VA como para VB) y para ESP01 y ESP18 de VA.

Los resultados indican que el área de Español fue relativamente fácil, lo cual indica la necesidad de ítems más complejos que permitan evaluar competencias superiores. Además, se encontraron ítems-hermano distantes en dificultad (ESP02 y ESP20). Si bien las distribuciones fueron similares, en el cálculo de los índices, existió una diferencia, a favor de VB, especialmente desde la TCT y no tanto en el modelo de Rasch. Probablemente, esto se deba al número de casos perdidos que no pudieron utilizarse para el cálculo de la correlación punto biserial de todos los reactivos. Por lo anterior se infiere que, según la TCT, las versiones tienen algunos problemas de isomorfismos; mientras que desde la TRI, estas diferencias no son tan marcadas.

Se obtuvo un mal desempeño de los ítems-hijo de ESP02, por lo tanto, se recomienda revisar su especificación con la plantilla y replantear los ejercicios. También se sugiere examinar ESP01 de VA y ESP18 de VB, y comparar con sus ítems-hermano de las otras versiones, para analizar posibles causas de sus deficiencias psicométricas.

4.1.2.4. Matemáticas

El área de Matemáticas (ver figura 4.18), correspondiente a la educación secundaria, contó con información de 19 de los 20 ítems evaluados, debido a que no se pudieron recuperar los datos de los reactivos de MAT14.

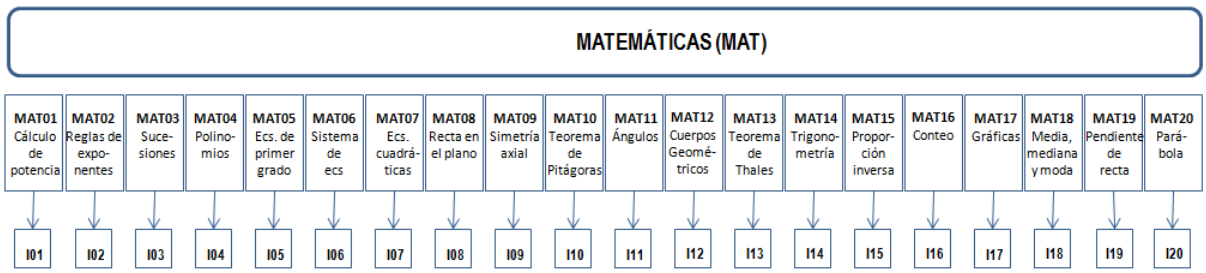


Figura 4.18. Esquema de contenidos del área de Matemáticas (VA y VB). Ecn = ecuación.

En la tabla 4.11 se sintetizan los cálculos efectuados a través de la TCT y del modelo de Rasch. La información en detalle se encuentra en el Anexo D, en el apartado de Matemáticas.

Tabla 4.11
Índices calculados a través de la TCT y de la TRI (modelo de Rasch) para los ítems de MAT de VA y VB

	TCT				TRI									
Item	dificultad		pbis		medida		Infit		outfit		pmed		Discriminación	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
MAT01	0.03	0.05	0.27	0.24	**2.22**	**1.48**	0.87	0.91	0.43	1.14	0.28	0.26	1.08	1.02
MAT02	0.03	0.03	0.32	0.30	2.22	2.11	0.89	0.86	0.36	0.31	0.28	0.32	1.08	1.09
MAT03	0.06	0.05	0.23	0.25	1.33	1.33	0.96	0.96	0.81	0.94	0.27	0.28	1.03	1.02
MAT04	0.05	0.01	0.22	0.23	**1.54**	**3.16**	0.99	0.89	0.81	0.47	0.25	0.21	1.03	1.06
MAT05	0.37	0.26	0.28	0.39	-1.27	-0.81	1.03	0.93	1.12	0.92	0.39	0.47	0.89	1.12
MAT06	0.15	0.13	0.26	0.36	0.18	0.25	0.97	0.89	1.23	0.88	0.33	0.41	1.00	1.08
MAT07	0.11	0.15	0.22	**0.08**	**0.59**	**-0.01**	0.99	0.98	1.03	1.13	0.29	0.36	0.99	1.00
MAT08	0.06	0.26	0.24	0.27	**1.43**	**-0.56**	0.98	1.09	0.65	1.14	0.28	0.44	1.04	0.85
MAT09	**0.92**	**0.60**	0.27	0.26	**-4.17**	**-1.84**	0.97	0.82	**2.31**	0.86	0.55	0.50	0.92	1.02
MAT10	0.04	0.06	0.22	0.39	**1.66**	**1.12**	0.93	0.87	**2.95**	0.43	0.21	0.40	1.00	1.11
MAT11	0.45	0.46	0.42	0.45	**-1.68**	**1.01**	0.87	0.85	0.83	**0.78**	0.54	0.57	1.37	1.32
MAT12	0.77	0.65	0.35	0.32	-3.03	-2.91	0.79	0.98	**2.94**	1.03	0.66	0.57	1.01	0.99
MAT13	**0.72**	**0.51**	0.37	0.31	**-3.18**	**-2.13**	0.92	0.96	0.87	0.98	0.52	0.47	1.14	1.10
MAT14														
MAT15	0.10	0.17	0.21	0.31	**0.74**	**-0.24**	0.99	1.01	0.80	1.10	0.31	0.37	1.03	0.99
MAT16	0.14	0.10	**0.16**	0.31	0.26	0.52	1.11	0.97	1.09	0.77	0.26	0.35	0.92	1.04
MAT17	0.69	0.74	**0.18**	0.25	-2.98	-3.36	1.11	1.01	1.30	1.03	0.37	0.39	**0.76**	0.98
MAT18	0.13	0.20	0.38	**0.16**	0.43	0.43	0.94	1.24	1.10	**1.54**	0.50	0.34	1.03	**0.61**
MAT19	0.16	0.00	**0.10**		0.11	-- [a]	1.14		1.28		0.23		0.85	
MAT20	0.01	0.17	**0.19**	0.22	**3.60**	**0.44**	0.99	1.12	0.20	1.06	**0.17**	0.38	1.05	0.86
Prom	0.26	0.24	0.26	0.27	0.00	0.00	0.97	0.91	1.16	0.87	0.35	0.37	1.01	0.96

Nota: pbis = índice de correlación punto biserial. pmed = índice de correlación punto medida. Prom = promedio.
[a] No se pudo estimar debido a que no hubo respuestas correctas.

Para los análisis a través de la TCT, se utilizaron 396 datos de VA y 296 de VB, previa eliminación de los casos perdidos. Ambas distribuciones resultaron leptocúrticas y aproximadamente simétricas, con una ligera cola hacia la derecha.

El promedio de dificultad del área fue de 0.261 para VA y de 0.242 para VB, con una desviación estándar de 0.29 y 0.23, respectivamente. Salvo tres ítems, los restantes presentaron dificultades inferiores a 0.60; y en algunos casos no llegó a 0.10 (e.g.: MAT01, MAT02, MAT03, MAT04 y el caso particular, MAT19 de VB, que no tuvo aciertos). En general, las dificultades de ítems-hermano resultaron similares en ambas versiones. Las mayores diferencias se encontraron en MAT09, MAT13 y MAT08 (mencionadas de manera decreciente), las tres entre 0.2 y 0.3.

De la información obtenida de los mapas de Wright (Anexo D, apartado de matemáticas), se infiere que la media de dificultad de los ítems fue mayor (en más de una desviación estándar) que la media de la habilidad de los examinados. Tanto en VA como en VB, menos del 10% tuvo la habilidad suficiente para contestar correctamente los reactivos MAT01, MAT02, MAT03, MAT04, MAT07, MAT10, MAT16 y MAT18. También se registraron ítems sencillos en ambos tests, tal es el caso de MAT09, MAT11, MAT13 y MAT17 con menos de -1 lógitos de dificultad. Los ítems-hermano más diferentes en cuanto a dificultad fueron: MAT20, MAT11, MAT09, MAT08, MAT04 y MAT13.

A diferencia de las áreas HV y ESP, en esta sección aparecieron algunos problemas de ajuste, solamente de *outfit*, lo cual indica que los desajustes se encontraron lejos de la zona de medición del ítem. En general, estos estadísticos se localizaron por encima de 1.3, lo que revela demasiada aleatoriedad en las respuestas. Estas desventajas se detectaron en ítems de crédito parcial, donde se solicita ubicar información en categorías y no, donde la respuesta es abierta. En cuanto a la discriminación, los índices fueron aceptables, con leves deficiencias en MAT17 para VA y MAT18 para VB (0.76 y 0.61, respectivamente).

En general, los ítems superaron el mínimo índice aceptable de correlación punto biserial; con la excepción de cuatro reactivos de VA y dos de VB (no fueron los mismos ítems en ambas versiones). Estos resultados se reflejaron en la confiabilidad de ambas pruebas (Alpha de Cronbach: VA = 0.655, VB = 0.686). Cabe aclarar que los problemas de correlación no persistieron en el modelo de Rasch, solamente se identificó un valor ligeramente menor (0.17), para la correlación punto medida de MAT20 de VA.

A través del AFC se evaluaron modelos diferentes, según la organización temática del plan de estudios de la materia. El primero de ellos, unidimensional reportó índices de ajuste insuficientes, por lo tanto se construyó un nuevo modelo de 3 factores: (1) sentido numérico y pensamiento algebraico, (2) forma, espacio y medida, y (3) manejo de la información); pero no mejoraron los ajustes. Se optó por unir el primer y el tercer factor, y así formar un nuevo modelo de dos factores. Esto proporcionó mejores resultados; sin embargo, todavía quedaron reactivos que no aportaron a los constructos, tal es el caso de MAT16 y MAT19 para VA y MAT07 para VB.

Tras el resumen de resultados, una característica que destaca es la alta dificultad de los ítems del área de Matemáticas. Esta particularidad, aunada a la escasa cantidad de datos que no ayudaron a estudiar las colas de la distribución, no permitió obtener un diagnóstico muy seguro. Con esta precaución, se infiere que, en general, las dos pruebas tienen características similares (distribuciones, promedios de dificultad, de correlaciones y de discriminación, e índices de confiabilidad). Las correlaciones fueron aceptables y los problemas de ajuste, escasos. Sin embargo, existieron desigualdades en algunos ítems-hijo, en ambas versiones. Se sugiere revisar el ítem MAT20 de VA, compararlo con el de VB y analizar las diferencias que ocasionaron poca correlación en uno de los reactivos. Lo mismo se aconseja para el MAT07 y el MAT19 de VB.

Para los ítems MAT09 y MAT12 se propone inspeccionar los elementos asociados a cada categoría para comprobar si pudieran prestarse a la adivinación.

4.1.2.5. Ciencias naturales

Para el área de Ciencias naturales, correspondiente a la educación secundaria, se contó con la información de los 20 ítems distribuidos en: 6 reactivos Biología, 6 de Física, 7 de Química y 1 de método científico (ver figura 4.19).

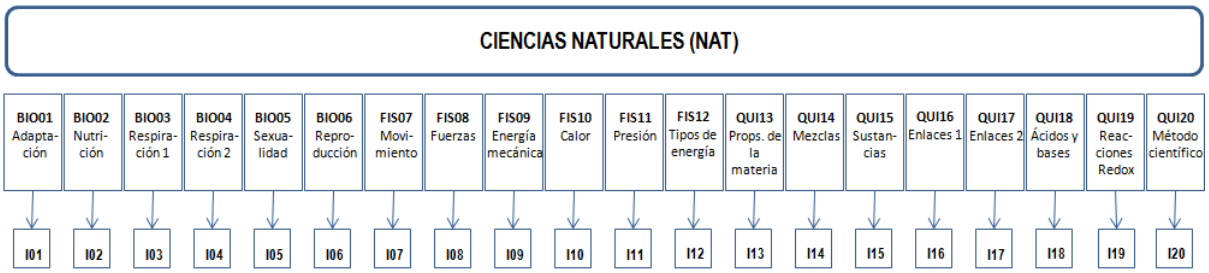


Figura 4.19. Esquema de contenidos del área de Ciencias naturales (VA y VB). Props. = Propiedades.

En la tabla 4.12 se presenta un resumen de los estadísticos calculados y se resaltan aquellos valores que no aportan a la calidad de los ítems o al isomorfismo entre ítems-hermano. La lista de tablas y figuras asociadas a estos resultados se puede consultar en el Anexo D, en el apartado de Ciencias naturales.

Tabla 4.12.
Índices calculados a través de la TCT y de la TRI (modelo de Rasch) para los ítems de NAT de VA y VB

Item	TCT				TRI									
	dificultad		pbis		medida		Infit		outfit		pmed		Discriminación	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
BIO01	0.50	0.36	**0.14**	**0.09**	-0.24	-0.41	1.09	1.06	1.11	1.07	0.37	0.33	0.86	0.95
BIO02	0.62	0.70	**0.17**	0.21	-0.73	-0.69	1.03	0.98	1.05	0.93	0.43	0.49	0.96	1.01
BIO03	0.56	0.60	**0.18**	0.26	-0.35	-0.43	1.03	0.95	1.02	0.95	0.42	0.47	0.97	1.05
BIO04	0.65	0.45	0.28	**0.04**	**-0.61**	**0.08**	0.96	1.04	0.94	1.06	0.49	0.37	1.05	0.95
BIO05	0.70	0.68	0.22	0.23	-0.83	-0.76	1.00	1.01	0.96	0.98	0.48	0.46	0.98	1.01
BIO06	0.52	0.50	0.26	**0.14**	-0.24	-0.43	0.96	1.03	0.94	1.03	0.45	0.38	1.03	0.96
FIS07	0.34	0.37	0.31	**0.14**	0.43	0.41	0.92	1.04	0.91	1.04	0.34	0.20	1.17	0.94
FIS08	0.62	0.78	0.35	**0.13**	-0.57	-0.88	0.86	1.03	0.86	0.95	0.56	0.49	1.07	0.97
FIS09	0.27	0.17	0.23	0.33	0.52	0.35	0.95	0.93	0.96	0.91	0.28	0.28	1.07	1.28
FIS10	0.27	0.39	0.28	**0.17**	0.61	0.27	0.96	0.97	0.94	0.96	0.23	0.22	1.10	1.16
FIS11	0.01	0.03	**0.16**	0.20	**4.07**	**2.78**	0.99	0.97	0.80	0.79	**0.09**	**0.15**	1.01	1.03
FIS12	0.52	0.60	0.27	0.23	**-0.14**	**-0.70**	0.98	0.97	0.98	0.97	0.41	0.39	1.04	1.05
QUI13	**0.30**	**0.80**	**-0.05**	0.25	**0.67**	**-1.11**	1.06	0.97	1.08	0.91	**0.14**	0.38	0.88	1.04
QUI14	**0.60**	**0.24**	0.27	**-0.10**	**-0.49**	**0.81**	0.97	1.15	0.96	1.16	0.47	**0.18**	1.07	**0.74**
QUI15	0.65	0.75	0.40	0.20	-0.57	-0.93	0.94	0.98	0.90	1.00	0.51	0.48	1.09	1.02
QUI16	**0.34**	**0.56**	**0.12**	**0.14**	0.02	0.04	1.12	0.93	1.22	0.92	0.31	0.46	0.86	1.08
QUI17	0.38	0.20	0.23	0.24	**0.00**	**0.63**	0.99	0.91	0.97	0.85	0.28	0.27	1.04	1.06
QUI18	0.56	0.45	0.28	**0.11**	-0.46	-0.12	0.96	1.04	0.96	1.06	0.47	0.37	1.08	0.96
QUI19	0.58	0.58	**0.15**	**0.12**	-0.44	-0.48	1.25	1.03	**1.86**	1.21	0.34	0.40	**0.75**	0.98
QUI20	**0.57**	**0.15**	0.22	**0.04**	**-0.63**	**1.56**	0.97	1.01	0.97	1.00	0.26	**0.11**	1.26	0.99
Prom	0.48	0.47	0.22	**0.16**	0.00	0.00	1.00	1.00	1.02	0.99	0.37	0.34	1.02	1.01

Nota: Pbis = correlación punto biserial. Pmed = correlación punto medida. Prom = promedio.

Para los análisis estadísticos a través de la TCT, se utilizaron 289 datos de VA y 216 de VB. Ambas distribuciones resultaron simétricas y ligeramente leptocúrticas. El promedio de dificultad fue de 0.478 para VA y de 0.468 para VB, con una desviación estándar de 0.31 y 0.27, respectivamente. En algunos casos, las dificultades fueron casi iguales para ítems-hermano (e.g.: BIO05, BIO06, FIS07). Las mayores diferencias se dieron en: QUI13, QUI20, QUI14 y QUI16, en ese orden de mayor a menor, en todos los casos superior a 0.20 (la mayor de 0.50). El ítem

más complejo fue el FIS11, en ambas versiones; su dificultad fue cercana a cero. Solamente hubo dos ítems fáciles, próximos al 80% de respuestas correctas, todos pertenecientes a VB.

En los mapas de Wright de ambas áreas se observa que la media de dificultad de los ítems fue mayor que la media de habilidad de los examinados (aproximadamente una desviación estándar más difíciles los reactivos que las medidas de las habilidades de los estudiantes). Tanto en VA como en VB, FIS11 estuvo fuera del alcance de los examinados; en VB también lo fue QUI20. Salvo esas dos excepciones, el resto de los ítems se localizaron entre -1 y 1 lógitos de dificultad. Los ítems-hijo más distantes en dificultad, al igual que en la TCT, resultaron QUI13, QUI20 y QUI14.

De acuerdo con el modelo de Rasch, en general, no se manifestaron problemas de ajuste. Los valores de *infit* quedaron dentro del rango establecido; algo similar ocurrió para los *outfit*, solamente QUI19 de VA superó 1.3. Los índices de discriminación señalaron ligeras deficiencias en QUI14 de VB (0.75) y QUI19 de VA (0.74).

La mayor debilidad de esta área se detectó en los índices de correlación punto biserial, especialmente en VB, donde el índice promedio no llegó al mínimo requerido de 0.20. Los casos más conflictivos fueron QUI13 de VA y QUI14 de VB (con registros negativos). Esta situación también se manifestó en la confiabilidad de ambas pruebas (Alpha de Cronbach de VA = 0.612, de VB = 0.502). Como en Ciencias naturales se evalúan tres materias, también se calcularon las correlaciones agrupadas por asignatura, para cotejar si de esta forma mejoraban los resultados. De acuerdo con esta clasificación, los coeficientes aumentaron, en general. En Biología solamente BIO06 de VB no llegó al mínimo de 0.2. En Física, se mantuvieron los mismos valores. En Química persistieron las deficiencias en QUI13 de VA, y en QUI14 y QUI20 de VB; aunque ya no resultaron negativas. La correlación punto medida que arrojó el análisis de Rasch

reafirmó los problemas en FIS11 (en este caso en ambas versiones), QUI13 de VA, y QUI14 y QUI20 de VB.

Según la organización temática del examen, para el AFC se construyeron tres modelos: modelo 1: unidimensional, modelo 2: de dos factores (uno donde se agruparon Biología y Química, y el otro, con Física) y modelo 3: de tres factores (Biología, Física y Química). VB no consiguió buenos índices de ajuste incremental en ninguno de los casos, esto se explica con la baja confiabilidad del área. En el caso de VA, el modelo de tres factores no se pudo ejecutar, debido a que las iteraciones no convergieron. De los resultados obtenidos, el modelo 2 es el que mejor se ajustó a la agrupación de los ítems; aunque con algunas deficiencias, especialmente en la asignatura de Química.

Estos resultados, con el recaudo de que solamente se pudo obtener información de dos tercios de la muestra (el resto fueron casos perdidos), sugieren que el área de Ciencias naturales necesita consolidarse como tal. Si bien las medias de las dificultades de VA y VB son parecidas, ambas pruebas no se comportaron de manera similar, puesto que se reflejaron problemas de isomorfismo, en cuanto al constructo, dentro de las familias de algunos ítems-hijo y del área, en general (e.g.: correlaciones pequeñas de algunos ítems, correlación punto biserial media de VB muy baja). De las tres asignaturas, Química se percibe como la más débil, en términos estadísticos. En particular, se propone revisar QUI13 de VA y cotejar con su homólogo de VB, lo mismo para QUI14 de VB, con su similar de VA. Otros reactivos, con fallas menos acentuadas, que necesitan examinarse son los de las familias de FIS11 y QUI20.

4.1.2.6. Ciencias Sociales

Para el área de Ciencias sociales de educación secundaria pudieron analizarse 19 de los 20 ítems correspondientes, debido a que, por problemas de diseño, no se administraron los ítems-hijo que evaluaban la competencia HIS06. Otra aclaración necesaria se refiere a la numeración de los reactivos. Al momento de elaborar las especificaciones, se presentaron serias dificultades conceptuales en el contenido FCYE15, por lo que el Comité Técnico del EXHCOBA decidió eliminar esta competencia y agregar una en Geografía (que se denominó GEO06); esta es la razón por la que en los contenidos de Ciencias sociales se repite el número 6 (uno para geografía y otro para historia) y falta el 15 (ver figura 4.20).

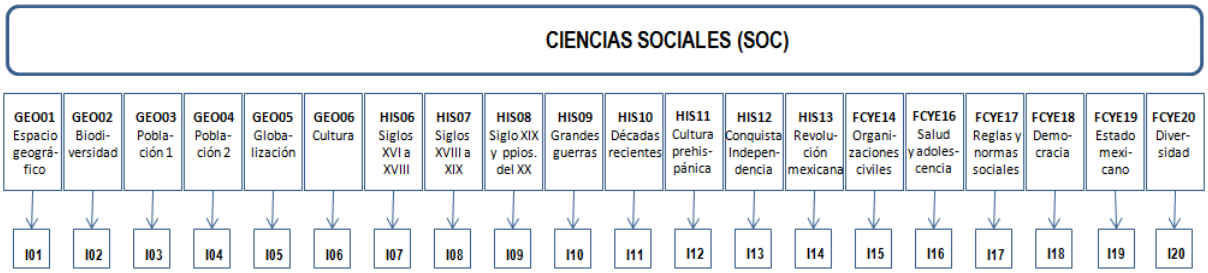


Figura 4.20. Esquema de contenidos del área de Ciencias sociales (VA y VB). ppios. = principios.

Un resumen de los resultados numéricos de los análisis psicométricos (TCT y TRI) se encuentran en la tabla 4.13. Para mayor información, en el anexo D, en el apartado de Ciencias sociales se encuentran todas las figuras y tablas de los estudios de esta área.

Tabla 4.13.
Índices calculados a través de la TCT y de la TRI (modelo de Rasch) para los ítems de SOC de VA y VB

Item	TCT				TRI									
	dificultad		pbis		medida		Infit		outfit		pmed		Discriminación	
	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB	VA	VB
GEO01	0.17	0.10	0.33	0.24	0.81	1.23	0.95	1.04	0.89	1.05	0.30	0.22	1.04	0.99
GEO02	0.42	0.52	0.48	0.44	0.13	0.04	1.03	1.16	1.02	1.16	0.55	0.51	0.99	0.85
GEO03	0.68	0.83	0.43	0.49	**-0.72**	**-1.26**	1.12	1.10	1.10	1.05	0.49	0.59	0.91	0.98
GEO04	0.60	0.69	0.53	0.35	-0.32	-0.71	1.08	**1.45**	1.08	**1.58**	0.59	0.49	0.97	**0.51**
GEO05	0.47	0.45	0.43	0.44	-0.05	0.34	1.11	1.12	1.22	1.12	0.47	0.47	0.81	0.85
GEO06	0.41	0.39	0.45	0.40	0.52	0.21	0.95	0.96	1.02	1.00	0.44	0.41	1.08	1.02
HIS06 [a]														
HIS07	0.41	0.24	0.56	0.39	0.72	0.69	0.86	1.00	0.84	1.05	0.52	0.39	1.25	0.99
HIS08	0.38	0.37	0.53	0.41	-0.13	0.05	0.91	1.07	0.90	1.10	0.54	0.45	1.14	0.86
HIS09	0.59	0.53	0.47	0.51	**-0.49**	**0.29**	1.11	1.04	1.13	1.05	0.52	0.55	0.81	0.98
HIS10	0.41	0.41	0.54	0.53	0.02	0.20	0.87	0.93	0.84	0.93	0.51	0.51	1.25	1.10
HIS11	**0.32**	**0.55**	0.43	0.56	**0.51**	**-0.31**	1.00	0.94	0.98	0.93	0.45	0.56	1.01	1.10
HIS12	0.59	0.58	0.53	0.65	-0.53	-0.31	1.01	0.81	1.00	0.79	0.55	0.59	1.01	1.33
HIS13	0.38	0.36	0.49	0.44	0.55	0.09	0.96	1.04	0.94	1.05	0.50	0.46	1.10	0.94
FCYE14	0.46	0.53	0.52	0.55	-0.40	-0.08	0.94	0.96	0.93	0.96	0.54	0.55	1.10	1.05
FCYE16	0.58	0.67	0.60	0.64	-0.44	-0.50	0.89	0.83	0.89	**0.76**	0.60	0.62	1.15	1.17
FCYE17	0.52	0.60	0.42	0.64	-0.12	-0.33	1.13	0.82	1.17	0.80	0.49	0.60	0.81	1.25
FCYE18	0.59	0.48	0.52	0.50	-0.40	-0.15	1.01	1.06	0.99	1.14	0.57	0.49	0.97	0.92
FCYE19	0.25	0.34	0.31	0.50	0.96	0.58	1.13	0.96	1.19	0.96	0.34	0.48	0.83	1.08
FCYE20	0.66	0.51	0.51	0.57	**-0.61**	**-0.07**	1.10	0.91	1.17	0.92	0.56	0.55	0.91	1.15
Prom	0.47	0.48	0.48	0.49	0.00	0.00	1.01	1.01	1.02	1.02	0.50	0.50	1.01	1.01

Nota: Pbis = correlación punto biserial. Pmed = correlación punto medida. Prom = promedio. [a] Sin datos.

Para los análisis a través de la TCT se utilizaron 397 datos de VA y 298 de VB, casi no se registraron casos perdidos. Ambas distribuciones presentaron una tendencia a ser platocúrticas, con una acentuada cola hacia la izquierda. El promedio de dificultad de los 19 ítems analizados fue de 0.468 para VA y de 0.482 para VB, con una desviación estándar de 0.28 y 0.29, respectivamente. Se observó una semejanza entre las dificultades de ítems-hermano. Los reactivos con mayor diferencia fueron los de HIS11, con 0.23.

Los resultados de los análisis desde la TRI también reflejaron un comportamiento similar en ambas versiones. En los mapas de Wright (ver anexo D, apartado de Ciencias sociales) se observa que, la media de dificultad de los ítems fue ligeramente mayor que la media de habilidad de los estudiantes examinados. En general, las dificultades de VA se correspondieron en VB, siendo un poco más amplio el rango de esta última versión (desde -1.5 a 1.5 lógitos). Al igual que en la TCT, HIS11 fue la familia cuya diferencia de dificultad fue mayor.

De acuerdo con la tabla 4.13, solamente se detectaron algunos problemas de ajuste en GEO04 de VB (*infit* = 1.45, *outfit* = 1.58 y discriminación = 0.51). Estos valores indican demasiada aleatoriedad cerca y lejos de la zona de medición del ítem. El *outfit* de FCYE16 de VB fue 0.76, ligeramente inferior a lo aceptado. Las correlaciones punto biserial y punto medida fueron aceptables para todos los reactivos en ambas versiones, en todos los casos superaron el mínimo de 0.20. Además, los promedios de los dos tipos de correlaciones resultaron altos y similares tanto en VA como en VB. Otro dato favorable es el de una buena consistencia interna, que se reflejó en la confiabilidad del área (Alpha de Cronbach, VA = 0.869 y VB = 0.877).

Para el AFC se construyeron dos modelos. En el modelo 1 se agruparon los ítems en un factor, Ciencias sociales, donde los errores covariaron por materia. El modelo 2 se organizó en tres factores (geografía, historia, y formación cívica y ética), con covarianza de factores y de errores. De acuerdo con los índices de ajuste, el modelo 1 presentó valores ligeramente mejores, en los dos exámenes; sin embargo, las cargas factoriales fueron superiores en la agrupación de tres factores, sobre todo para geografía, y formación cívica y ética. En todos los casos, se observó una ligera desventaja, aunque los números fueron aceptables, de GEO01 para ambas versiones.

Los análisis estadísticos efectuados al área de Ciencias sociales ofrecen evidencias de una

buena calidad de los reactivos y de un isomorfismo entre ítems-hermano; esto se manifiesta en dos versiones similares del área (en dificultad, ajuste, correlación y confiabilidad) y con buenas propiedades psicométricas, en general.

Después de la lectura e interpretación de los resultados obtenidos a través de la TCT, del modelo de Rasch y del AFC, se elaboró la tabla 4.14 donde se resumen los ítems con problemas para cada área. Las consideraciones que se tomaron en cuenta fueron: (a) la diferencia de dificultades entre ítems-hijo de una misma familia, mayor que 0.2 (en el caso de p) y, a su vez, la diferencia de medidas mayor que 1 lógito, (b) los índices de ajuste (*infit* o *outfit*) fuera del rango de aceptación, (c) el índice de correlación punto biserial menor que 0.10, (d) el índice de correlación punto medida menor que 0.15, (e) el índice de discriminación según Rasch, menor que 0.60 y (f) las cargas factoriales menores que 0.20.

Tabla 4.14.

Resumen de los ítems con deficiencias psicométricas análisis estadísticos por área, de las versiones A y B

		VA					Dific.	VB					
	Ítem	Pb	Pm	Aj	Dis	E.I.	VA - VB	Pb	Pm	Aj	Dis	E.I.	Ítem
HV	HV02					x							
								x				x	HV03
												x	HV05
	HV07						x						HV07
	HV08					x							
	HV13					x							
	HV15	x	x			x	x	x	x		x	x	**HV15**
	HV16					x							
HC	HC01						x						HC01
	HC03			x						x			HC03
	HC07	x		x	x	x				x	x	x	**HC07**
	HC09						x	x [a]				x	HC09
	HC11												
ESP	ESP01	x [a]				x							
	ESP02	x				x	x	x	x		x	x	**ESP02**
	ESP18					x							
	ESP20						x						ESP20
MAT								x				x	MAT07
	MAT08						x						
	MAT09			x			x						
	MAT10			x									
										x			MAT11
	MAT12			x									
	MAT13						x						
	MAT16					x							
										x			MAT18
	MAT19					x							MAT19
	MAT20						x						MAT20
NAT	BIO01					x		x				x	BIO01
								x				x	BIO04
	FIS11		x										FIS11
	QUI13	x	x			x	x						
							x	x				x	**QUI14**
	QUI16					x							
												x	QUI18
	QUI19					x						x	QUI19
	QUI20						x	x				x	**QUI20**
SOC										x	x		GEO04
										x			FCYE16
		4	3	5	1	15	13	8	2	6	4	13	

Nota: Aj = ajuste (*infit*, *outfit*), Pb = índice de correlación punto biserial, Pm = índice de correlación punto medida, Dis = discriminación, E.I. = cargas factoriales del AFC, Dific VA – VB = Diferencia de dificultades entre ítems de VA y VB.

[a] Índice de correlación punto biserial en el límite de aceptación.

De acuerdo con la tabla 4.14, los ítems que mostraron serias deficiencias en ambas versiones fueron: HV15, HC07 y ESP02, por lo cual, se revisaron las tres especificaciones en busca de posibles causas. Se encontró que HV15 es un reactivo dicotómico de tres opciones donde los distractores resultan confusos en muchos casos. La revisión del HC07 permitió detectar que, a diferencia del resto de los ítems de HC que son de aplicación, este reactivo apela a la memoria, a reconocer los nombres de elementos en una circunferencia; por lo tanto, no se agrupa con el resto de los reactivos del área. En el caso de ESP02, se trata de un ítem dicotómico de cuatro opciones; en él se solicita reconocer la oración principal de un párrafo de cuatro oraciones. Hasta ese momento, el sistema permitía seleccionar más de una oración, lo cual iba a ser modificado para que solamente se pudiera marcar una opción. Acorde con esta intención de funcionamiento del ítem, se decidió anular los casos con más de una respuesta. Así, en VA se debieron cancelar 104 de 401 datos, en VB, 69 de 301. Probablemente, esta falla en el sistema también pudo alterar los resultados.

Los ítems QUI13 y QUI14 también revelaron más de dos problemas, aunque solamente en una de sus versiones. Sería apropiado comparar con sus ítems-hermano eficientes y analizar las causas del mal funcionamiento. Esta desventaja de unos ítems frente a sus similares parece repetirse en los demás casos. Por lo tanto, se recomienda revisar estas especificaciones en busca de diferencias y posibles errores.

En términos generales, se observa que la mayor debilidad de la prueba, analizada por áreas, se concentra en la correlación punto biserial, asociada estrechamente con las cargas factoriales del AFC. Los problemas de ajuste son menores, y en su mayoría de *outfit*, es decir en la zona de medición lejana a la dificultad del reactivo. Tampoco se observan inconvenientes severos de discriminación. Estos resultados aportan evidencias de que la prueba puede funcionar

bien como un examen normativo; pero que se necesitan afinar ciertas especificaciones con sus plantillas, para mejorar las propiedades de isomorfismo de los ítems de igual familia (dificultades similares y pertenencia al constructo) y así producir exámenes equivalentes, a través de la GAI.

4.2. Nivel de familia de ítems: análisis de las muestras HV, HC, ESP, MAT, NAT y SOC

A continuación, se introducen los resultados de las seis muestras tomadas por área: HV, HC, ESP, MAT, NAT y SOC. En primer lugar se exponen los hallazgos de los análisis efectuados a las seis muestras completas. En segundo lugar, se reporta el comportamiento de los 6 ítems-hijo de cada una de las 20 familias que conforman cada área. Finalmente, se exponen resultados de análisis psicométricos a elementos de algunos ítems de crédito parcial. Los dos primeros apartados se ilustran con tablas y figuras para la muestra HC; para el resto de las áreas, se realizan los comentarios pertinentes y los datos puntuales se incluyen en los Anexos E y F. En el tercer apartado, se presentan los resultados de las propiedades psicométricas de los elementos de una familia por área del examen, es decir, un total de seis familias.

4.2.1. De cada muestra completa del área

Estos resultados incluyen los análisis de cada muestra de 120 ítems efectuadas por área, con seis ítems-hijo por contenido. Se presenta una descripción general de cada área, cómo se distribuyeron los datos y las propiedades psicométricas de los ítems que componen las muestras.

4.2.1.1. Habilidades matemáticas (muestra HC)

El examen del área de Habilidades matemáticas (figura 4.21) fue aplicado en dos instituciones, primeramente en la Universidad Autónoma de Ciudad Juárez (UACJ), a 99 estudiantes, y posteriormente en la Universidad Estatal de Sonora, sede Hermosillo (CESUES, Hmo), con 56

evaluados. Se presentaron dificultades tanto, en la administración de los ítems como en la recuperación de la información, por lo tanto la cantidad de datos para efectuar los análisis fue reducida.

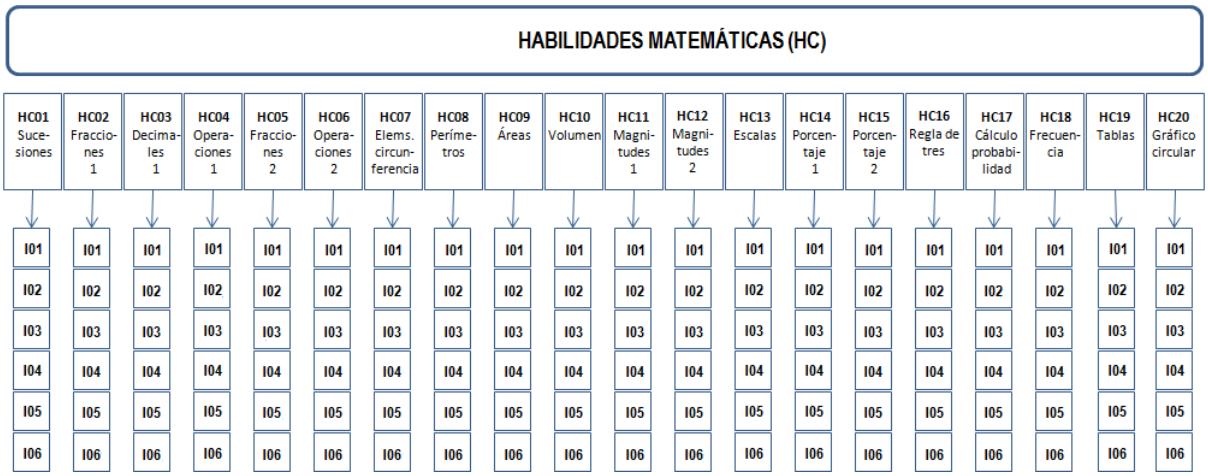
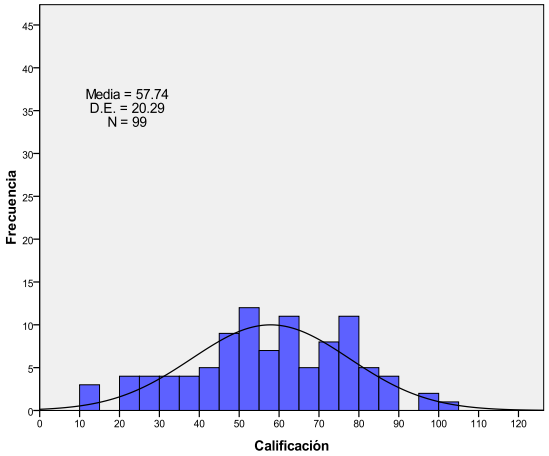


Figura 4.21. Esquema de la muestra HC.

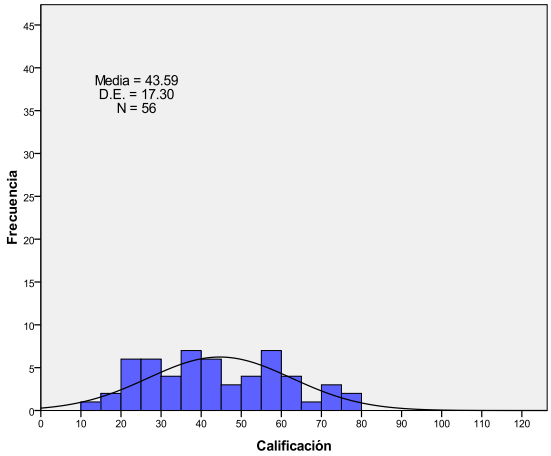
En la figura 4.22, se muestran las distribuciones de las dos aplicaciones, por separado, ya que al juntarlas la intersección de respuestas de todos los ítems daba vacía y no se podían obtener índices estadísticos. De acuerdo con la TCT, de la prueba aplicada en UACJ se obtuvo información de 99 examinados para un total de 102 ítems (no se registró información de seis reactivos de HC02, seis de HC07, tres de HC05 y tres de HC19). La dificultad media fue de 57.74 puntos sobre 102 (p = 0.57) con una desviación estándar de 20.29. El Alpha de Cronbach fue de 0.968. En el caso de CESUES, fueron 56 evaluados, para 117 reactivos (faltaron 3 ítems de HC05). La dificultad media fue de 43.59 sobre 117 (p = 0.37) y la desviación, 17.30. El Alpha de Cronbach resultó 0.943. Ambas distribuciones se aproximaron, dentro de los valores aceptados, a la de una curva normal (simétrica y mesocúrtica).

UACJ



Simetría	-.201
Error estándar de simetría	.243
Curtosis	-.476
Error estándar de Curtosis	.481
Mínimo	12.48
Máximo	102.00

CESUES



Simetría	.286
Error estándar de simetría	.319
Curtosis	-.799
Error estándar de Curtosis	.628
Mínimo	14.15
Máximo	79.46

Figura 4.22. Distribución de las calificaciones de las muestras HC de UACJ y CESUES, sede Hermosillo.

Para el modelamiento de Rasch, se consideró toda la población, ya que el programa *Winsteps* permitía estimar los estadísticos solicitados. De acuerdo con este modelo, la media de las dificultades de los ítems coincidió con la media de las habilidades de los estudiantes evaluados (ver figura 4.23). En general, tanto las habilidades de las personas como las dificultades de los ítems recorrieron un rango de -3 a 4 lógitos; aunque sería aconsejable tener más reactivos entre 0 y 0.5 lógitos, según la ausencia que marca la gráfica.

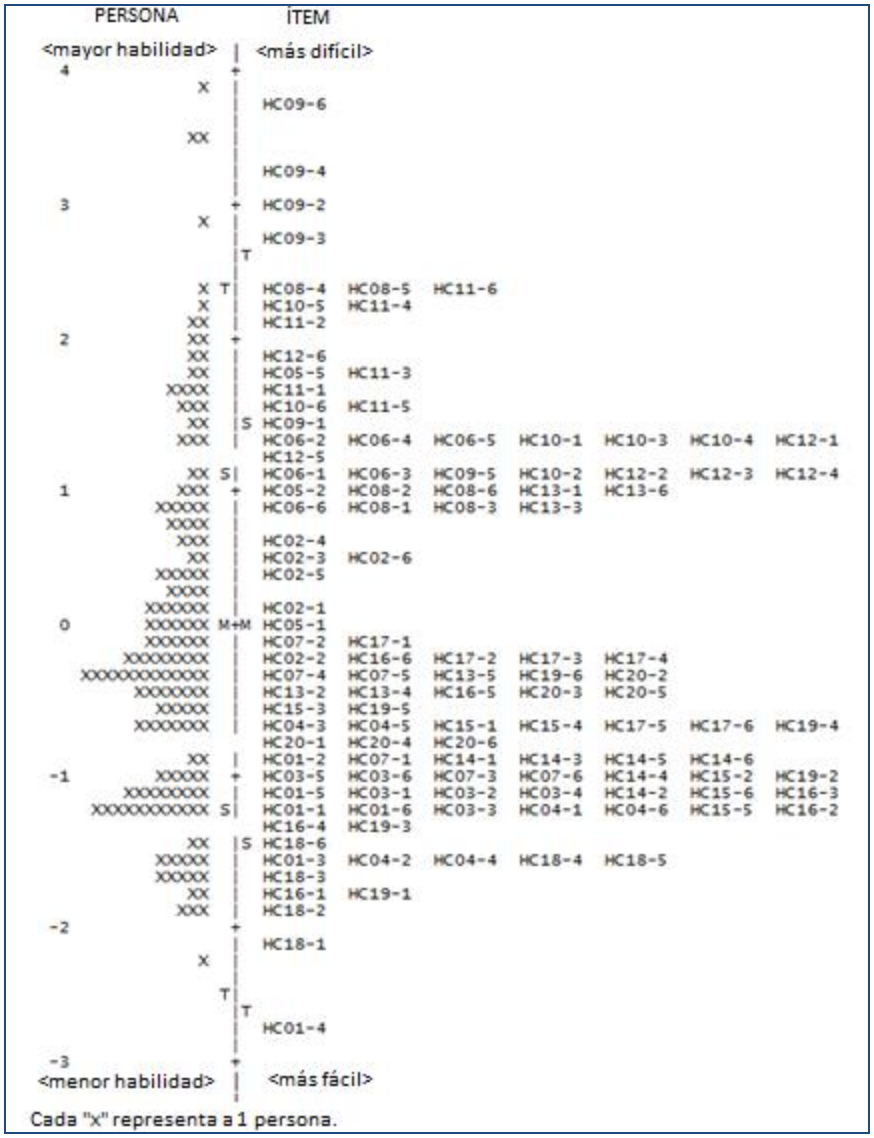


Figura 4.23. Mapa de Wright de la muestra HC aplicada a estudiantes de CESUES, sede Hermosillo, y de la UACJ.

De la información obtenida en la tabla 4.15 se desprende que las correlaciones de los ítems fueron altas; solamente dos de ellas resultaron menores que 0.20. Doce reactivos presentaron correlaciones entre 0.20 y 0.30 y el resto, superaron estos valores (los índices se agruparon entre 0.40 y 0.60). Estos números se reflejaron en un buen promedio, de 0.45.

Los problemas de ajuste se concentraron en dos familias: HC03 y HC07. En ambas, los valores de *infit* y de *outfit* sobrepasaron el rango de aceptación. Esto indica que, tanto cerca como

lejos de la medida de los ítems-hijo, se observó demasiada aleatoriedad. En general, estos desajustes por exceso estuvieron asociados a índices bajos de discriminación.

Un dato extraño es el comportamiento del sexto ítem de HC19, puesto que presentó correlación y discriminación negativas. Una posible causa de este resultado es una clave de calificación incorrecta.

Tabla 4.15.
Infit, outfit, correlación punto medida y discriminación según el modelo de Rasch para la muestra HC aplicada a CESUES (Hermosillo) y a Universidad Autónoma de Ciudad Juárez

IT	M	IN	OUT	PM	DIS	IT	M	IN	OUT	PM	DIS	IT	M	IN	OUT	PM	DIS
C1.1	-1.27	1.07	1.08	.32	0.90	C08.1	0.89	0.94	0.88	.53	1.11	C14.1	-0.83	0.92	0.97	.46	1.16
C1.2	-0.82	0.97	0.93	.43	1.05	C08.2	0.97	0.98	0.96	.49	1.03	C14.2	-1.08	0.84	0.77	.52	1.36
C1.3	-1.49	0.99	1.00	.35	1.00	C08.3	0.93	0.87	0.84	.57	1.20	C14.3	-0.93	0.82	0.74	.55	1.44
C1.4	-2.76	0.85	0.57	.34	1.11	C08.4	2.42	0.93	0.79	.47	1.07	C14.4	-0.97	0.83	1.26	.51	1.33
C1.5	-1.07	0.89	0.98	.44	1.12	C08.5	2.42	0.90	0.81	.48	1.08	C14.5	-0.90	**0.78**	**0.67**	.58	1.55
C1.6	-1.20	0.89	1.00	.44	1.12	C08.6	1.01	0.93	0.90	.52	1.10	C14.6	-0.93	**0.79**	0.67	.57	1.52
C2.1	0.11	0.73	0.50	.72	1.15	C09.1	1.39	1.16	**1.44**	.32	**0.75**	C15.1	-0.69	1.04	**1.49**	.36	**0.78**
C2.2	-0.28	1.07	1.05	.58	0.90	C09.2	3.05	1.17	1.29	.24	0.88	C15.2	-1.00	0.83	0.86	.52	1.36
C2.3	0.44	0.86	0.76	.59	1.12	C09.3	2.75	1.16	1.58	.27	0.87	C15.3	-0.63	0.81	0.74	.57	1.52
C2.4	0.59	0.71	0.76	.63	1.07	C09.4	3.28	1.08	0.98	.29	0.95	C15.4	-0.73	0.85	0.86	.53	1.36
C2.5	0.35	0.89	0.76	.67	1.07	C09.5	1.09	1.08	1.12	.41	0.87	C15.5	-1.22	0.82	0.68	.53	1.38
C2.6	0.51	**0.57**	0.48	.76	1.19	C09.6	3.72	0.88	0.34	.42	1.10	C15.6	-1.08	0.84	0.71	.52	1.38
C3.1	-1.17	**1.66**	**2.24**	.39	**0.48**	C10.1	1.30	0.82	0.71	.61	1.25	C16.1	-1.80	1.00	**1.88**	.28	0.92
C3.2	-1.13	**1.74**	**3.96**	.35	**0.56**	C10.2	1.14	0.86	0.91	.56	1.16	C16.2	-1.30	0.97	1.03	.39	1.04
C3.3	-1.20	**1.44**	1.66	.42	**0.79**	C10.3	1.21	0.85	0.90	.56	1.17	C16.3	-1.19	0.83	0.73	.52	1.37
C3.4	-1.11	**1.55**	**2.88**	.40	**0.65**	C10.4	1.26	0.79	0.76	.62	1.27	C16.4	-1.30	0.99	0.99	.38	1.02
C3.5	-1.02	**1.55**	**9.90**	.40	**0.46**	C10.5	2.21	0.92	0.89	.47	1.06	C16.5	-0.46	0.92	0.84	.51	1.25
C3.6	-1.06	**1.59**	2.25	.42	**0.67**	C10.6	1.44	**0.74**	**0.60**	.65	1.32	C16.6	-0.30	0.97	0.93	.48	1.10
C4.1	-1.22	1.04	1.00	.36	0.93	C11.1	1.58	1.00	1.07	.45	0.99	C17.1	-0.14	0.91	0.86	.53	1.25
C4.2	-1.46	1.10	1.04	.29	0.85	C11.2	2.14	0.97	1.48	.39	0.95	C17.2	-0.20	0.92	0.87	.52	1.23
C4.3	-0.73	1.04	0.99	.40	0.92	C11.3	1.79	0.85	0.99	.53	1.12	C17.3	-0.23	0.88	0.83	.54	1.33
C4.4	-1.50	1.05	1.46	.28	0.82	C11.4	2.28	0.97	1.12	.41	0.99	C17.4	-0.29	0.92	0.88	.51	1.22
C4.5	-0.73	1.03	1.06	.40	0.91	C11.5	1.44	0.80	**0.70**	.61	1.24	C17.5	-0.73	0.80	**0.70**	.58	1.53
C4.6	-1.22	0.93	0.97	.43	1.12	C11.6	2.35	0.88	1.13	.47	1.07	C17.6	-0.69	0.89	0.77	.52	1.34
C5.1	0.04	0.97	0.91	.49	1.10	C12.1	1.30	1.27	**1.61**	.25	**0.60**	C18.1	-2.10	0.99	**2.12**	.26	0.94
C5.2	0.96	0.98	0.96	.48	1.04	C12.2	1.17	1.16	1.30	.35	**0.76**	C18.2	-1.85	0.95	0.93	.36	1.06
S/D[a]						C12.3	1.09	**1.33**	**1.38**	.25	**0.55**	C18.3	-1.66	0.95	1.10	.36	1.03
S/D						C12.4	1.09	1.13	1.21	.37	0.80	C18.4	-1.50	0.89	0.78	.44	1.20
C5.5	1.80	1.01	1.10	.42	0.97	C12.5	1.26	1.14	1.13	.38	0.83	C18.5	-1.46	0.84	0.71	.49	1.29
S/D						C12.6	1.90	1.16	1.28	.32	0.83	C18.6	-1.34	0.89	0.82	.44	1.14
C6.1	1.17	0.81	0.70	.61	1.27	C13.1	0.97	1.30	**1.43**	.26	**0.53**	C19.1	-1.77	1.16	**1.73**	.20	**0.76**
C6.2	1.30	0.88	0.79	.56	1.17	C13.2	-0.46	1.10	1.12	.36	**0.73**	C19.2	-1.00	0.94	0.94	.46	1.10
C6.3	1.17	0.80	**0.65**	.63	1.30	C13.3	0.89	**1.32**	**1.38**	.25	**0.50**	C19.3	-1.29	1.00	1.16	.47	0.95
C6.4	1.21	0.92	0.83	.54	1.13	C13.4	-0.46	1.06	1.07	.39	0.83	C19.4	-0.76	1.24	1.26	**.18**	**0.45**
C6.5	1.26	0.90	0.73	.56	1.17	C13.5	-0.36	1.01	0.97	.44	0.99	C19.5	-0.67	0.96	0.95	.47	1.09
C6.6	0.93	0.86	0.82	.57	1.21	C13.6	0.97	1.16	**1.36**	.34	**0.72**	C19.6	-0.36	**1.53**	**1.60**	**-.03**	**-.21**
C7.1	-0.92	**1.34**	**2.12**	.42	**0.37**							C20.1	-0.76	0.85	0.80	.53	1.39
C7.2	-0.12	**1.35**	**1.36**	.47	**0.45**							C20.2	-0.39	**0.79**	**0.73**	.60	1.58
C7.3	-0.97	1.23	1.25	.44	**0.65**							C20.3	-0.56	0.82	0.80	.57	1.47
C7.4	-0.41	**1.32**	**1.31**	.48	**0.46**							C20.4	-0.73	**0.78**	0.71	.59	1.57
C7.5	-0.33	**1.38**	**1.99**	.44	**0.24**							C20.5	-0.53	**0.74**	**0.65**	.63	1.71
C7.6	-1.02	**1.37**	**1.54**	.42	**0.44**							C20.6	-0.69	**0.78**	**0.67**	.59	1.59

Nota: IT = Item, M = medida, IN = infit, OUT = outfit, PM = índice de correlación punto medida, DIS = índice de discriminación. C = HC.
[a] S/D = sin datos, no se pudo obtener información de tres ítems de HC05 (C05.3, C05.4 y C05.6) de las bases de datos originales.

4.2.1.2. Habilidades del lenguaje (muestra HV)

La muestra de Habilidades del lenguaje (figura 4.24) fue aplicada en la universidad de Guanajuato y consistió en 108 ítems, ya que no se incluyó la familia de HV01 (debido a fallas técnicas) y no se pudo calificar la de HV19 (por problemas de diseño), con lo cual se eliminaron 12 reactivos (seis de cada familia). El total de participantes fue de 167 estudiantes. Las tablas y figuras de los análisis estadísticos se encuentran en el Anexo E, en el apartado de Habilidades del lenguaje.

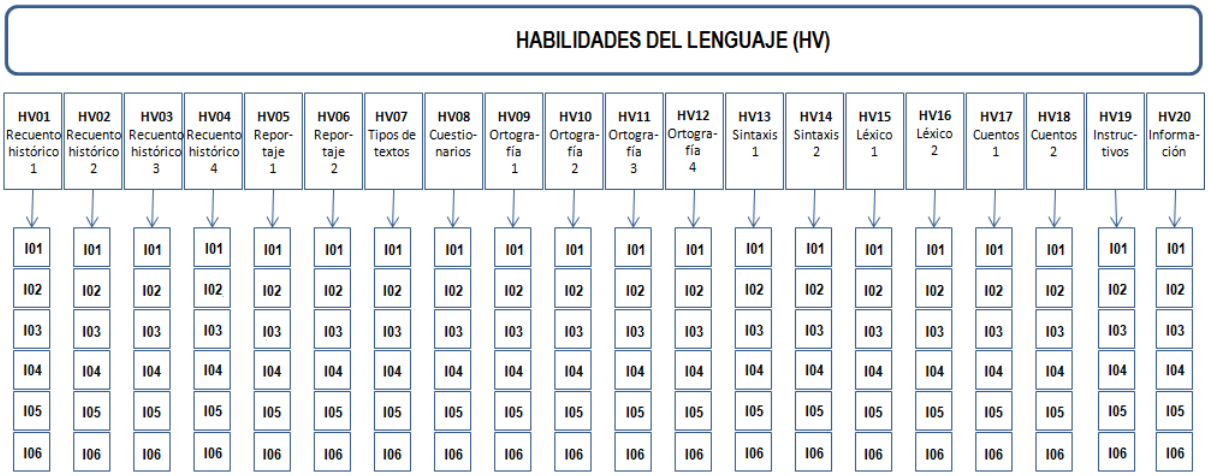


Figura 4.24. Esquema de la muestra HV.

De acuerdo con la TCT, el examen presentó un índice de dificultad media de 84.10 de un total de 104 puntos (p = 0.809) con una desviación estándar de 7.97. El Alpha de Cronbach fue de 0.904, aunque si se eliminaran los seis ítems correspondientes a HV15, el valor subiría a 0.910. La distribución de las calificaciones se ajustó, en general, a la distribución normal. De acuerdo con el modelo Rasch, la media de las dificultades de los ítems resultó inferior a la media de las habilidades de los estudiantes evaluados, en dos desviaciones estándar. Según el mapa, la mayoría de los ítems están en el rango de dificultad [-1.5; 2], excepto dos reactivos muy difíciles (HV15-3 y HV15-4).

El cálculo de los índices de ajuste arrojó todos los valores de infit, y casi todos los de outfit, dentro del rango de aceptación. Hubo un único valor de discriminación, según el modelo de Rasch, por debajo de 0.8 (HV15-2, con 0.52), el resto de las cifras fueron adecuadas.

Al igual que el área de habilidades del lenguaje evaluada en VA y VB, esta prueba manifestó problemas de correlación. Del total, 29 ítems obtuvieron índices inferiores a 0.20, de los cuales 11 fueron iguales o menores que 0.10. El promedio de las correlaciones punto medida fue de 0.28. La familia de HV15 fue la más afectada, y le siguieron HV16 y HV06.

4.2.1.3. Español (muestra ESP)

La muestra de Español (figura 4.25) fue aplicada en la universidad de Querétaro y constó del total planificado, los 120 ítems, seis por cada familia. La muestra contó con la información de 217 estudiantes. En el Anexo E, apartado de Español, se pueden observar las tablas y figuras de los análisis psicométricos de la muestra.

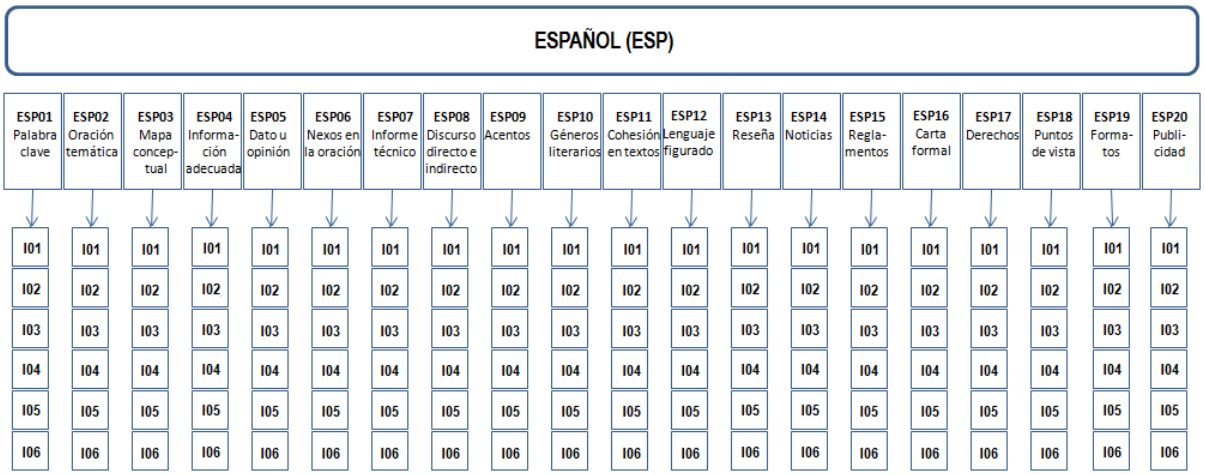


Figura 4.25. Esquema de la muestra ESP

De acuerdo con la TCT, la prueba presentó un índice de dificultad media de 88.13 puntos sobre 120 (p = 0.73) con una desviación estándar de 15.32. El Alpha de Cronbach fue de 0.959.

La distribución de las calificaciones resultó leptocúrtica, con casos aislados hacia la izquierda de la gráfica, esto último se manifestó en un coeficiente de asimetría de -2.194.

De acuerdo con el modelo de Rasch, la media de las dificultades de los ítems resultó inferior, en más de una desviación estándar, a la media de las habilidades de los evaluados. Según el mapa, los ítems se distribuyeron en un rango de -1 a 1 lógitos; excepto ESP02-3, ESP02-1, ESP14-5 y ESP02-5, que quedaron entre 1 y 2, y ESP06-1 con -1.78.

Se identificaron algunos problemas aislados de ajuste. La familia con mayores desajustes fue ESP04, en su mayoría por encima de 1.3, lo que indica demasiada aleatoriedad. El resto de los valores deficientes se encontraron, generalmente, por debajo del rango [0.8; 1.3]; es decir, mostraron determinismo en las respuestas. En cuanto a la discriminación, no se observaron índices importantes de alerta.

En general, los resultados indican buenas correlaciones (promedio de 0.40). Los reactivos con mayores fallas en sus propiedades psicométricas fueron: ESP02-1, con una correlación negativa, y ESP05-1, con una correlación pequeña sumada a problemas de *infit* y *outfit*.

4.2.1.4. Matemáticas (muestra MAT)

La muestra de Matemáticas (figura 4.26) fue aplicada en CESUES, sede Hermosillo. En la administración participaron 239 estudiantes, sin embargo en cinco familias de ítems solamente se recuperaron los datos de 132 personas. Así que, para el análisis desde la TCT se consideró esta última cantidad. Otra aclaración necesaria es que la familia de MAT14 no pudo decodificarse, por lo que no se incluye su análisis. En el Anexo E (apartado de Matemáticas) se pueden observar tablas y figuras con los resultados de los análisis psicométricos de la muestra.

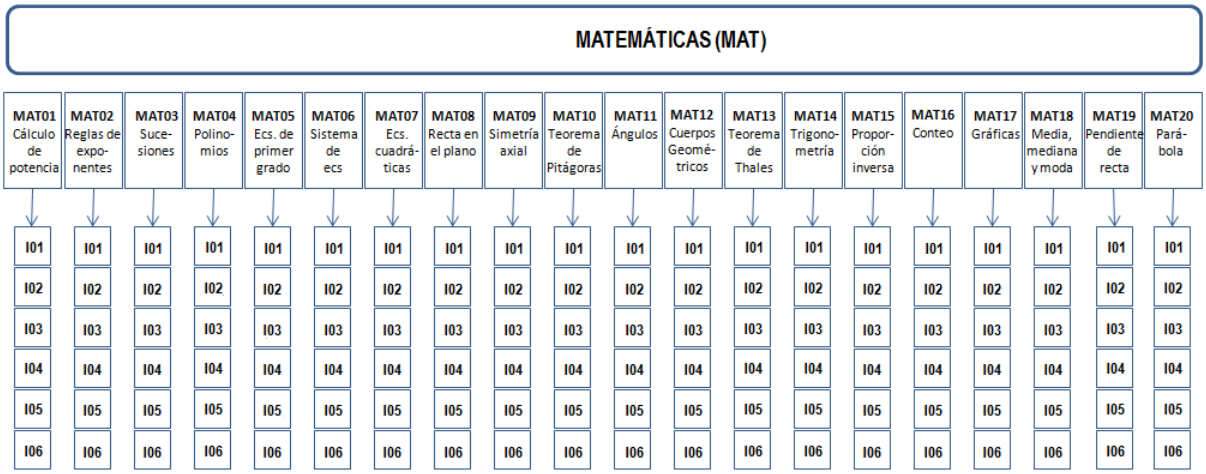


Figura 4.26. Esquema del la muestra MAT

De acuerdo con la TCT, esta prueba de 114 ítems presentó un índice de dificultad media de 24.60 puntos sobre 114 (p = 0.22) con una desviación estándar de 12.27. El Alpha de Cronbach fue de 0.934. La distribución de las calificaciones se acercó aproximadamente a una normal. Cabe destacar la gran dificultad de la prueba, ya que para los 130 evaluados, los ítems 1, 5 y 6 de MAT03, los ítems 4, 5 y 6 de MAT08, los ítems 1 y 5 de MAT19, y los ítems 1 y 2 de MAT20 no tuvieron respuestas correctas.

Para el análisis de Rasch se utilizó el total de la muestra. La media de las dificultades de los ítems fue dos desviaciones superior a la media de las habilidades de las personas evaluadas. Estas mostraron muy poca habilidad matemática, el promedio se ubicó cerca de -2 lógitos. Para este modelo, a seis ítems no se les pudo estimar la dificultad, debido a que no se obtuvieron respuestas correctas. Es importante aclarar que si bien la prueba fue difícil, también hubo un gran grupo de reactivos que apelaron a la habilidad de estos estudiantes.

Las cinco primeras familias presentaron *outfit* debajo de 0.8, lo cual indica determinismo en las respuestas, lejos de la zona de medición del ítem. Los *outfit* fuera de rango también se

manifestaron en varios ítems de las familias restantes. Por el contrario, los índices de discriminación fueron aceptables en todos los reactivos del examen.

Los problemas de correlación se concentraron en MAT20 y en MAT08 (los seis reactivos de la primera familia y cinco de la segunda registraron índices próximos a cero). También se observaron algunos valores bajos en MAT01, MAT02, MAT03, MAT07 y MAT19. Estos resultados pudieron estar influenciados por la extrema dificultad de dichas familias.

4.2.1.5. Ciencias naturales (muestra NAT)

La muestra de Ciencias naturales (figura 4.27) fue aplicada en dos instituciones: UACJ y CESUES, sede San Luis Río Colorado. De la primera se obtuvo información de 60 estudiantes y de la segunda, de 100. Sin embargo, no se pudieron juntar ambas bases de datos para todos los ítems, debido a que en ciertos reactivos se utilizaron ítems-hijo diferentes en una universidad con respecto a la otra. Además, debido a ciertos problemas de funcionamiento del editor de reactivos, se efectuó una aplicación extra de FIS12 en CESUES, sede Hermosillo. Para revisar los resultados puntuales de los estudios psicométricos, ver el Anexo E, apartado de Ciencias naturales

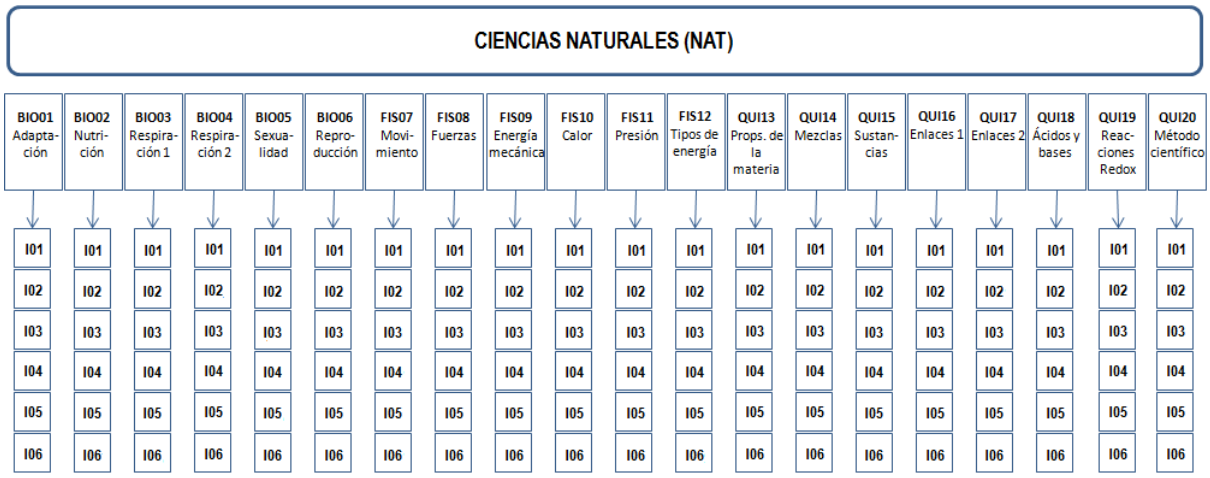


Figura 4.27. Esquema del la muestra NAT

Para los análisis, desde la TCT, se utilizó la muestra de CESUES (San Luis Río Colorado), por ser la que contenía mayor cantidad de examinados. De esta muestra se recuperaron los datos de 102 ítems (si bien se aplicó el examen completo, no se pudo rescatar la información de las familias de FIS10, FIS12 Y QUI13). Esta prueba presentó un índice de dificultad media de 52.33 puntos sobre 102 (p = 0.51) con una desviación estándar de 8.14. El Alpha de Cronbach fue de 0.894. La distribución de las calificaciones se acercó, aproximadamente, a una normal.

Para el modelamiento según Rasch se ingresaron todos los datos, puesto que el programa *Winsteps* permite ejecutar los análisis. La media de las dificultades de los ítems coincidió con la media de las habilidades de los evaluados. La distribución de los ítems se aproximó a una normal, donde los reactivos más complejos fueron: FIS07, FIS09 y FIS11, y los más fáciles, en general, coincidieron con las familias de Biología.

Existieron problemas de ajuste escasos y aislados, que se caracterizaron porque el *outfit* superó el máximo aceptado de 1.3 (aleatoriedad en zonas lejanas a la dificultad del ítem). Llaman la atención dos reactivos cuyas correlaciones resultaron negativas, estos fueron: FIS10-2 y FIS10-5 (con la salvedad de que FIS10 solo contó con 60 datos). Otras familias con correlaciones inferiores a 0.2 fueron FIS11 (en cinco ítems) y QUI16 (en dos ítems).

4.2.1.6. Ciencias sociales (muestra SOC)

La muestra de Ciencias sociales (figura 4.28) fue aplicada en la universidad de Querétaro y constó de 114 ítems, ya que no se incluyó HIS06 (con lo cual se eliminaron 6 ítems). En este pilotaje participaron 206 estudiantes. En el Anexo E, apartado de Ciencias sociales, se pueden revisar las tablas y figuras de los resultados de los análisis psicométricos de SOC.

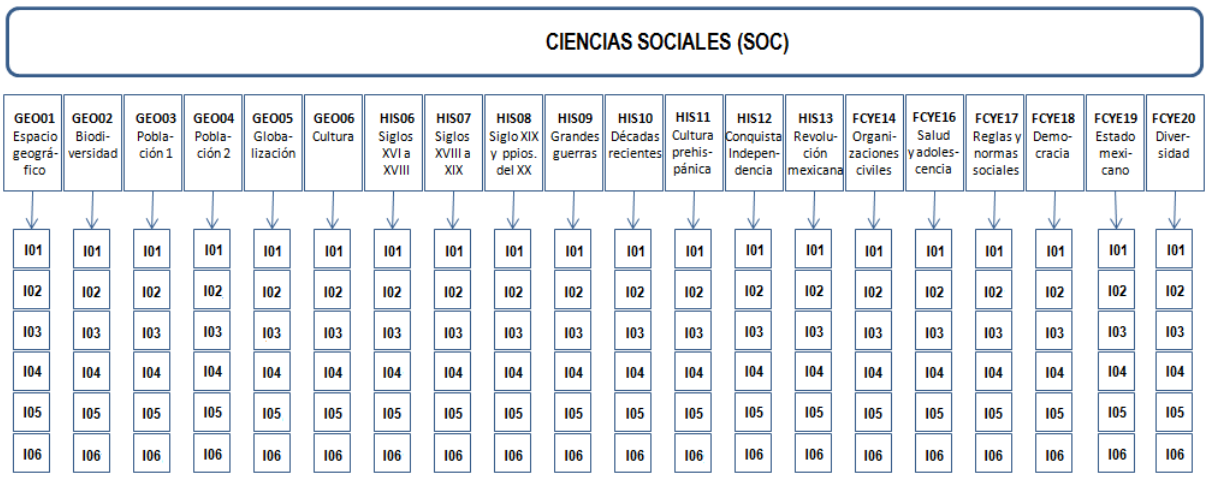


Figura 4.28. Esquema del la muestra SOC

Según la TCT, el índice de dificultad media fue de 73.28 (p = 0.61), con un Alpha de Cronbach: 0.960 y la distribución de las calificaciones se ajustó a la distribución normal. En el mapa de Wright se observa una distribución compacta de los ítems. La media de las dificultades de los ítems resultó inferior a la media de las habilidades de los evaluados; aunque, en general, los reactivos recorrieron el rango de habilidades de todos los estudiantes examinados (de -1 a 1 lógitos).

Las deficiencias de ajuste fueron pocas y aisladas. Los desajustes lejanos a la zona de medición del reactivo se dieron, por partes iguales, tanto por exceso como por defecto. El menor número de problemas se encontró en los *infit*; estos, en gran parte resultaron bajos; por lo tanto, marcan determinismo cercano a la zona de dificultad del ítem. También fueron pocos los índices de discriminación menores que 0.8 y todos coincidieron con valores altos de *infit-outfit*. Las correlaciones resultaron, en general, aceptables; en la mayoría de los casos superaron a 0.30. Toda la información permite inferir un buen comportamiento, en general, de esta evaluación parcial del área de Ciencias sociales.

4.2.2. De los ítems de cada contenido

En este apartado se incluyen los resultados de cada familia de 6 ítems, para cada una de las muestras: HV, HC, ESP, MAT, NAT y SOC. Los hallazgos referentes al área de Habilidades matemáticas se ilustran con gráficas y tablas, que se describen de manera detallada. De las familias de las áreas restantes, se expone un resumen y el resto de la información se adjunta en el anexo F, en los apartados correspondientes.

4.2.2.1. Familias de HC

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 Habilidades matemáticas (ver figura 4.29). Para tal efecto se consideró la población de CESUES, si bien era menor que la de UACJ, evaluó todas las familias, mientras que de UACJ no se pudo recuperar la información de HC02, ni de HC07, ni de tres reactivos de HC19. De acuerdo con la figura 4.29, las medias de las dificultades por familia se distribuyeron entre 0 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. Se encontraron dos casos en que los índices de confiabilidad estaban por debajo de 0.5 (ver figura 4.30). El primero corresponde a HC05 donde solo se analizaron 3 ítems. El segundo se refiere a HC09, con una confiabilidad cercana a cero; para contrastar este índice, se calculó el Alpha de Cronbach con el total de 155 datos y el valor subió a 0.648.

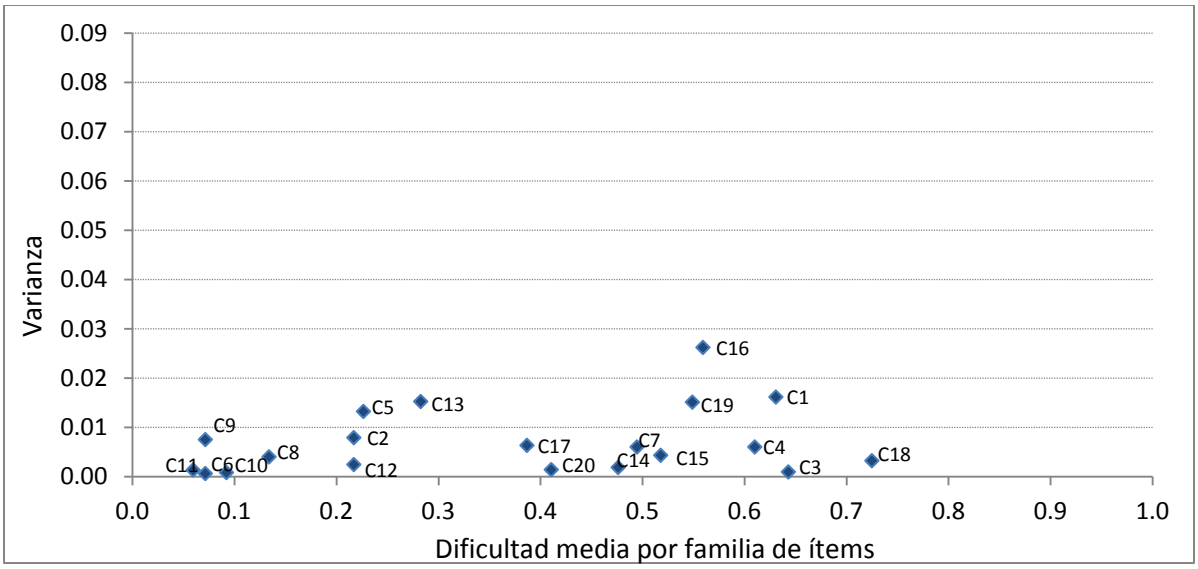


Figura 4.29. Gráfica de la dificultad media por familia de 6 ítems de la muestra HC vs. Varianza.

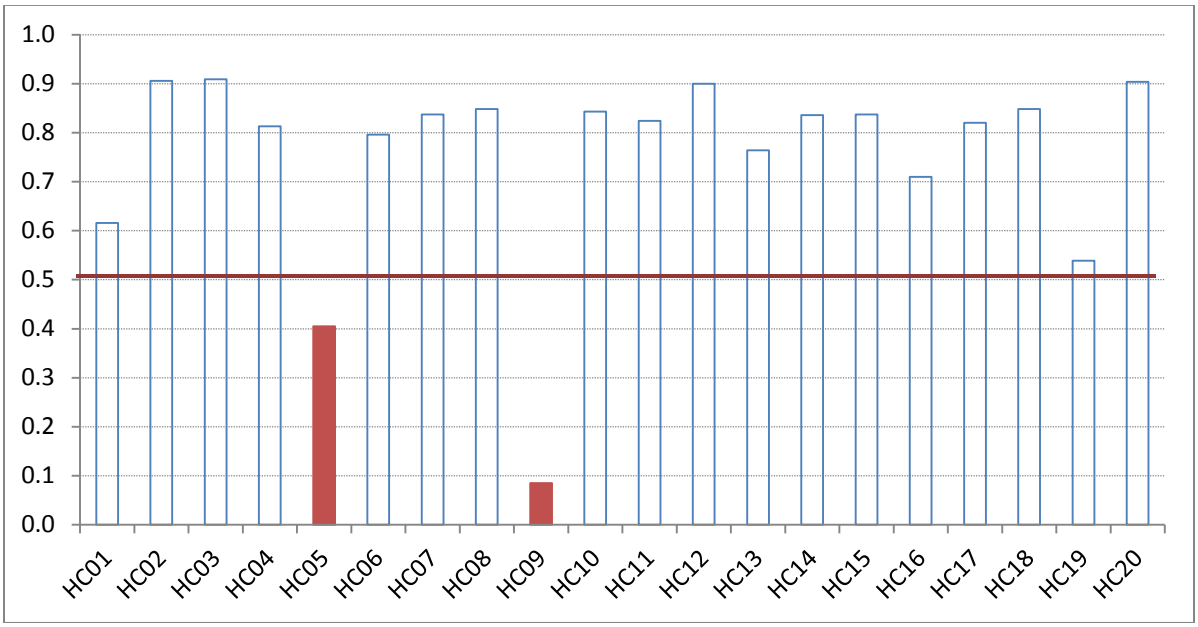


Figura 4.30. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra HC

Los resultados de la aplicación de la TRI revelaron ciertos problemas en los índices de ajuste, no concentrados particularmente en ninguna familia (ver tabla 4.16). Generalmente, a *infit* pequeños le correspondieron *outfit* también bajos; en coherencia, valores altos de *infit* se asociaron a *outfit* también altos. Esto indica que el comportamiento de los ítems fue similar, tanto cerca como lejos de su nivel de dificultad. Curiosamente, HC07 que presentó problemas de

ajuste al analizarlo dentro del área, no los manifestó dentro de la familia. También se registraron los índices de discriminación, en la tabla 4.17 se resaltaron los valores por debajo de 0.80. El caso más notorio es el ítem 6 de HC06, cuyo índice fue negativo.

Tabla 4.16.
Infit y Outfit para cada ítem de cada familia de la muestra HC

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
HC01	1.04	1.09	0.89	1.23	0.87	0.97	1.05	1.07	0.91	1.60	0.85	0.92
HC02	1.46	1.00	**0.54**	1.04	**0.48**	0.63	1.43	1.14	0.68	2.35	**0.46**	1.19
HC03	1.16	1.09	0.72	0.76	1.16	1.13	1.14	1.07	0.70	0.71	1.15	1.01
HC04	1.08	1.11	0.92	1.30	**0.79**	0.89	1.04	1.21	0.83	1.32	**0.70**	0.91
HC05	1.00	0.78	---	---	1.22	---	0.99	0.78	---	---	1.61	---
HC06	1.00	0.78	**0.58**	0.91	0.94	**1.91**	1.00	0.68	**0.56**	0.82	0.88	**2.08**
HC07	1.01	1.27	0.84	0.95	0.90	0.96	0.98	1.28	0.84	0.97	0.91	0.94
HC08	1.04	1.09	0.92	1.10	0.58	0.91	**3.04**	1.73	0.76	1.69	0.32	0.75
HC09	0.92	0.83	0.85	0.99	1.13	1.21	0.98	0.83	0.86	1.03	1.18	1.59
HC10	0.91	1.02	0.97	1.04	1.28	0.74	0.97	1.14	0.92	0.99	1.83	0.64
HC11	1.09	1.14	0.94	1.08	0.82	0.93	1.19	1.29	0.89	1.27	0.73	0.71
HC12	1.12	1.01	1.21	0.85	**0.74**	1.15	1.03	0.92	1.37	0.78	0.68	0.99
HC13	**1.34**	0.96	1.16	**0.70**	**0.70**	0.86	**2.29**	0.86	**1.67**	0.47	0.53	1.20
HC14	1.07	1.23	1.01	0.82	0.83	1.03	1.14	1.31	1.05	0.78	0.76	1.02
HC15	1.16	1.06	**0.75**	0.78	0.99	1.22	1.27	0.99	0.75	0.74	1.05	1.35
HC16	**1.33**	0.96	0.90	1.01	0.90	0.89	**2.02**	0.98	0.77	1.04	0.80	0.79
HC17	0.88	0.91	1.00	1.18	0.92	1.06	0.94	0.90	1.10	1.18	0.86	1.12
HC18	**1.41**	1.15	0.79	0.98	0.86	**0.76**	**1.79**	1.19	0.74	0.98	0.89	**0.72**
HC19	1.06	0.98	0.77	0.97	1.05	1.26	0.96	0.94	0.75	0.92	1.13	**1.48**
HC20	**1.54**	0.96	1.13	**0.69**	0.89	0.76	**1.68**	0.98	1.14	0.69	0.89	0.74

Nota: las celdas sombreadas indican los *infit* y *outfit* fuera de rango tales que, los índices estandarizados también quedan fuera de [-2; 2]. I1, I2, I3, I4, I5, I6 refieren a item 1, item 2, item 3, item 4, ítem 5 e ítem 6, respectivamente.

Tabla 4.17.
Índice de discriminación de los ítems de cada una de las familias de la muestra HC

	I1	I2	I3	I4	I5	I6
HC01	0.92	0.76	1.15	0.83	1.32	1.09
HC02	0.64	0.92	1.26	0.64	1.39	1.13
HC03	0.88	0.95	1.15	1.17	0.84	0.92
HC04	0.91	0.85	1.22	0.67	1.53	1.15
HC05	1.01	1.46	---	---	0.73	---
HC06	1.00	1.40	1.61	1.17	1.12	**-0.16**
HC07	1.00	0.68	1.18	1.03	1.13	1.06
HC08	0.72	0.79	1.13	0.81	1.43	1.16
HC09	1.20	1.12	1.12	1.01	0.43	0.88
HC10	1.10	0.93	1.07	0.96	0.52	1.37
HC11	0.73	0.78	1.14	0.87	1.57	1.12
HC12	0.81	1.03	0.49	1.34	1.55	0.84
HC13	0.24	1.05	0.63	1.39	1.38	1.15
HC14	0.81	0.59	0.96	1.35	1.38	0.95
HC15	0.62	0.94	1.52	1.44	0.98	0.62
HC16	0.39	1.07	1.27	0.97	1.23	1.22
HC17	1.21	1.17	0.95	0.65	1.14	0.88
HC18	0.42	0.76	1.04	1.28	1.57	1.45
HC19	0.97	1.05	1.34	1.07	0.87	0.42
HC20	0.18	1.08	0.76	1.46	1.21	1.37

En cuanto a la correlación punto medida, en la figura 4.31 se observa que todas las familias presentaron correlaciones por encima de 0.3, y en la mayoría de los casos, mayores que 0.6; lo cual indica que cada familia contiene ítems isomorfos, en el sentido de que evalúan una misma habilidad.

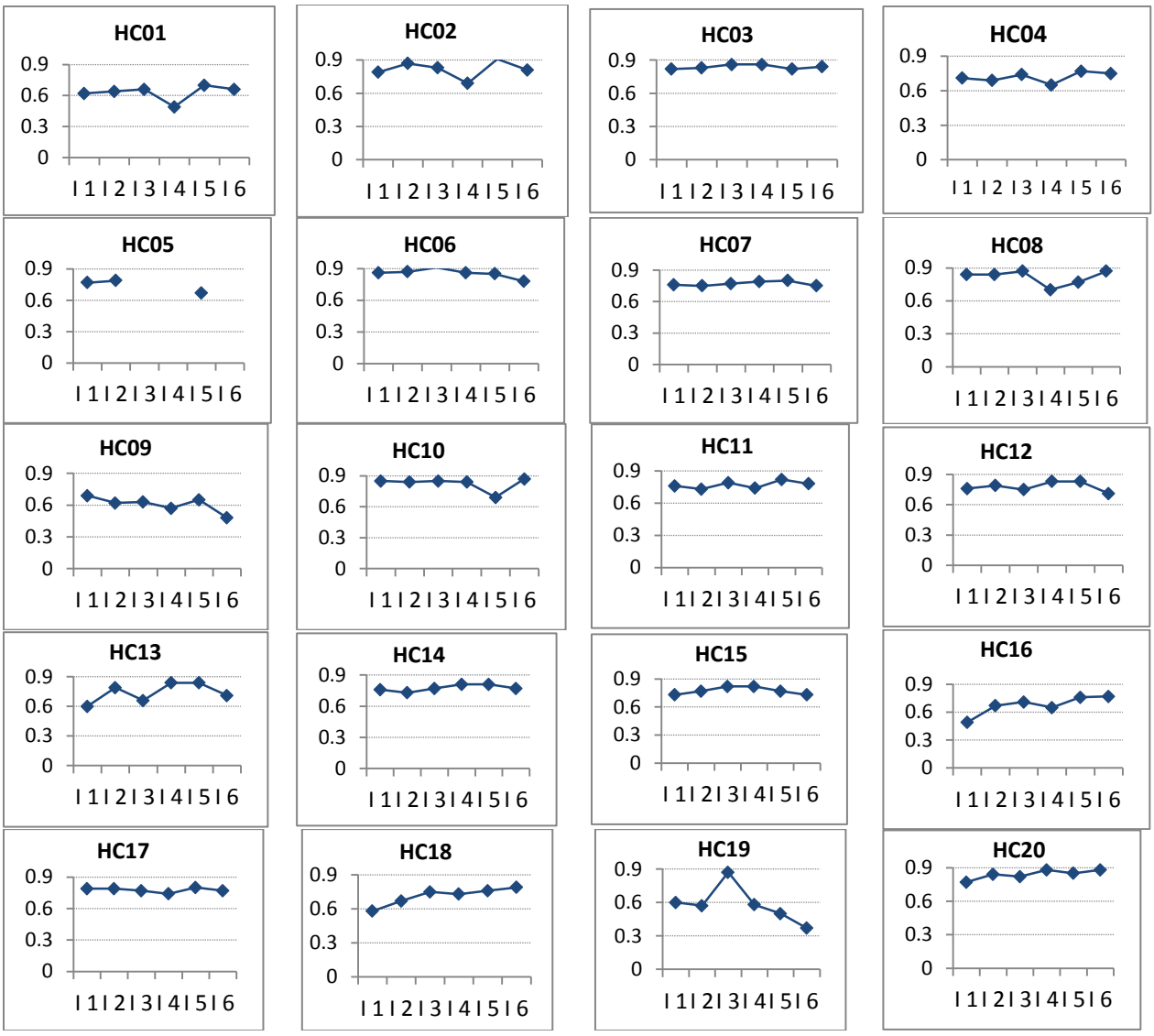


Figura 4.31. Gráfica de correlaciones ítem medida por familia, de la muestra HC.

También se efectuaron AFC unidimensionales para cada familia de reactivos, en total fueron 20 análisis (ver tabla 4.18). Las cargas obtenidas en cada uno, en general, fueron muy buenas y además, estuvieron respaldadas por índices de ajuste también aceptables. Solamente un ítem de la familia HC19, el sexto, no cargó al constructo correspondiente y su valor fue negativo. Este último resultado no se corresponde con las correlaciones obtenidas a través del modelamiento de Rasch, ya que el índice del sexto ítem, si bien menor que el de los cinco reactivos restantes, fue de 0.37.

Tabla 4.18.
Índices de ajuste de AFC por familia de la muestra HC, con sus respectivas cargas factoriales

	Indices de ajuste						Cargas factoriales[a]					
	Chi	Lib	p	NNFI	CFI	RMSEA	I1	I2	I3	I4	I5	I6
HC01UACJ	0.111	4	0.998	1.117	1.000	0.000	0.862	0.677	0.699	0.257[b]	0.421	0.323[b]
HC02CESUES	0.616	1	0.432	1.024	1.000	0.000	0.659	0.575	0.968	0.636	0.916	0.610
HC03	8.470	6	0.205	0.990	0.996	0.058	0.858	0.881	0.926	0.817	0.676	0.681
HC04	14.371	6	0.025	0.924	0.970	0.095	0.714	0.662	0.690	0.424	0.636	0.586
HC05(3 ítems)	0.000	0	--	--	--	--	0.520	0.787	--	--	0.535	--
HC06	5.834	9	0.756	1.008	1.000	0.000	0.826	0.860	0.926	0.843	0.831	0.672
HC07	13.210	9	0.153	0.951	0.971	0.085	0.707	0.610	0.777	0.729	0.741	0.678
HC08	5.731	5	0.333	0.996	0.999	0.031	0.847	0.827	0.858	0.595	0.673	0.850
HC09	2.765	5	0.736	1.046	1.000	0.000	0.566	0.592	0.611	0.550	0.374	0.469
HC10	1.921	2	0.382	1.001	1.000	0.000	0.896	0.786	0.782	0.822	0.608	0.818
HC11	4.792	5	0.441	1.001	1.000	0.000	0.718	0.629	0.706	0.635	0.842	0.720
HC12	0.374	1	0.541	1.016	1.000	0.000	0.759	0.838	0.621	0.730	0.753	0.619
HC13	5.863	9	0.753	1.015	1.000	0.000	0.439	0.770	0.539	0.851	0.856	0.615
HC14	24.527	9	0.003	0.936	0.962	0.106	0.662	0.615	0.695	0.829	0.831	0.712
HC15	7.343	8	0.500	1.003	1.000	0.000	0.666	0.706	0.819	0.809	0.698	0.632
HC16	8.264	7	0.309	0.988	0.994	0.034	0.355	0.568	0.695	0.592	0.674	0.670
HC17	7.003	6	0.320	0.994	0.998	0.033	0.814	0.756	0.754	0.620	0.702	0.656
HC18UACJ	12.728	8	0.121	0.932	0.964	0.078	0.264	0.402	0.712	0.561	0.747	0.756
HC19CESUES	8.475	9	0.487	1.040	1.000	0.000	0.496	0.650	0.703	0.318[b]	0.239[b]	**-.056**[b]
HC20	18.697	9	0.027	0.975	0.985	0.084	0.685	0.775	0.770	0.895	0.838	0.884

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. [a] Para los datos dicotómicos, se utilizó la matriz de correlaciones de Pearson para ejecutar los análisis. [b] Cargas no significativas al nivel de 0.05.

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 a considerar. 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 subsanado al incluir más datos en la muestra. 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.

4.2.2.2. Familias de HV

Para el caso de Habilidad el lenguaje (HV), se analizaron 18 familias. Se contó con los datos de 167 estudiantes de la Universidad de Guanajuato. En la tabla 4.19 se incluyen los resultados psicométricos, por familia, con características que van en detrimento de la calidad de los ítems. Los criterios utilizados fueron: escasa dificultad ($p > 0.8$), varianza mayor que 0.3, Alpha de Cronbach < 0.5, infit y outfit fuera de rango (menor que 0.8 o mayor que 1.3, y que, además, en sus versiones estandarizadas escapaban de [-2; 2]), discriminación menor que 0.8 y cargas factoriales inferiores a 0.2. Para los datos exactos, revisar el Anexo F, en el apartado de HV.

Tabla 4.19.
Familias de reactivos de HV y sus deficiencias en los diferentes índices psicométricos

	Dific	**Varianza**	**Alpha**	Infit	Outfit	Disc	**Pmed**	**AFC**
HV02	xx							
HV03		x		xx	xx	x		
HV04	x	x	x			x	x	x
HV05								
HV06		x	x				x	
HV07	x	x						
HV08								
HV09	x		x					xx
HV10					x			
HV11	x					x		
HV12	x				x			x
HV13				x	x	x		
HV14	x							
HV15		xx	x					xx
HV16	x	x	x				x	x
HV17	x	xx	x				x	xx
HV18	xx		x				x	x
HV20	x		x				x	x

Nota: Para Dific (dificultad), "x" indica *p* media > 0.8, "xx" indica *p* media > 0.9.
Para varianza, "x" indica varianza > 0.03, "xx" indica varianza > 0.06.
Para Alpha de Cronbach, "x" indica Alpha < 0.5.
Para *infit* y *outfit*, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para la discriminación, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para Pmed (índice de correlación punto medida), "x" indica de uno a dos ítems con correlación menor que 0.3, "xx" indica tres o más ítems con correlación menor que 0.3.
Para el AFC, "x" indica de uno o dos ítems con carga menor que 0.2, "xx" indica más de dos ítems con carga menor que 0.2. Las familias HV01 y HV19 no fueron analizadas.

Si se observa la tabla 4.19 por columna, se aprecia que 11 familias presentaron un índice de dificultad media superior a 0.8, y entre ellas dos estuvieron por encima de 0.9. Si bien, es aceptable la presencia de ítems sencillos dentro de una prueba, en este caso la cantidad superó la mitad de las familias que conforman el área. También se registraron dos grupos de reactivos con mucha varianza (HV15 y HV17). Otra debilidad del área fue la escasa confiabilidad de ocho familias de ítems, que repercutió en el constructo de cada una.

Tras un análisis de la tabla 4.19 por fila, se percibe que las familias HV02, HV05, HV07, HV08, HV10, HV11, HV12, HV13 y HV14 no presentaron serios problemas en los ítems-hijo analizados (HV05 y HV08 no mostraron irregularidades). El grupo HV03 manifiestó desajustes en cuatro de sus seis reactivos. HV04, HV06, HV09, HV15, HV16, HV17, HV18 y HV20 tuvieron escasa confiabilidad, que se reflejaron en problemas de correlación o de AFC (los más serios en HV09, HV15 y HV17).

Estos resultados permiten inferir que las familias de HV tienen problemas de isomorfismo en dificultad por familia y también deficiencias en agrupación por constructo. La familia más comprometida es HV15, le siguen HV17, HV04 y HV03.

4.2.2.3. Familias de ESP

La muestra ESP, administrada en la universidad de Querétaro, contó con 217 datos de las 20 familias que conforman el área. En la tabla 4.20 se muestran, únicamente, aquellos resultados que disminuyen la calidad psicométrica de los ítems, por familia. Para una información precisa, revisar el Anexo F, en el apartado de ESP.

En cinco casos, la media de las dificultades superó 0.8, lo cual representa el 25% de los ítems. Las medias se concentraron en la franja de 0.6 a 0.85, con una familia aislada (ESP02), de dificultad 0,42. Estos datos permiten inferir que se trata de un área de escasa dificultad. Las varianzas, en general, estuvieron por debajo de 0.03. Solamente ESP02 superó esta barrera, con 0.034.

Dentro de la TRI, se registraron ligeros desajustes (más de *outfit* que de *infit*), en casi todos los casos fue un ítem el que quedó fuera del rango de aceptabilidad. Lo mismo ocurrió con la discriminación, siempre se trató de un reactivo el que discriminó menos de lo aceptado (y en

muchos casos estuvo en el límite). Una excepción fue el sexto reactivo de ESP02, con índice negativo.

Tabla 4.20.
Familias de reactivos de ESP y sus deficiencias en los diferentes índices psicométricos

	Dific	Varianza	Alpha	Infit	Outfit	Disc	Pmed	AFC
ESP01								
ESP02		x		x	x	x	x	x
ESP03	x							
ESP04					x	x		
ESP05								
ESP06	x			x		x		
ESP07	x					x		
ESP08	x							
ESP09								
ESP10						x		
ESP11						x		
ESP12					x			
ESP13								x
ESP14					x			
ESP15	x			x	x			
ESP16						x		x
ESP17								
ESP18								
ESP19					x	x		
ESP20				x	x			

Nota: Para Dific (dificultad), "x" indica p media > 0.8, "xx" indica p media > 0.9.
Para varianza, "x" indica varianza > 0.03, "xx" indica varianza > 0.06.
Para Alpha de Cronbach, "x" indica Alpha < 0.5.
Para *infit* y *outfit*, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para la discriminación, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para Pmed (índice de correlación punto medida), "x" indica de uno a dos ítems con correlación menor que 0.3, "xx" indica tres o más ítems con correlación menor que 0.3.
Para el AFC, "x" indica de uno o dos ítems con carga menor que 0.2, "xx" indica más de dos ítems con carga menor que 0.2.

Para cada familia los ítems correlacionaron bien, excepto el ítem 6 de ESP02. Esto también se reflejó en el AFC. Este último análisis también arrojó problemas en el sexto ítem de ESP13 y de ESP16, situación que no se manifestó en la correlación punto medida.

De estos resultados se infiere que, si bien el área es relativamente fácil (faltan ítems cuya dificultad media se encuentre en el rango [0.2; 0.6]), tiene un buen comportamiento psicométrico, en general. La familia con mayores deficiencias es ESP02, especialmente concentradas en el sexto reactivo. Esta área reflejó un comportamiento similar en los análisis de VA y VB.

4.2.2.4. Familias de MAT

Se analizaron 19 familias de la muestra MAT, ya que no se decodificaron las bases de datos de MAT14. Se contó con la información de 239 participantes de CESUES, sede Hermosillo. Las gráficas y tablas con los valores en detalle se encuentran en el anexo F, en el apartado de MAT. En la tabla 4.21 se describen algunos casos particulares de ítems que presentaron algún conflicto dentro de la familia a la cual pertenecen.

Tabla 4.21.
Familias de reactivos de MAT y sus deficiencias en los diferentes índices psicométricos

	Dific	Varianza	Alpha	Infit	Outfit	**Disc**	Pmed	AFC
MAT01	xx					x	x	x
MAT02	xx					x		
MAT03	xx							
MAT04	xx					x	x	
MAT05	x					x		
MAT06	x					x		
MAT07	xx							x
MAT08	x					x	x	xx
MAT09								
MAT10	xx			x	x	x		
MAT11				x	x	x		
MAT12					x	x		
MAT13					x	x		
MAT15					x	x		
MAT16	x			xx	xx	x		x
MAT17								
MAT18	x			x	x	x		
MAT19	xx				x	x		
MAT20	xx				x	x		

Nota: Para dific (dificultad), "x" indica p media < 0.2, "xx" indica p media < 0.1.
Para varianza, "x" indica varianza > 0.03, "xx" indica varianza > 0.06.
Para Alpha de Cronbach, "x" indica Alpha < 0.5. Para *infit* y *outfit*, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para la disc (discriminación), "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para Pmed (índice de correlación punto medida), "x" indica de uno a dos ítems con correlación menor que 0.3, "xx" indica tres o más ítems con correlación menor que 0.3.
Para el AFC, "x" indica de uno o dos ítems con carga menor que 0.2, "xx" indica más de dos ítems con carga menor que 0.2.
La familia MAT14 no fue analizada. Para la familia MAT03 no pudieron ejecutarse los análisis Rasch ni el AFC.

De acuerdo con la tabla 4.21, las familias se concentraron en la franja de 0 a 0.20 de *p* de dificultad (con 13 de 19 familias en ese intervalo). El grupo más fácil tuvo un promedio de 0.66 (MAT17). Los grupos MAT01, MAT02, MAT03, MAT04, MAT19 y MAT20 registraron una dificultad media menor que 0.04. Cabe aclarar que algunos reactivos no fueron contestados correctamente por ninguna persona evaluada, este fue el caso de los ítems 1, 5 y 6 de MAT03, el

ítem 5 de MAT08 y los ítems 1 y 5 de MAT19. En coherencia con estos valores bajos, las varianzas también resultaron pequeñas (cercanas a cero), la mayor correspondió a MAT08, que superó ligeramente a 0.02.

Debido a la extrema dificultad de MAT03 no se pudieron ejecutar los análisis Rasch correspondientes. Por razones obvias, tampoco se obtuvo información de los ítems con porcentaje de aciertos igual a cero. Los resultados muestran que, en general, los índices de ajuste fuera de rango estuvieron por debajo del valor mínimo, lo cual indica determinismo, ya sea cercano o lejano a la zona de dificultad del reactivo. Además, en muchos casos, los problemas de *infit* se repitieron en el *outfit* de los mismos ítems. Los índices de discriminación marcaron deficiencias aisladas en algunos ítems de 15 familias; los casos extremos fueron valores negativos en un ítem-hijo de las familias MAT16 y MAT19, y un índice positivo cercano a cero para un reactivo de MAT10.

En cuanto a la correlación punto medida, se halló un grupo, MAT08, donde dos de sus ítems presentaron correlaciones cercanas a cero. Esta última situación se reflejó en la confiabilidad, ya que el Alpha de Cronbach de MAT08 no llegó mínimo de 0.5. Se detectaron dos familias con un ítem cada una, donde el índice de correlación se acercó al mínimo aceptable (sus valores fueron 0.28 y 0.29). El resto de las familias presentaron correlaciones altas y muy similares (por ejemplo: MAT13, MAT15, MAT17).

Los AFC también reflejaron el problema de la alta dificultad de los ítems. Para las familias de MAT03 y de MAT04 no se pudieron ejecutar los análisis. En MAT08 y MAT19 se excluyeron los reactivos con dificultad máxima ($p = 0$). En MAT20, primeramente se corrieron los análisis con los seis ítems, al dar matrices no definidas positivas, se eliminaron los dos reactivos de respuesta correcta casi nula y así se obtuvieron índices y cargas factoriales. En el

resto de las familias se efectuaron los AFC, cada uno de 6 ítems. Las cargas de tres reactivos de la familia de MAT08 fueron inferiores a 0.2, en los casos de MAT01 y MAT07 dos ítems no cargaron al constructo; para MAT16 fue uno el ítem defectuoso. El resto de las familias no presentaron problemas.

Como resumen, y considerando las dificultades para ejecutar los análisis, debido a las altas dificultades de los ítems, se infiere que sería necesario repetir los análisis psicométricos con una población más numerosa, y así obtener resultados más fidedignos.

Con la salvedad mencionada, se indica que las familias MAT02, MAT05, MAT06, MAT09, MAT10, MAT11, MAT12, MAT13, MAT15, MAT17 y MAT18 (donde se analizaron los seis reactivos) parecen comportarse de manera que sus ítems-hermano resultaron isomorfos. El grupo de MAT08 es el que necesita mayor atención, en cuanto a constructo; le sigue MAT07. En MAT16, un ítem-hijo desajustó con demasiada aleatoriedad, reportó escasa carga factorial y correlación baja.

4.2.2.5. Familias de NAT

A continuación, se describen los resultados por familia de la muestra NAT. Debido a que la aplicación se realizó en dos instituciones y hubo cambios en algunos de los ítems, para los cálculos de los estadísticos desde la TCT se optó por la muestra administrada en CESUES (SLRC). Para el modelamiento de Rasch y el AFC, se consideró como base la de CESUES (SLRC) y en el caso de que coincidieran los reactivos, se agregaron los datos de UACJ. Cuando no había información de CESUES de SLRC, se utilizó la base de UACJ. Además, para FIS12 se analizó la aplicación especial efectuada en CESUES de Hermosillo. En la tabla 4.22 se muestran aquellos resultados que perjudicaron la calidad psicométrica de las familias; para ver la información al detalle, consultar el anexo F, en el aparado de NAT.

Las medias de las dificultades por familia se concentraron en la franja de 0.5 a 0.7 (diez familias), y en las restantes ninguna superó la media de 0.8. La familia FIS11 es la más difícil con el 5% de respuestas correctas, le siguen FIS07, FIS09 y QUI17, todas inferiores a 0.4. Las varianzas resultaron pequeñas, solamente BIO04 superó el valor 0.02 (varianza = 0.027). Fuera de la aplicación de CESUES, se contó con la información de FIS10 (administrado en UACJ) y de FIS12 (en CESUES, Hermosillo). Los resultados de la familia de FIS10 dieron una media de 0.64 con una varianza de 0.04; en el caso de FIS12, la media fue de 0.56 con una varianza de 0.01.

Cuatro familias no superaron los índices de confiabilidad de 0.50, estas son: BIO01, BIO04, QUI16 y QUI18. En los casos particulares de FIS10 y FIS12, los Alpha de Cronbach resultaron: 0.55 y 0.84, respectivamente.

Según el modelo de Rasch, los resultados revelaron algunos índices de ajuste fuera de rango, entre ellos la familia con mayores desajustes fue FIS07, ya que tres de sus ítems presentaron, tanto en *infit* como en *outfit*, valores fuera del intervalo de aceptación. Si bien la mayoría de las familias exhibieron ítems con índices de discriminación inferiores a lo deseado; en general, se trató de un reactivo por grupo y los números se acercaron al mínimo necesario. Los casos más destacados fueron tres ítems de FIS11, y le siguieron con dos ítems, FIS07 y FIS10.

Los ítems correlacionaron bien dentro de cada familia; solamente existieron dos reactivos, uno perteneciente a BIO01 y otro a FIS10, que estuvieron ligeramente por debajo del mínimo aceptado de 0.30. Esta situación también se manifestó en los AFC, ya que los mismos ítems presentaron cargas muy bajas. A esta problemática se agregaron las familias de QUI16 (con cargas pequeñas en tres ítems) y QUI20 con un reactivo deficiente.

Tabla 4.22.
Familias de reactivos de NAT y sus deficiencias en los diferentes índices psicométricos

	CESUES (SLRC)			Mayor muestra posible					
	Dific	Varianza	Alpha	Infit	Outfit	Disc	Pmed	AFC	
BIO01			x		x	x	x	x	BIO01
BIO02					x				BIO02
BIO03									BIO03
BIO04			x						BIO04
BIO05				x	x				BIO05
BIO06				x	x	x			BIO06
FIS07				x	x	x			FIS07
FIS08						x			FIS08
FIS09				x	x	x			FIS09
FIS10				x		x	x	x	FIS10
FIS11					x	x			FIS11
FIS12						x			FIS12
QUI14						x			QUI14
QUI15						x			QUI15
QUI16			x					x	QUI16
QUI17				x		x			QUI17
QUI18			x						QUI18
QUI19				x	x	x			QUI19
QUI20						x		x	QUI20

Nota: Para dific (dificultad), "x" indica p media > 0.8, "xx" indica p media > 0.9.
Para varianza, "x" indica varianza > 0.03, "xx" indica varianza > 0.06.
Para Alpha de Cronbach, "x" indica Alpha < 0.5.
Para *infit* y *outfit*, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para la disc (discriminación), "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para Pmed (índice de correlación punto medida), "x" indica de uno a dos ítems con correlación menor que 0.3, "xx" indica tres o más ítems con correlación menor que 0.3.
Para el AFC, "x" indica de uno o dos ítems con carga menor que 0.2, "xx" indica más de dos ítems con carga menor que 0.2.
Los índices de TCT fueron calculados para la muestra de CESUES (SLRC). Los índices según el modelo de Rasch y las cargas factoriales se realizaron con las muestras de mayor tamaño.
La familia de QUI13 no pudo ser analizada.

Después de una interpretación de la información estadística obtenida, se infiere que la familia de FIS10 tiene problemas tanto de ajuste como de pertenencia al constructo; aunque cabe aclarar que estos resultados deben considerarse con precaución, debido al tamaño de la muestra (55 datos). En QUI16, si bien no se presentaron desajustes, la confiabilidad fue pequeña y mostró

serios conflictos en la agrupación en torno al constructo. En BIO01, un reactivo reflejó disonancias en casi todos los índices; por lo tanto, se sugiere revisar y comparar con sus ítems-hermano. El resto de las familias no se identificaron serios conflictos.

4.2.2.6. Familias de SOC

La muestra SOC, administrada en la universidad de Querétaro, contó con 206 datos de 19 de las 20 familias que conforman el área (faltó HIS06). En la tabla 4.23 se muestran aquellos resultados que disminuyen la calidad psicométrica de los ítems, por familia. La información precisa con tablas y figuras se encuentra en el Anexo F, en el apartado de SOC.

De acuerdo con la tabla 4.23, desde la TCT se observa que las dificultades medias se distribuyeron entre 0.4 y 0.87, solamente en dos familias superaron a 0.8. Tampoco se encontraron varianzas superiores a 0.02. Quien más varianza observó fue el grupo de GEO03, con 0.014. En todos los casos el Alpha de Cronbach fue superior a 0.50; la confiabilidad más baja se localizó en FCYE14, con 0.59.

Los problemas de *infit* y *outfit* fueron aislados, en general, en la familia donde se identificaron desajustes solo se trató de un ítem. En su mayoría se encontró cierto determinismo, ya sea cerca o lejos de la zona de medición del ítem. En cuanto a la discriminación, tampoco se encontraron deficiencias serias, los índices fuera de rango estuvieron alrededor de 0.70.

Si bien se localizaron dos ítems en el grupo de GEO03 y un ítem en HIS13 y FCYE14 con correlaciones menores a 0.30, los resultados de los AFC indican que, en todas las familias, los índices de ajuste fueron muy buenos y las cargas factoriales superaron 0.20.

Tabla 4.23.
Familias de reactivos de SOC y sus deficiencias en los diferentes índices psicométricos.

	Dific	Varianza	Alpha	Infit	Outfit	Disc	Pmed	AFC
GEO01						x		
GEO02					x			
GEO03					x	x	x	
GEO04								
GEO05				x	x			
GEO06					x			
HIS07						x		
HIS08								
HIS09								
HIS10						x		
HIS11					x	x		
HIS12				x	x	x		
HIS13				x			x	
FCYE14							x	
FCYE16	x				x			
FCYE17				x				
FCYE18	x							
FCYE19				x		x		
FCYE20						x		

Nota: Para dific (dificultad), "x" indica p media > 0.8, "xx" indica p media > 0.9.
Para varianza, "x" indica varianza > 0.03, "xx" indica varianza > 0.06.
Para Alpha de Cronbach, "x" indica Alpha < 0.5.
Para *infit* y *outfit*, "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para la disc (discriminación), "x" indica de uno a tres ítems fuera de rango, "xx" indica más de tres ítems fuera de rango.
Para Pmed (índice de correlación punto medida), "x" indica de uno a dos ítems con correlación menor que 0.3, "xx" indica tres o más ítems con correlación menor que 0.3.
Para el AFC, "x" indica de uno o dos ítems con carga menor que 0.2, "xx" indica más de dos ítems con carga menor que 0.2.
La familia MAT14 no fue analizada. Para la familia MAT03 no pudieron ejecutarse los análisis Rasch ni el AFC.

Estos resultados ofrecen evidencias de familias de ítems isomorfos en dificultad y buena pertenencia a cada uno de los 19 constructos analizados del área de Ciencias sociales.

Finalmente, la tabla 4.24 resume los hallazgos de las diferentes familias agrupadas por área.

Tabla 4.24.
Resumen de las propiedades psicométricas de las familias de ítems de cada una de las seis áreas del EXHCOBA-R/MS

Área	Dific.	Varianza	α	Ajuste	Discrim.	Pmed/AFC	Comentarios
HV	↓	2 ☹ 5 😐	8 ☹	Desajustes aislados	☺	9 ☹	• Revisar plantillas • Replicar
HC	↑	☺	1 ☹	Desajustes aislados	😐	1 ☹	• Buenas propiedades psicométricas
ESP	↓	☺	☺	Desajustes aislados	☺	3 ☹	• Elevar complejidad
MAT	↑↑	☺	1 ☹	Desajustes ± aislados	😐	4 ☹	• Considerar una familia por contenido. • Replicar
NAT	↑↓	☺	1 ☹	Desajustes aislados	😐	4 ☹	• Revisar plantillas • Replicar
SOC	↑↓	☺	☺	☺	☺	1 ☹	• MB propiedades psicométricas

Nota: Dific. = dificultad media por familia. Discrim. = índice de discriminación. Pmed/AFC = índice de correlación punto medida y resultados del AFC.

4.2.3. De los elementos que conforman los ítems

Como ya se explicó, muchos tipos de reactivos admiten crédito parcial, es decir, cada ítem-hijo tiene, a su vez, elementos que lo componen y cada elemento contestado correctamente se computa, en partes iguales, para conformar el total del puntaje del reactivo. Debido a la extensión de efectuar un análisis por elemento de todas las familias cuyos ítems son de crédito parcial, se seleccionó una por cada muestra. Las familias analizadas son: HV17, HC02, ESP18, MAT09, FIS12 e HIS07. A continuación, se presentan los resultados.

4.2.3.1. Elementos de la familia HV17

Dentro de los tipos de ítems que se utilizaron para generar los reactivos de HV, el más frecuente es *selección de elementos*, con 8 familias de ítems. Por tal motivo, se eligió al azar una para analizar sus elementos, esta es HV17 (figura 4.32). En la figura 4.33 se muestra un ejemplo de reactivo. En él aparece un texto literario con tres símiles (analogías) subrayados y se solicita elegir, por cada símil, una de tres frases que lo interprete correctamente. Por lo tanto, para los análisis estadísticos se consideraron 6 ítems, cada uno con tres elementos. Por cada elemento hay una respuesta correcta y dos distractores.

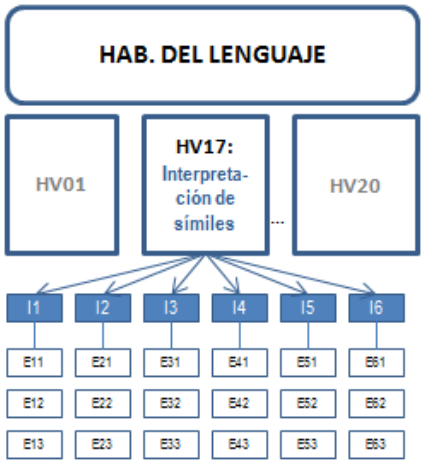


Figura 4.32. Esquema de los elementos de seis ítems de la familia HV17

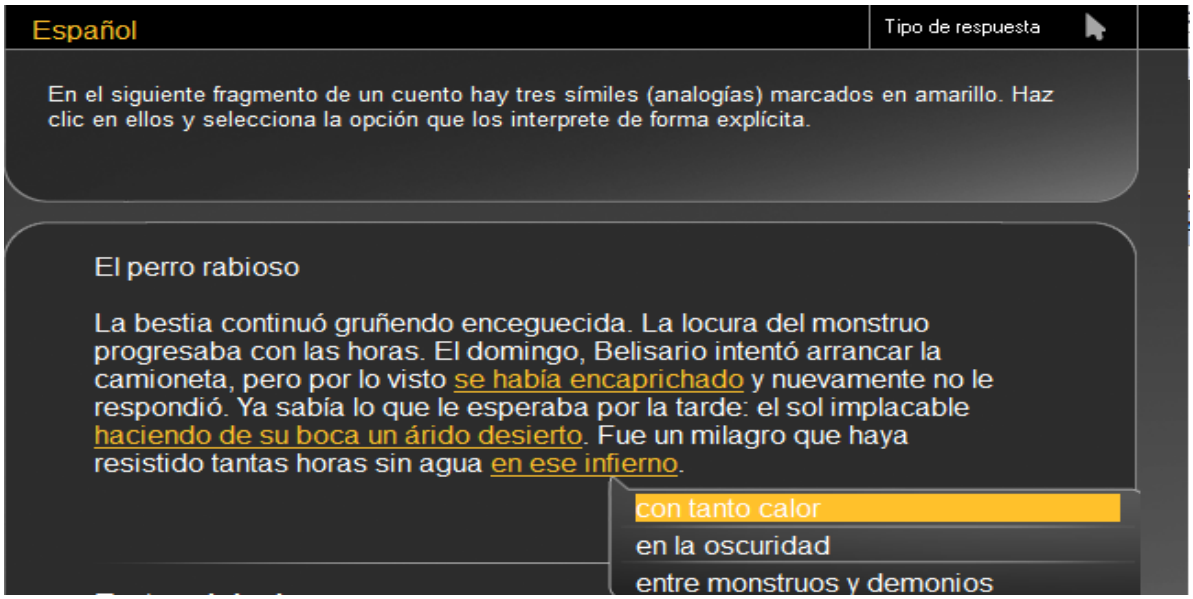


Figura 4.33. Ejemplo de ítem de la familia de reactivos HV17.

Como se aprecia en la tabla 4.25, los reactivos son bastante sencillos y en muchos casos, los distractores no cumplen con su función de *distraer*, ya que sus frecuencias son muy bajas. Los reactivos con mayor *presencia* de sus elementos incorrectos son los ítems 2, 3 y 4.

Tabla 4.25.
Lista de elementos correctos y distractores, con sus respectivas frecuencias, de los tres símiles para cada uno de los 6 ítems de HV17

	Ítem 1			**Ítem 2**			**Ítem 3**		
	texto	frec	p	texto	frec	p	texto	frec	p
Símil 1	de complexión delgada	157	0.94	muy negro	150	0.90	era cada vez mejor	137	0.83
	de mente cerrada	9		escondido en la sombra	9	0.05	mejoraba con el tiempo	21	0.13
	de estatura baja	1		misterioso	5	0.03	se recuperaba con el tiempo	5	0.03
	no contesta				3			4	
Símil 2	de piel clara	165	0.99	estaba muy enojado	88	0.53	tenía mal motor	114	0.69
	de piel fría	1		agitaba la cabeza	39	0.23	decidió no hacerlo	21	0.13
	de piel rosada	1		estaba muy enojado	37	0.23	se enojó con él	28	0.17
	no contesta			no contesta	3		no contesta	4	

Símil 3	de mal carácter	152	0.91	parados	96	0.57	causándole mucha sed	155	0.93
	de mejillas rojizas	10		filosos y tiesos	50	0.30	causando mal sabor en su boca	4	0.02
	de mal olor	5		rizados y ásperos	18	0.11	provocando que se callara	4	0.02
	no contesta				3			4	
P total			**.95**			**.67**			**.81**

	ítem 4			**ítem 5**			**ítem 6**		
	texto	frec	p	texto	frec	p	texto	frec	p
Símil 1	celestes	144	0.86	sin parar	157	0.94	sin parar	162	0.97
	con mirada perdida	1		de forma inesperada	3		de forma muy ruidosa	3	
	llenos de vida	20	0.12	de forma muy ruidosa	5		de forma inesperada	0	
	no contesta	2		no contesta	2		no contesta	2	
Símil 2	con dentadura blanca	96	0.57	provocaban inundaciones	162	0.97	no se veían	158	0.95
	dulce y hermosa	15	0.09	hacían salir a la gente	1		eran cometas	4	
	elegante y resplandeciente	55	0.33	llenaban de gente las calles	1		eran meteoros	2	
	no contesta	1		no contesta	3		no contesta	3	
Símil 3	güero y largo	131	0.78	con facilidad se hacían pedazos	155	0.93	grandes cantidades de lluvia	156	0.93
	delgado y fino	25	0.15	eran muy grandes y fuertes	3		muchos animales acuáticos	8	
	elegante y delicado	11	0.07	se veían indestructibles	7		víboras de cascabel	0	
	no contesta				2			3	
P total			**.74**			**.94**			**.95**

En el análisis Rasch, los valores de *infit* y *outfit* estandarizados no se escaparon de los rangos de aceptabilidad (figura 4.34). Se detectaron tres elementos con correlación inferior a 0.20, dos del primer ítem y uno del sexto (figura 4.35). Un comportamiento extraño es que, según el modelo de Rasch, los mejores ítems fueron el 2, 3 y 4, tanto como reactivos completos,

como por elementos. No obstante, las cargas factoriales no se agruparon en torno a estos ítems. Los índices de discriminación por elemento fueron muy buenos en general, solamente uno presentó discriminación menor que 0.80 (ver figura 4.36).

ENTRY NUMBER	MEASURE - +	INFIT STANDARDIZED -3 -2 -1 0 1 2 3	OUTFIT STANDARDIZED -3 -2 -1 0 1 2 3	ITEM
1	*	: * :	: . * :	ITEM 1.1
2	*	: .* :	: . * :	ITEM 1.2
3	*	: . * :	: . * :	ITEM 1.3
4	*	: * . :	: * . :	ITEM 2.1
5	*	: . * :	: . * :	ITEM 2.2
6	*	: * :	: * :	ITEM 2.3
7	*	: . * :	: .* :	ITEM 3.1
8	*	: .* :	: . * :	ITEM 3.2
9	*	: * . :	: * . :	ITEM 3.3
10	*	:* . :	: * . :	ITEM 4.1
11	*	: * . :	: *. :	ITEM 4.2
12	*	: * . :	: * . :	ITEM 4.3
13	*	: . * :	: . * :	ITEM 5.1
14	*	: * :	: *. :	ITEM 5.2
15	*	: .* :	: .* :	ITEM 5.3
16	*	: .* :	: . * :	ITEM 6.1
17	*	: *. :	: * . :	ITEM 6.2
18	*	: * :	: . * :	ITEM 6.3

Figura 4.34. Índices de dificultad, *infit* y *outfit* de los elementos de la familia HV17.

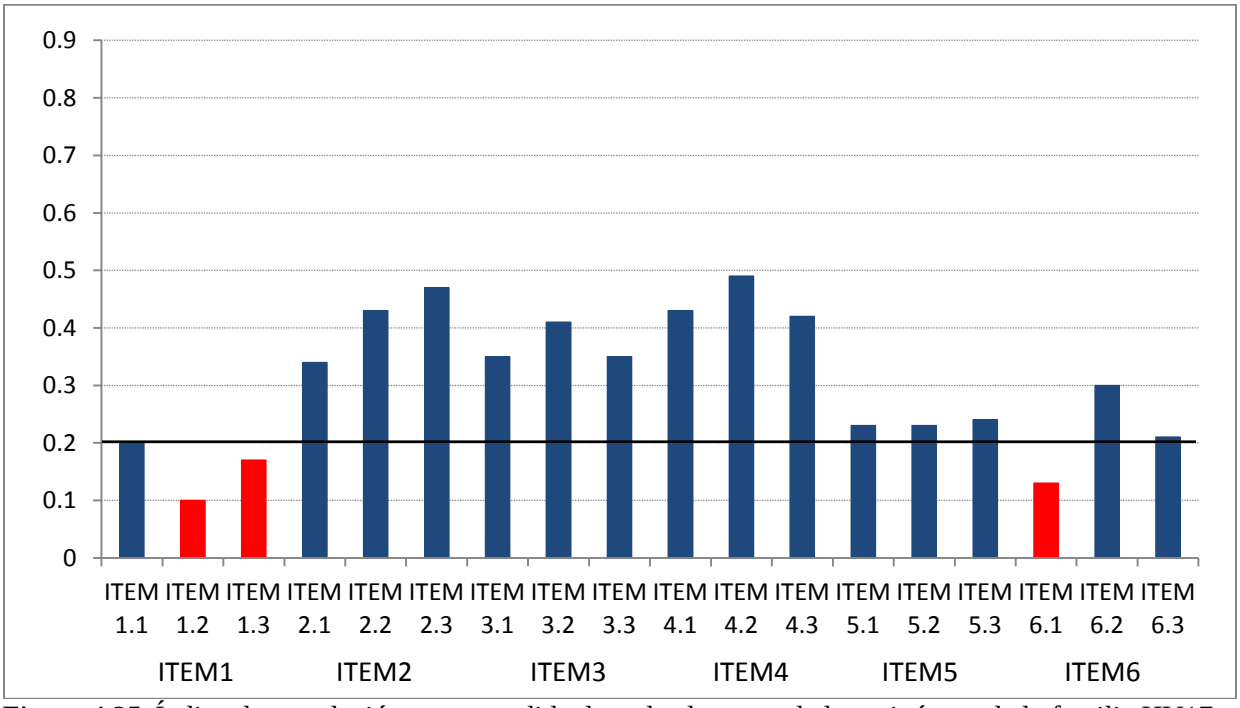


Figura 4.35. Índice de correlación punto medida de cada elemento de los seis ítems de la familia HV17

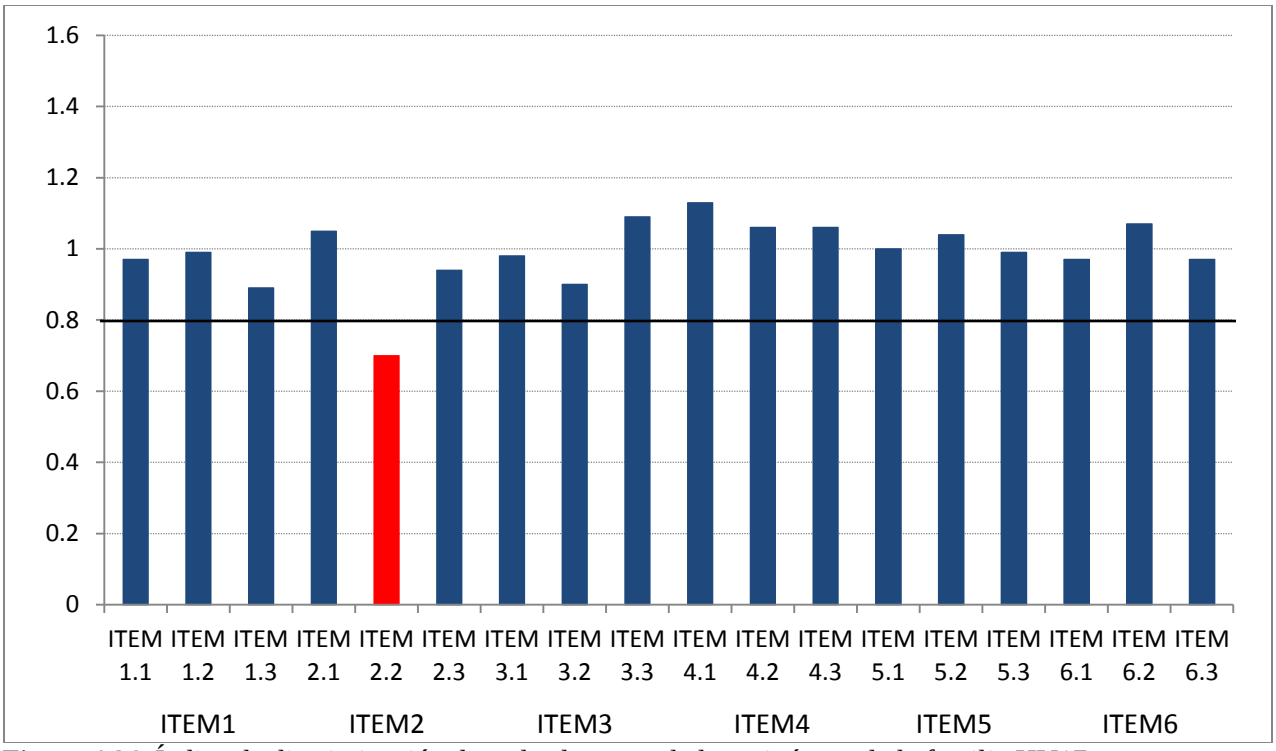


Figura 4.36. Índice de discriminación de cada elemento de los seis ítems de la familia HV17.

4.2.3.2. Elementos de la familia HC02

En el área de Habilidades matemáticas solamente tres ítems admitieron respuestas parciales, de ellos se seleccionó HC02 del tipo elemento-imagen, por ser el que presentaba mayor variedad de elementos disponibles y posibles respuestas. El reactivo consiste en ubicar tres fracciones en la recta numérica. Los elementos que conformaron los seis ítems hijos de la familia HC02 son 18, tres por cada ítem-hijo (ver figura 4.37).

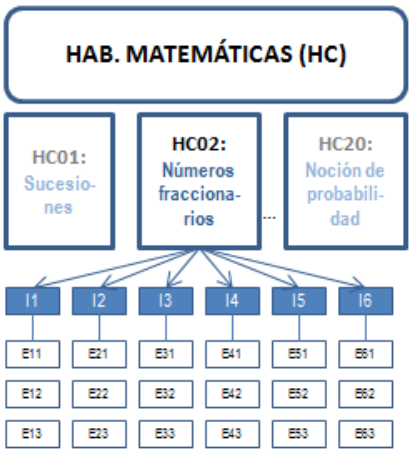


Figura 4.37. Esquema de los elementos de seis ítems de la familia HC02

En la figura 4.38 se puede apreciar que los *infit* y *outfit* estandarizados de todos los elementos se ubicaron en la franja deseada. También se observa la medida de cada elemento. Por ejemplo, la fracción más sencilla de ubicar fue $\frac{13}{10}$, seguramente se debe a que la unidad que se presenta en el ítem está dividida en décimos. La siguiente, en dificultad, es $\frac{2}{2}$, y luego $\frac{4}{2}$, $\frac{2}{2}$ y $\frac{1}{6}$. La fracción más difícil fue $\frac{4}{7}$ (lo cual es predecible, ya que dividir por 7 es más complejo). Todos los elementos correlacionaron por encima de 0.3 (ver figura 4.39). Los índices de discriminación por ítem afectaron a los reactivos 1 y 4; los índices de discriminación por elemento mostraron valores bajos en los elementos de estos ítems y también en dos del reactivo 2 y uno del reactivo 6 (ver tabla 4.17 y figura 4.40).

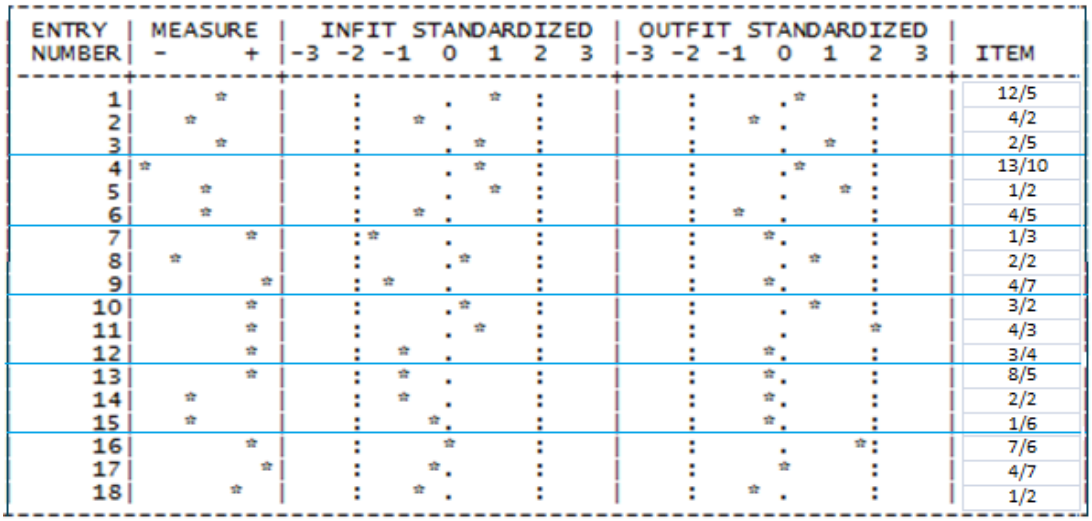


Figura 4.38. Medida, *Infit* y *outfit* estandarizados de los elementos de seis de la familia HC02.

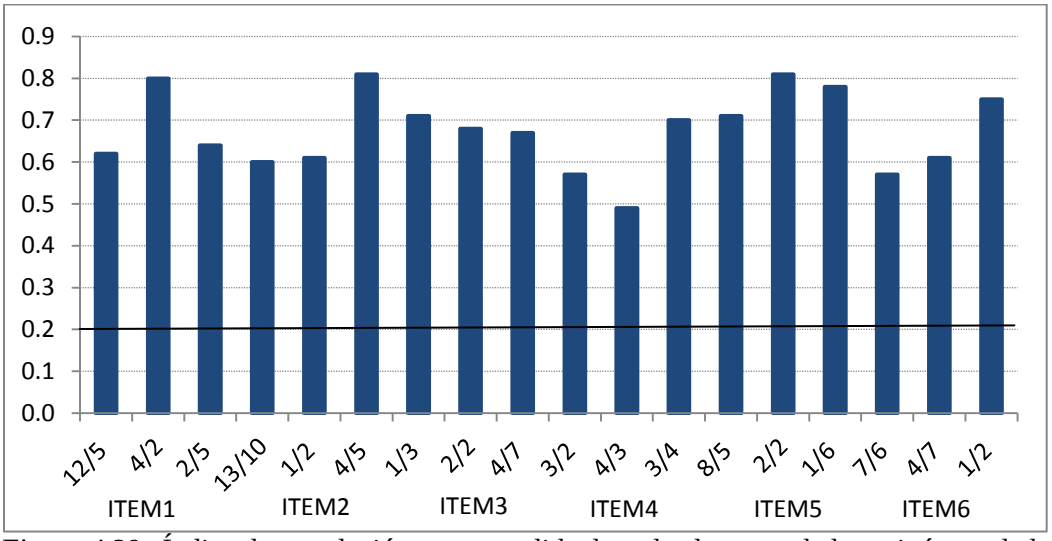


Figura 4.39. Índice de correlación punto medida de cada elemento de los seis ítems de la familia HC02

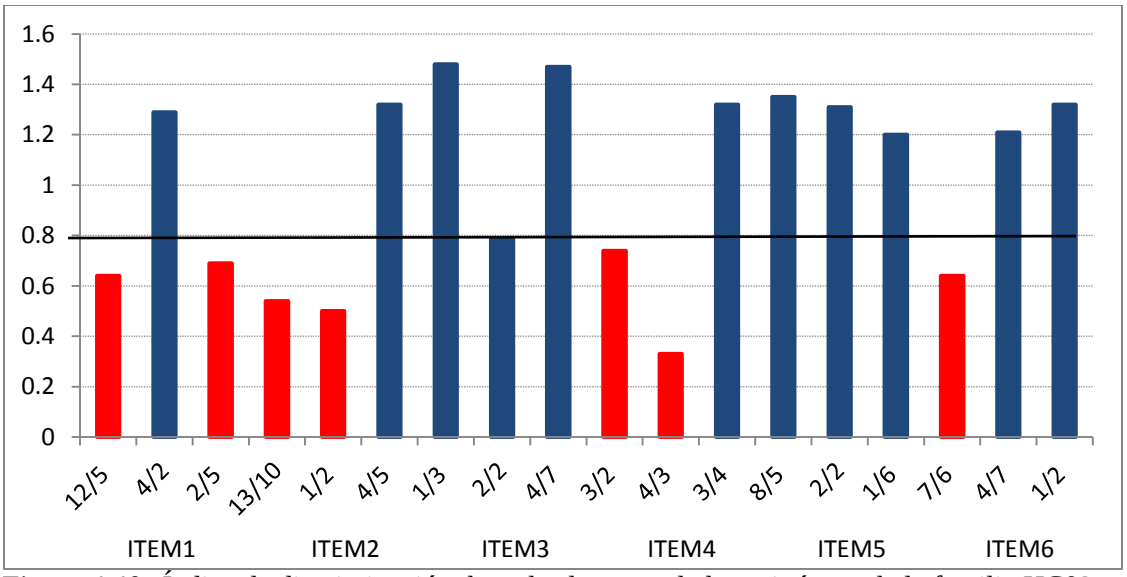


Figura 4.40. Índice de discriminación de cada elemento de los seis ítems de la familia HC02

4.2.3.3. Elementos de la familia ESP18

El área de Español cuenta con 11 familias del tipo elemento-categoría, 6 de selección-elemento, 2 de frase-imagen y 1 de selección-frase. De los 20 reactivos, solamente uno es dicotómico, el resto solicitan dos o más respuestas. Por lo tanto, se decidió tomar una familia del tipo más numeroso. Con este criterio, se seleccionó, al azar la de ESP18, del tipo elemento-categoría. Los elementos que conformaron los seis ítems-hijo de la familia ESP18 son 18, tres por cada ítem hijo (ver figura 4.41).

Figura 4.41. Esquema de los elementos de seis ítems de la familia ESP18

En este reactivo se presenta un artículo de opinión y tres enunciados referentes a dicho artículo. La tarea consiste en decidir si representan o no, la opinión del autor del texto. En la figura 4.42 se pueden apreciar dos elementos fuera del rango de aceptación: el elemento 2 del ítem 1 (tanto *infit* como *outfit*) y el elemento 1 del ítem 3 (en su *outfit*). También se observa que los elementos del ítem 3 fueron más sencillos que el resto y los más difíciles se concentraron en el ítem 6. En cuanto a la correlación punto medida, todos los elementos superaron la barrera de 0.20 (ver figura 4.43). Mientras que la discriminación por ítem no presentó ningún valor inferior a 0.80 en la familia ESP18 (ver anexo F, apartado de Español), los índices de discriminación por elemento mostraron valores bajos en dos de ellos: el segundo del ítem 1 y el primero del ítem 3 (ver figura 4.44).

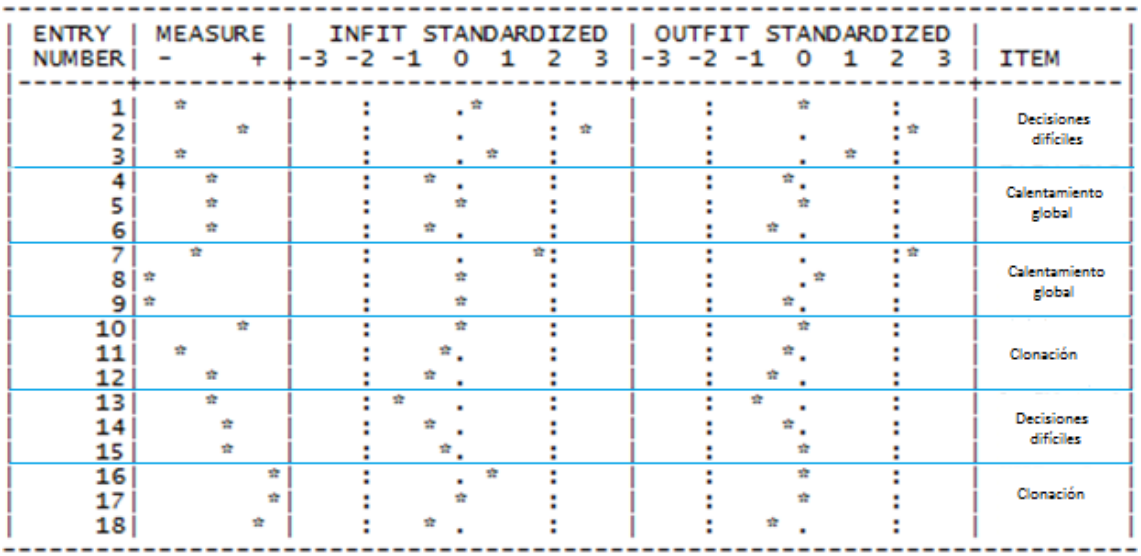


Figura 4.42. Medida, *Infit* y *outfit* estandarizados de los elementos de seis ítems de la familia ESP18.

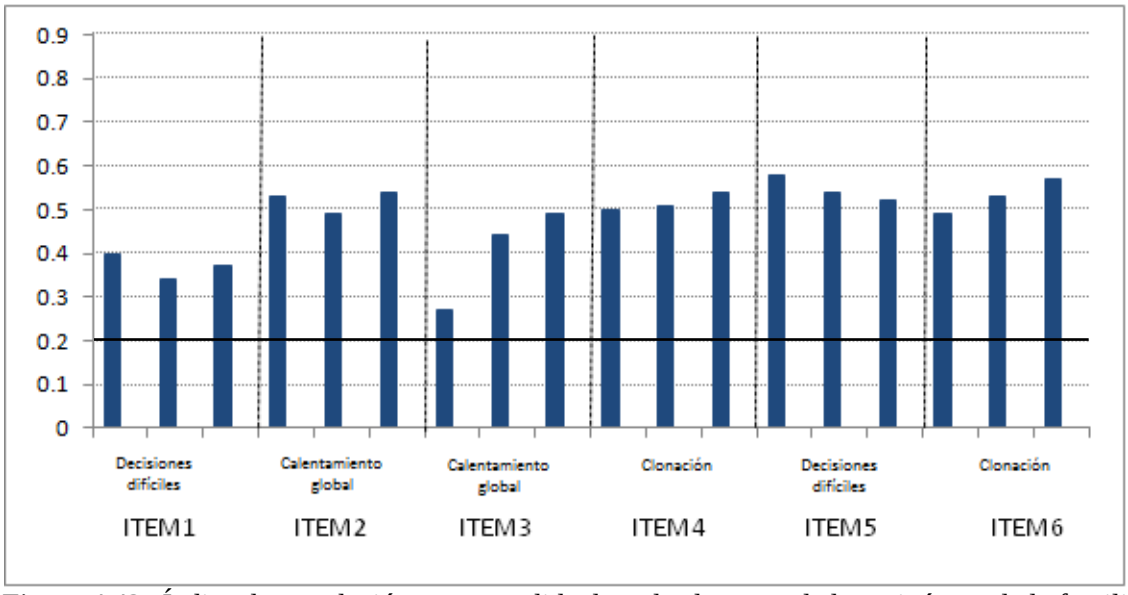


Figura 4.43. Índice de correlación punto medida de cada elemento de los seis ítems de la familia ESP18

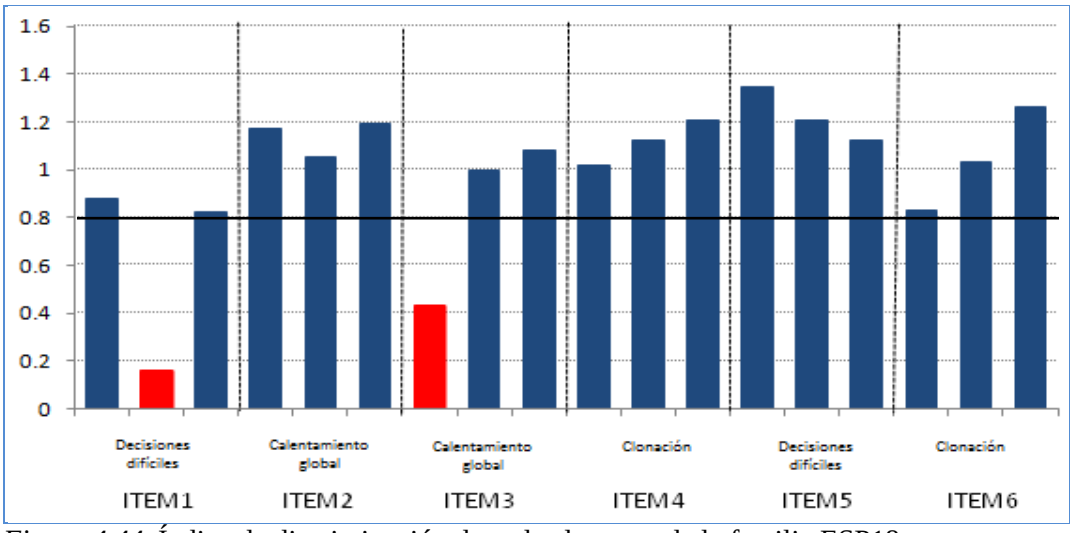


Figura 4.44. Índice de discriminación de cada elemento de la familia ESP18.

4.2.3.4. Elementos de la familia MAT09

El área de Matemáticas posee una gran diversidad de tipos de ítems; sin embargo, de todos ellos, el único que contiene ítems de crédito parcial es elemento-categoría. De las dos familias disponibles, se seleccionó al azar, MAT09 y se analizaron los 18 elementos correspondientes a seis ítems (ver figura 4.45). En este reactivo se presentan tres imágenes, cada una cortada por una recta, se solicita determinar qué figuras son simétricas respecto a la recta que las divide.

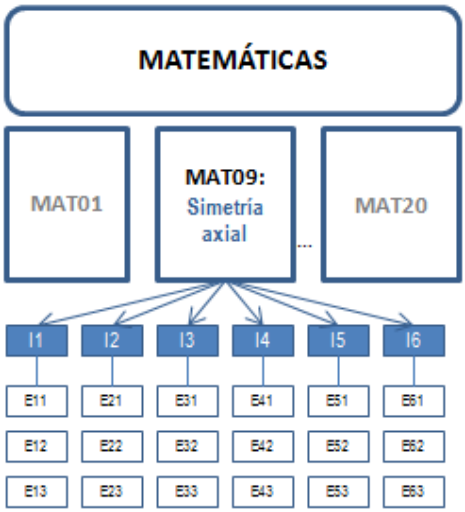


Figura 4.45. Esquema de los elementos de seis ítems de la familia MAT09

En la figura 4.46 se puede apreciar que los *infit* y *outfit* estandarizados de todos los elementos se ubicaron en la franja deseada, excepto el *infit* del segundo elemento del primer ítem ( ). También se observa la medida de cada elemento, por ejemplo, el primer elemento del ítem 1 fue el más sencillo ( ), y los elementos de mayor dificultad fueron el tercero del ítem 5 ( ) y el tercero del ítem 2 ( ).En cuanto a la correlación punto medida, se detectaron dos elementos con correlación inferior a 0.20 (ítem 1.2 e ítem 5.3) (ver figura 4.47). Los índices de discriminación por ítem no afectaron a los reactivos en general, sin embargo, cinco elementos mostraron valores bajos (ver Anexo F, apartado de Matemáticas y figura 4.48).

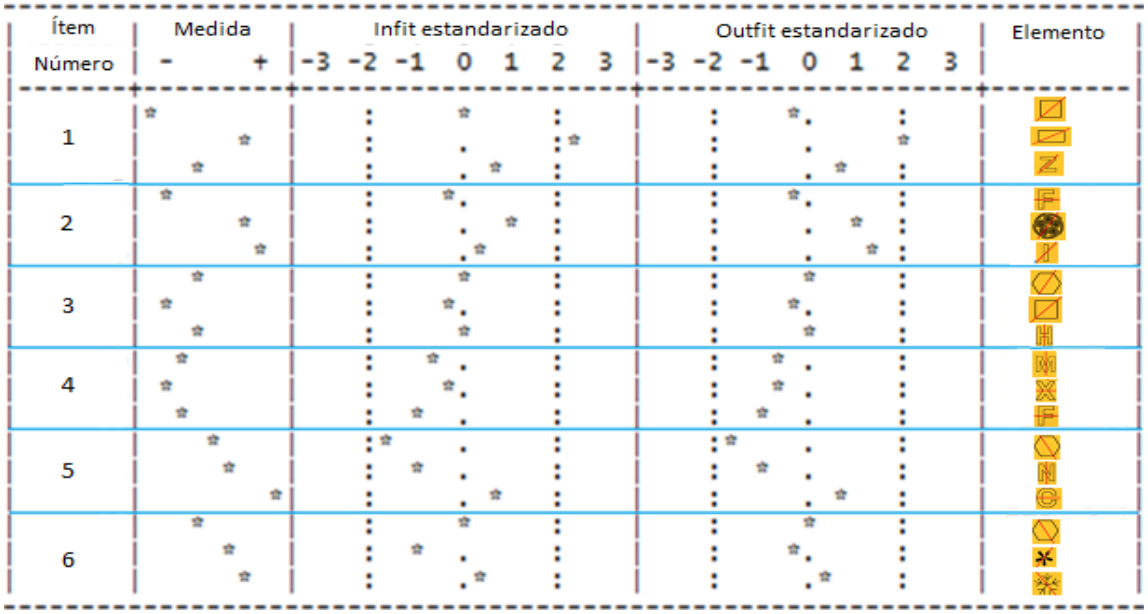


Figura 4.46. Medida, *Infit* y *outfit* de los elementos de la familia MAT09.

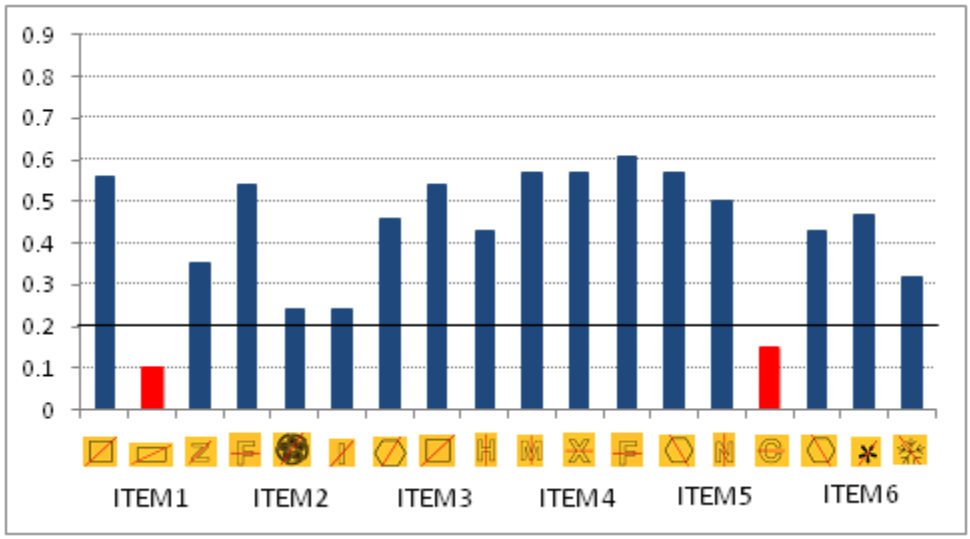


Figura 4.47. Índice de correlación punto medida de cada elemento de los seis ítems de la familia MAT09

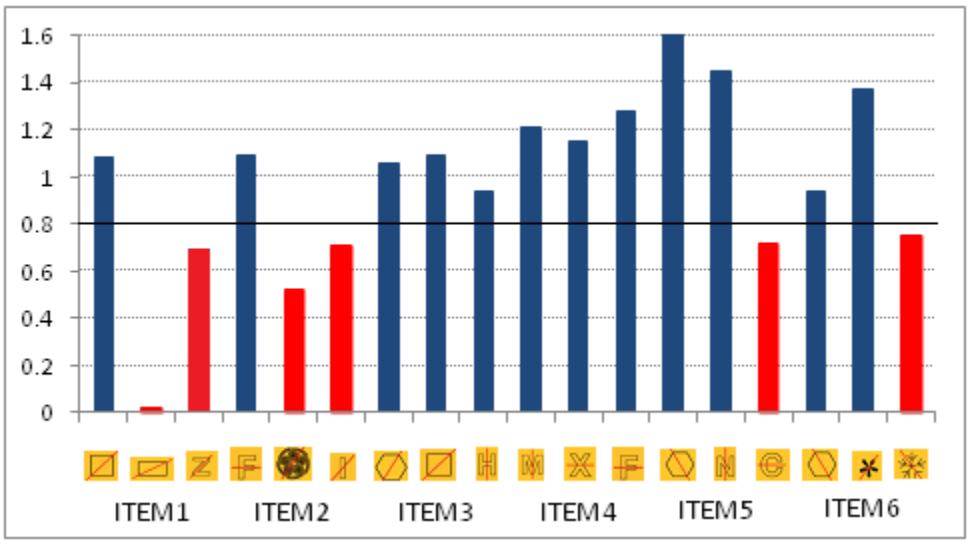


Figura 4.48. Índice de discriminación de cada elemento de seis ítems de la familia MAT09.

4.2.3.5. Elementos de la familia FIS12

El área de Ciencias naturales cuenta con tres tipos de reactivos que se califican con crédito parcial: elemento-categoría, selección de elementos y selección de elementos múltiples. De estos, la selección de elementos múltiples es propia del área de naturales. De uno de ellos, QUI13, no se pudo decodificar la información. Por lo tanto, se decidió estudiar la familia de FIS12. Para cada ítem son 6 elementos, con lo cual, para los 6 ítems corresponden 36 elementos (ver figura

4.49). En este reactivo se presentan tres imágenes de aplicaciones domésticas y etiquetas con nombres de tipos de energía; para cada aplicación se solicita elegir el tipo de energía que este requiere para funcionar y la que libera.

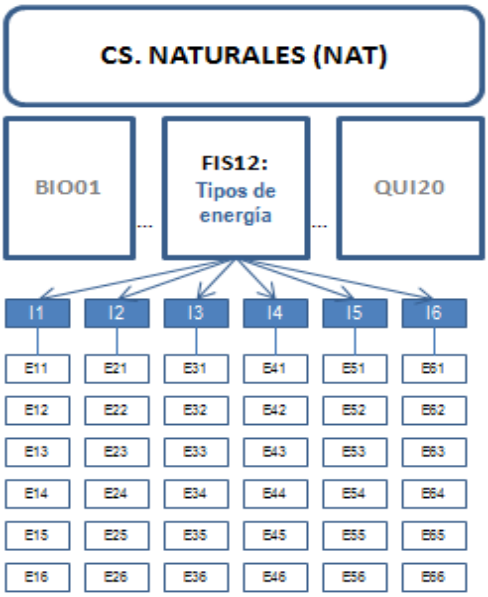


Figura 4.49. Esquema de los elementos de seis ítems de la familia FIS12

El análisis Rasch arrojó buenos ajustes de *infit* y *outfit* (figura 4.50). En cuanto a la dificultad, en general, los elementos más fáciles coinciden con la *energía eléctrica*, sobre todo cuando se trata de la energía necesaria para funcionar. Se observa que los elementos del cuarto ítem son todos sencillos, mientras que en los demás reactivos, existen unos elementos más fáciles que otros. Tras el análisis de correlaciones se detectaron dos elementos con coeficiente inferior a 0.20 (el primer elemento del ítem 2 y el segundo del ítem 5) (ver figura 4.51). Los índices de discriminación efectuados por ítem no afectaron a los reactivos en general; sin embargo, por elemento mostraron valores bajos en cuatro elementos del ítem 1, en tres del ítem 2 y en uno de los 3, 5 y 6 (ver Anexo F, apartado de Ciencias naturales y figura 4.52).

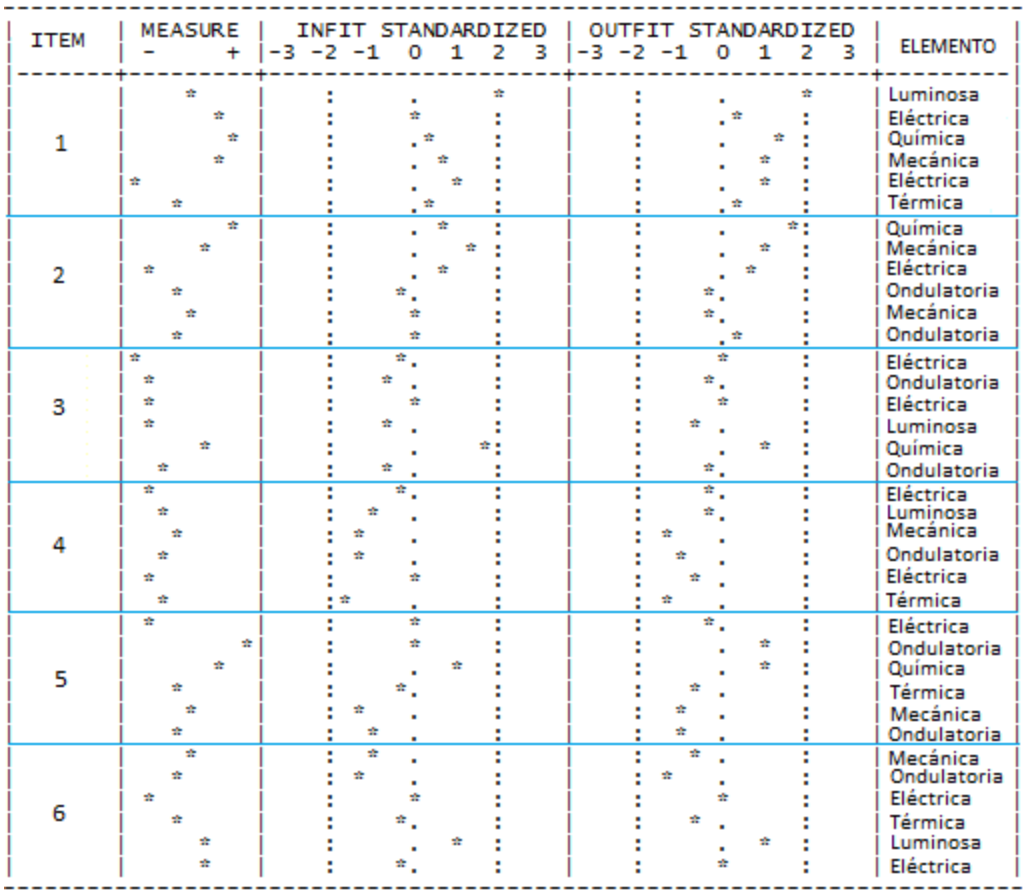


Figura 4.50. Medida, *Infit* y *outfit* de los elementos de seis ítems de la familia FIS12.

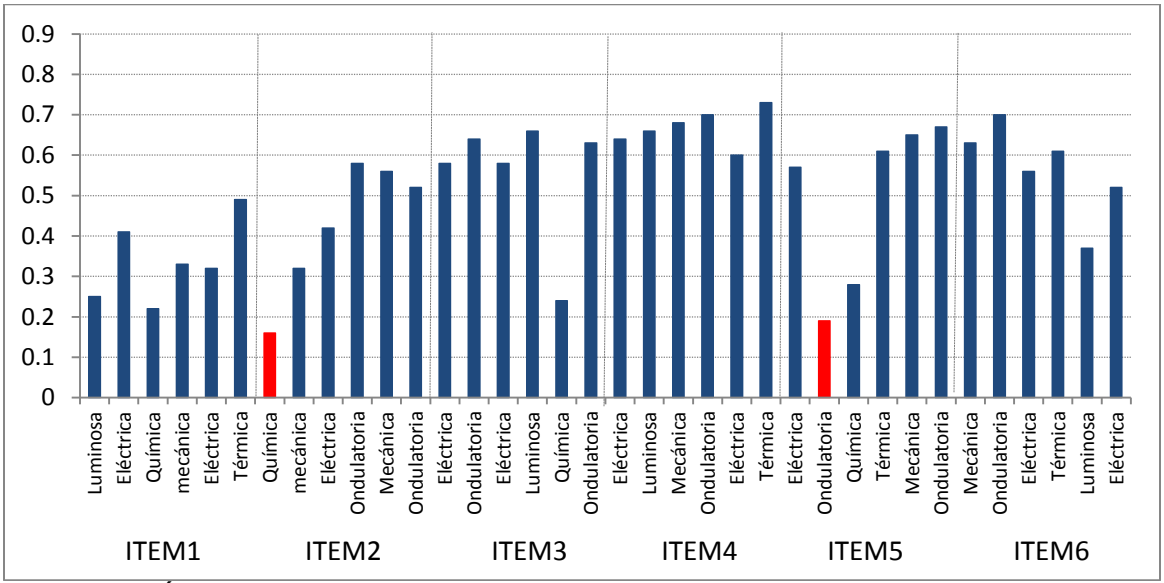


Figura 4.51. Índice de correlación punto medida de cada elemento de seis ítems de la familia FIS12

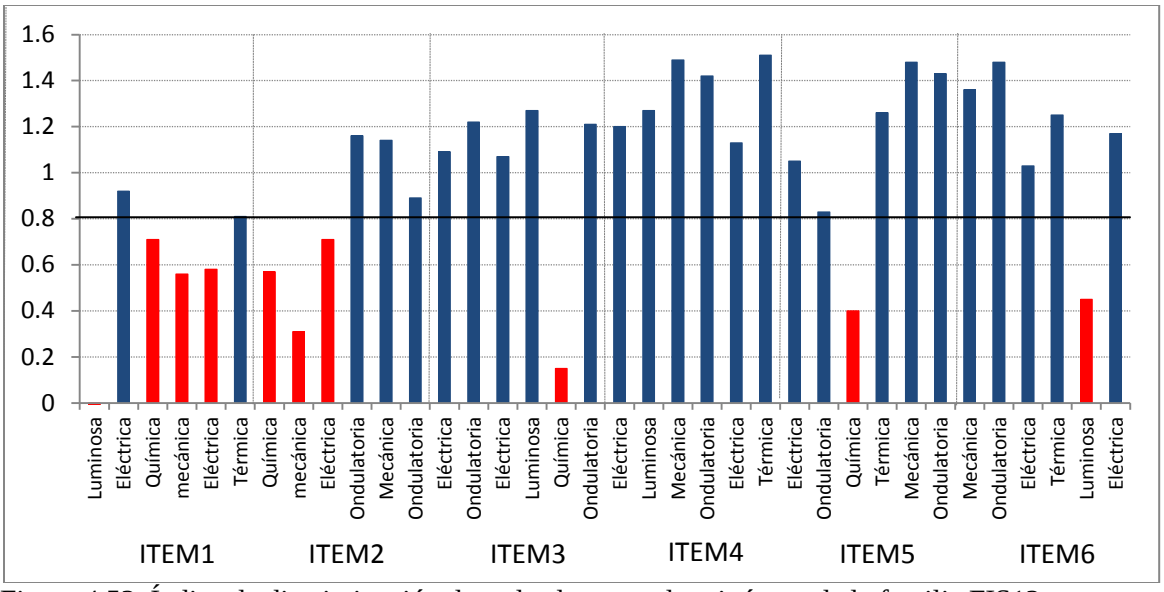


Figura 4.52. Índice de discriminación de cada elemento de seis ítems de la familia FIS12.

4.2.3.6. Elementos de la familia HIS07

En el caso de Ciencias sociales, se utilizan únicamente dos tipos de reactivos: tres familias son del tipo elemento-imagen y el resto, de elemento-categoría, todos son de crédito parcial. Se seleccionó, al azar, HIS07 que pertenece a elemento-categoría. En este reactivo se presentan cinco acontecimientos (cinco elementos) que transcurrieron entre 1750 y 1850, y se solicita ubicarlos en el tipo de revolución atlántica al que pertenecieron (se muestran tres de cuatro revoluciones: francesa, hispanoamericana, inglesa y norteamericana). Por lo tanto, los elementos que conformaron los 6 ítems hijos de la familia HIS07 son 30, cinco por cada ítem hijo (ver figura 4.53).

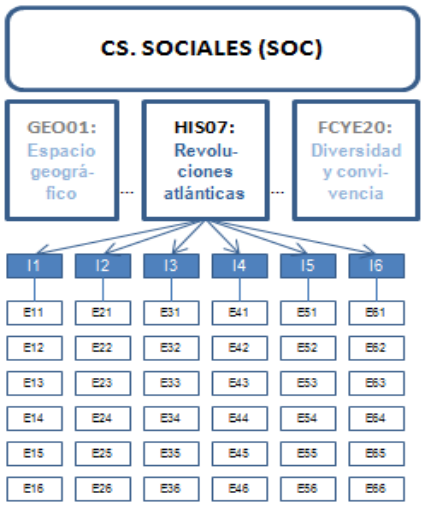


Figura 4.53. Esquema de los elementos de seis ítems de la familia HIS07

En el análisis Rasch se detectó que el elemento 5 del ítem 3 (ítem3.5) presentaba un *infit* estandarizado fuera del rango (-2,2), según se muestra en la figura 4.54. En cuanto a la correlación punto medida, se detectaron dos elementos con correlación inferior a 0.20, el peor de todos, el quinto elemento del tercer ítem (ver figura 4.55). Los índices de discriminación por ítem solamente mostraron una ligera deficiencia en el reactivo 2; sin embargo, siete elementos presentaron valores bajos: uno del ítem 2, uno del 3, dos del 4 y tres del 5 (ver Anexo F, apartado de Ciencias sociales y figura 4.56). Una revisión de los contenidos permitió verificar que el elemento 3.5 no correspondía al periodo histórico que se evaluaba en el contenido y, por lo tanto, les provocaba ruido a las personas examinadas.

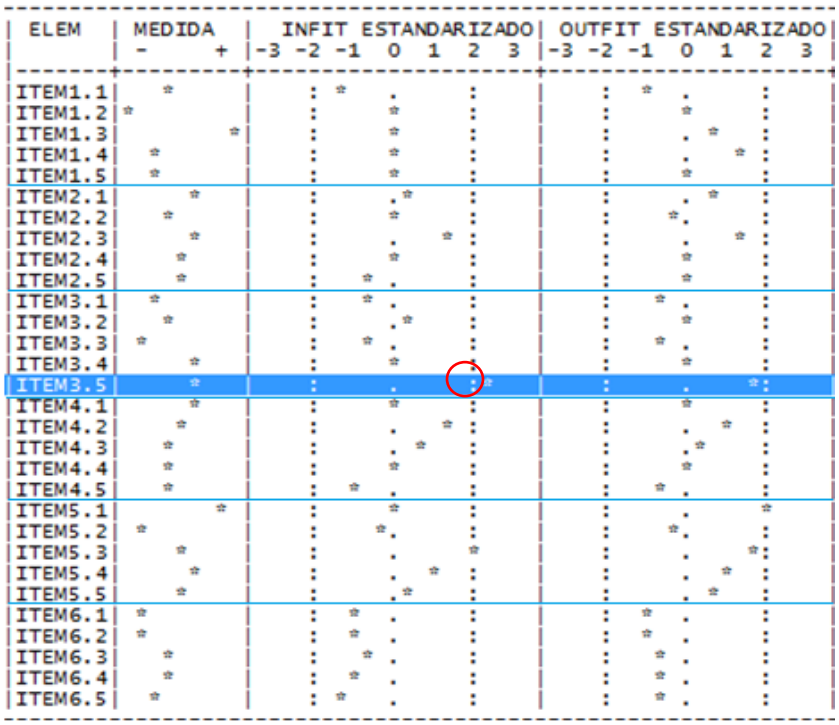


Figura 4.54. Medida, *infit* y *outfit* de los 30 elementos correspondientes a los 6 ítems de la familia HIS07 de la muestra SOC.

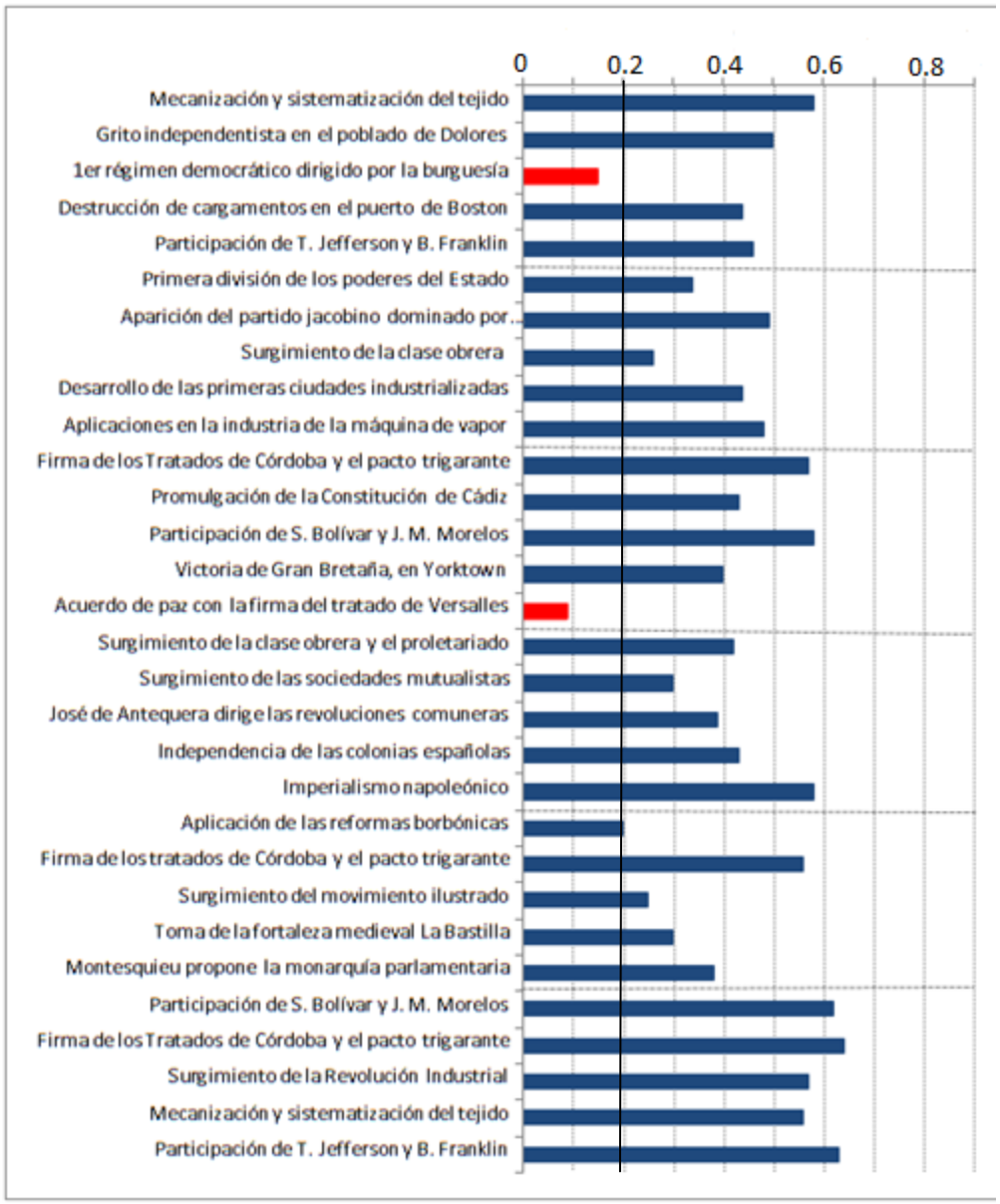


Figura 4.55. Índice de correlación punto medida de cada elemento de seis ítems de la familia HIS07

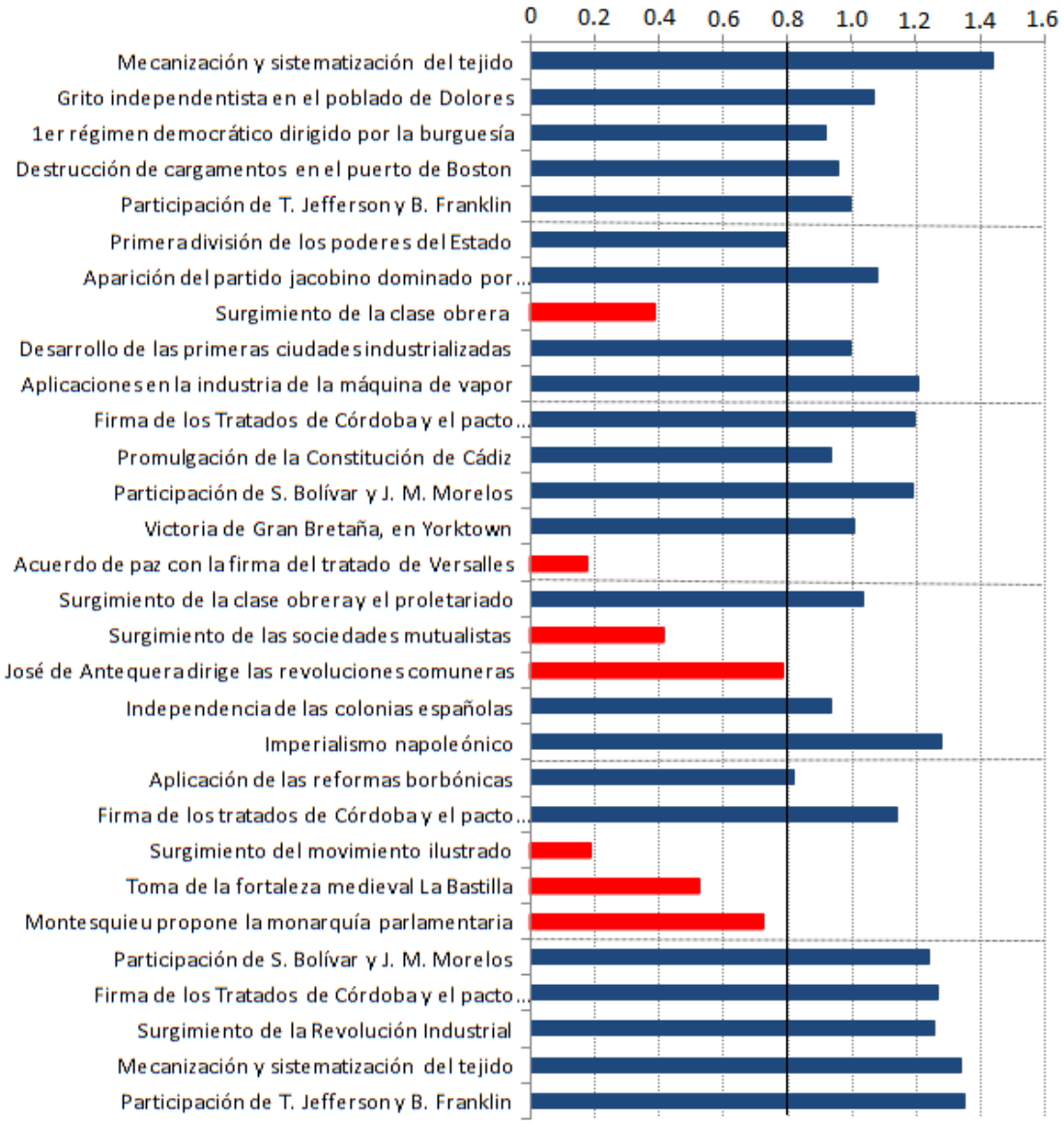


Figura 4.56. Índice de discriminación de cada elemento de seis ítems de la familia HIS07.

5

Discusión y conclusiones

La pregunta de investigación que guió el presente trabajo de tesis es ¿cómo obtener evidencias de validez, basadas en la estructura interna, de exámenes producidos a través de la Generación Automática de Ítems? La respuesta fue dada a través de la propuesta de una metodología, de su aplicación al EXHCOBA-R/MS y de los resultados obtenidos tras su implementación. Se desprende, entonces, la necesidad de discutir acerca de los beneficios de este método, de su alcance, de sus limitaciones y de cómo continuar el proceso de validación en próximas investigaciones.

Primeramente, es necesario recordar qué tipo de examen es el EXHCOBA-R/MS. Se trata de una prueba de alto impacto, que se aplica a gran escala, para evaluar a quienes aspiran ingresar a la EMS. Cabe aclarar que es un examen de conocimientos de educación básica íntegramente alineado a los planes de estudios de la educación primaria y secundaria mexicanas. No se trata de un test adaptativo; en cada administración se evalúan los mismos contenidos a través de una cantidad fija de ítems (120 en total).

En segundo lugar, se requiere ubicar al examen dentro del marco teórico sobre el cual se originó. Para su estructuración se utilizó el modelo de desarrollo de exámenes propuesto por el Instituto Nacional para la Evaluación de la Educación (INEE, 2005) con algunas modificaciones del Comité Técnico del EXHCOBA. La estructura consistió en grupos colegiados que trabajaron en forma escalonada. Cada grupo cumplió con una función específica (e.g.: estructura del examen, selección de contenidos, elaboración de especificaciones, entre otras) y complementaria en el proceso de construcción. Así, los productos de un equipo en una etapa se convirtieron en insumos de otro en la siguientes etapa; por lo que el proceso de desarrollo de la prueba se

consideró, en parte, el inicio de su validación (Contreras, 2000; Contreras, Backhoff y Larrazolo, 2003).

Hasta aquí, las características del nuevo examen no difieren del tradicional EXHCOBA. La tercera característica es la que implica mayor novedad y desafíos en la validación; para cada contenido se desarrollaron especificaciones con modelos de reactivos que permiten construir una gran cantidad de ítems similares, agrupados en familias, y que evalúan dicho contenido. De cada familia se puede elegir un ítem de manera aleatoria y, así, conformar un examen único de 120 reactivos. En otras palabras, el EXHCOBA-R/MS puede producir diferentes versiones similares, como resultado de la GAI.

La GAI puede desarrollarse desde dos aproximaciones teóricas: una fuerte y la otra débil (Drasgow, Luecht y Bennett, 2006). La primera teoría está sustentada en modelos cognitivos precisados en *modelos de tareas* que determinan las características cognitivas del contenido, establecen un registro teórico de los elementos que afectan el nivel de dificultad de los reactivos generados y de las habilidades requeridas para resolverlos. La segunda teoría carece del soporte cognitivo, en ella se utilizan guías de diseño (*especificaciones*) con el objetivo de crear modelos de ítems que generen reactivos isomorfos. En el caso específico del EXHCOBA-R, la GAI está fundada en la teoría débil, ya que —como señalaron Gierl y Lai (2012) — hay pocas teorías cognitivas para guiar el desarrollo de reactivos para la gran cantidad de contenidos que se necesitan en los exámenes educativos.

Dentro de los exámenes producidos por GAI se plantearon en el capítulo dos, del Marco de referencia, tres metodologías diferentes de validación con recursos estadísticos, las cuales fueron propuestas por Sinharay y Johnson (2012). La primera metodología consiste en el análisis a través de modelos componenciales como el LLTM, de Fischer, previa aplicación de modelos

de la TRI (el de Rasch, para el caso del LLTM). Para ello se necesita que la GAI se sustente en una teoría fuerte, con habilidades subyacentes a cada ítem, definidas en modelos de tareas. La segunda metodología apunta al estudio de las familias de ítems y no al examen en general; dicho estudio está enfocado en controlar que los ítems-hermano sean isomorfos. Este método está especialmente dirigido hacia los tests adaptativos, ya que aseguraría un correcto reemplazo de un ítem por su clon, sin alterar la dificultad del reactivo ni la habilidad evaluada, y no tanto a la estructura completa del examen. La tercera propuesta combina las dos metodologías anteriores.

Por lo anterior, surge la pregunta acerca de cuál de las tres aproximaciones expuestas es la que se adecua al EXHCOBA-R/MS. En primer lugar, por tratarse de un examen basado en teoría débil, no se cuenta con un modelo 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. En segundo lugar, se trata de analizar diferentes versiones de un mismo examen, es decir, no es suficiente conocer el comportamiento de las familias de ítems, debe asegurarse que las estructuras de las diferentes versiones sean también semejantes. Por lo tanto, si las condiciones muestrales lo permitiesen podría implementarse la segunda metodología, aunque el análisis estadístico se reduciría a las familias de ítems y no contemplaría el examen completo. En cuanto a la tercera metodología que implica la unión de las dos primeras, tampoco es factible por las razones mencionadas en los dos casos anteriores.

Por lo tanto, se planteó la necesidad de implementar un método que evalúe, por un lado, el examen completo con sus posibles versiones (a pesar de no contar con un modelo cognitivo que lo sustente) y, por otro lado, que evalúe cada familia de ítems para verificar qué tan isomorfos son los ítems-hermano. La consigna fue utilizar las teorías psicométricas conocidas y

emplear los modelos más parsimoniosos para explicar ambos comportamientos estadísticos, del examen y de las familias. Otra condición para la tarea fue que, como se trataba de las primeras aplicaciones, estas deberían ser a modo de pilotaje y con muestras pequeñas.

Se partió de dos supuestos: (1) el modelo colegiado para la estructuración del examen, propuesto por el INEE, que le proporcionó un sostén orgánico para comenzar el proceso de validación, y (2) los modelos de ítems explicados por Gierl, Zhou y Alves (2008). La metodología utilizada para obtener evidencias de estructura interna del EXHCOBA-R/MS descansa, principalmente, en el supuesto de que la evaluación de cada contenido está estructurada por los modelos de ítems. Cada modelo manipula y controla tanto el contenido como la dificultad de los ítems, de modo que reproduce tareas evaluativas idénticas para un mismo contenido (Bejar, 2002; Gierl y Lai, 2011; Gierl, Zhou y Alves, 2008). Cada modelo es independiente y es desarrollado por un único experto, lo cual aporta también uniformidad a una familia de ítems. No ocurre como en los exámenes tradicionales donde cada ítem, aunque evalúe un mismo constructo, es producido por elaboradores diferentes, según la versión de la prueba que se desarrolle (Drasgow et al., 2006). En consecuencia, los modelos de ítems tienen la función de *precalibrar* los reactivos; de este modo, eliminan la necesidad de tratar cada ítem de manera individual como se realiza en los tests tradicionales (Drasgow et al., 2006; Gierl, Zho y Alves, 2008).

De lo expuesto se infiere que, si bien no se requería analizar todos los exámenes generados por la GAI ni cada ítem-hijo de un modelo de ítem, era necesario aplicar un método estadístico que respaldara los desarrollos de esta GAI de teoría débil. Para ello, se utilizaron dos tipos de muestras tomadas al azar: un tipo correspondiente a un examen tal cual se aplicaría para el ingreso a la EMS y el otro tipo donde se incluyeran seis reactivos por familia de ítems. Se

consideró apropiado efectuar los análisis de los dos tipos de muestras desde la teoría más sencilla (la TCT) y desde el modelo más parsimonioso de la TRI (el modelo de Rasch para ítems dicotómicos y para ítems de crédito parcial). Ambos análisis permitieron aportar solidez a los resultados y, a su vez, complementarse. También se juzgó pertinente sustentar, a través del AFC, la agrupación de los ítems en constructos, tanto por área como por familia.

Una vez explicadas las razones del método elegido y sus características, a continuación se procede a *evaluar la eficacia de la metodología propuesta para la validación de la GAI* en el caso particular del EXHCOBA-R/MS. Para tal fin, el capítulo se desarrolla en tres apartados. En el primer apartado se expone una síntesis de los resultados obtenidos tras la aplicación de la metodología propuesta para la validación del EXHCOBA-R/MS y en qué medida el método utilizado respondió a las preguntas planteadas en la introducción de la presente tesis. En el segundo apartado se describen los alcances y las limitaciones del estudio. Estas últimas se clasifican en dos tipos: (a) prácticas, propias de las diferentes aplicaciones de los exámenes, y (b) inherentes a la metodología. Finalmente, en el tercer apartado se incluyen nuevas líneas de investigación que complementen y que consoliden la metodología para obtener evidencias de validez de estructura interna de exámenes obtenidos a través de la GAI.

5.1. Síntesis de los resultados obtenidos tras la aplicación de la metodología al EXHCOBA-R/MS

A continuación se presentan las preguntas de investigación *replanteadas* —ya no como la búsqueda de un método, sino como su aplicación para el caso concreto del EXHCOBA-R/MS— con sus respectivas respuestas. De este modo, se pretende establecer en qué medida la metodología aplicada aportó evidencias de validez del EXHCOBA-R/MS.

- *¿Cuáles son las propiedades psicométricas de los exámenes que se generan a través de la GAI?*

De los análisis estadísticos efectuados a las dos versiones —VA y VB del EXHCOBA-R/MS— se sintetiza que el examen (con la aplicación de 117 de los 120 ítems) tiene una buena confiabilidad en ambas versiones; además, los niveles de dificultad son similares (según se muestra en la última fila de la tabla 5.1). Sin embargo, aunque el comportamiento general es parecido en las dos pruebas, no todas las áreas reflejan la misma calidad de los ítems. Habilidades del lenguaje y Español, según lo mostraron las respuestas de los examinados, fueron las áreas más fáciles de todo el examen y sus índices de confiabilidad resultaron bajos. Habilidades matemáticas y Matemáticas fueron las más difíciles, los ítems de educación primaria tuvieron un mejor comportamiento que los de educación secundaria; probablemente se deba a que las dificultades de estos últimos son todavía mayores. Ciencias naturales mostró poca confiabilidad (particularmente en VB), mientras que Ciencias sociales arrojó muy buenos índices.

El análisis desde la TRI de VA y VB reflejó un comportamiento de los ítems similar al obtenido a través de la TCT. El modelo de Rasch para ítems de crédito parcial explicó el 38.5% y 37.3%, para ambas versiones, respectivamente, a través de sus medidas. Lo cual implica que, con una diferencia del 1.2%, ambas pruebas reflejan similitud y explican en la misma proporción, aproximadamente.

Además, una comparación de estas versiones con las aplicaciones del EXHCOBA tradicional, a los mismos estudiantes, mostraron correlaciones significativas al nivel de 0.01 entre las calificaciones de las dos modalidades. Lo mismo ocurrió entre las puntuaciones de la educación primaria y de la educación secundaria de dichos alumnos con su equivalente del

EXHCOBA-R/MS.

Tabla 5.1.
Dificultades media y confiabilidad del EXHCOBA-R/MS para VA y VB, examen completo y por áreas

Área	Ítems	VA		VB	
		Dificultad	Confiabilidad	Dificultad	Confiabilidad
HV	19	0.68	0.633	0.63	0.547
HC	20	0.38	0.784	0.38	0.788
ESP	20	0.69	0.587	0.66	0.706
MAT	19	0.26	0.655	0.24	0.691
NAT	20	0.48	0.612	0.47	0.502
SOC	19	0.47	0.869	0.48	0.877
EXHCOBA-R/MS	117	**0.52**	**0.902**	**0.50**	**0.897**

Nota: HV = Habilidades del lenguaje, HC = Habilidades matemáticas, ESP = Español, MAT = Matemáticas, NAT = Ciencias naturales, SOC = Ciencias sociales

- *¿En qué grado la estructura conceptual de los exámenes generados a través de la GAI concuerda con su estructura empírica?*

Analizadas VA y VB, por áreas, a través del modelo de Rasch, se identificaron algunos ítems con serios problemas de correlación, de ajuste o de discriminación, en ambas versiones (un ítem en HV, uno en HC, uno en ESP). Estas características se repitieron en los resultados desde la TCT, lo cual permite otorgar mayor certeza acerca del funcionamiento del examen.

Se identificaron ítems con problemas únicamente en una de las versiones (en particular, QUI13 de VA y QUI14 de VB). Esto es indicio de algunas deficiencias en las familias de estos ítems, lo cual indica falta de isomorfismo en los grupos y, en consecuencia, también en las versiones producidas por la GAI.

Para la agrupación por áreas (ver tabla 5.2), los AFC arrojaron mejores ajustes para los siguientes modelos: Ciencias sociales, Habilidades matemáticas y Español, cada uno en modelos de un factor; para las tres áreas restantes, cada una con modelos de dos factores. Se encontraron

desajustes para algunos ítems, estos problemas ya se habían detectado en el análisis completo de ambas pruebas.

Tabla 5.2.
Modelos de agrupación de ítems de cada una de las seis áreas del EXHCOBA-R/MS, para VA y VB

	HV	HC	ESP	MAT	NAT	SOC
Modelo	2 factores que covarían	1 factor	1 factor	2 factores que covarían	2 factores que covarían	1 factor
Factores	F1: Lectura y comprensión de textos F2: Gramática y ortografía	Habilidades matemáticas de educación primaria	Conocimientos de Español de educación secundaria	F1: sentido numérico, pensamiento algebraico y manejo de la información F2: forma, espacio y medida	F1: Biología y Química F2: Física	Ciencias Sociales
Ítems que no cargan al modelo en ambas versiones	HV15	HC07	ESP02	-- [a]	BIO01 QUI19	-- [b]

Nota: [a] No se detectaron problemas compartidos en ambas versiones. [b] No se detectaron problemas en ninguna versión.

Además, resulta necesario aclarar la separación del estudio en seis áreas del conocimiento y no en cuatro, puesto que Habilidades del lenguaje podría haberse fusionado con Español y algo equivalente hubiera sido unir Habilidades matemáticas con Matemáticas. Sin embargo, se optó por respetar la división original diseñada por el Comité Técnico y los distintos Comités Académicos del EXHCOBA, especialmente con la intención de agrupar por áreas según el nivel de aprendizaje asociado (educación primaria y educación secundaria).

- *Los ítems, agrupados en familias, ¿poseen propiedades psicométricas similares y si se asocian en el constructo que los define?*

Como se detalló en los capítulos 3 (del Método) y 4 (de Resultados), se aplicaron seis muestras, una por cada área del EXHCOBA-R/MS, y donde cada contenido estuvo representado por seis ítems-hijo de una misma familia12. Los análisis estadísticos efectuados se resumen en la tabla 5.3. De esta se deduce que las familias del área de Ciencias sociales presentaron buenas propiedades psicométricas y se agruparon en torno a la habilidad evaluada en cada caso. En un segundo lugar se ubicó Habilidades cuantitativas con una familia de baja confiabilidad y correlaciones menores. En tercer lugar se situaron Matemáticas y Español. Para Matemáticas, se considera que las principales causas de desajustes fueron la dificultad, las deficiencias en el editor de reactivos y el empleo de dos familias diferentes para evaluar un mismo contenido. En el caso de Español, el área resultó demasiado sencilla, con algunas deficiencias leves de constructo en algunas familias. Las áreas más débiles fueron Habilidades del lenguaje y Ciencias naturales. Habilidades del lenguaje presentó problemas de varianza, constructo y confiabilidad dentro de las familias. En Ciencias naturales, una gran limitación fue el número escaso de datos que se pudieron analizar; bajo esta consideración, cabe señalar que emergieron problemas de confiabilidad y de constructo.

Es importante destacar que el análisis de esta sección permitió identificar dos características interesantes. La primera es que aunque un reactivo no aporte carga factorial al área donde pertenece, puede ocurrir que la familia a la cual pertenezca esté conformada por ítems-hijo isomorfos (similares en dificultad y en la habilidad evaluada), lo cual refleja la necesidad de efectuar los dos niveles de análisis (del examen y de las familias). La segunda es

12 Hubo tres contenidos tales que, cada uno estuvo representado por dos familias de tres ítems cada una. Estos contenidos son: HC05, MAT08 y MAT20.

que hubo dos contenidos que se evaluaron, cada uno, desde dos familias diferentes (apuntando a habilidades intelectuales distintas); esta situación se vio reflejada en los resultados de la agrupación de los ítems por constructo (los contenidos fueron MAT08 y MAT20).

Tabla 5.3
Propiedades psicométricas de cada familia de ítems del EXHCOBA-R/MS, agrupadas por área

	HV	HC	ESP	MAT	NAT	SOC
Datos						
Min	132	100	184	130	55	206
Max	167	155	220	237	155[a]	206
Flias. Analizadas	18	20	20	19	20	19
Hallazgos						
Grado de dificultad	Fácil	MB	Fácil	Muy difícil	B	MB
Varianza de dificultad	2 familias con mucha varianza	Menor que 002 [b]	Menor que 003 [c]	Menor que 002 [b]	Menor que 003	Menor que 002
Confiabi-lidad[d]	Escasa en 8 familias	Escasa en 1 familia	No se detectaron problemas	Escasa en 1 familia	Escasa en 2 familias	No se detectaron problemas
Ajuste	Problemas Leves	Problemas Leves	Problemas Leves	Algunos problemas	Problemas Leves	No se detectaron problemas
Discrimi-nación[e]	No se detectaron problemas	Problemas Leves	No se detectaron problemas	Problemas Leves	Problemas Leves	No se detectaron problemas
Constructo (correla-ciones)[f]	Problemas en 9 familias	Problemas en 1 familia	Problemas leves en 3 familias	Problemas en 4 familias	Problemas en 4 familias	Problemas leves en 1 familia
Conclusiones y/o sugerencias	Revisar los modelos de ítems que presentaron deficiencias. Proponer ítems más complejos.	En general, buenas propiedades psicométricas de los ítems por familia	Revisar los modelos de ítems para proponer ítems más complejos	Considerar una familia por contenido. Replicar con muestra de mayor tamaño	Revisar los modelos de ítems que presentaron deficiencias. Replicar con muestras de mayor tamaño	En general, buenas propiedades psicométricas de los ítems por familia

Nota: Min = mínimo, Max = máximo, B = bueno, MB = muy bueno.
[a] En una familia se analizaron 237 casos. [b] En una familia la varianza superó levemente 0.02. [c] En una familia la varianza superó levemente 0.03. [d] Alpha de Cronbach. [e] índice de discriminación, según el modelo de Rasch. [f] AFC, correlación punto biserial y correlación punto medida.

- *¿Cuáles son las propiedades psicométricas de los elementos que componen los reactivos de crédito parcial?*

Otra novedad del EXHCOBA-R/MS es la inclusión de ítems de crédito parcial. La mitad del examen incluye reactivos que contienen en su estructura dos, tres, cuatro o cinco *mini ítems* dicotómicos. Cada respuesta a estos *mini ítems* aporta a la calificación total del reactivo. Dada esta composición de *mega* reactivos, se consideró necesario revisar los *mini* ítems o elementos que los componen.

Se analizaron seis familias, una de cada área. Los tipos estudiados fueron *elemento-categoría* (se tomaron 2 de 46 familias), *selección-elementos* (1 de 15), *elemento-imagen* (1 de 6) y *selección elementos múltiple* (1 de 2). Los tres primeros, los más representativos del examen, se utilizan para evaluar el 56% de la prueba.

Un resultado interesante fue que, si bien los ítems de crédito parcial no mostraban problemas de correlación ni de discriminación, algunos de sus elementos sí lo presentaron. Podría ocurrir que la calidad superior de algunos de sus componentes compense las deficiencias de otros. Otra característica valiosa del análisis fue la identificación de elementos más sencillos y más difíciles, estos podrían combinarse y así graduar, con mayor precisión, la dificultad de los ítems de crédito parcial. Esta revisión también permitió identificar la calidad de los distractores que en algunos casos no funcionaron como tales.

Otro hallazgo fue que las familias más deficientes en propiedades estadísticas fueron aquellas donde se conservó el prototipo clásico de opción múltiple con tres o con cuatro opciones, lo cual respalda el cambio del EXHCOBA-R con ítems de respuesta construida o semi-construida con crédito parcial. Finalmente, cabe aclarar que, si bien estos análisis aportan hallazgos interesantes, es necesario estudiar con mayor profundidad los ítems de crédito parcial,

para poder describir con mayor precisión y certeza el comportamiento estadístico de este tipo de reactivos.

5.2. Alcances y limitaciones de la metodología utilizada

Tras la aplicación de la metodología al EXHCOBA-R/MS y del análisis de los resultados obtenidos, se obtuvo un buen diagnóstico inicial, ya que con muestras relativamente pequeñas se pudieron detectar problemas que después de ajustarse permitirán un mejor comportamiento de las diferentes versiones producidas por GAI. A continuación, se citan los beneficios más relevantes.

Se obtuvo una descripción de fácil lectura e interpretación de las propiedades básicas de los exámenes generados y de las familias de ítems. Quedaron puntualizados algunos problemas y la revisión pudo enfocarse en determinados modelos de ítems. Esto fue beneficioso, ya que permitió subsanar falencias de manera inmediata.

Debido a que se realizaron tres tipos de análisis: TCT, Rasch y AFC, los resultados coincidentes proporcionaron mayor certeza en los resultados y aquellos que no concordaron se consideraron como menos relevantes.

Se comprobó la necesidad de efectuar análisis en los dos niveles: del examen y de cada familia. Esto se verificó, por ejemplo, en el caso del contenido de HC07. El modelo de ítems generó una familia de reactivos isomorfos (propiedades psicométricas similares y buenas correlaciones entre sí). Sin embargo, el ítem no aportó al constructo definido por el área, no funcionó en sintonía con el resto de Habilidades matemáticas.

Como el análisis se efectuó en dos versiones, se compararon las propiedades de los ítems y la estructura por áreas. Los hallazgos similares en los dos tests proporcionaron mayor solidez a la GAI y las diferencias indicaron focos de atención a remediar.

Se detectó la estabilidad de la dificultad de cada área. En otras palabras, la dificultad reflejada en un área del examen (VA y VB) se repitió en las muestras por áreas. Por ejemplo, tanto en VA como en VB, Matemáticas fue el área más difícil, lo cual se repitió al contrastar la muestra MAT con el resto de las muestras por área.

Se descubrió que es más conveniente evaluar un contenido desde una única familia de reactivos y no, desde más familias, ya que cambia la habilidad (constructo) y podría modificarse también la dificultad. Ocurrió que para el contenido MAT08, las dos familias de ítems definidas pertenecían al constructo del área, pero los diferentes ítems-hijo no resultaron isomorfos ni en dificultad ni en habilidad evaluada (no reflejaron una misma competencia). Lo mismo ocurrió para el contenido MAT20.

Se identificó un peor comportamiento del tipo de ítem tradicional, dicotómico de 3 ó 4 opciones. Lo cual indica que el método funciona mejor para ítems de respuesta construida o de crédito parcial (múltiples selecciones), que no contemplan a la adivinación como un parámetro.

A través del análisis Rasch de los elementos que conforman los distintos ítems-hijo de una familia de ítems de crédito parcial quedaron expuestos aquellos elementos de calidad (con buenas propiedades estadísticas) que colaboran a la buena ejecución de los ítems, de los que provocan desajustes. Además, se constató que los ítems politómicos pueden funcionar correctamente, aunque contengan algún elemento de menor calidad (en correlación, ajuste o discriminación), ya que los elementos más *fuertes* compensarían a los más *débiles*.

Si bien la metodología propuesta para validar el EXHCOBA-R/MS reportó beneficios para la mejora del examen, es necesario aclarar que también reflejó ciertas limitaciones. Por un lado, aquellos inconvenientes que se dieron en la práctica de la aplicación de las muestras y por

otro, las restricciones conceptuales del propio método. A continuación se resumen y se explican estas limitaciones.

En la práctica, la primera desventaja es que todas las muestras provinieron de aplicaciones que se realizaron a modo de pilotaje, no fueron administraciones reales ni con alto impacto para los evaluados. Los nuevos ítems no podían formar parte de una administración del EXHCOBA porque los formatos y el tipo de reactivos son muy diferentes de los ítems de opción múltiple utilizados para la prueba tradicional. Así que, se generaron los ocho tipos de pruebas; algunos se aplicaron en una única oportunidad, otros, como HC y NAT debieron efectuarse en dos ocasiones. Con el fin de promover el interés por obtener buenas calificaciones, en cada institución participante se invitó a las 100 mejores ejecuciones a participar en un sorteo de una laptop.

Otra limitación fue el tamaño de las muestras. Según lo manifestado en el capítulo tres del Método, la recomendación fue una exigencia mínima de 600 evaluados para VA y VB (Gorsuch, 1983) y de 200 para los análisis por áreas y por familias (Wright y Stone, 1979). Se buscó una forma de estimular la participación y el buen desempeño en la resolución del nuevo examen. Sin embargo, los incentivos no fueron suficientes, ya que en algunas ocasiones solamente asistieron 50 estudiantes.

Otra acotación fue el tipo de participantes. Salvo en el caso de VA y VB, donde se trató de aspirantes a ingresar a la EMS, las otras muestras fueron aplicadas a estudiantes que ya se encontraban estudiando en la universidad. No se tuvo información acerca de qué carreras estaban cursando ni de sus habilidades cognitivas; según se indagó, se convocó a personas de los primeros semestres.

Tampoco se tienen registros acerca de la información que se les dio a los evaluados, previa a la resolución del examen; por ejemplo, qué explicación se les ofreció acerca de cómo resolver los ítems, si se les indicó cómo acceder y usar las herramientas auxiliares (calculadora y formulario). Sí se tienen datos de una encuesta a los examinados acerca de qué tan fácil fue la interacción con la nueva interfaz del examen. Sus respuestas apuntaron a que, en general, fue sencilla.

Al tratarse de la primera implementación del EXHCOBA-R, todavía existían problemas técnicos del editor. Las deficiencias de funcionamiento en los tipos de reactivos ocurrieron, principalmente, en Habilidades del lenguaje y Español. Por ejemplo, HV01 no pudo evaluarse por familias, aunque sí se hizo en VA y VB. En Matemáticas, la limitación se manifestó al escribir expresiones algebraicas. Así, para la expresión *2x*, debía anotarse *2*x*, conceptualmente correcta, aunque un tanto artificial, si se compara con la forma utilizada en los libros de texto[13].

También se presentaron dificultades para la recuperación de los datos: información incompleta o ausente. Esta característica fue más notoria en Habilidades matemáticas y en Matemáticas. Entre los incidentes más destacados figuran: el caso de HC05 donde solamente se obtuvo la información de una de las dos familias evaluadas, para MAT07 se registró una de las dos respuestas solicitadas y el tipo de reactivos de MAT09 no se decodificó.

Como consecuencia de las insuficiencias en el funcionamiento del editor y de la recuperación de datos, entre otros, se generaron casos perdidos. Si bien, en muchas situaciones como esta se utilizan técnicas de imputación simple, se tomó la decisión de no imputar, ya que este procedimiento puede generar sesgos en los estimadores (Backhoff, Andrade, Sánchez, Peon & Bouzas, 2006). En todos los casos se dejó que cada programa estadístico eliminara los casos con datos faltantes.

13 Se debe aclarar que en la pantalla de este tipo de ítems figura la advertencia de cómo escribir expresiones algebraicas.

Por último, no se puede omitir la presencia de posibles errores en la depuración de las bases de datos, de la calificación de los ítems y de los posteriores análisis estadísticos. Si bien en todos los casos se dedicó el mayor esfuerzo y atención, y se revisó minuciosamente cada etapa del trabajo, es probable que se hayan introducido errores involuntarios, inherentes a la condición humana.

La metodología también conlleva limitaciones conceptuales, propias de los modelos empleados. La aplicación de la TCT impone restricciones a las muestras, si se desean establecer comparaciones. Por ejemplo, para poder efectuar una descripción de las dificultades entre familias, todos los ítems deben ser resueltos por los mismos individuos. Lo mismo ocurre para los AFC, los ítems-hijos de una misma familia deben aplicarse a la misma población.

Como consecuencia, estas sucesivas repeticiones de reactivos de una misma familia (ítems-hermano) aplicados al mismo estudiante pudieron generar cansancio en el examinado. Efectivamente, se constató en varias familias, especialmente para la muestra de ESP, que el último ítem de la prueba aparecía sin respuesta o con problemas de correlaciones, de ajuste o de discriminación. Una propuesta para subsanar este inconveniente es reducir el número de ítems-hermano a cinco, ya que no se identificaron anomalías para estos casos.

En cuanto a la TRI, se utilizó un modelo de un parámetro, la dificultad. No se consideraron como parámetros ni la discriminación ni la adivinación. Un modo de subsanar esta limitación fue obtener índices de discriminación que, si bien no alteran el modelo basado en la dificultad, ofrecen información acerca del comportamiento de los reactivos. Los modelos de tres parámetros son particularmente útiles para ítems de opción múltiple donde la adivinación juega un rol importante. Dadas las características de los reactivos del EXHCOBA-R/MS, no se planteó la necesidad de incluir este tercer parámetro. La decisión de utilizar el modelamiento de Rasch se

justifica porque en un primer análisis se deben buscar los modelos más parsimoniosos, además de que requieran el menor costo en cantidad de evaluados.

Para el caso del AFC, es necesario aclarar que los análisis se efectuaron sobre los datos crudos (para ítems politómicos) y sobre las matrices de correlaciones de Pearson (para ítems dicotómicos). Para los casos de Habilidades cuantitativas y Matemáticas de VA y VB, donde hubo mayor cantidad de reactivos dicotómicos, se calcularon las matrices tetracóricas y posteriormente se efectuaron AFE, a fin de comparar con los resultados del AFC. Ambos análisis arrojaron cargas e índices similares. En realidad, el AFC fue aún más exigente en las cargas factoriales.

Es importante puntualizar que, si bien el método se planteó como un recurso para aportar evidencias de validez de estructura interna de exámenes producidos por GAI en forma genérica, en la práctica se ajustó a las características propias del EXHCOBA-R/MS. Esto es, se respetó el número de áreas, de contenidos, el tipo de reactivos, entre otras particularidades, del examen que se debía validar. Sin embargo, estas condiciones se podrían modificar para cada situación específica, sin cambiar la esencia del método. Lo fundamental es que se trata de una metodología dirigida a obtener las primeras evidencias de validez de estructura interna de exámenes de conocimientos de alto impacto, no adaptativos, producidos por GAI con teoría débil, cuyos ítems son de respuesta construida breve o semiconstruida (de múltiples opciones).

Otra acotación necesaria es que se analizaron los ítems, únicamente, desde dos puntos de vista: (a) sus propiedades psicométricas (dificultad, discriminación, correlaciones) y (b) su pertenencia al constructo definido por el currículo y seleccionado por los especialistas; no se propuso un análisis desde la teoría cognitiva, por ejemplo, a través de modelos componenciales.

Para ello hubiera sido necesario, previamente contar con modelos cognitivos para cada área del EXHCOBA-R/MS (en otras palabras, con una GAI de teoría fuerte).

Por último, si bien los resultados representan una imagen estática de un proceso dinámico que es la propia generación de ítems-hijo. Este tipo de análisis es una opción viable que permite obtener una proyección acerca del funcionamiento de la GAI. Otros estudios de validación de exámenes producidos por GAI constan desde 2 a 10 ítems-hijo de una misma familia para la ejecución de los diferentes análisis (Arendasy, Sommer, Gittler, Hergovich, 2006; Arendasy, Sommer y Mayr, 2012; Geerlings, Glass y Van der Linden, 2011; Sinharay y Johnson, 2012).

5.3. Nuevas líneas de investigación

Sin duda, la metodología propuesta tiene muchas facetas que pueden mejorarse. Desde la TRI, sería interesante incluir Análisis Diferencial del Ítem (Análisis DIF), puesto que añadiría evidencia a favor de la fiabilidad y robustez de los grupos de ítems generados por los GAI. Otro aspecto a desarrollar sería la aplicación de modelos TRI de dos parámetros, es decir, incluir la discriminación como una variable que regule el funcionamiento de los ítems. Desde los Modelos de Ecuaciones Estructurales, sería importante realizar Pruebas de Invarianza de Parámetros a través de grupos (Invariance of Parameters across groups). Esto le agregaría solidez, puesto que se trata de garantizar la comparabilidad de las propiedades de las familias de reactivos que genera cada GAI. Otra actividad a realizar es incluir en la aplicación del examen, un sistema de monitoreo que detecte, en las administraciones, aquellos ítems con deficiencias.

Desde otra línea de investigación, la Ingeniería de la Evaluación (IE) propone cuatro pasos para el desarrollo de pruebas a través de la GAI: (1) modelos cognitivos que guíen el desarrollo de ítems (GAI desde la teoría fuerte), (2) modelos de ítems que sirvan para producir ítems isomorfos, (3) procedimientos computacionales que permitan generar estos ítems clones y

(4) modelos psicométricos que, de manera confirmatoria, evalúen el ajuste entre teoría y datos (Luecht, 2012).

Si bien, el desarrollo de la IE es novedoso, los avances en lo que respecta a sus modelos estadísticos son aún más recientes y escasos. Algunos aportes destacados pertenecen a Geerlings, Glas & van der Linden (2011), Glas & van der Linden (2003), Sinharay & Johnson (2008, 2012), y Sinharay, Johnson, & Williamson (2003).

Desde la IE, la ventaja fundamental de las teorías fuertes es el soporte teórico que aportan a los análisis estadísticos de validez de los tests producidos por GAI. Las teorías cognitivas permiten desarrollar los modelos de tareas que se emplean para delimitar contenidos y habilidades cognitivas que provean delineamientos precisos para el desarrollo de los modelos de ítems. Estos modelos producen grandes cantidades de ítems isomorfos perfectamente graduados desde un marco cognitivo.

A finales del siglo xx comenzó una tendencia a incorporar elementos cognitivos en los modelos de medición. Si bien en los exámenes normativos el aprovechamiento de los estudiantes se interpreta en términos de la ejecución de otros estudiantes en el mismo test, sería muy útil incorporar modelos cognitivos que provean de bases sustantivas y de un marco de referencia más preciso para establecer estándares de ejecución. En otros términos, haría los números más interpretables (Pellegrino, Chudowsky & Glaser, 2001).

Como los modelos cognitivos son escasos y casi todos pertenecen a procesos psicológicos (Drasgow. Luecht y Bennett, 2006), sería necesario iniciar la tarea de construir un modelo cognitivo para cada una de las seis áreas del EXHCOBA-R/MS para luego evaluar desde modelos componenciales. Con respecto a este tema, existe un estudio para el área de Habilidades matemáticas de la educación primaria (Pérez-Morán, 2014). Si bien, el trabajo es arduo, se

dificulta más aún para el lenguaje y las ciencias sociales. Backhoff, Larrazolo y Rosas (2000) expusieron una razón por la cual encontrar habilidades cognitivas que establezcan jerarquías en ciencias sociales es más complejo que en matemáticas, por ejemplo.

> Las matemáticas evalúan competencias que tienen una estrecha relación lógica e inclusiva; es decir, sus conceptos fundamentales se enlazan y se construyen unos sobre otros, en forma progresiva, de tal manera que no es posible entender algún concepto o resolver algún problema sin entender los anteriores […] En sentido opuesto, las disciplinas sociales se construyen sin esta estrecha relación entre sus conceptos básicos (Backhoff et al., 2000, p.16).

La otra línea de investigación, enfocada en una GAI de teoría débil, es conseguir una mejor aproximación de los parámetros de los ítems por familia, a través de los modelos ISM y RSM. Los trabajos de Sinharay & Johnson (2008) y los de Sinharay, Johnson, & Williamson (2003) son ejemplos interesantes que sirven de referencia para próximas investigaciones en torno al EXHCOBA-R.

Sinharay y Johnson (2008, 2012) estudiaron una parte de la sección cuantitativa de una aplicación real del GRE a través de la GAI. Para cada modelo de ítem (familia de reactivos) se consideraron 10 ítems-hermano isomorfos originados por el generador automático propio del GRE. Todos los reactivos eran de opción múltiple con cinco alternativas de respuesta. Los ítems se administraron a 32,921 estudiantes; en promedio, cada ítem fue evaluado 821 veces. Los autores aclararon que su trabajo se incluye entre los pocos conocidos, producto de la GAI para exámenes de gran escala y de alto impacto, como lo es el GRE14. Ellos refirieron a otros dos trabajos con GAI, ambos publicados en 2002, uno realizado por Wright con 1273 examinados,

14 El estudio se efectuó con el GRE adaptativo.

correspondiente a una batería de tests para reclutar personas para la armada de Gran Bretaña, y el otro realizado por Hornke, con pruebas de inteligencia de bajo impacto.

Lo interesante de las investigaciones de Sinharay y Johnson es que esta metodología se puede aplicar a la GAI para exámenes pertenecientes a teoría débil (Drasgow, Luetch y Bennett, 2006). Los datos se analizaron desde dos modelos: el ISM, que consiste en tratar indistintamente a los ítems-hijo, de la misma familia, como si fueran el mismo reactivo, y el otro modelo, RSM, que considera una relación entre los ítems-hermanos. Este último es un modelo jerárquico que asume una función de respuesta al ítem para cada reactivo, pero relaciona a los ítems-hermano a través de un componente jerárquico. Ambos modelos pertenecen a la TRI y suponen funciones de 2 o 3 parámetros (dificultad y discriminación, para el primer caso, y agrega un tercer parámetro de adivinación, para el segundo caso).

Sinharay, Johnson y Williamson (2003) definieron una función que permite establecer los análisis de las familias de ítems, esta se denomina Función de Respuesta Esperada por Familia (*Family Expected Response Function*, FERF). Esta función resume la probabilidad de respuesta correcta a un ítem generado, aleatoriamente, de una familia de ítems. Lo destacable es que las FERFs proveen las bases para un gráfico que sintetiza los resultados del RSM.

Sin embargo, estos análisis estadísticos implican una mayor inversión, tanto de trabajo de investigación como del tamaño de las muestras. Por lo tanto, se sugiere revisar y perfeccionar el EXHCOBA-R/MS con la información obtenida como producto de esta tesis y, posteriormente, continuar el estudio desde los modelos ISM y RSM. Lo cual proporcionaría una estructura aún más sólida que respalde la calidad del examen.

Epílogo

Tras la aplicación, a muestras del EXHCOBA-R/MS, de la metodología propuesta en el presente trabajo, se puede concluir que el área más consolidada es la de Ciencias sociales con buen desempeño tanto por área como por familias. En un segundo lugar se encuentra Habilidades cuantitativas con un ítem que no aporta al constructo en general (HC07) y otro con baja confiabilidad y correlaciones pequeñas (HC09). En tercer lugar se ubican las áreas de Matemáticas y Español. Para la primera, se considera que las principales causas de problemas son la dificultad (se necesitó una muestra de mayor tamaño), la falta de claridad para escribir expresiones algebraicas y el empleo de dos familias para evaluar un mismo contenido (MAT08 y MAT20). En el caso de Español, el área presenta un buen comportamiento, aunque los ítems, en general, demandan poca habilidad. Los ítems con comportamiento deficiente son dos (ESP01 y ESP02). Las áreas más débiles son Habilidades del lenguaje y Ciencias naturales. La primera manifiesta problemas de varianza, constructo y confiabilidad dentro de las familias, lo cual repercute en el área. Se recomienda corregir el funcionamiento del tipo de reactivos *seleccionar frase*, para evitar múltiples selecciones. Se sugiere reconstruir HV15 y HV17, y revisar HV06, HV16, HV18 y HV20. En cuanto a Ciencias naturales, no se puede obtener una información certera, debido a la escasa cantidad de datos por familia evaluada; sin embargo, los análisis en VA y VB marcan una deficiencia en la organización del área por constructos. Se sugiere revisar las especificaciones de: BIO01, BIO04, FIS10, QUI13, QUI14 y QUI16.

En exámenes producidos por GAI a través de teorías débiles, como el EXHCOBA-R/MS alineado a contenidos curriculares, se estima conveniente obtener evidencias basadas en la estructura interna desde dos ámbitos: el test, como instrumento completo, y las familias de ítems, como una necesidad creada por la propia GAI. Además, el método debe sostenerse sin necesidad

de un modelo cognitivo. La metodología propuesta en la presente tesis permitió obtener una buena descripción del funcionamiento del generador, con muestras relativamente pequeñas y con herramientas estadísticas básicas. Los resultados se complementarían muy bien con un análisis cualitativo puntual de las especificaciones detectadas con deficiencias.

Como ya se anunció, el proceso de validación es constante e inacabado, y los desarrollos en el campo de la GAI son novedosos; por lo tanto, se divisa un ámbito muy interesante de investigaciones en torno a los métodos de validación interna de los generadores.

Referencias

Abad, F., Garrido, J., Olea, J. y Ponsoda, V. (2006). *Introducción a la psicometría: teoría clásica de los tests y teoría de respuesta al ítem.* Madrid: Universidad Autónoma de Madrid, Facultad de Psicología.

American Educational Research Association-American Psychological Association-National Council on Measurement in Education (1985). *Standards for Educational and Psychological Testing*. Washington, DC: American Psychological Association.

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

Arendasy, M., Sommer, M., Gittler, G. y Hergovich, A. (2006). Automatic generation of quantitative reasoning items. A pilot study. *Journal of Individual Differences, 27* (1), 2-14.

Arendasy, M., Sommer, M. y Mayr, F. (2012). Using automatic item generation to simultaneously construct German and English versions of a word fluency test. *Journal of Cross-Cultural Psychology*, *43*(3), 464-479. DOI: 10.1177/0022022110397360. Recuperado el 20 de noviembre de 2013, de http://jcc.sagepub.com/content/43/3/464

Ausubel, D. (2002). *Adquisición y retención del conocimiento: una perspectiva cognitiva*. Barcelona: Paidós.

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

Backhoff, E., Andrade, E., Sánchez, A., Peon, M. y Bouzas, A. (2006). *El aprendizaje del español y las matemáticas en la educación básica en México: sexto de primaria y tercero de secundaria*. México: Instituto Nacional para la Evaluación de la Educación.

Backhoff, E., Ibarra, M. y Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología, 12* (1), 55-62.

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. Recuperadoel 20 de noviembre de 2010, de *http://www.cesu.unam.mx/iresie/Revistas/REVISTAS/MX/REVEDUSUP/1996/V25N1(97)A1996/MX.REVEDUSUP.1996.V25N1(97).P41-54.HTM*

Backhoff, E. y Larrazolo, N. (2001). Validación de contenido del Examen de Habilidades y Conocimientos Básicos. *Revista Mexicana de Psicología, 18* (1), 9.

Backhoff, E., Larrazolo, N. y Rosas, M. (2000). Niveles de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa, 2* (1). Recuperado el 20 de noviembre de 2010, de http://redie.uabc.mx/contenido/vol2no1/contenido-backhoff.pdf

Backhoff, E., Sánchez, A., Peon, M., Monroy, L. y Tamachi, M. L. (2006). Diseño y desarrollo de los exámenes de la calidad y el logro educativos. *Revista Mexicana de Investigación Educativa, 11* (29), 617-638.

Backhoff, E. y Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos. *Revista de la Educación Superior, 21* (3), 95-118. Recuperado el 7 de diciembre de 2010, de http://www.exhcoba.mx/pdf/1992_Desarrollo_del_EXHCOBA.pdf

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederikson, R. J. Mislevy e I. I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 323-359). Mahwah, NJ: Erlbaum.

Bejar, I. I. (2002). Generative testing: From conception to implementation. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 199-217). Mahwah, NY: Erlbaum.

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107,* 238-246.

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

Bentler, P. M. y Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. *Psychological Bulletin, 88,* 588-606.

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Zand Scholten, A. y Franic, S. (2009). The end of construct validity. En R. W. Lissitz, *The concept of validity. Revisions, new directions, and applications* (pp. 135-170). Charllote, NC: Age Publishing.

Borsboom, D., Mellenbergh, G. J. y van Heerden, J. (2004). The concept of validity. *Psychological Review, 3* (4), 1061–1071.

Browne, M. W. y Cudeck, R. (1989). Single sample cross-validation indices for covariance structure. *Multivariate behavioral Research, 24,* 445-455.

Carnoy, M., Elmore, R. y Siskin, L. (Eds). (2003). The new accountability: high schools and high-stakes testing. Nueva York: Routledge Falmer.

Carretero, M. (2004). *Constructivismo y educación* (8va. ed.). Buenos Aires: Aique Grupo Editor.

Coll, C. (2001). Constructivismo y educación: la concepción constructivista de la enseñanza y el aprendizaje. En C. Coll, J. Palacios y A. Marchesi (Comps), *Desarrollo psicológico y educación. Vol. 2. Psicología de la educación escolar* (pp. 157-188). Madrid: Alianza.

College Board. (2008). *Guía de estudio para presentar las pruebas de ingreso al nivel de educación media superior PIENSE II*. Recuperado el 17 de noviembre de 2010, de http://www.escolar.udg.mx/escolar2000/Guia_PIENSE_II.pdf

Contreras, L. A. (2000). *Desarrollo y pilotaje de un examen de Español para la educación primaria de Baja California* (Tesis de maestría no publicad). Universidad Autónoma de Baja California, Ensenada, México.

Contreras, L. A., Backhoff, E. y Larrazolo, N. (2003). *Curso taller para la elaboración de exámenes criteriales: manual para el comité diseñador del examen.* Manuscrito mimeografiado, Universidad Autónoma de Baja California, Instituto de Investigación y Desarrollo Educativo en Ensenada, México.

Corral-Verdugo, V. y Obregón-Salido, F. J. (1998). Aplicaciones del modelamiento de variables latentes a la Teoría de la conducta. *Acta Comportamentalia, 6,* 73-86. Recuperado el 24 de marzo de 2011, de http://www.journals.unam.mx/index.php/acom/article/view/18249/17346

Cortada de Kohan, N. (2004). Teoría de Respuesta al Ítem: Supuestos básicos. *Evaluar, 4*. Recuperado el 30 de noviembre de 2010, de http://www.revistaevaluar.com.ar/45.pdf

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16* (3), 297-334).

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational Measurement* (2a. ed., pp. 443-507). Washington, DC: American Council on Education-The Oryc Press.

Cronbach, L. J. y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281-302.

Diseth, A. (2007). Approaches to learning, course experience and examination grade among undergraduate psychology students: testing of mediator effects and construct validity. *Studies in Higher Education, 32* (3), 373–388.

Drasgow, F., Luecht, R. M. y Bennett, R. E. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement* (pp. 471-515). Westport, CT: ACE-Praeger.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika, 64* (4), 407-433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NY: Erlbaum.

Embretson, S. E. y Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NY: Lawrence Erlbaum.

Field, A. (2005). *Discovering statistics using SPSS* (2da. ed.). Londres: Sage.

Freund, P. A., Hofer, S. y Holling, H. (2008). Explaining and controlling for the psychometric properties or computer-generated figural matrix items. *Applied Psychological Measurement, 32*, 195-210.

García, R. y Castañeda, S. (2006). Validación de constructo en la comprensión de lectura en inglés como lengua extranjera. *Razón y Palabra, 51*. Recuperado el 14 de enero de 2014, de http://www.razonypalabra.org.mx/anteriores/n51/garciacastaneda.html

Geerlings, H., Glass, C. A. W. y van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika, 76* (2), 337-359.

Gierl, J. y Haladyna, T. M. (2012). Automatic item generation: an introduction. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice*. Nueva York: Routledge.

Gierl, M. J. y Lai, H. (2011, abril). The role of item models in automatic item generation. Trabajo presentado en *Annual Meeting of the National Council on Measurement in Education*, Nueva Orleans, LA.

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.

Gierl, M. J., Zhou, J. y Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning, and Assessment, 7* (2), 1-50.

Glas, C. A. W. Y van der Linden, W. J. (2003). Computarized adaptative Testing with item cloning. *Applied Psychological Measurement, 27*, 247-261.

González-Montesinos, M. (2004). *Defining and measuring academic standards for higher education: a formative study at the University of Sonora.* **(**Tesis de doctorado no publicada). The University of Arizona, Tucson, AZ, Estados Unidos de América.

González-Montesinos, M. (2008). *El análisis de reactivos con el modelo Rasch. Manual técnico A. Serie: Medición y metodología*. Hermosillo: Universidad de Sonora-Instituto Nacional para la Evaluación de la Educación.

Gorsuch, R. L. (1983). *Factor analysis* (2a. ed.). Hillsdale, NJ: Lawrence Erlbaum.

Hair, J. F., Anderson, R. E., Tatham, R. L. y Black, W. C. (1992). *Multivaria data analysis* (3a. ed.). Nueva York, NY: Macmillan.

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation*: *Theory and practice*. Nueva York: Routledge.

Haladyna, T. M. y Shindoll, R. R. (1989). Shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, *12*, 97-104.

Herrero, J. (2010). El análisis factorial confirmatorio en el estudio de la estructura y estabilidad de los instrumentos de evaluación: Un ejemplo con el Cuestionario de Autoestima CA-14. *Intervención Psicosocial, 19* (3), 289-300.

Heubert, J. P. y Hauser, R. M. (Ed.). (1999). *High stakes: testing for tracking, promotion and graduation.* Washington: National Academy of Sciences-National Research Council.

Hively, W., Patterson, H. L. y Page, S. H. (1968). A "universe-defined" system for arithmetic achievement tests. *Journal of Educational Measurement, 5*, 275-290.

Holling, H., Berling, J. P. y Zeuch, N. (2009). Automatic item generation for probability word problems. *Studies in Educational Evaluation*, *35*, 71-76.

Hombo, C. y Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Trabajo presentado en Annual Meeting of the National Council on Measurement in Education, Seatle, WA, Estados Unidos de América.

Instituto Nacional para la Evaluación de la Educación. (2005). *Manual técnico para el diseño de exámenes de la calidad y el logro educativos*. México: Autor. Recuperado el 7 de diciembre de 2010, de http://www.inee.edu.mx/index.php/bases-de-datos/bases-de-datos-excale/marcos-de-referencia/3551

Irvine, S. H. & Kyllonen, P. C. (Eds.). (2002). *Item generation for test development.* Mahwah, NY: Erlbaum.

Janssen, R., Schepers, J. y Peres, D. (2004). Models with item group predictors. En P. De Boeck y M. Wilson (Eds.), *Explanatory item response models: a generalized linear and non linear approach* (pp. 189-212). Nueva York, NY: Springer.

Kane, M. T. (2006). Validation. En R. L. Brennan (Ed.), *Educational measurement* (4a. ed., pp. 17-64). Westport, CT: American Council on Education, Praeger.

Keller, G., Deneen, J. R. y Magallán, R. J. (Eds.). (1991). *Assessment and access: Hispanics in higher education*. Albany, NY: State University of New York.

Linacre, J. M. (2002). What do infit and outfit, Mean-square and Standardized mean? *Rasch Measurement Transactions, 16* (2), 878.

Linacre, J. M. (2010b). Winsteps® (Version 3.70.0.2) [Software]. Beaverton, OR: Winsteps.com. Recuperado de http://www.winsteps.com/

Linacre, J. M. (2010a). *A user´s guide to winsteps ministep Rasch-Model computer programs*. Recuperado el 7 de noviembre de 2011, de http://www.winsteps.com/a/winsteps-manual.pdf

Li, Y. H. y Tompkins, L. J. (2004). Examining the construct validity for the multiple-content testing programs. *International Journal of Testing, 4* (3), 217-238.

Luecht, R. M. (1998). Computer assisted test assembly using optimization heuristics. *Applied Psychological Measurement, 22*, 224-236.

Luecht, R. M. (2012). Connecting theory and practice. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation*: *Theory and practice*. Nueva York: Routledge.

Magnusson, D. (1967). *Tests theory*. Massachusets, MA: Addison-Wesley.

Martínez-Arias, M. R., Hernández-Lloreda, M. V. y Hernández-Lloreda, M. J. (2006). *Psicometría*. Madrid: Alianza Editorial.

Martínez-Rizo, F., Backhoff, E., Castañeda, S., De la Orden, A., Schmelkes, S., Solano-Flores, G. *et al.* (2000). *Estándares de calidad para instrumentos de evaluación educativa*. México: Ceneval. Recuperado el 4 de abril de 2011, de http://archivos.ceneval.edu.mx/archivos_portal/2758/EstandaresCalidad.pdf.pdf

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

Masters, G. N. (1988) Item discrimination: When more is worse. *Journal of Educational Measurement, 25* (1), 15.

Messick, S. (1993). Validity. En R. L. Linn (Ed.), *Educational measurement* (3a. ed., pp. 13-103). Phoenix, AZ: American Council on Education-The Oryc Press.

Messick, S. (1998). Test validity: A matter of consequences. *Social Indicators Research, 45*, 35-44.

Mortimer**,** T**.,** Stroulia**,** E. y Yazdchi, M. V. (2012). IGOR: A web-based automated assessment generation tool. En M. I. Gierl y T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 217-230). Nueva York: Routledge.

Muñiz, J. (2010). Las teorías de los tests: teoría clásica y teoría de respuesta a los ítems. *Papeles del Psicólogo 31* (1), 57-66. Recuperado el 30 de noviembre de 2010, de http://www.cop.es/papeles

Muñiz, J. y Hambleton, R. K. (1992). Medio siglo de Teoría de Respuesta a los Ítems. *Anuario de psicología, 52*, 41-66. Recuperado el 16 de marzo de 2011, de http://www.raco.cat/index.php/AnuarioPsicologia/article/viewFile/64681/88708

National Council on Measurement in Education. (1999). *Testing memo 8: Realiability of Test Scores*. Autor.

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.

Nichols, S. L y Berliner, D. C. (2007). *Collateral damage: how high-stakes testing corrupts America's schools*. Cambridge, MA: Harvard Education Press.

Novak, J. D. (1982): *Teoría y práctica de la educación.* Madrid: Alianza Editorial.

Noddings, N. (2004). High stakes testing: why? *Theory and Research, 2* (3), 263-269. DOI: 10.1177/1477878504046520.

Nunnally, J. C. y Berstein, I. H. (1995). *Teoría psicométrica* (2a. ed.). México: McGraw-Hill.

Oliveira, E., Almeida, L., Ferrándiz, C., Ferrando, M., Sainz, M. y Prieto, M. D. (2009). Tests de pensamiento creativo de Torrance (TTCT): elementos para la validez de constructo en adolescentes portugueses. *Psicothema, 21* (4), 562-567.

Osburn, H. G. (1968). Item sampling for achievement tests. *Educational and Psychological Measurement, 28*, 95-104.

Pellegrino, J. W., Chudowsky, N. y Glaser, R. (Eds.). (2001). *Knowing what students know: the science and design of educational assessment*. Board on Testing and Assessment, Center for Education, National Research Council. Recuperado el 2 de diciembre de 2010, de http://www.nap.edu/catalog/10019.html

Pérez-Morán, J. C. (2014). *Análisis del aspecto sustantivo de la validez de constructo de una prueba de habilidades cuantitativas* (Tesis de doctorado no publicada). Universidad Autónoma de Baja California, Ensenada, México.

Phakiti, A. (2008). Construct validation of Bachman and Palmer's (1996) strategic competence model over time in EFL reading tests. *Language Testing, 25* (2), 237–272.

Pozo, J. I. (1996). *Teorías cognitivas del aprendizaje* (4ta. ed.). Madrid: Morata.

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.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema, 10* (3), 709-716.

Roid, G. H. y Haladyna, T. M. (1978). The use of domains and item formsin the formative evaluation of instructional materials. *Educational and Psychological Measurement, 38*, 19-28.

Rojas-Tejada, A. J. (2001). Pasado, presente y futuro de los tests adaptativos informatizados: entrevista con Isaac Bejar. *Psicothema, 13* (4), 685-690.

Rosas, M., Ramírez, J. L. y Larrazolo, N. (2009, septiembre). *Examen de selección: Sistema computarizado de exámenes Sicodex version 3*. Trabajo presentado en el X Congreso Nacional de Investigación Educativa, Veracruz, México.

Sackett, P. R., Borneman, M. J. y Connelly, B. S. (2008). High-stakes testing in higher education and employment. Appraising the evidence for validity and fairnes. *American Psychologist, 63*, (4) 215-227. DOI: 10.1037/0003-066X.63.4.215.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometric Monograph, 18.*

Sánchez-Hernández, B. A., Bazán-Ramírez, A. y Corral-Verdugo, V. (2009). Análisis de características morfológicas y funcionales de competencias de lectura y escritura en niños de primaria. *Revista Mexicana de Análisis de la Conducta, 35* (1), 37-57.

Sands, W. A., Waters, B. K. y McBride, J. R. (1997). *Computarized adaptative testing: from inquiry to operation.* Washington, DC: American Psychological Association.

Santos, D. M. y Castañeda, S. (2008). Objetivación de información en aprendizaje matemático autorregulado. Validez empírica de constructo. *Revista Mexicana de Investigación Educativa, 13* (38), 713-736.

Secretaría de Educación Pública. (2006). *Educación básica. Secundaria. Plan de estudios 2006.* México: Autor.

Secretaría de Educación Pública. (2009). *Plan de estudios 2009. Educación básica. Primaria*. México: Autor.

Secretaría de Educación Pública. (2011). *Documento base del bachillerato general*. México: Autor.

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

Sinharay, S. y Johnson, M. (2008). Use of item models in large-scale admissions test: a case of study. *International Journal of Testing, 8,* 209-236. DOI: 10.1080/15305050802262019.

Sinharay, S. y Johnson, M. (2012). Statistical modeling of Automatic Item Generation. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation*: *Theory and practice*. Nueva York: Routledge.

Sinharay, S., Johnson, M. S. y Williamson, D. M. (2003). Calibrating item families and summarizing the results using family expected response functions. *Journal of Education and Behavioral Statistics. 28*, 295. DOI: 10.3102/10769986028004295.

Solano-Flores, G., Jovanovic, J. Shavelson, R. J. y Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessment. *International Journal of Science Education, 21* (3), 293-315.

Solano-Flores, G. y Shavelson, R. J. (1997). Development of performance assessment in science: conceptual, practical and logistical issues. *Educational Measurement: Issues and Practice, 16* (3), 13-22.

Solano-Flores, G., Shavelson, R. J. y Schneider, S. A. (2001). Expanding de notion of assessment shell: from task development tool to instrument for guiding the process of science assessment development. *Revista Electrónica de Investigación Educativa, 3* (1). Recuperado el 19 de noviembre de 2010, de http://redie.uabc.mx/contenido/vol3no1/contents-solano.pdf

SPSS Inc. (2008). *SPSS Statistics for Windows, Version 17.0.* Chicago: SPSS Inc.

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavioral Research, 25* (2), 173-180.

Tirado, F. y Backhoff, E. (1999). La compleja elaboración de exámenes, 16 razones para utilizar la opción "No sé". *Revista Mexicana de Investigación Educativa, 4* (7), 13-26. Recuperado el 20 de noviembre de 2010, de http://redalyc.uaemex.mx/redalyc/src/inicio/ArtPdfRed.jsp?iCve=14000702

Tristán-López, A. (2001). *Análisis de Rasch para todos. Una guía simplificada para evaluadores educativos*. México: Ceneval.

Wainer, H. & Kiely, G. L. (1987). Item clusters and computerized adaptive testing: a case of testlets. *Journal of Educational Measurement, 24*, 185-201.

Williams, F. y Monge, P. (2001). *Reasoning with statistics: How to read quantitative research* (5a. ed.). Belmont, CA: Thomson Higher Education.

Wright, B. D. y Stone, M. (1998). *Diseño de mejores pruebas utilizando la técnica de Rasch* (Trads. R. Vidal-Uribe y A. Díaz-Cadena). México: Ceneval.

Wright, B. D. y Stone, M. (1999). *Measurement essentials* (2a. ed.). Wilmington, DE: Wide Range.

Yen, W. M. y Fitzpatrick, A. R. (2006). Ítem response theory. En R. L. Brennan (Ed.), *Educational Measurement* (4a. ed., pp. 111-153). Westport, CT: American Council on Education-Praeger.

Zamora-Muñoz, S., Monroy-Cazorla, L. y Chávez-Álvarez, C. (2010*). Análisis factorial: una técnica para evaluar la dimensionalidad de las pruebas. Cuaderno técnico 6*. México: Ceneval. Recuperado el 20 de febrero de 2011, de http://archivos.ceneval.edu.mx/archivos_portal/7492/CuadernoTecnico061aed.pdf

Anexo A

Ejemplo de especificación de reactivos

I. DATOS DEL ELABORADOR Y DE LOS REVISORES

Redactor de la especificación	**Fecha de redacción inicial**
(Se omiten nombres)	06 de Diciembre del 2009
Revisores (Se omiten nombres)	
Revisiones y correcciones	**Fechas de envío**
Revisión 1 Corrección 1 Revisión 2 Revisión 3 Revisión 4 Corrección 2 Corrección 3	Sesión presencial (7 dic. 2009) 10 diciembre 2009 27 enero 2010 16 febrero 2010 11 de marzo de 2010 24 de marzo de 2010 14 de abril de 2010

II. DESCRIPCIÓN DEL CONTENIDO

II. a) Datos de identificación del contenido a evaluar

Asignatura		**Nivel educativo**	
Español		Primaria	
Clave	**Ámbito**	**Tema**	**Subtema**
HV09	Estudio	Ortografía y Puntuación	Mayúsculas y punto
Contenido	**Nombre**	Mayúsculas y punto.	
	Definición	Separación de oraciones con mayúsculas y punto.	

II. b) Características del contenido a evaluar

Importancia (justificación) del contenido a evaluar
Es un contenido básico de la ortografía. Permite la correcta comprensión de los textos. Es un recurso esencial para la producción de textos, ya que permite organizar la escritura en párrafos estructurados.
Delimitación del contenido
Se utilizarán textos breves (máximo 20 palabras). Se identificará el uso de mayúsculas al inicio de oración y en nombres propios. Se identificará el uso del punto seguido y punto final.
Conocimientos y habilidades involucrados en la solución correcta del reactivo
Los sustentantes deberán tener habilidades: • para conocer, identificar y aplicar las reglas ortográficas del uso de las mayúsculas y del punto. • de lectura mecánica (identificación de los signos). • de comprensión lectora.

III. REGLAS Y ELEMENTOS CONCEPTUALES PARA ELABORAR UN CONJUNTO DE REACTIVOS

Estrategia de evaluación

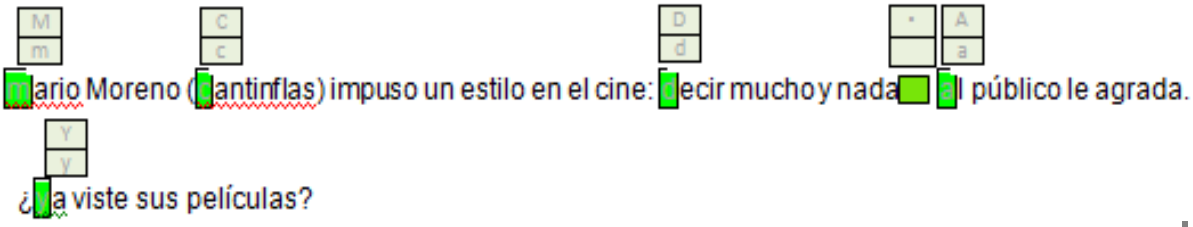
Se utilizarán textos breves, en los que el sustentante colocará las mayúsculas en las palabras que las necesiten, así como los puntos necesarios.

Base del reactivo con su plantilla

El siguiente fragmento de un texto requiere corrección ortográfica en el uso de mayúsculas o minúsculas y de puntuación. Haz clic en los espacios vacíos y selecciona la opción adecuada.

(Texto)

Reactivo ejemplo



Respuesta correcta

Mario Moreno (Cantinflas) impuso un estilo en el cine: decir mucho y nada. Al público le agrada. ¿Ya viste sus películas?

Datos para el programador

Generador de reactivos:

- Seleccionar un texto al azar.
- Cada texto cuenta con una cantidad de 7 a 10 elementos (recuadros verdes con una letra o nada) que se pueden utilizar para aleatorizar y construir distintas versiones de los reactivos. Algunos elementos implican el uso de mayúscula/minúscula y otros el uso correcto del punto.
- Del texto seleccionado por el sistema, el mismo deberá elegir 6 elementos a resaltar que incluyan, al menos, una regla para uso de punto y una regla para uso de mayúscula.
- Los elementos (letra o nada) que el sistema seleccione para que el alumno observe y responda deben estar visibles dentro de los recuadros en verde, pero en color gris tenue para indicar que aún no han sido respondidos (ver ilustración en la parte de arriba de este formato). Cuando se trate de una letra, deberá aparecer siempre en minúscula. En el caso del punto/nada, el cuadro deberá aparecer siempre vacío.

Para las respuestas:

- Cuando el sustentante pase el mouse sobre cada recuadro verde, se deberá iluminar un recuadro pequeño en la parte superior, dentro del cual aparecerán 2 opciones de respuesta:
- En el caso de que sea letra: Mayúscula arriba, y minúscula abajo
- En el caso de que sea punto: Punto arriba y vacío abajo.

- El sistema deberá permitir al sustentante seleccionar una de las 2 opciones mediante un clic.
- Al ser seleccionada una de las opciones, automáticamente la letra que se encuentra en el cuadro verde, se tornará color negro, indicando que ya fue contestado ese elemento del reactivo.
- El sistema deberá permitir al sustentante corregir y modificar su respuesta.
- Se le otorgará un punto si selecciona correctamente el 100% de los casos.

Tabla de textos para aleatorizar

Texto	Presentación en pantalla	Respuesta correcta
1	ario oreno (antinflas) impuso un estilo en el cine: ecir mucho y nada l público le agrada ¿ a viste sus películas?	Mario Moreno (Cantinflas) impuso un estilo en el cine: decir mucho y nada. Al público le agrada. ¿Ya viste sus películas?
2	Texto 2	
3	Texto 3	
4	Texto 4	
5	Texto 5	
6	Texto 6	
7	Texto 7	
8	Texto 8	
9	Texto 9	
10	Texto 10	
11	Texto 11	

Anexo B

Lista de reactivos organizados por áreas, por tipo y por su calificación

HV	Tipo	Número de elementos	Puntos
01	Orden oraciones	5 con 5 opciones cada uno	5
02	Frase imagen	4 con 4 opciones cada uno	4
03	RN y selección	2 abierto [a]	2
04	Selección frases	3	3
05	Selección frases	3	3
06	Elemento categoría	3	3
07	Selección elementos	3 con 3 opciones cada uno	3
08	Elemento categoría	4 con 2 opciones cada uno	4
09	Selección elementos	5 con 2 opciones cada uno	5
10	Selección elementos	5 abierto [a]	5
11	Selección elementos	5 con 2 opciones cada uno	5
12	Selección elementos	5 con 2 opciones cada uno	5
13	Selección frases	3 con 4 opciones cada uno	3
14	Selección elementos	3 con 3 opciones cada uno	3
15	Selección elementos	1 con 3 opciones cada uno	1
18	Frase imagen	5 con 5 opciones cada uno	5
16	Elemento categoría	3 con 2 opciones cada uno	3
17	Selección elementos	3 con 3 opciones cada uno	3
19	Frase imagen	-	-
20	Elemento categoría	5 con 3 opciones cada uno	5
HC			
01	RN/sucesiones	1	1
02	Elemento Imagen	3	3
03	Orden Números	4	3
04	RN/fórmula	1	1
05	RN/iluminación	1	1
06	RN/fórmula	1	1
07	Elemento Imagen	3	3
08	RN/fórmula	1	1
09	RN/etiquetas	1	1
10	RN/etiquetas	1	1
11	RN/fórmula	1	1
12	RN/fórmula	1	1
13	RN/rangos	1	1
14	RN/fórmula	1	1
15	RN/fórmula	1	1
16	RN/fórmula	1	1
17	RN/fórmula	1	1

Continúa tabla

18	RN/fórmula	1	1
19	RN/fórmula	2	2
20	RN/fórmula	1	1
ESP			
01	Elemento categoría	5 con 2 opciones cada uno	5
02	Selección frases	1 con 4ó5 opciones cada uno	1
03	Elemento categoría	3 con 2 opciones cada uno	3
04	Elemento categoría	5 con 2 opciones cada uno	5
05	Elemento categoría	4 con 2 opciones cada uno	4
06	Selección elementos	3 con 2 opciones cada uno	3
07	Selección elementos	3 con 2 opciones cada uno	3
08	Selección elementos	4 con 3 opciones cada uno	4
09	Selección elementos	5 con 2 opciones cada uno	5
10	Elemento categoría	5 con 3 opciones cada uno	5
11	Selección elementos	3 con 3 opciones cada uno	3
12	Selección elementos	2 con 3 opciones cada uno	2
13	Elemento categoría	3 con 2 opciones cada uno	3
14	Frase imagen	13 ó 14 abierto [a]	4
15	Elemento categoría	5 con 3 opciones cada uno	5
16	Elemento categoría	3 con 2 opciones cada uno	3
17	Elemento categoría	3 con 2 opciones cada uno	3
18	Elemento categoría	3 con 2 opciones cada uno	3
19	Frase imagen	19, 20 ó 21 con 2 opciones c/u	4
20	Elemento categoría	3 con 2 opciones cada uno	3
MAT			
01	R Algebraica	1	1
02	R Algebraica	1	1
03	R Algebraica	1	1
04	R Algebraica	1	1
05	R Algebraica	1	1
06	RN/ecuaciones	1	1
07	RN/ecuaciones	2	2
08	RN R Algebraica	1 o 2	1 o 2 [b]
09	Elemento categoría	3	3
10	RN/etiquetas	1	1
11	RN/etiquetas	1	1
12	Elemento categoría	3	3
13	RN/etiquetas	1	1
14	RN/triángulos	1	1
15	RN/fórmulas	1	1
16	RN/sucesiones	1	1
17	RN/gráficas	1	1

Continúa tabla

18	RN/fórmulas	3	3
19	RN/pendiente	1	1
20	RN R Algebraica	1 o 2	1 o 2 [b]
NAT			
01	Elemento categoría	5 con 2 opciones cada uno	5
02	Elemento categoría	5 con 3 opciones cada uno	5
03	Elemento categoría	5 con 3 opciones cada uno	5
04	Elemento categoría	5 con 2 opciones cada uno	5
05	Elemento categoría	5 con 2 opciones cada uno	5
06	Elemento categoría	5 con 2 opciones cada uno	5
07	RN/fórmulas	2	2
08	Elemento categoría	5 con 2opciones cada uno	5
09	RN/fórmulas	2	2
10	RN/rangos	1	1
11	RN/fórmulas	1	1
12	Orden elementos múltiple	6	6
13	Orden elementos múltiple	3	2
14	Elemento categoría	5 con 3opciones cada uno	5
15	Elemento categoría	5 con 3opciones cada uno	5
16	Elemento categoría	5 con 2opciones cada uno	5
17	Elemento categoría	2 con 10opciones cada uno	2
18	Elemento categoría	5 con 2opciones cada uno	5
19	Elemento categoría	5 con 2opciones cada uno	5
20	Selección frase	1 con 4opciones cada uno	1
SOC			
01	Elemento imagen	3 con 30 opciones cada uno	3
02	Elemento categoría	5 con 3opciones cada uno	5
03	Elemento categoría	3 con 2opciones cada uno	3
04	Elemento categoría	5 con 2opciones cada uno	5
05	Elemento categoría	5 con 3opciones cada un	5
06	Elemento Imagen	3 con 7opciones cada uno	3
06	Elemento Imagen	3 con n opciones cada uno	3
07	Elemento categoría	5 con 3opciones cada uno	5
08	Elemento categoría	5 con 3opciones cada uno	5
09	Elemento categoría	5 con 2opciones cada uno	5
10	Elemento categoría	5 con 3opciones cada uno	5
11	Elemento categoría	5 con 3opciones cada uno	5
12	Elemento categoría	5 con 3opciones cada uno	5
13	Elemento categoría	5 con 3opciones cada uno	5
14	Elemento categoría	5 con 2opciones cada uno	5
16	Elemento categoría	5 con 3opciones cada uno	5
17	Elemento categoría	5 con 3opciones cada uno	5

Continúa tabla

18	Elemento categoría	5	con 2opciones cada uno	5
19	Elemento categoría	5	con 3opciones cada uno	5
20	Elemento categoría	5	con 3opciones cada uno	5

Nota: [a] Abierto significa que son más de 10 opciones. [b] son dos familias, una de ellas requiere una respuesta, la otra solicita dos respuestas.

Anexo C

Índices psicométricos de VA y VB, según la TCT y el modelo de Rasch

Tabla C.1.
Índice de dificultad con su desviación estándar y correlación punto biserial para los ítems de VA y VB del EXHCOBA-R/MS

Item	VA			VB		
	Media	DE	R_{pb}	Media	DE	R_{pb}
HV01	0.58	0.37	0.387	0.35	0.42	0.244
HV02	0.67	0.34	0.233	0.65	0.35	0.293
HV03	0.29	0.33	0.153	0.36	0.38	0.248
HV04	0.85	0.27	0.364	0.62	0.20	0.249
HV05	0.73	0.29	0.395	0.49	0.30	**0.192**
HV06	0.64	0.23	0.212	0.64	0.25	**0.101**
HV07	0.73	0.31	0.341	0.29	0.38	0.295
HV08	0.88	0.28	**0.081**	0.81	0.30	0.212
HV09	0.91	0.14	0.393	0.86	0.16	0.237
HV10	0.63	0.22	0.260	0.42	0.28	**0.128**
HV11	0.76	0.21	0.241	0.58	0.21	0.245
HV12	0.68	0.21	0.425	0.78	0.21	0.266
HV13	0.56	0.31	**-0.045**	0.51	0.30	**0.131**
HV14	0.94	0.15	**0.189**	0.87	0.23	**0.124**
HV15	0.11	0.31	**0.086**	0.55	0.50	**-0.112**
HV16	0.54	0.21	**0.108**	0.92	0.23	0.243
HV17	0.90	0.22	**0.165**	0.74	0.30	0.314
HV18	0.95	0.15	**0.170**	0.95	0.18	**0.142**
HV19	--	--	--	--	--	--
HV20	0.92	0.15	0.347	0.86	0.19	0.281
HC01	0.77	0.42	**0.167**	0.54	0.50	0.338
HC02	0.24	0.34	0.281	0.30	0.42	0.465
HC03	0.65	0.41	0.344	0.56	0.44	0.363
HC04	0.31	0.46	0.309	0.43	0.50	0.401
HC05	0.64	0.48	0.496	0.57	0.50	0.318
HC06	0.25	0.43	0.398	0.31	0.46	0.517
HC07	0.31	0.35	**0.071**	0.53	0.40	**0.067**
HC08	0.29	0.45	0.437	0.30	0.46	0.326
HC09	0.46	0.50	0.327	0.04	0.20	**0.189**
HC10	0.34	0.48	0.447	0.32	0.47	0.512
HC11	0.03	0.17	0.289	0.19	0.40	0.346
HC12	0.07	0.25	**0.186**	0.04	0.20	0.262
HC13	0.41	0.49	0.229	0.56	0.50	0.263
HC14	0.51	0.50	0.377	0.49	0.50	0.480
HC15	0.38	0.49	0.353	0.33	0.47	0.547
HC16	0.48	0.50	0.364	0.43	0.50	0.385
HC17	0.47	0.50	0.254	0.40	0.49	0.405
HC18	0.67	0.47	0.322	0.69	0.47	0.349

Continúa tabla

Item	VA			VB		
	Media	DE	R_{pb}	Media	DE	R_{pb}
HC19	0.55	0.38	0.329	0.64	0.34	0.214
HC20	0.41	0.49	0.385	0.55	0.50	0.456
ESP01	0.46	0.14	**0.151**	0.76	0.28	0.229
ESP02	0.12	0.33	**0.177**	0.53	0.50	0.082
ESP03	0.76	0.33	0.306	0.80	0.26	0.260
ESP04	0.82	0.26	**0.139**	0.65	0.16	**0.155**
ESP05	0.91	0.18	**0.096**	0.76	0.24	0.273
ESP06	0.86	0.23	0.296	0.89	0.24	**0.114**
ESP07	0.90	0.17	**0.172**	0.74	0.24	**0.107**
ESP08	0.86	0.23	0.260	0.75	0.24	0.292
ESP09	0.77	0.17	**0.083**	0.61	0.24	**0.186**
ESP10	0.47	0.25	0.288	0.67	0.20	**0.084**
ESP11	0.81	0.22	0.313	0.70	0.25	0.357
ESP12	0.72	0.35	0.203	0.50	0.38	**0.142**
ESP13	0.71	0.23	**0.152**	0.45	0.31	**0.060**
ESP14	0.56	0.19	0.226	0.57	0.15	0.364
ESP15	0.96	0.13	0.262	0.78	0.18	0.209
ESP16	0.87	0.24	**0.046**	0.77	0.24	**-0.023**
ESP17	0.86	0.28	0.224	0.76	0.32	0.342
ESP18	0.63	0.34	**0.043**	0.59	0.35	0.242
ESP19	0.77	0.24	0.292	0.77	0.22	0.351
ESP20	0.60	0.18	**0.115**	0.93	0.20	**0.116**
MAT01	0.04	0.19	0.374	0.05	0.22	0.344
MAT02	0.04	0.19	0.278	0.03	0.18	0.339
MAT03	0.06	0.24	0.290	0.06	0.24	0.285
MAT04	0.04	0.19	0.212	0.01	0.09	0.242
MAT05	0.45	0.50	0.339	0.29	0.45	0.342
MAT06	0.17	0.38	0.336	0.08	0.28	0.326
MAT07	0.13	0.34	0.248	0.16	0.37	**0.079**
MAT08	0.06	0.24	0.386	0.28	0.32	0.286
MAT09	0.92	0.23	**0.142**	0.58	0.21	0.203
MAT10	0.08	0.27	0.283	0.06	0.24	0.412
MAT11	0.50	0.50	0.460	0.48	0.50	0.529
MAT12	0.81	0.36	0.260	0.64	0.26	0.235
MAT13	0.80	0.40	0.243	0.53	0.50	0.261
MAT14	--	--	--	--	--	--
MAT15	0.15	0.36	0.261	0.22	0.42	0.298
MAT16	0.16	0.37	0.245	0.10	0.30	0.398
MAT17	0.67	0.47	0.209	0.80	0.40	0.002
MAT18	0.15	0.22	0.376	0.21	0.22	**0.193**
MAT19	0.15	0.36	0.209	0.00	0.00	--
MAT20	0.01	0.08	**0.155**	0.19	0.26	**0.220**
BIO01	0.54	0.33	**0.121**	0.37	0.18	**-0.022**
BIO02	0.63	0.24	**0.092**	0.70	0.20	**0.189**
BIO03	0.59	0.21	**0.101**	0.59	0.23	0.278
BIO04	0.72	0.29	**0.119**	0.45	0.14	**0.083**
BIO05	0.73	0.22	**0.042**	0.67	0.29	0.214

Continúa tabla

Item	VA			VB		
	Media	DE	R_{pb}	Media	DE	R_{pb}
BIO06	0.56	0.26	**0.121**	0.49	0.22	**0.071**
FIS07	0.38	0.33	0.403	0.35	0.28	**0.160**
FIS08	0.68	0.20	0.373	0.78	0.23	**0.045**
FIS09	0.26	0.34	0.287	0.19	0.24	0.382
FIS10	0.31	0.47	0.257	0.37	0.49	**0.134**
FIS11	0.01	0.11	**0.195**	0.02	0.13	0.245
FIS12	0.55	0.21	0.439	0.61	0.22	0.223
QUI13	0.29	0.28	0.061	0.82	0.35	0.259
QUI14	0.64	0.29	0.304	0.23	0.17	**-0.049**
QUI15	0.71	0.26	0.228	0.75	0.30	0.310
QUI16	0.35	0.27	**-0.048**	0.58	0.22	0.221
QUI17	0.42	0.45	0.244	0.23	0.37	0.287
QUI18	0.61	0.29	0.388	0.47	0.21	0.266
QUI19	0.60	0.41	**0.162**	0.55	0.44	**0.151**
QUI20	0.66	0.48	**0.169**	0.13	0.33	**0.061**
GEO01	0.19	0.30	0.407	0.10	0.21	**0.125**
GEO02	0.44	0.17	0.270	0.55	0.24	0.409
GEO03	0.73	0.34	0.395	0.88	0.25	0.399
GEO04	0.64	0.25	0.424	0.69	0.30	0.251
GEO05	0.48	0.31	0.281	0.52	0.27	**0.162**
GEO06	0.41	0.29	0.502	0.40	0.28	0.389
HIS06	--	--	--	--	--	--
HIS07	0.46	0.30	0.444	0.25	0.23	**0.094**
HIS08	0.40	0.22	0.298	0.39	0.23	0.262
HIS09	0.63	0.30	0.325	0.56	0.25	0.284
HIS10	0.46	0.31	0.397	0.41	0.28	0.322
HIS11	0.34	0.22	0.252	0.57	0.31	0.351
HIS12	0.63	0.34	0.341	0.57	0.34	0.461
HIS13	0.40	0.25	0.296	0.36	0.23	0.294
FCYE14	0.50	0.27	0.362	0.55	0.31	0.283
FCYE16	0.61	0.30	0.452	0.70	0.33	0.345
FCYE17	0.53	0.30	0.212	0.64	0.32	0.264
FCYE18	0.63	0.31	0.272	0.50	0.35	0.282
FCYE19	0.24	0.25	0.229	0.34	0.24	0.351
FCYE20	0.67	0.36	0.320	0.56	0.31	0.404

Nota: Para VA se consideraron 163 datos, para VB fueron 118. DE = desviación estándar. Pbs = coeficiente de correlación biserial.

Tabla C.2.
Infit, outfit, correlación punto medida y discriminación según el modelo de Rasch para VA y VB

Item	VA				VB				
	Infit	*Outfit*	Pmed	Discr.	*Infit*	*Outfit*	Pmed	Discr.	Item
HV01	1.16	1.19	0.31	**0.76**	1.22	**1.41**	0.34	0.87	HV01
HV02	1.18	1.24	**0.18**	**0.73**	1.07	1.20	0.30	0.88	HV02
HV03	0.97	0.97	0.27	1.05	0.98	1.07	0.24	1.01	HV03
HV04	0.98	0.95	0.31	1.02	1.05	1.02	**0.19**	0.97	HV04
HV05	0.97	0.95	0.32	1.06	1.06	1.07	0.21	0.86	HV05
HV06	1.03	1.01	**0.18**	0.97	1.03	1.02	0.23	0.97	HV06
HV07	1.00	0.97	0.30	1.02	1.04	1.12	0.25	0.93	HV07
HV08	1.15	**1.51**	**0.18**	0.93	1.17	1.28	0.22	0.89	HV08
HV09	1.06	1.07	0.34	1.04	1.15	1.22	0.28	0.88	HV09
HV10	1.00	1.00	0.33	1.00	1.22	1.24	**0.19**	**0.73**	HV10
HV11	1.01	1.06	0.30	0.99	1.02	1.03	0.32	0.98	HV11
HV12	1.01	0.99	0.33	0.99	1.09	1.18	0.29	0.92	HV12
HV13	1.11	1.11	**0.11**	**0.72**	1.03	1.03	0.23	0.92	HV13
HV14	0.95	0.89	0.29	1.03	0.99	0.99	0.31	1.01	HV14
HV15	1.01	1.07	**0.04**	0.99	1.06	1.06	**0.03**	**-0.01**	**HV15**
HV16	1.03	1.03	**0.13**	0.96	1.02	0.86	0.34	1.01	HV16
HV17	0.98	1.02	0.26	1.01	1.01	0.98	0.30	1.00	HV17
HV18	1.02	1.30	0.26	0.99	1.19	**1.89**	0.28	0.97	HV18
HV19	---	---	---	---	---	---	---	---	HV19
HV20	1.09	1.15	0.23	0.98	1.13	1.13	0.28	0.93	HV20
HC01	0.99	0.97	**0.19**	1.04	0.97	0.97	0.24	1.54	HC01
HC02	1.05	1.06	0.24	0.96	1.02	1.09	0.31	0.99	HC02
HC03	1.03	1.07	0.30	0.96	1.05	1.14	0.30	0.93	HC03
HC04	0.97	0.97	0.23	1.06	0.96	0.96	0.26	1.36	HC04
HC05	0.93	0.93	0.36	1.89	0.95	0.94	0.30	1.68	HC05
HC06	0.94	0.91	0.34	1.11	0.91	0.85	0.40	1.16	HC06
HC07	1.15	1.16	**0.09**	**0.70**	1.13	1.19	**0.18**	**0.71**	HC07
HC08	0.94	0.92	0.32	1.12	0.96	0.94	0.25	1.07	HC08
HC09	0.95	0.95	0.30	1.39	1.00	1.04	**0.04**	1.00	**HC09**
HC10	0.92	0.89	0.39	1.20	0.92	0.87	0.38	1.18	HC10
HC11	0.97	0.66	0.23	1.03	0.95	0.87	0.27	1.06	HC11
HC12	0.98	0.92	**0.17**	1.02	0.98	0.82	**0.18**	1.02	HC12
HC13	1.00	0.99	**0.17**	1.05	0.96	0.96	0.27	1.69	HC13
HC14	0.95	0.94	0.32	1.88	0.95	0.95	0.29	1.87	HC14
HC15	0.95	0.94	0.30	1.19	0.90	0.87	0.41	1.30	HC15
HC16	0.93	0.93	0.36	2.02	0.95	0.96	0.28	1.77	HC16
HC17	0.96	0.95	0.29	1.28	0.92	0.90	0.38	1.55	HC17
HC18	0.97	0.96	0.26	1.27	1.00	0.99	**0.18**	1.02	HC18
HC19	0.96	0.96	0.30	1.14	1.00	1.00	0.24	1.00	HC19
HC20	0.95	0.94	0.31	1.27	0.91	0.90	0.40	2.59	HC20
ESP01	1.07	1.06	**0.07**	0.95	1.18	1.27	0.25	0.85	ESP01
ESP02	0.98	0.98	**0.19**	1.02	1.08	1.08	**-0.01**	**-0.25**	**ESP02**
ESP03	1.02	1.00	0.26	0.99	0.97	0.95	0.34	1.04	ESP03
ESP04	1.13	**1.37**	**0.16**	0.88	1.05	1.08	0.25	0.97	ESP04
ESP05	1.02	1.36	**0.18**	0.99	1.04	1.11	0.30	0.95	ESP05
ESP06	0.98	0.98	0.24	1.02	0.98	0.98	0.32	1.01	ESP06

Continúa tabla

	VA					VB				
Item	*Infit*	*Outfit*	Pmed	Discr.		*Infit*	*Outfit*	Pmed	Discr.	Item
ESP07	0.99	0.96	0.23	1.02		1.08	1.13	**0.14**	0.86	ESP07
ESP08	1.06	1.04	0.22	0.97		1.02	1.06	0.31	0.97	ESP08
ESP09	1.08	1.11	**0.15**	0.95		1.04	1.05	0.33	0.94	ESP09
ESP10	1.03	1.02	0.31	0.96		1.05	1.08	0.31	0.94	ESP10
ESP11	0.99	0.99	0.24	1.02		0.98	0.97	0.32	1.03	ESP11
ESP12	1.01	1.01	**0.19**	0.99		1.05	1.05	**0.15**	0.81	ESP12
ESP13	1.00	1.00	0.20	1.00		1.09	1.16	**0.17**	0.92	ESP13
ESP14	0.98	0.98	0.26	1.02		1.04	1.03	0.25	1.00	ESP14
ESP15	1.00	0.97	0.26	1.01		0.97	1.08	0.38	1.01	ESP15
ESP16	1.03	1.01	0.22	0.98		1.00	1.03	0.26	0.98	ESP16
ESP17	1.03	1.07	**0.16**	0.99		0.98	0.96	0.33	1.02	ESP17
ESP18	1.13	1.24	**0.08**	0.71		1.06	1.07	0.25	0.81	ESP18
ESP19	1.00	1.03	0.28	1.01		0.93	0.99	0.41	1.07	ESP19
ESP20	1.00	1.00	**0.17**	1.00		0.98	1.04	0.29	1.00	ESP20
MAT01	0.96	0.67	0.29	1.04		0.97	0.77	0.23	1.03	MAT01
MAT02	0.98	0.70	0.23	1.03		0.98	0.70	0.20	1.03	MAT02
MAT03	0.98	0.87	**0.19**	1.02		0.98	0.85	**0.17**	1.02	MAT03
MAT04	0.97	0.79	0.25	1.03		0.99	0.72	**0.14**	1.02	MAT04
MAT05	0.95	0.94	0.31	1.27		0.96	0.95	0.24	1.08	MAT05
MAT06	0.96	0.91	0.27	1.05		0.93	0.86	0.31	1.06	MAT06
MAT07	0.98	0.91	0.21	1.02		1.02	1.04	**0.05**	0.98	**MAT07**
MAT08	0.97	0.86	0.23	1.03		0.98	0.98	0.23	1.03	MAT08
MAT09	1.00	1.09	0.20	0.99		0.97	0.90	0.28	1.01	MAT09
MAT10	0.97	0.94	**0.18**	1.02		0.95	0.70	0.31	1.05	MAT10
MAT11	0.94	0.93	0.35	1.87		0.93	0.92	0.35	2.15	MAT11
MAT12	1.02	1.06	0.29	0.99		0.97	0.97	0.32	1.03	MAT12
MAT13	0.95	0.92	0.31	1.13		0.96	0.95	0.28	1.81	MAT13
MAT14	---	---	---	---		---	---	---	---	MAT14
MAT15	0.98	0.89	0.21	1.03		0.96	0.89	0.26	1.06	MAT15
MAT16	0.98	0.91	0.23	1.03		0.96	0.83	0.26	1.04	MAT16
MAT17	0.99	0.99	**0.18**	1.04		0.97	0.95	0.25	1.06	MAT17
MAT18	0.94	0.90	0.33	1.05		1.01	1.09	**0.16**	0.97	MAT18
MAT19	1.00	0.99	0.14	1.01		---	---	---	---	MAT19
MAT20	0.99	0.66	**0.13**	1.02		0.99	0.98	0.20	1.02	MAT20
BIO01	1.19	1.21	0.25	**0.70**		1.12	1.13	**0.14**	0.88	BIO01
BIO02	1.08	1.11	0.27	0.89		1.01	0.97	0.35	0.98	BIO02
BIO03	1.06	1.05	0.25	0.94		0.98	0.98	0.38	1.01	BIO03
BIO04	1.19	1.22	0.27	**0.75**		1.02	1.02	0.24	0.98	BIO04
BIO05	1.13	1.13	0.22	0.83		1.03	1.06	0.37	0.94	BIO05
BIO06	1.11	1.12	0.26	0.88		1.12	1.15	0.20	0.87	BIO06
FIS07	0.92	0.91	0.37	1.16		1.00	1.00	**0.19**	1.00	FIS07
FIS08	0.91	0.92	0.46	1.05		1.06	1.09	0.28	0.93	FIS08
FIS09	0.99	0.98	0.25	1.01		0.92	0.90	0.36	1.32	FIS09
FIS10	0.96	0.94	0.27	1.09		0.97	0.96	0.23	1.13	FIS10
FIS11	0.99	0.87	**0.13**	1.01		0.98	0.82	**0.18**	1.02	FIS11
FIS12	0.96	0.95	0.39	1.05		0.99	0.99	0.33	1.01	FIS12
QUI13	1.07	1.08	**0.06**	0.87		0.93	0.87	0.37	1.06	QUI13
QUI14	0.96	0.96	0.45	1.08		1.14	1.15	**0.08**	**0.76**	**QUI14**
QUI15	0.93	0.90	0.46	1.10		0.96	0.93	0.46	1.05	QUI15
QUI16	1.29	**1.34**	**0.13**	**0.67**		1.02	1.04	0.35	0.98	QUI16
QUI17	1.01	0.99	0.26	0.99		0.95	0.89	0.29	1.03	QUI17
QUI18	0.92	0.92	0.48	1.18		1.07	1.09	0.27	0.93	QUI18

Continúa tabla

Item	VA				VB				
	Infit	*Outfit*	Pmed	Discr.	*Infit*	*Outfit*	Pmed	Discr.	Item
QUI19	**1.34**	**1.66**	0.24	**0.68**	**1.33**	**1.59**	0.28	0.81	**QUI19**
QUI20	0.96	0.95	0.29	1.51	1.02	1.01	**0.07**	0.98	**QUI20**
GEO01	0.89	0.81	0.44	1.07	0.96	0.81	0.30	1.03	GEO01
GEO02	0.90	0.89	0.43	1.06	0.87	0.87	0.49	1.17	GEO02
GEO03	0.88	0.85	0.46	1.18	0.83	**0.73**	0.49	1.08	GEO03
GEO04	0.84	0.81	0.52	1.14	0.99	1.06	0.39	0.97	GEO04
GEO05	1.03	1.04	0.37	0.94	0.99	0.99	0.38	1.02	GEO05
GEO06	0.87	0.87	0.48	1.24	0.89	0.88	0.44	1.22	GEO06
HIS06	---	---	---	---	---	---	---	---	HIS06
HIS07	0.87	0.87	0.53	1.28	1.01	1.01	0.32	0.99	HIS07
HIS08	0.90	0.91	0.45	1.18	1.02	1.02	0.32	0.97	HIS08
HIS09	0.98	0.98	0.42	1.01	0.92	0.95	0.44	1.11	HIS09
HIS10	0.89	0.90	0.52	1.23	0.90	0.89	0.48	1.18	HIS10
HIS11	0.96	0.96	0.37	1.10	0.90	0.89	0.48	1.20	HIS11
HIS12	0.95	0.97	0.47	1.09	0.80	**0.78**	0.57	1.42	HIS12
HIS13	0.94	0.94	0.43	1.13	0.99	0.98	0.36	1.03	HIS13
FCYE14	0.97	0.96	0.40	1.06	0.93	0.92	0.46	1.12	FCYE14
FCYE16	0.87	0.87	0.51	1.23	0.91	0.84	0.49	1.11	FCYE16
FCYE17	1.05	1.07	0.33	0.91	0.89	0.88	0.49	1.18	FCYE17
FCYE18	1.00	1.01	0.38	0.97	1.00	1.00	0.44	1.01	FCYE18
FCYE19	1.09	1.13	0.26	0.88	0.95	0.95	0.41	1.09	FCYE19
FCYE20	1.02	1.06	0.40	0.97	0.90	0.90	0.49	1.23	FCYE20

Al momento de elaborar las especificaciones de reactivos, se presentaron serias dificultades conceptuales en el contenido FCYE15, por lo que el Comité Técnico del EXHCOBA decidió eliminar este contenido y agregar uno más en Geografía (GEO06); esta es la razón por la que en los contenidos de Ciencias sociales se repite el número 6 (uno para geografía y otro para historia) y falta el 15. Los ítems HV19 y HV06 no se examinaron y el ítem MAT14 no fue decodificado, por dichos motivos no aparecen los resultados en la tabla.

Anexo D

Resultados de los análisis psicométricos de VA y VB, por área

1. Habilidades del lenguaje

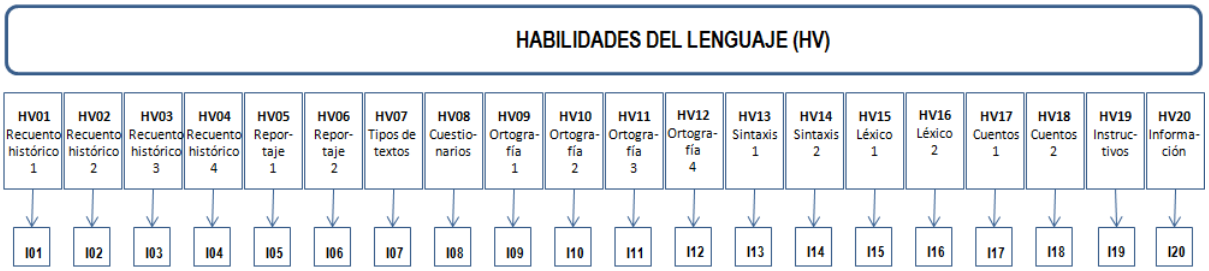
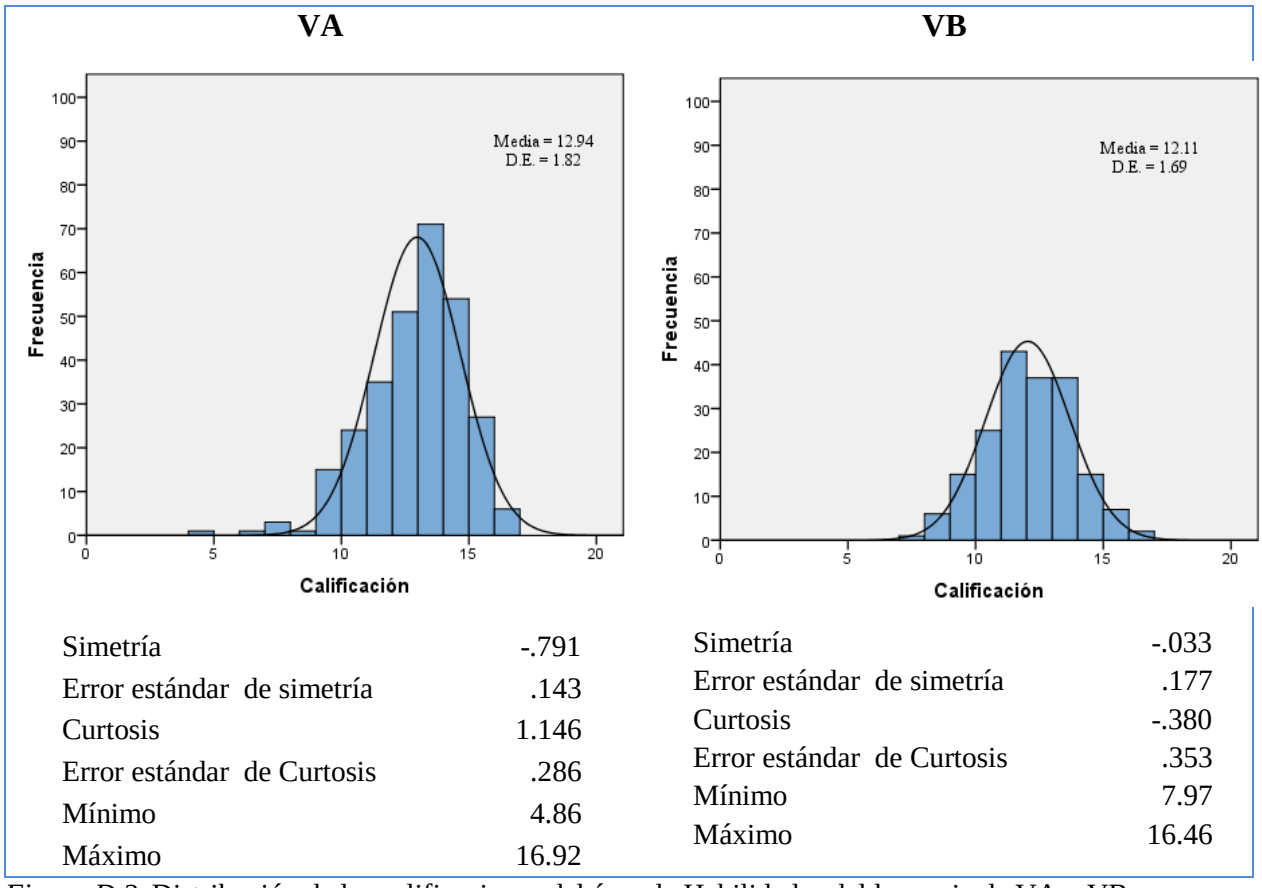


Figura D.1. Esquema del área de Habilidades del lenguaje de VA y VB.



Simetría	-.791
Error estándar de simetría	.143
Curtosis	1.146
Error estándar de Curtosis	.286
Mínimo	4.86
Máximo	16.92

Simetría	-.033
Error estándar de simetría	.177
Curtosis	-.380
Error estándar de Curtosis	.353
Mínimo	7.97
Máximo	16.46

Figura D.2. Distribución de las calificaciones del área de Habilidades del lenguaje de VA y VB.

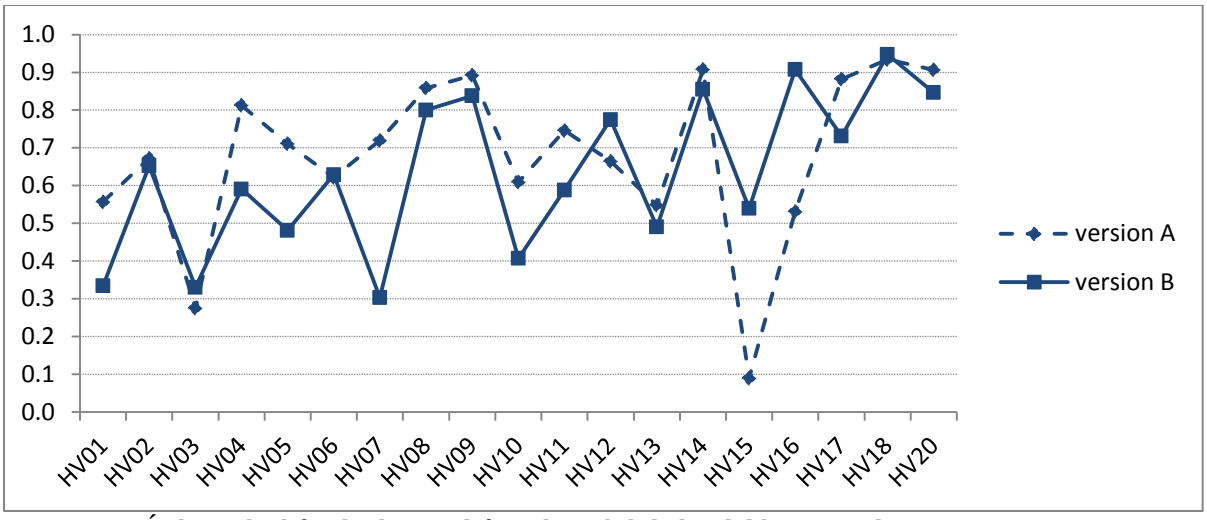


Figura D.3. Índices de dificultad para el área de Habilidades del lenguaje de VA y VB.

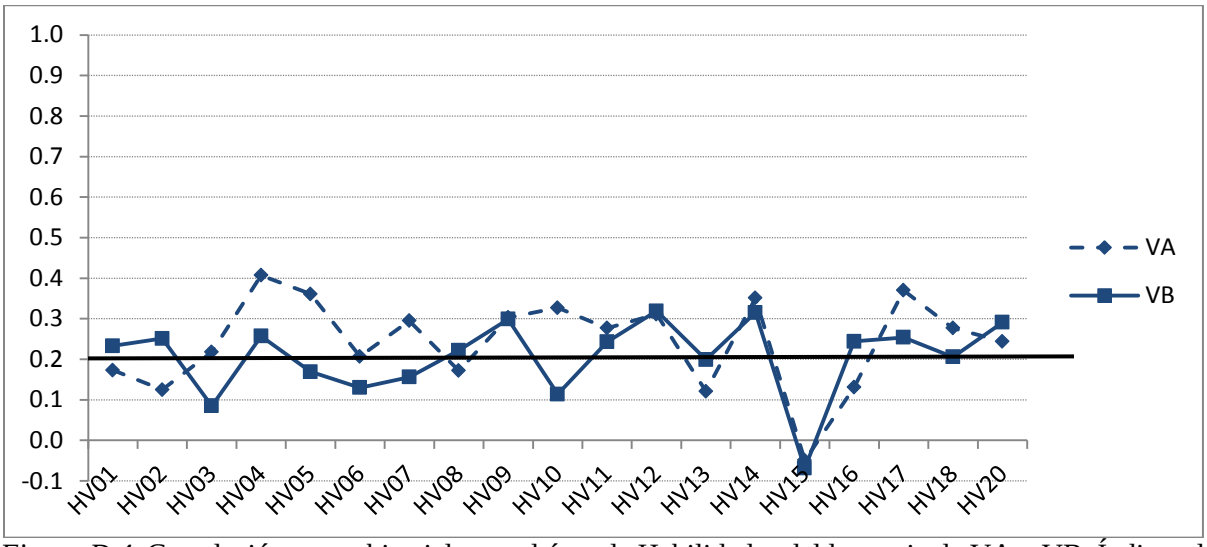
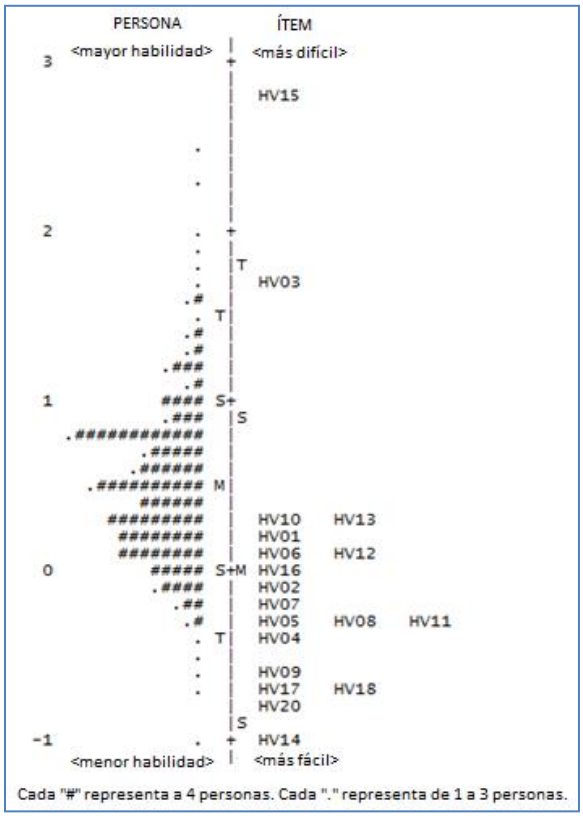


Figura D.4. Correlación punto biserial para el área de Habilidades del lenguaje de VA y VB. Índices de confiabilidad, alpha de Cronbach: VA = 0.633, VB = 0.547.

VA



VB

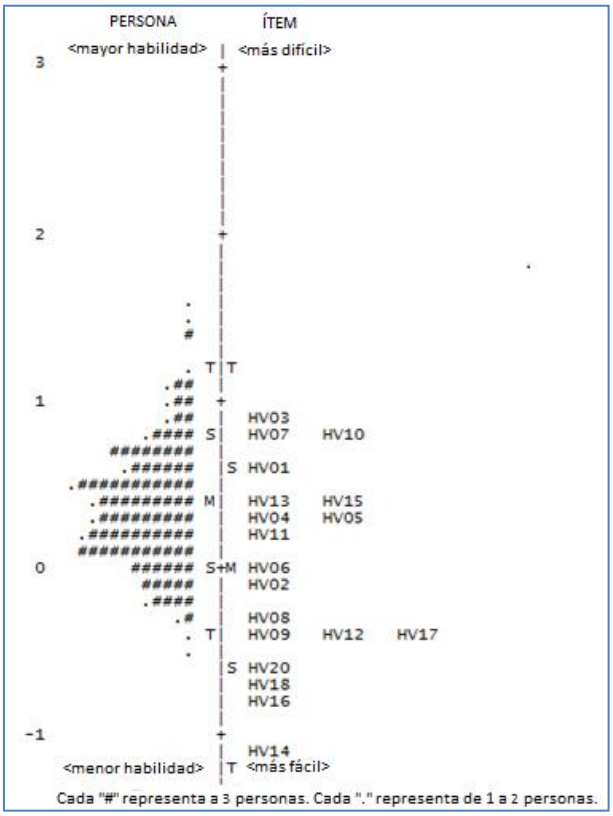


Figura D.5. Mapas de Wright de Habilidades del lenguaje de VA y VB.

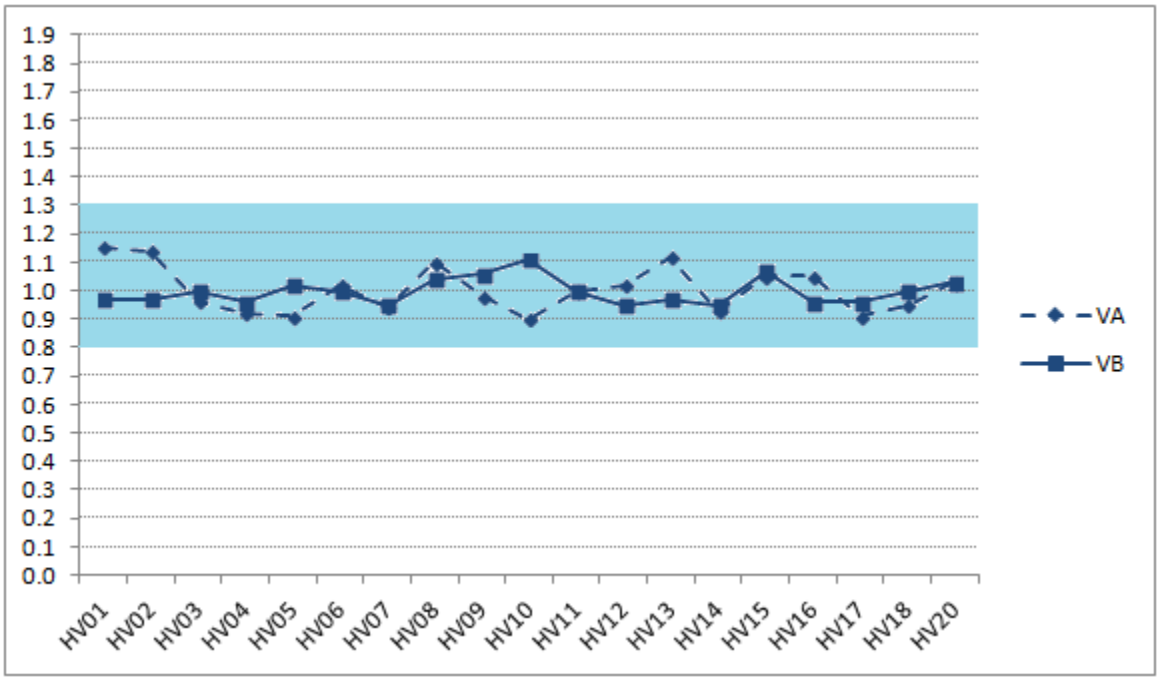


Figura D.6. Valores de *infit* para cada ítem del área de Habilidades del lenguaje en VA y VB.

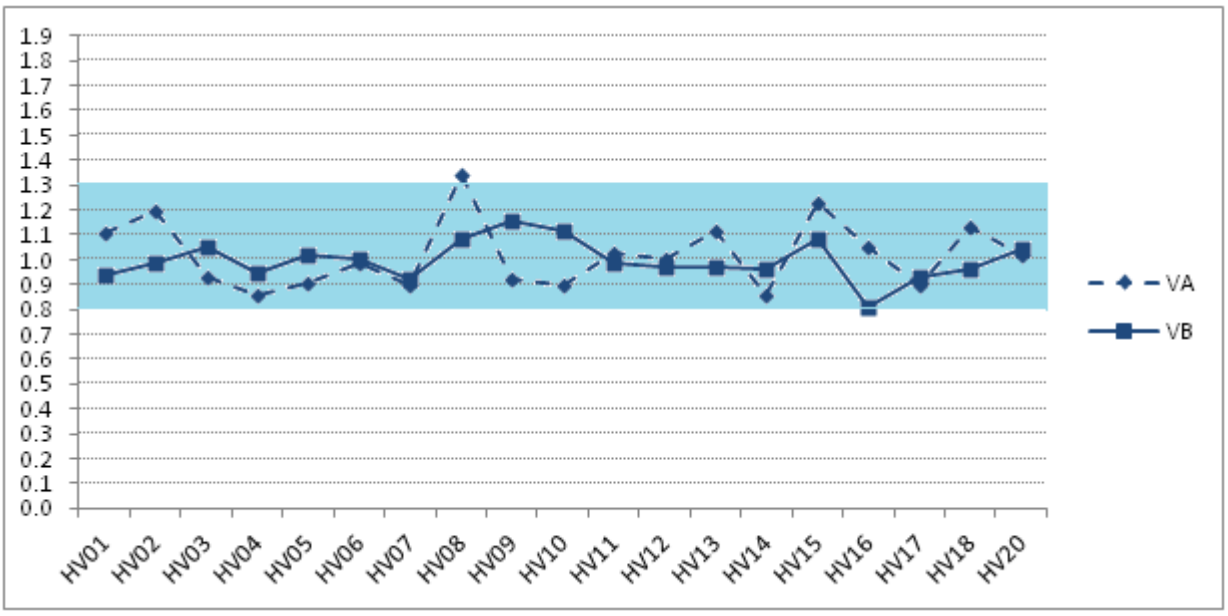


Figura D.7. Valores de *outfit* de cada ítem del área de Habilidades verbales de VA y VB.

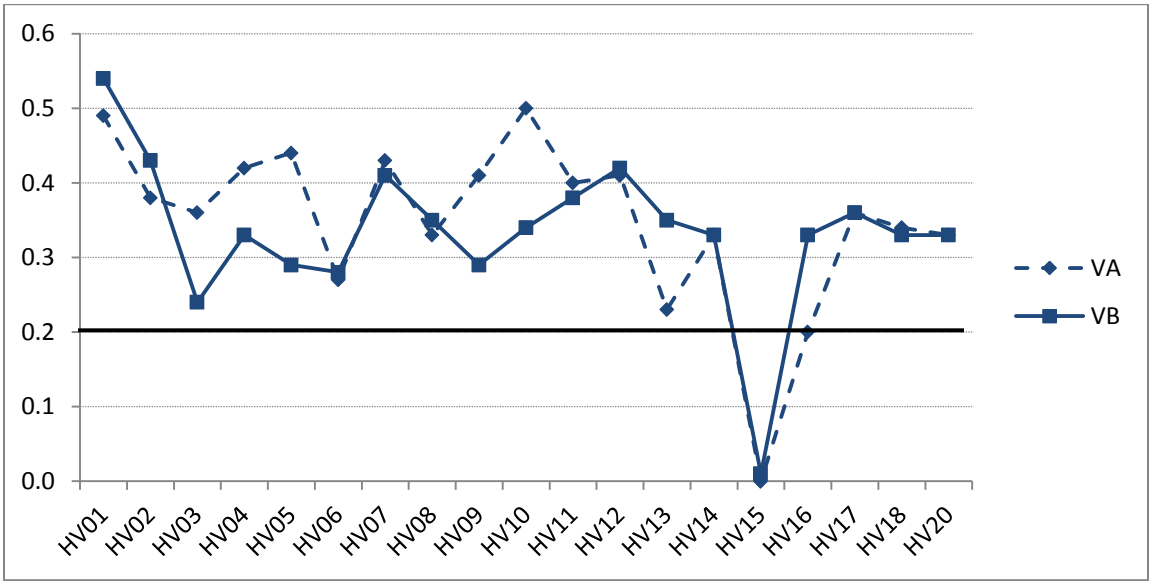


Figura D.8. Correlación punto medida para el área de Habilidades del lenguaje de VA y VB.

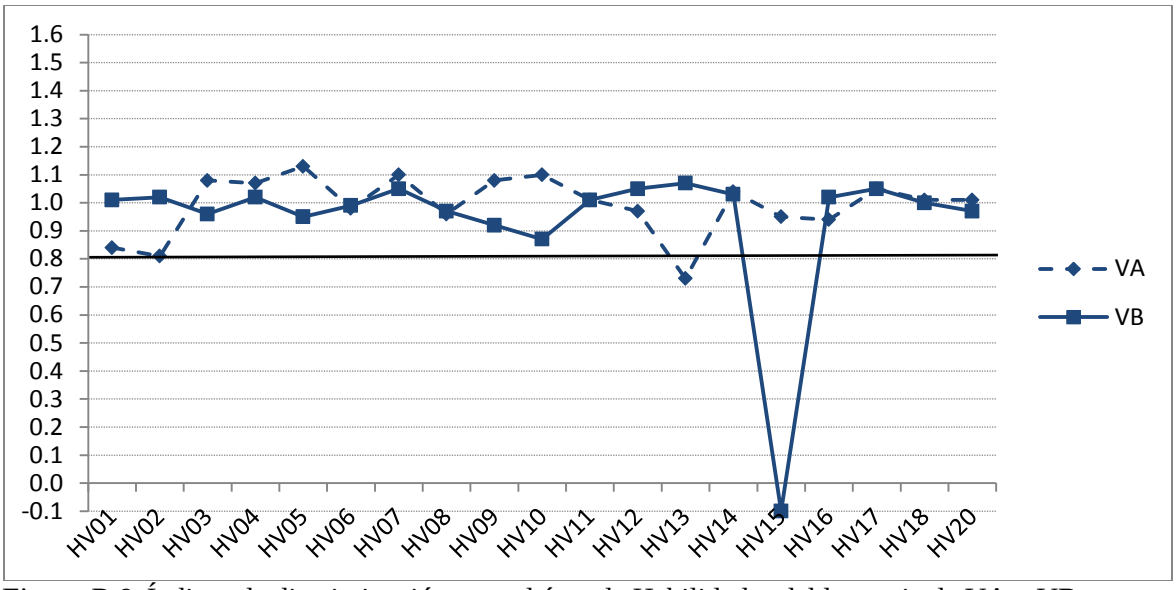


Figura D.9. Índices de discriminación para el área de Habilidades del lenguaje de VA y VB.

Tabla D.1.
Área de Habilidades del lenguaje. Índices de ajuste de los AFC para VA y VB, por modelo propuesto.

Índices	Versión A		Versión B	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Chi cuadrado	116.793	108.972	163.571	119.833
Grados libertad	125	122	133	123
p	.687	.794	.037	0.563
NNFI	1.030	1.048	**0.796**	1.023
CFI	1.000	1.000	**.833**	1.000
RMSEA	.000	.000	.035	0.000
Alpha de Cronb.	.657	.657	.609	.609
Covarianza F1F2		.815		.976

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*.

Ítem	VA (- -)	VB (--)
HV01	.278	.276
HV02	**.146**	.265
HV03	.259	**.113**
HV04	.362	.299
HV05	.384	.200
HV06	.301	**.183**
HV07	.370	.217
HV08	**.160**	.317
HV09	.431	.357
HV10	.397	**.151**
HV11	.351	.298
HV12	.402	.441
HV13	**.198**	.263
HV14	.470	.458
HV16	**.176**	.336
HV17	.483	.377
HV18	.362	.367
HV20	.283	.410

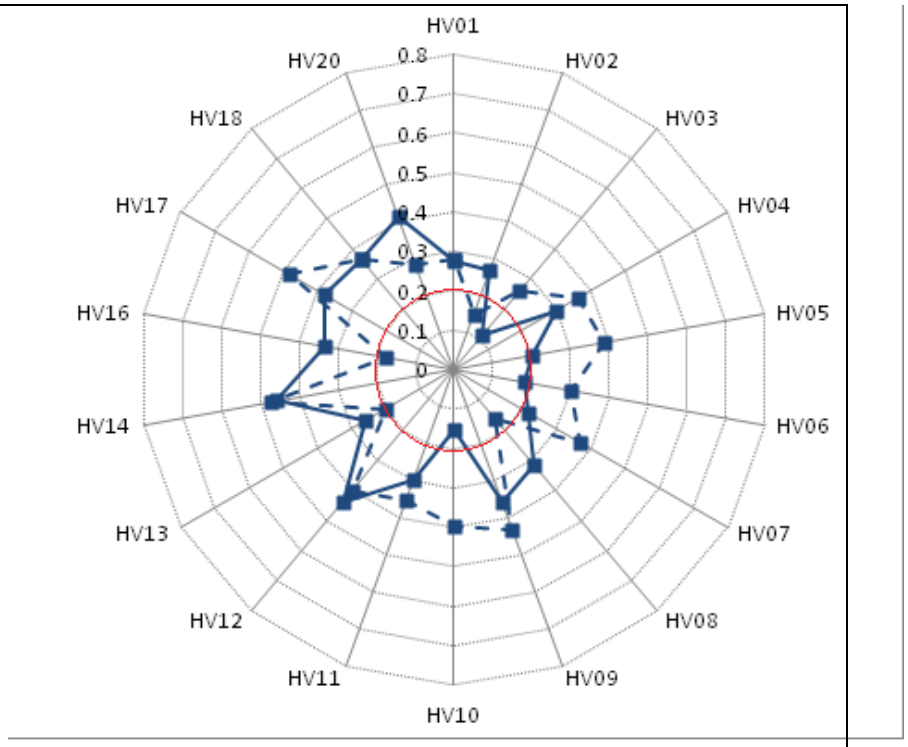


Figura D.10. Cargas factoriales estandarizadas para Habilidades del lenguaje de VA y VB. Modelo 1: Un factor con errores que covarían.

Ítems	VA (- -)	VB (--)
HV01	.328	.371
HV02	**.149**	.249
HV03	.282	**.101**
HV04	.406	.313
HV05	.374	**.175**
HV06	.321	.205
HV07	.370	.232
HV08	**.164**	.351
HV16	**.127**	.275
HV17	.567	.389
HV20	.235	.409
HV09	.477	.259
HV10	.424	.217
HV11	.473	.286
HV12	.450	.442
HV13	**.172**	.228
HV14	.444	.477
HV18	.431	.333

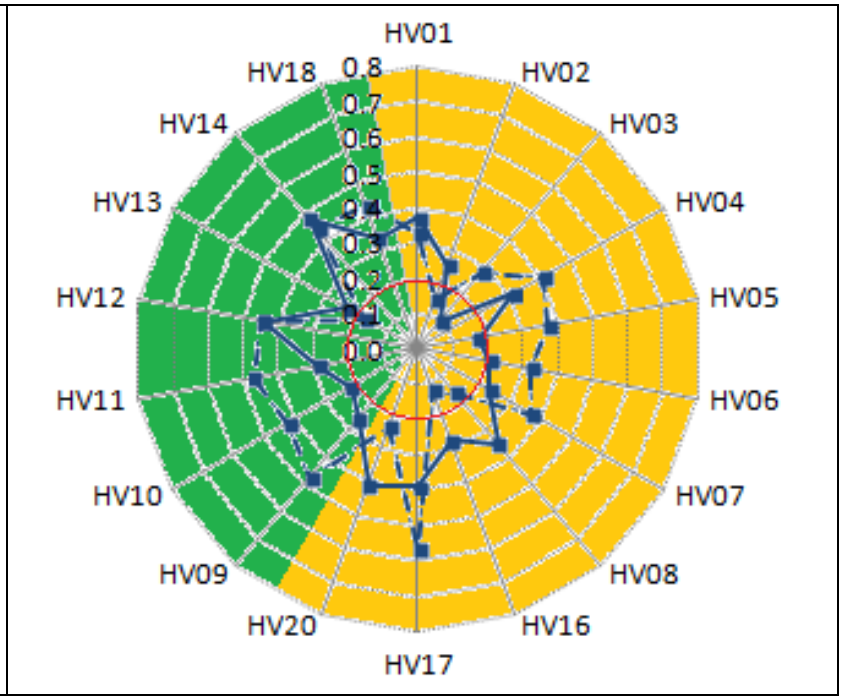


Figura D.11. Cargas factoriales estandarizadas para Habilidades del lenguaje de VA y VB. Modelo 2: dos factores con errores que covarían.
Estadísticos significativos al nivel de 0.05, excepto para las cargas de HV02 y HV16 de VA, y HV03, HV05, HV06, HV10 y HV13 de VB.

2. Español

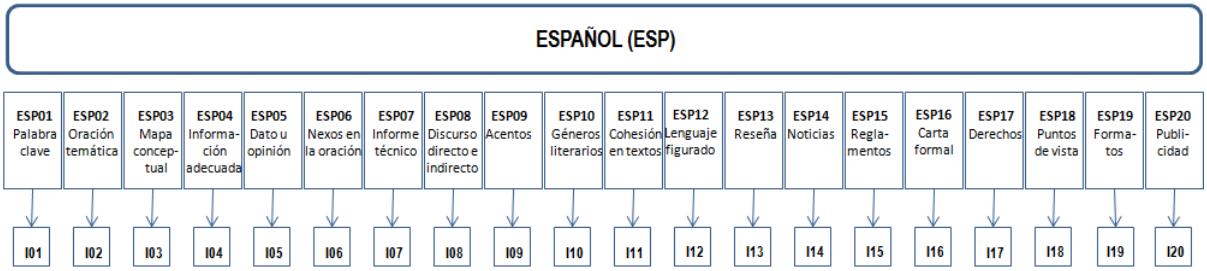
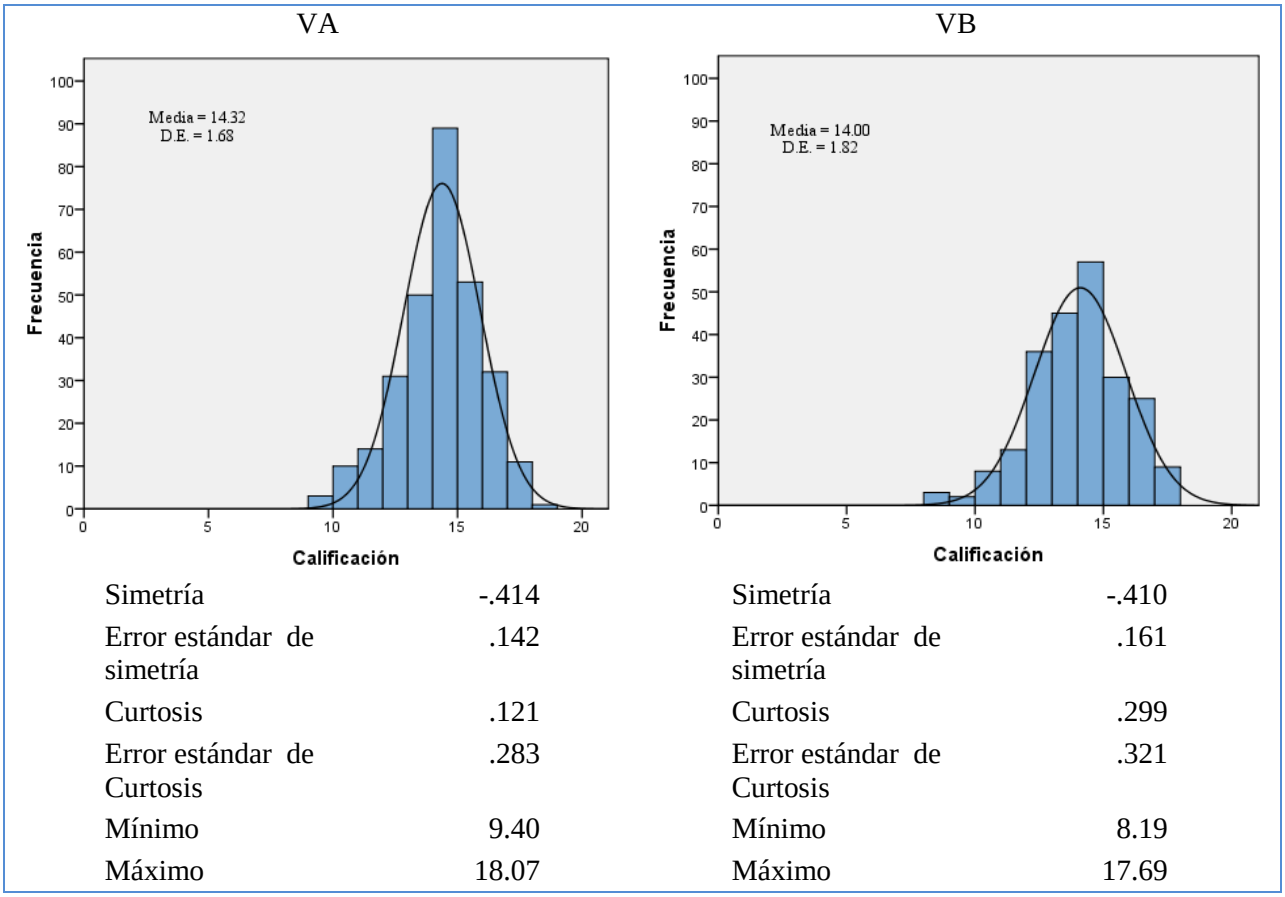


Figura D.12. Esquema del área de Español de VA y VB



Simetría	-.414	Simetría	-.410
Error estándar de simetría	.142	Error estándar de simetría	.161
Curtosis	.121	Curtosis	.299
Error estándar de Curtosis	.283	Error estándar de Curtosis	.321
Mínimo	9.40	Mínimo	8.19
Máximo	18.07	Máximo	17.69

Figura D.13. Distribución de las calificaciones del área de Español de VA y VB.

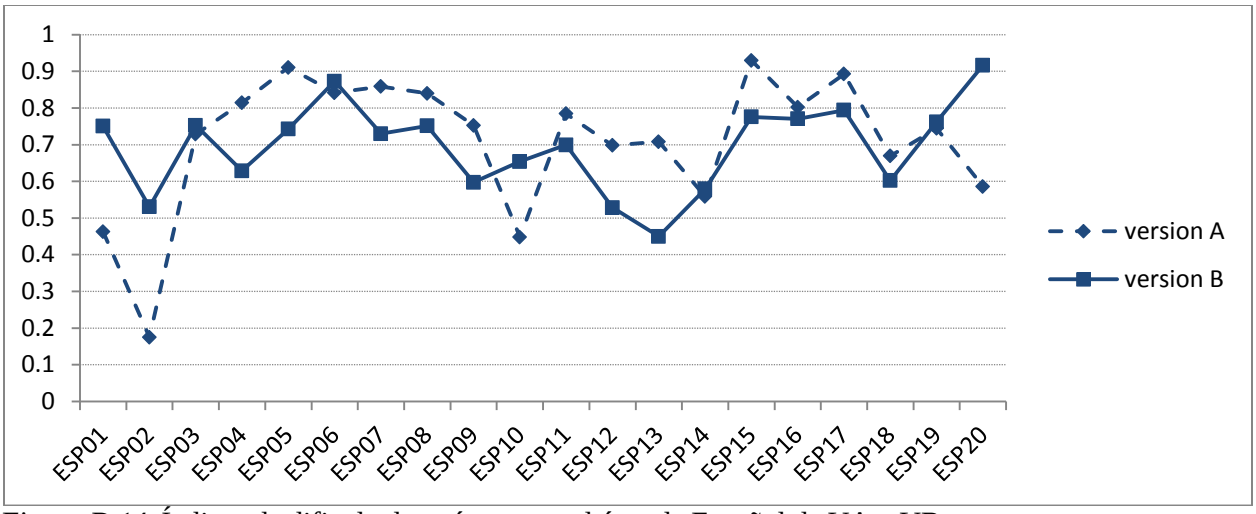


Figura D.14. Índices de dificultad por ítem para el área de Español de VA y VB.

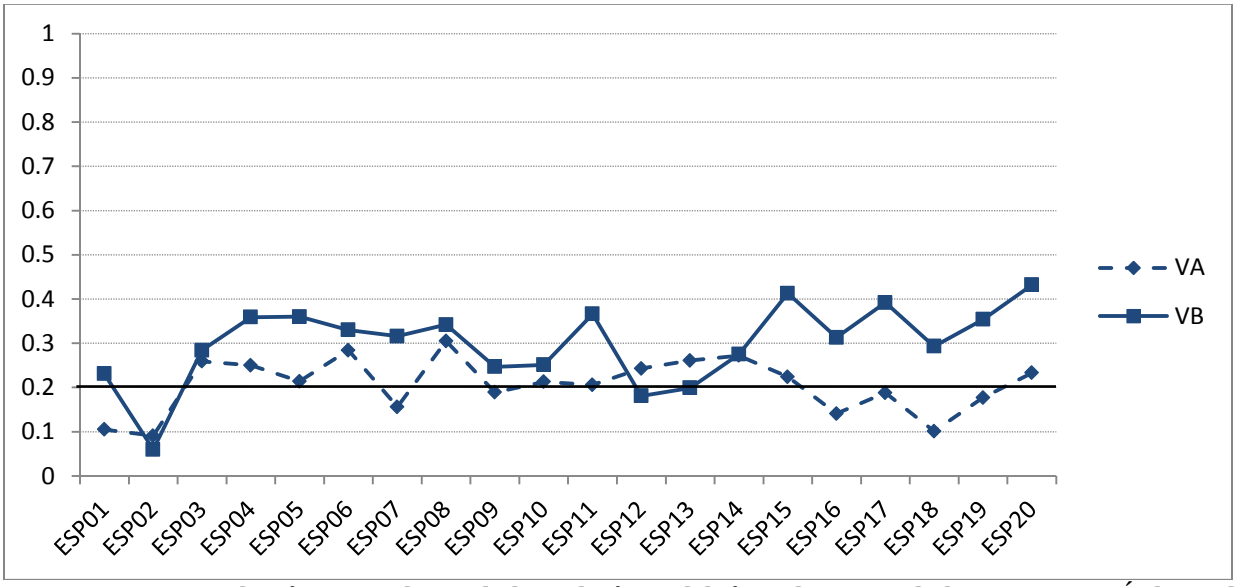
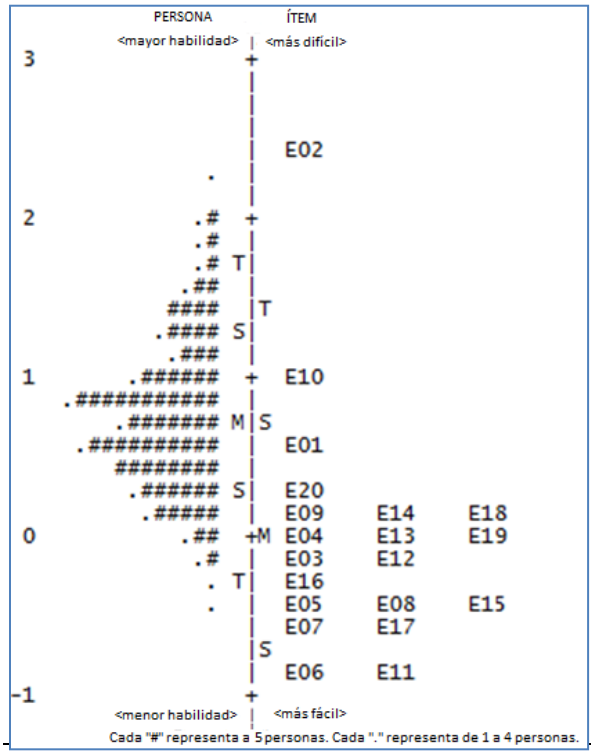


Figura D.15. Correlación punto biserial de cada ítem del área de Español de VA y VB. Índices de confiabilidad, Alpha de Cronbach: VA = 0.587, VB = 0.706.

VA



VB

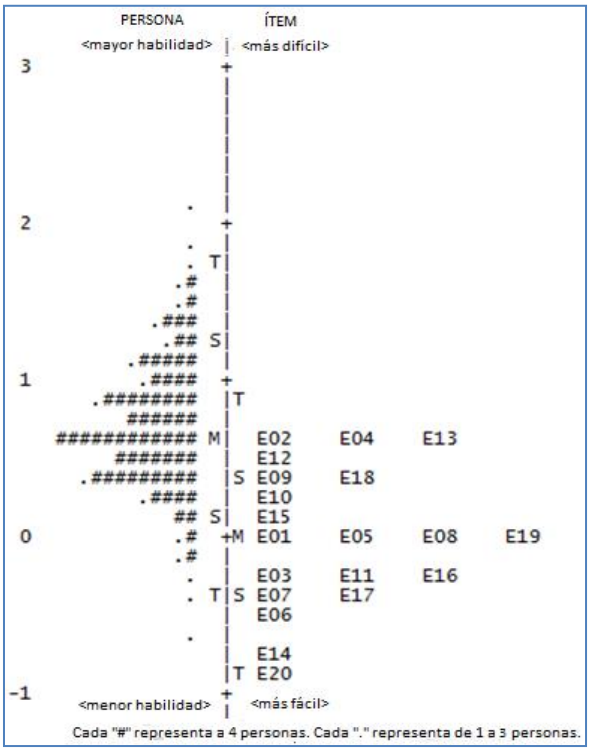


Figura D.16. Mapas de Wright para el área de Español, de VA y VB.

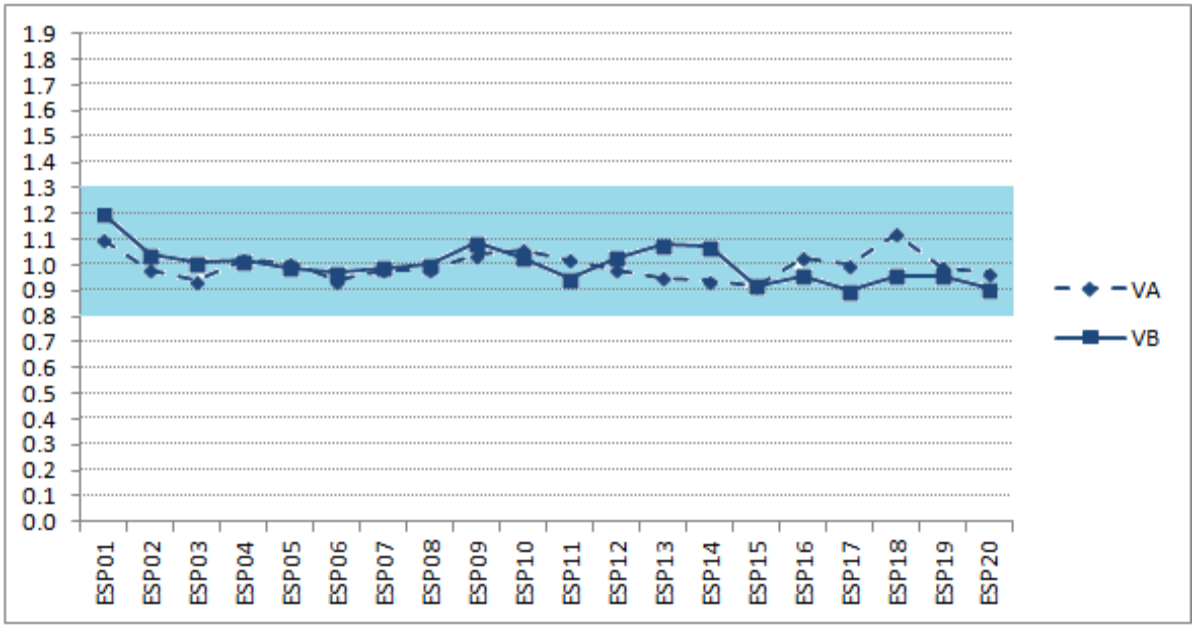


Figura D.17. Valores de *infit* para cada ítem del área de Español, de VA y VB.

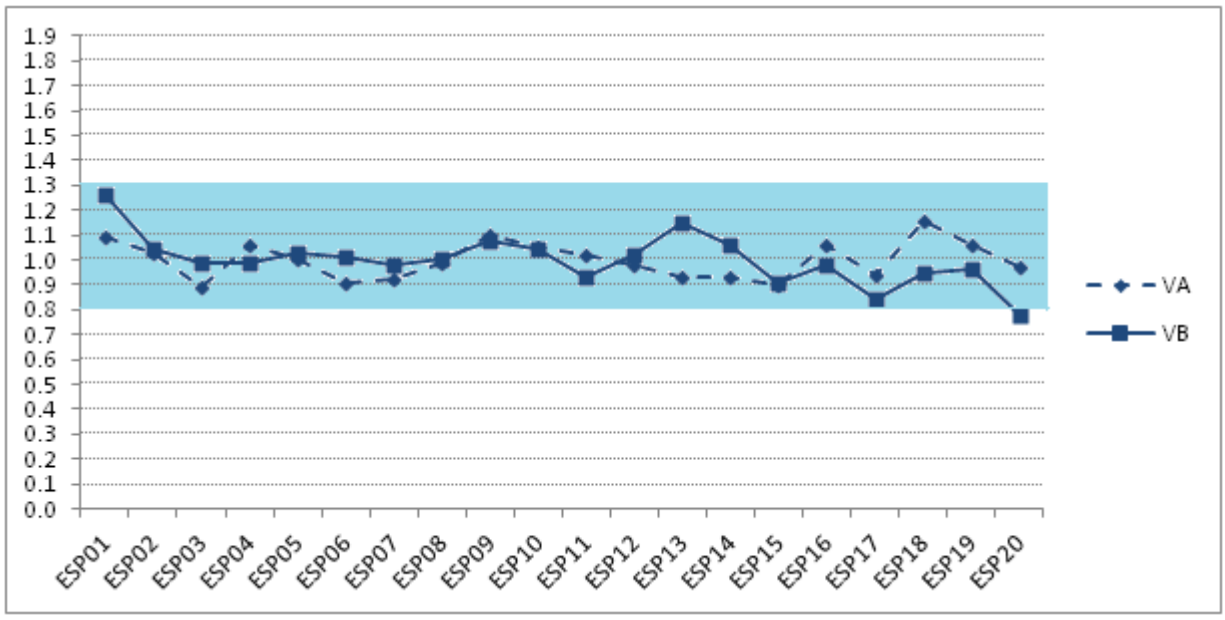


Figura D.18. Valores de *outfit* de cada ítem del área de Español de VA y VB.

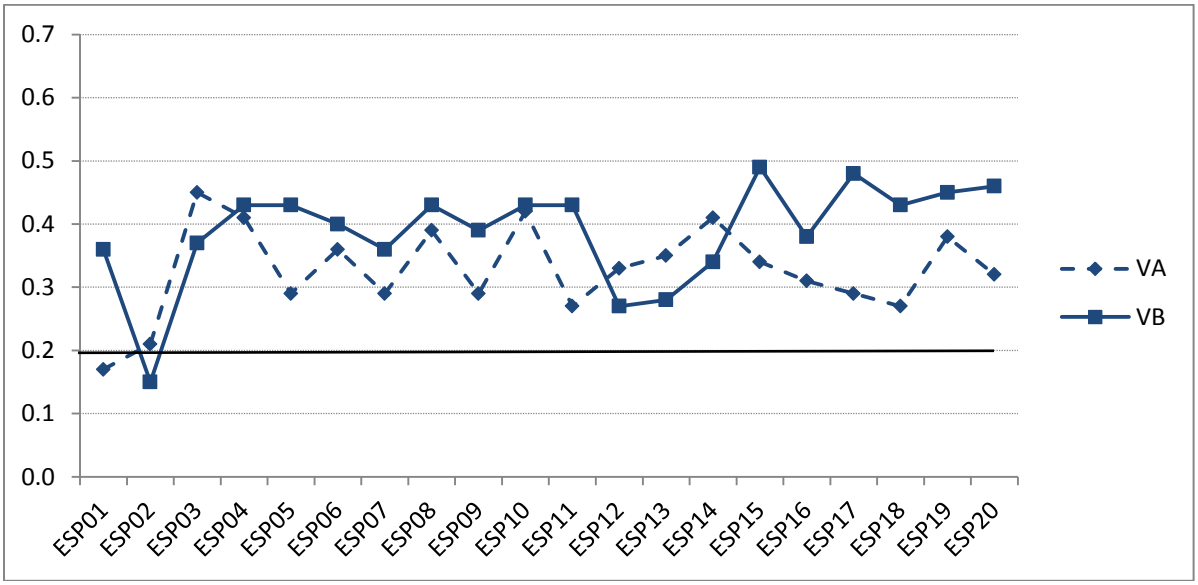


Figura D.19. Correlación punto medida para cada ítem del área de Español, de VA y VB.

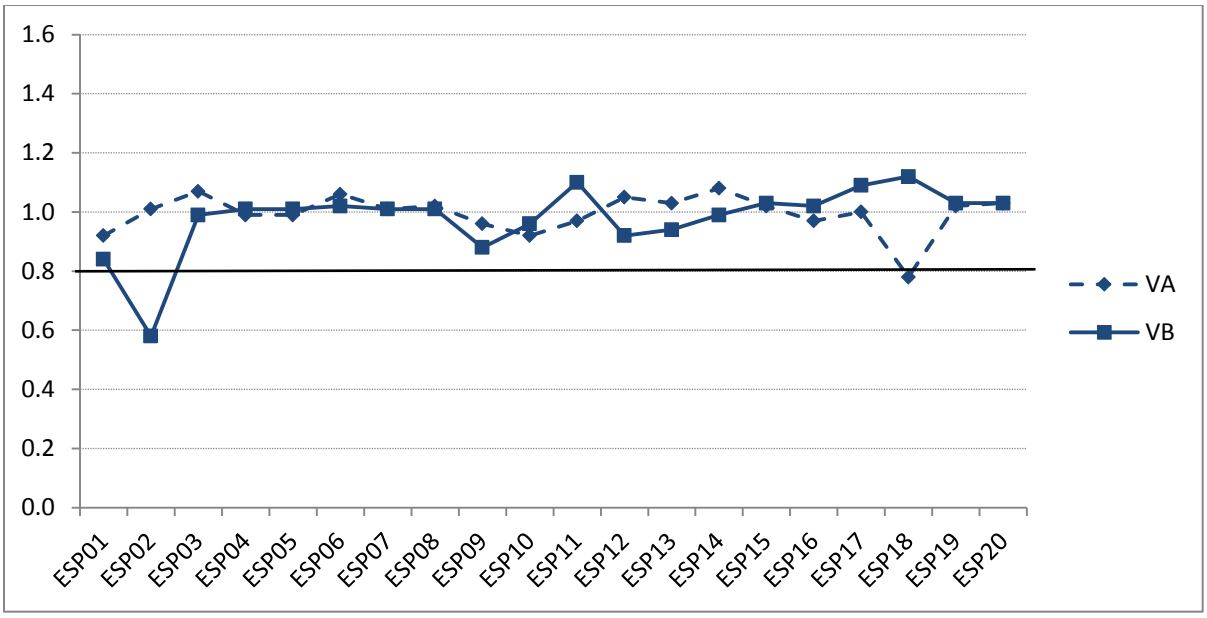


Figura D.20. Discriminación para cada ítem del área de Español, de VA y VB.

Tabla D.2.
Área de Español. Índices de ajuste de los AFC para VA y VB

Índices	Versión A	Versión B
Chi cuadrado	168.494	173.730
Grados libertad	164	165
p	.388	.305
NNFI	.976	.975
CFI	.979	.978
RMSEA	.010	.015

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*.

ítem	VA	VB
ESP01	**.091**	.253
ESP02	**.093**	**.023**
ESP03	.302	.319
ESP04	.322	.392
ESP05	.310	.396
ESP06	.414	.371
ESP07	.249	.339
ESP08	.411	.408
ESP09	.320	.326
ESP10	.223	.319
ESP11	.298	.366
ESP12	.304	.196
ESP13	.322	.226
ESP14	.308	.371
ESP15	.267	.572
ESP16	.206	.381
ESP17	.263	.511
ESP18	**.099**	.325
ESP19	.257	.496
ESP20	.293	.535

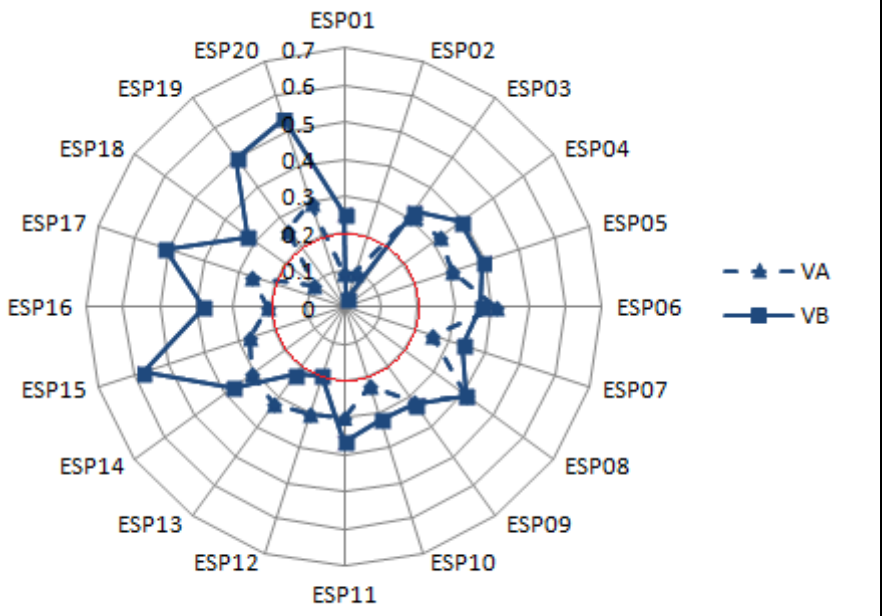


Figura D.21. Cargas factoriales estandarizadas del área de Español, de VA y V B. Modelo de un factor con errores que covarían.
Estadísticos significativos al nivel de 0.05; excepto para ESP01, ESP02 y ESP18 de VA, y ESP02 de VB.

3. Matemáticas

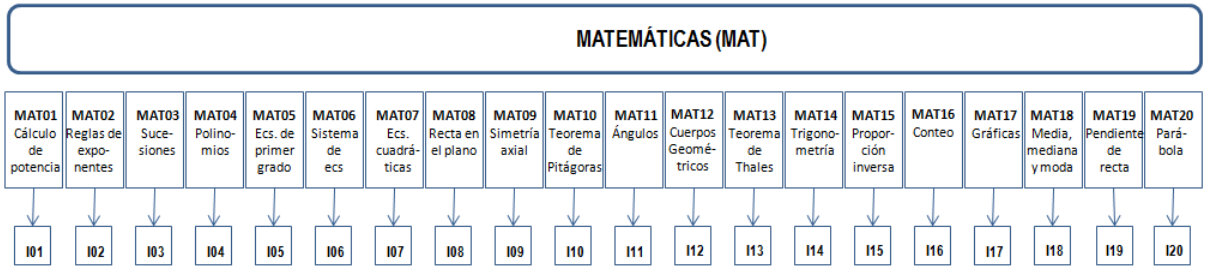
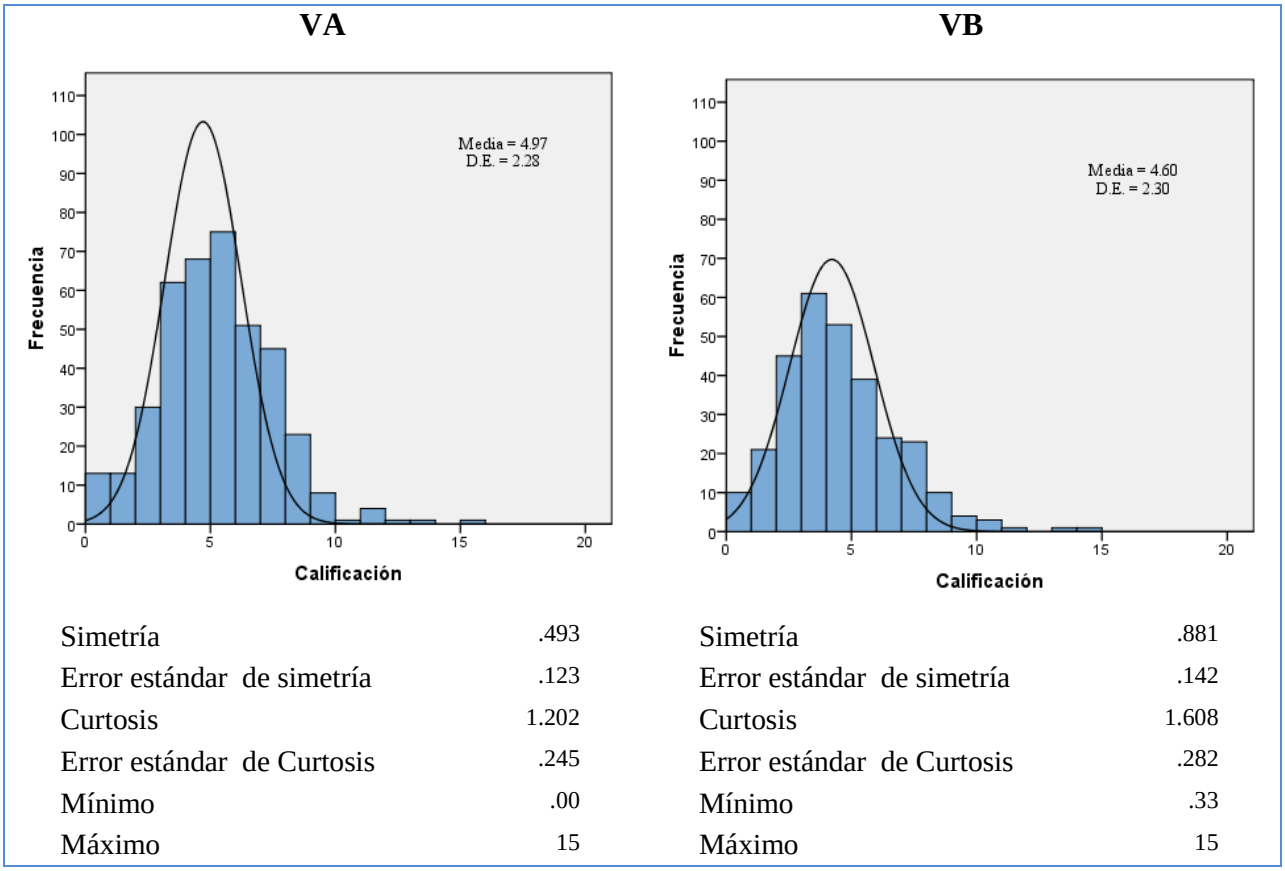


Figura D.22. Esquema del área de Matemáticas de VA y VB



Simetría	.493	Simetría	.881
Error estándar de simetría	.123	Error estándar de simetría	.142
Curtosis	1.202	Curtosis	1.608
Error estándar de Curtosis	.245	Error estándar de Curtosis	.282
Mínimo	.00	Mínimo	.33
Máximo	15	Máximo	15

Figura D.23. Distribución de las calificaciones del área de Matemáticas de VA y VB.

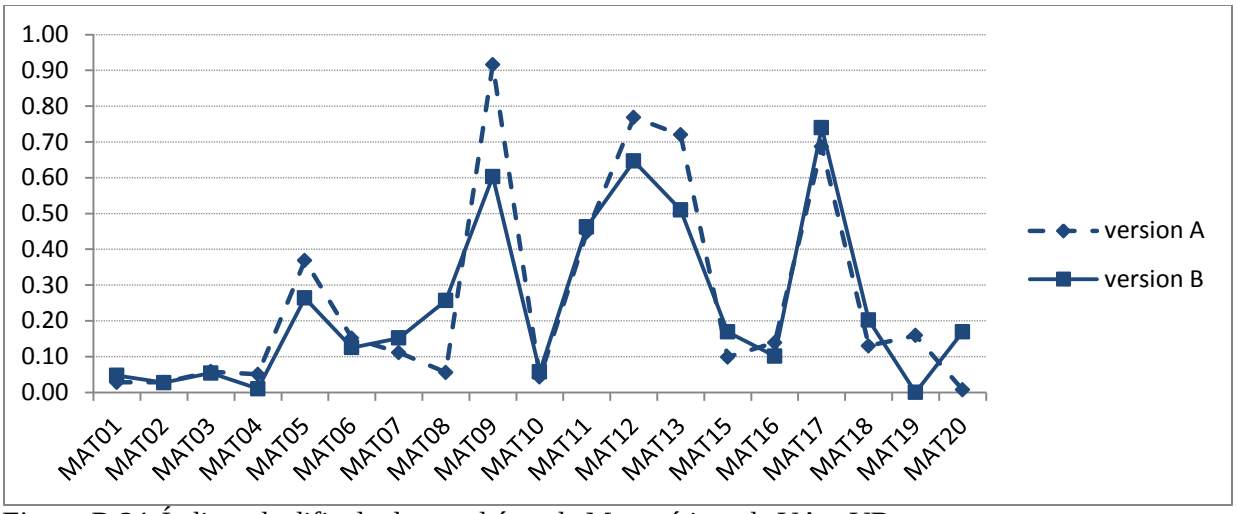


Figura D.24. Índices de dificultad para el área de Matemáticas de VA y VB.

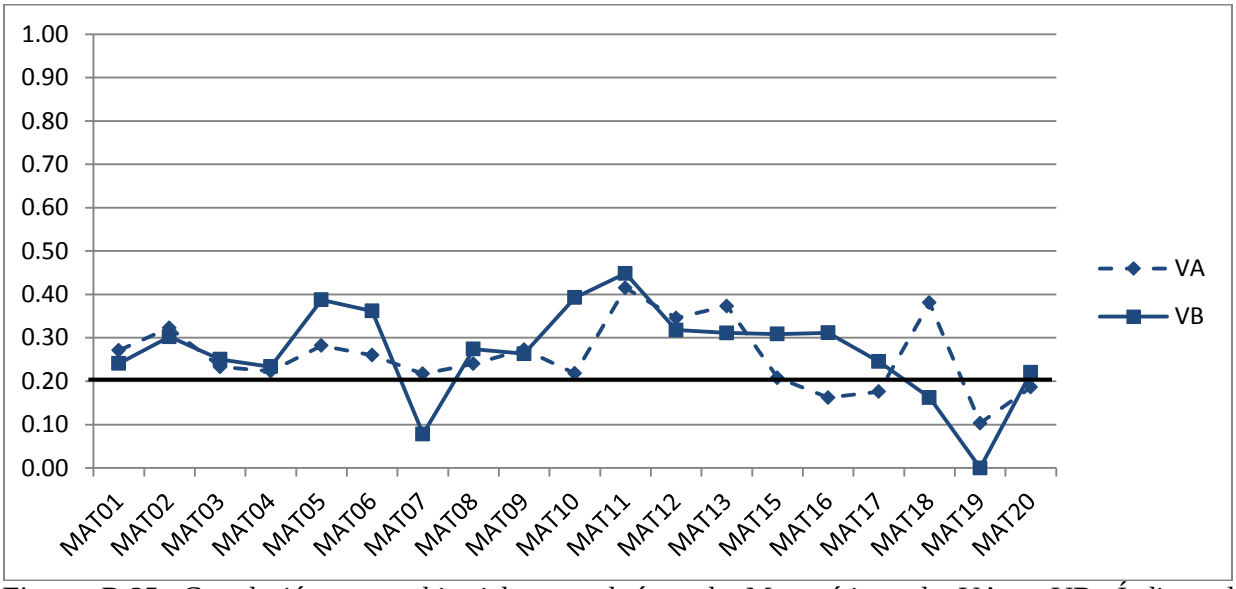
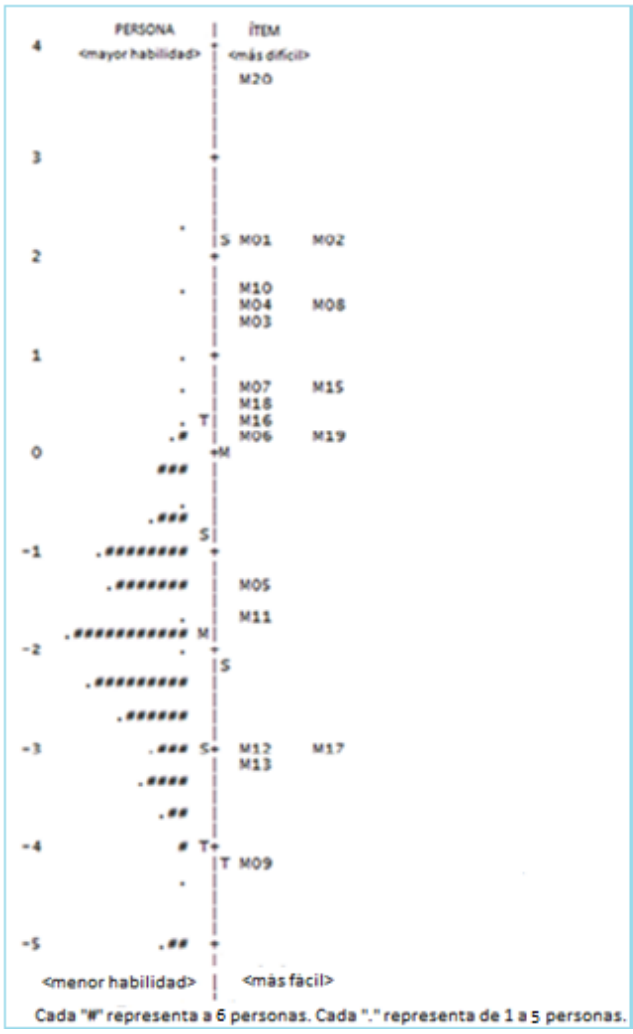


Figura D.25. Correlación punto biserial para el área de Matemáticas de VA y VB. Índices de confiabilidad, alpha de Cronbach: VA = 0.655, VB = 0.686.

VA



VB

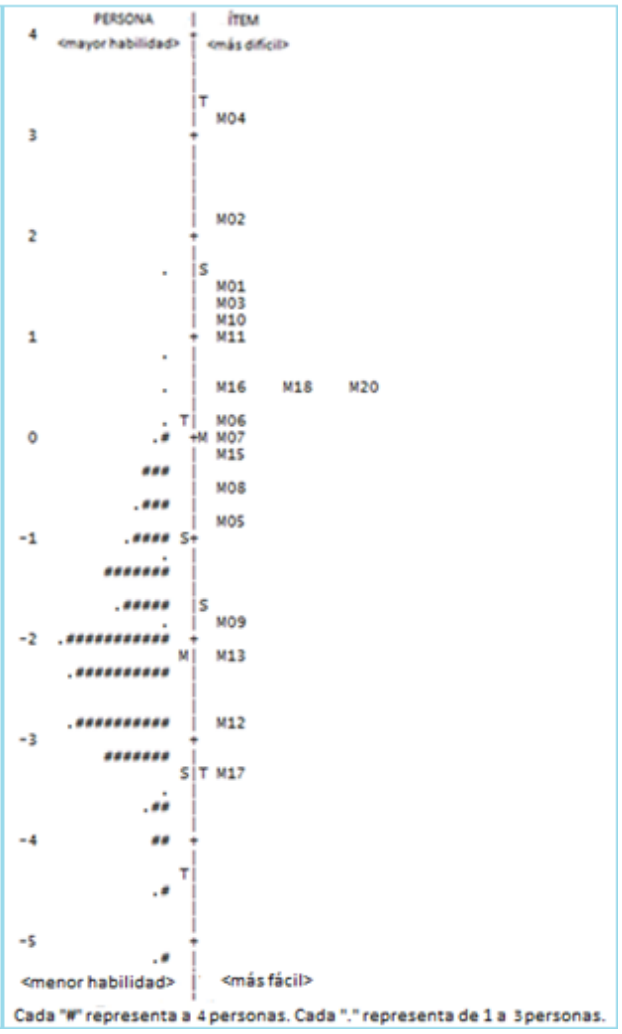


Figura D.26. Mapas de Wright para el área de Matemáticas, de VA y VB.

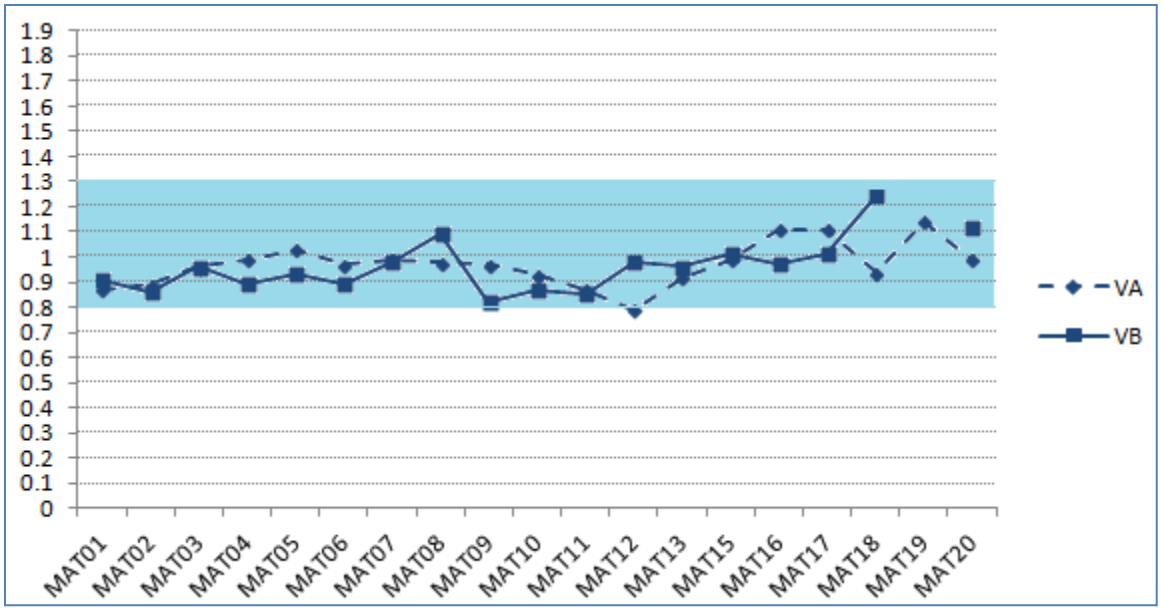


Figura D.27. Valores de *infit* para cada ítem del área de Matemáticas en VA y VB.

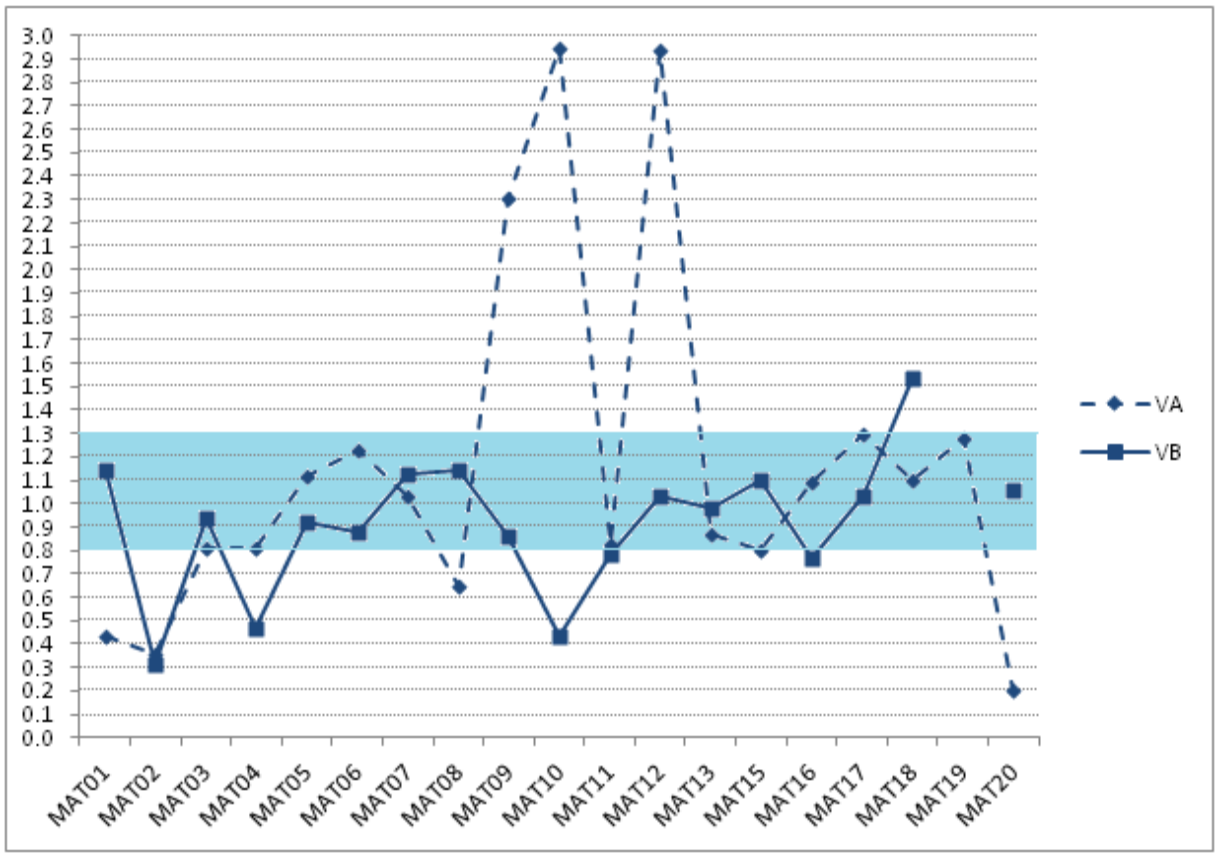


Figura D.28. Valores de *outfit* de cada ítem del área de Matemáticas de VA y VB.

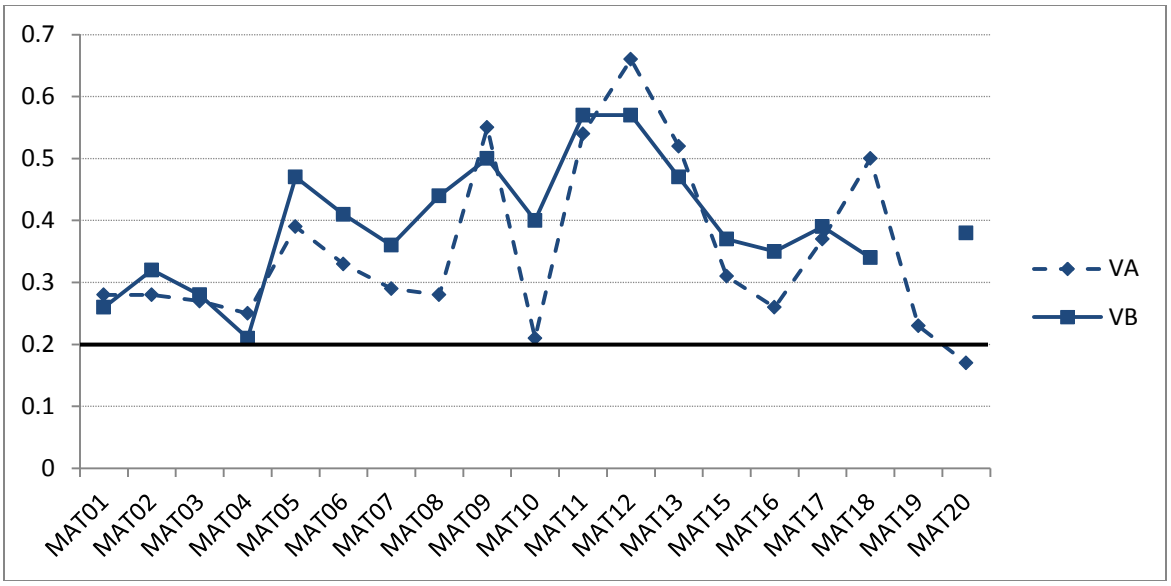


Figura D.29. Correlación punto medida para el área de Matemáticas en VA y VB.

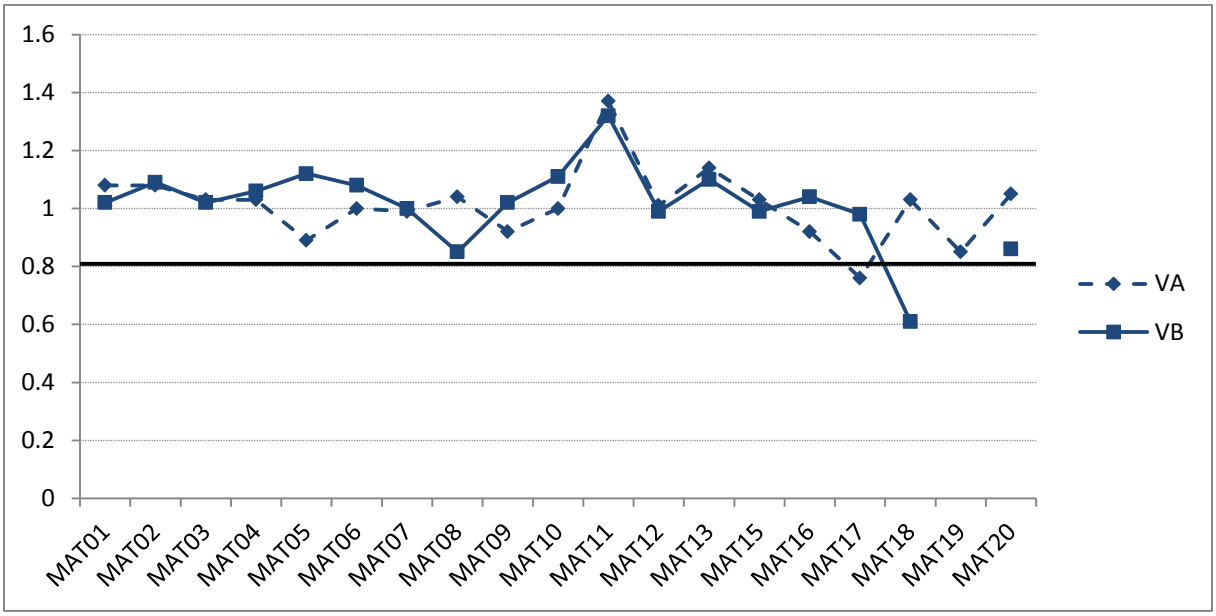


Figura D.30. Discriminación para el área de Matemáticas en VA y VB.

Tabla D.3.
Área de Matemáticas. Índices de ajuste de los AFC para VA y VB, modelos unidimensional, dos y tres factores

Índices	Versión A			Versión B		
	Modelo 1	Modelo 2	Modelo 3	Modelo 1	Modelo 2	Modelo 3
Chi cuadrado	233.909	239.456	146.11	166.394	184.648	106.598
Grados libertad	137	135	123	128	127	85
p	**.000**	**.000**	.076	**.012**	**.000**	.056
NNFI	**0.836**	**0.820**	0.956	**.899**	**0.847**	0.914
CFI	**.868**	**.858**	.969	.915	**.880**	.952
RMSEA	.042	.044	.022	.032	.039	.029
Covarianzas						
F1-F2		.927	.519		.826	.407
F1-F3		.928			.611	
F2-F3		.853			.838	

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. Alpha de Cronbach: VA = .655, VB = .707. Modelo 1: unidimensional. Modelo 2: tres factores que covarían. Modelo 3: dos factores que covarían.

Ítems	VA	VB
MAT01	.356	.345
MAT02	.447	.411
MAT03	.307	.563
MAT04	.298	.302
MAT05	.361	.599
MAT06	.396	.700
MAT07	.298	**.089**
MAT08	.367	.429
MAT09	.353	.494
MAT10	.790	1.000
MAT11	.902	.964
MAT12	.484	.591
MAT13	.651	.624
MAT14	S/D	S/D
MAT15	.284	.385
MAT16	**.170**	.492
MAT17	.219	.281
MAT18	.445	.296
MAT19	**.113**	---[a]
MAT20	.256	.374

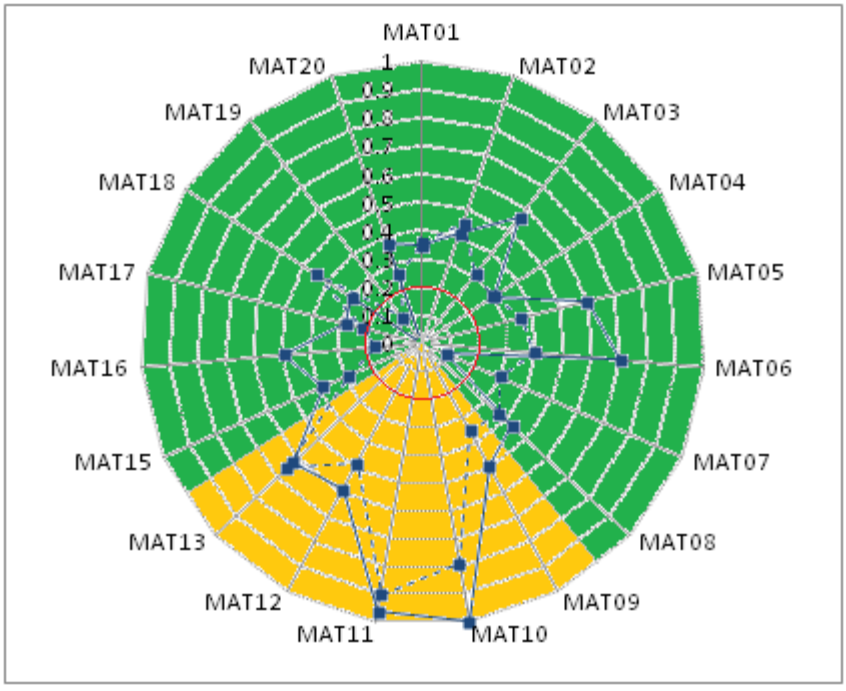


Figura D.31. Cargas factoriales estandarizadas para Matemáticas de VA y VB. Modelo 3: dos factores que covarían. S/D = sin datos. [a] No se pudo calcular debido a que no hubo respuestas correctas. Estadísticos significativos al nivel de 0.05; excepto para MAT19 de VA y MAT07 de VB.

4. Ciencias naturales

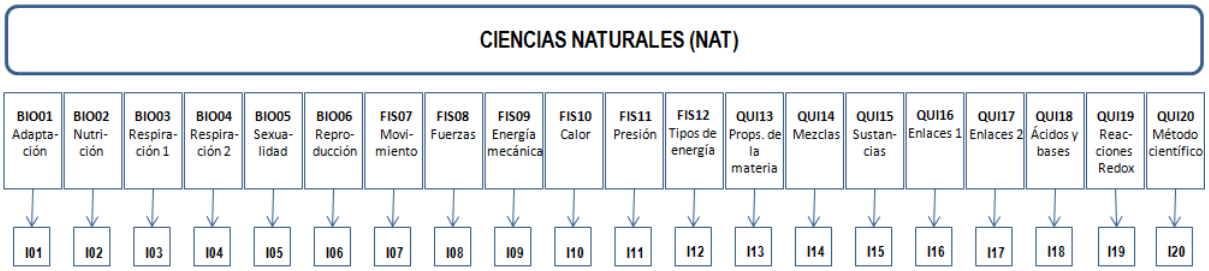
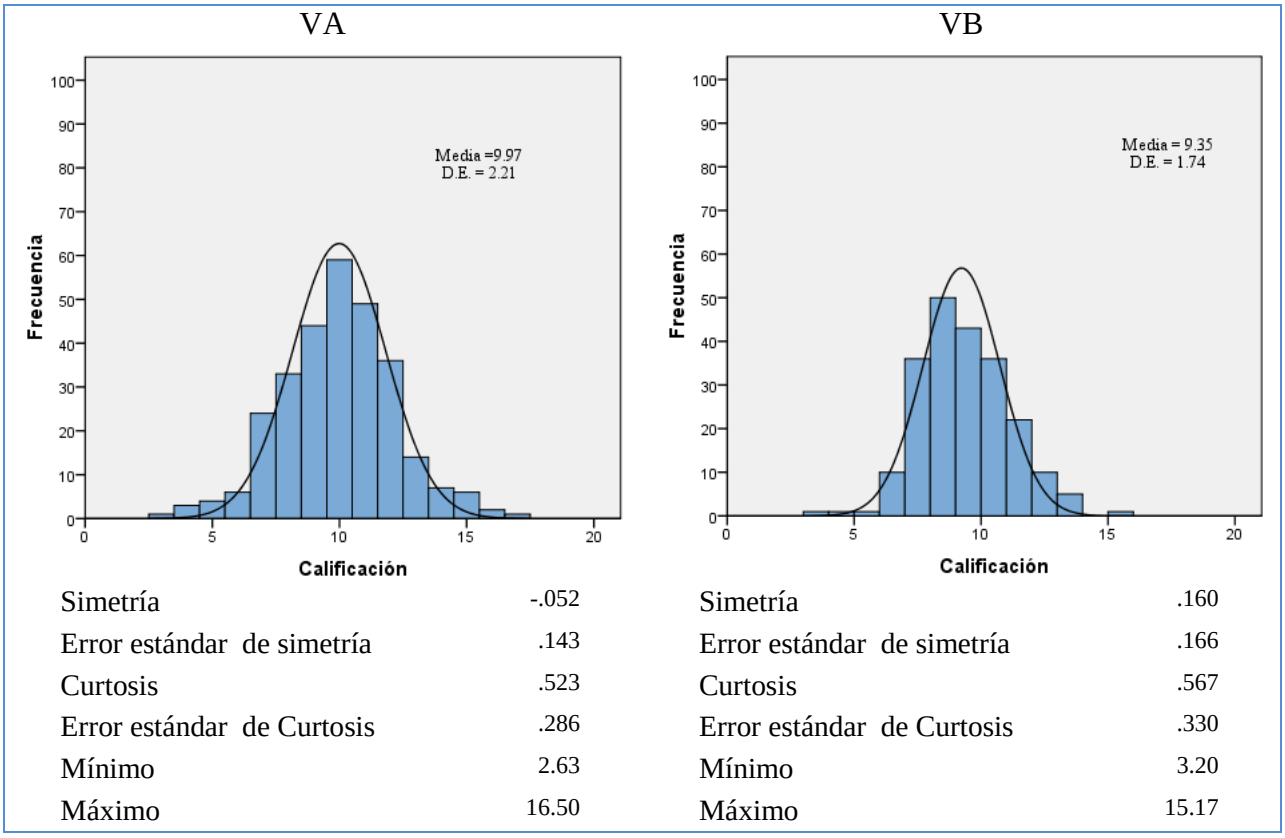


Figura D.32. Esquema del área de Ciencias naturales de VA y VB



Simetría	-.052	Simetría	.160
Error estándar de simetría	.143	Error estándar de simetría	.166
Curtosis	.523	Curtosis	.567
Error estándar de Curtosis	.286	Error estándar de Curtosis	.330
Mínimo	2.63	Mínimo	3.20
Máximo	16.50	Máximo	15.17

Figura D.33. Distribución de las calificaciones del área de Ciencias naturales de VA y VB

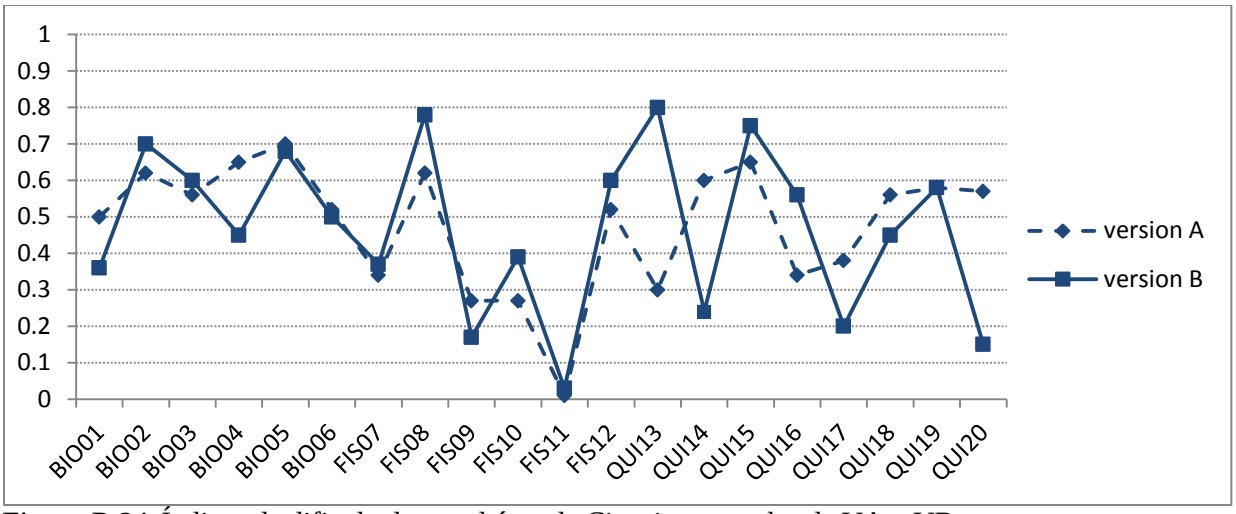


Figura D.34. Índices de dificultad para el área de Ciencias naturales de VA y VB.

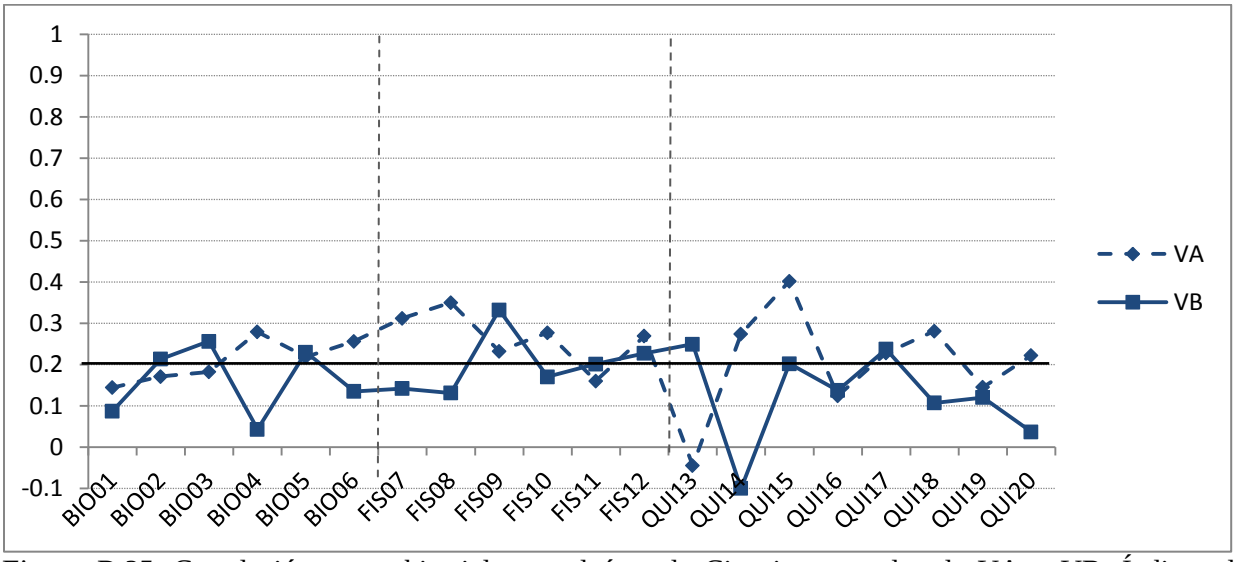


Figura D.35. Correlación punto biserial para el área de Ciencias naturales de VA y VB. Índices de confiabilidad, Alpha de Cronbach: VA = 0.612, VB = 0.502.

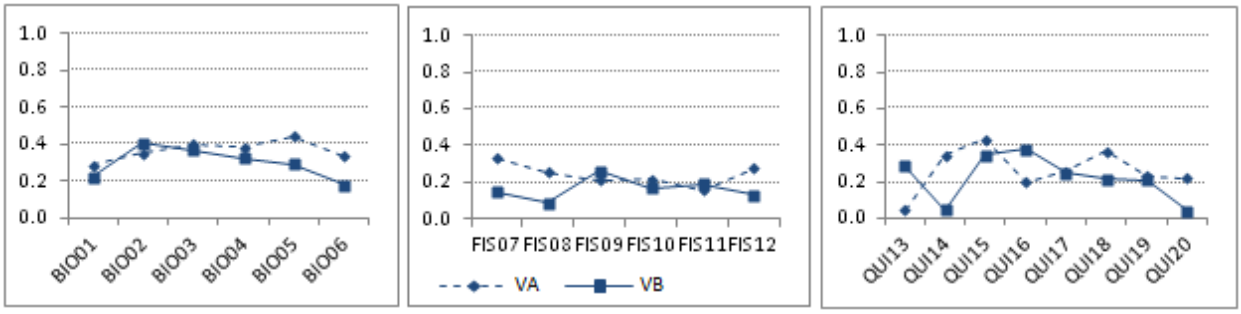
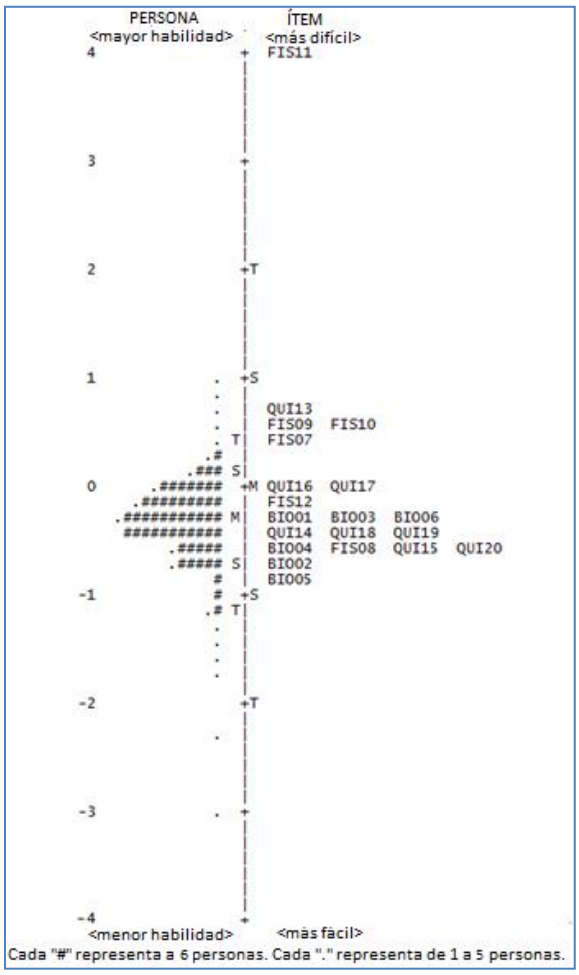


Figura .D.36. Gráficas de correlaciones punto biserial por materias del área de Ciencias naturales de VA y VB.

VA



VB

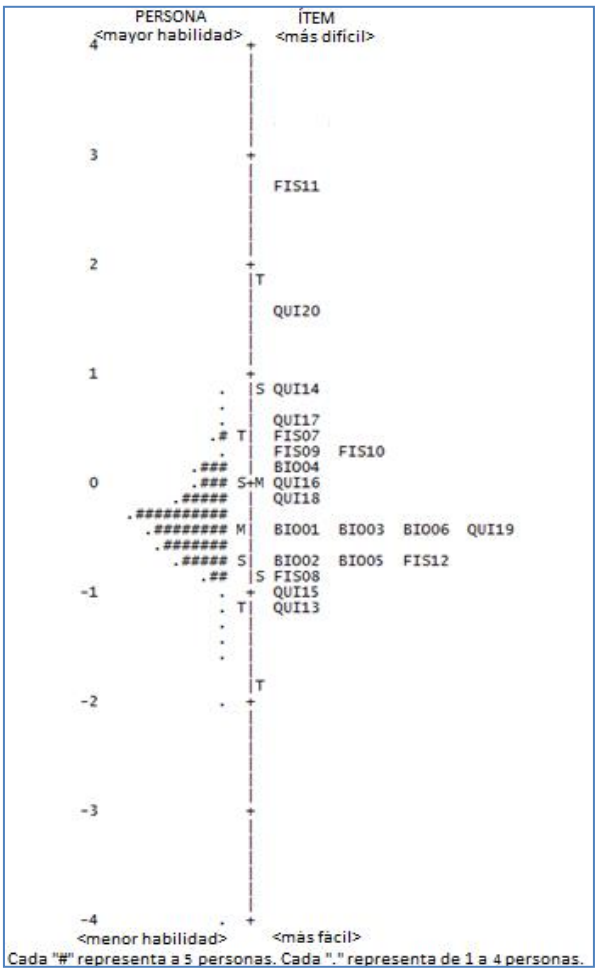


Figura D.37. Mapas de Wright de VA y VB para el área Ciencias naturales.

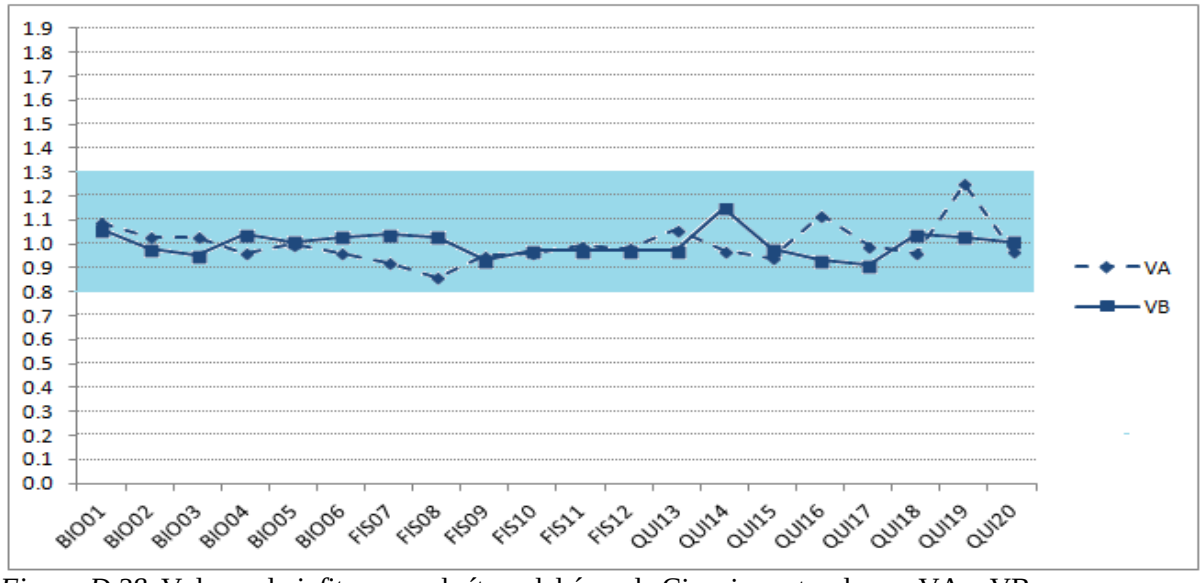


Figura D.38. Valores de *infit* para cada ítem del área de Ciencias naturales en VA y VB.

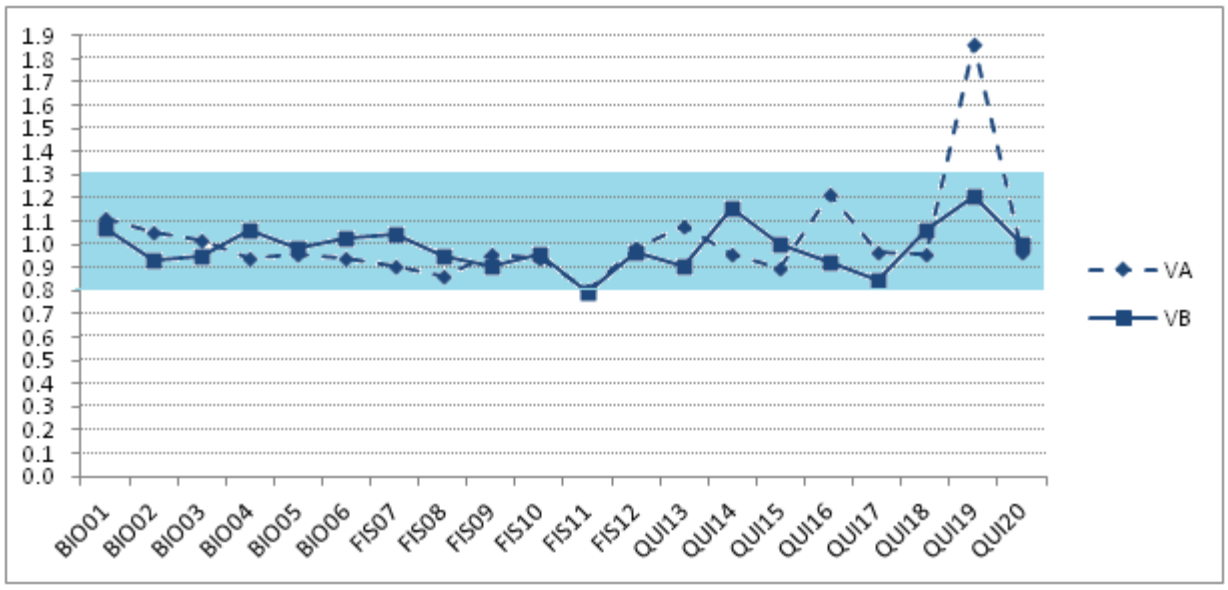


Figura D.39. Valores de *outfit* de cada ítem del área de Ciencias naturales de VA y VB.

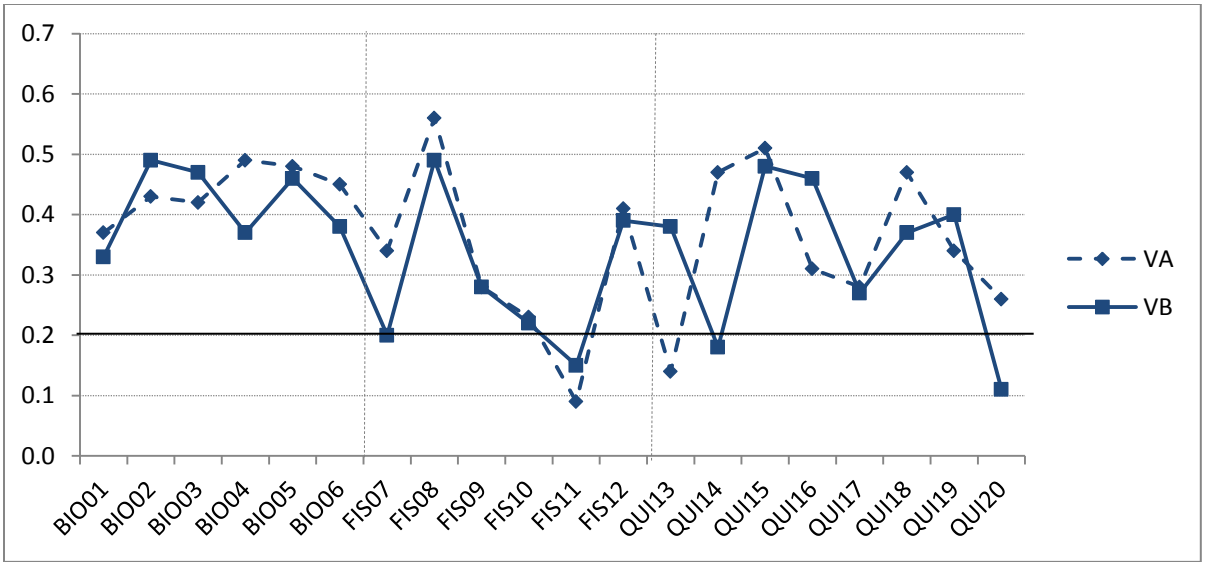


Figura D.40. Correlación punto medida para el área de Ciencias naturales en VA y VB.

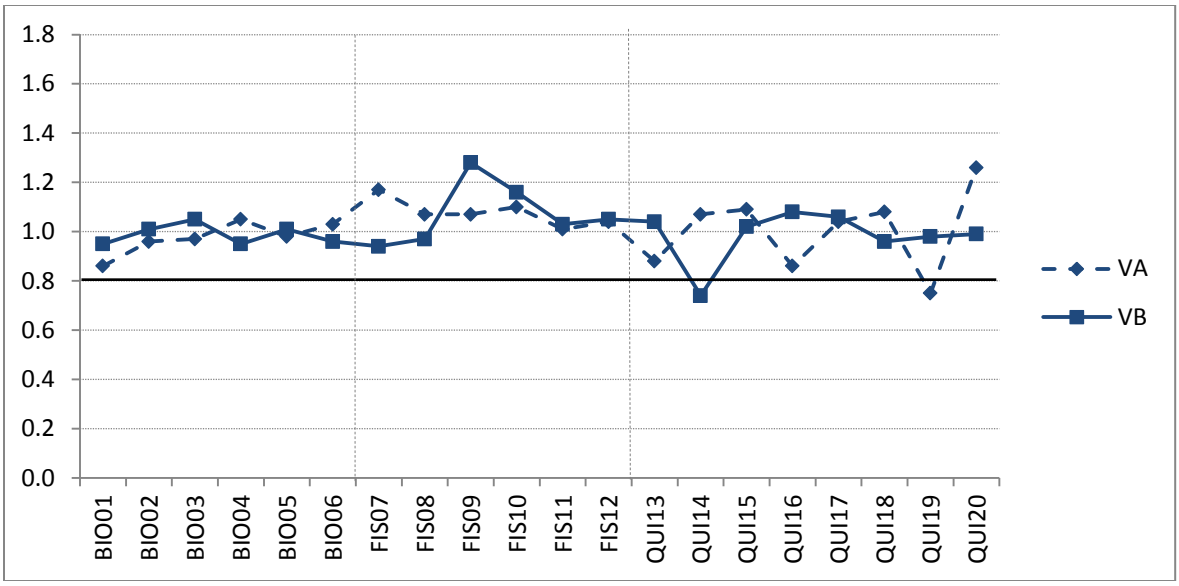


Figura D.41. Discriminación para el área de Ciencias Naturales en VA y VB.

Tabla D.4.

Área de Ciencias naturales. Índices de ajuste de los AFC para VA y VB, modelos unidimensional, dos y tres factores

Parámetros	Versión A			Versión B		
	Modelo 1	Modelo 2	Modelo 3	Modelo 1	Modelo 2	Modelo 3
Chi cuadrado	148.594	146.776	No converge	201.028	480.961	200.807
Grados libertad	165	163	--	164	157	162
p	.815	.814	--	.025	.092	.020
NNFI	1.069	1.069	--	**0.717**	**0.808**	**0.699**
CFI	1.000	1.000	--	**.755**	**.842**	**.744**
RMSEA	.000	.000	--	.032	.027	.033
Covarianzas						
F1-F2		.885			.929	.592
F1-F3						.548
F2-F3						.643

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. Modelo 1: Un factor. Modelo 2: Dos factores que covarían. Modleo 3: Tres factores que covarían.

Ítem	VA	VB
BIO01	**.192**	**.070**
BIO02	.229	.526
BIO03	.234	.370
BIO04	.336	**.125**
BIO05	.302	.308
BIO06	.324	.117
FIS07	.479	.194
FIS08	.477	.271
FIS09	.322	.425
FIS10	.322	.115
FIS11	.231	.149
FIS12	.473	.377
QUI13	**-.069**	.387
QUI14	.319	**-.115**
QUI15	.521	.430
QUI16	**.130**	.200
QUI17	.254	.412
QUI18	.340	**.159**
QUI19	**.149**	**.093**
QUI20	.273	**.013**

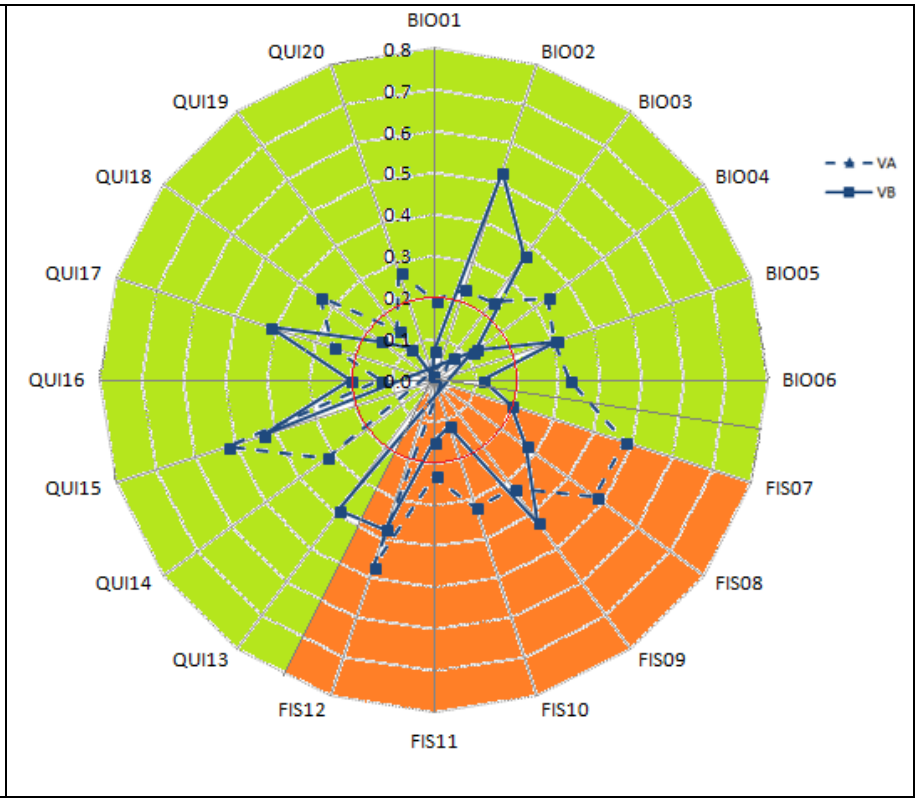


Figura D.42. Cargas factoriales estandarizadas para Ciencias naturales de VA y VB. Modelo 2: dos factores que covarían y covarianzas de errores.
Estadísticos significativos de VA al nivel de 0.05; excepto para las cargas de QUI13, QUI16 y QUI19. En VB, los estadísticos significativos al nivel de 0.05 se presentan únicamente en las cargas de FIS09 y FIS12.

5. Ciencias Sociales

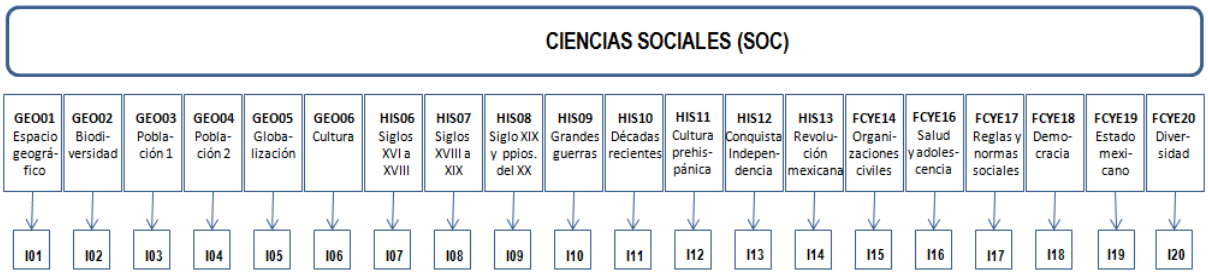
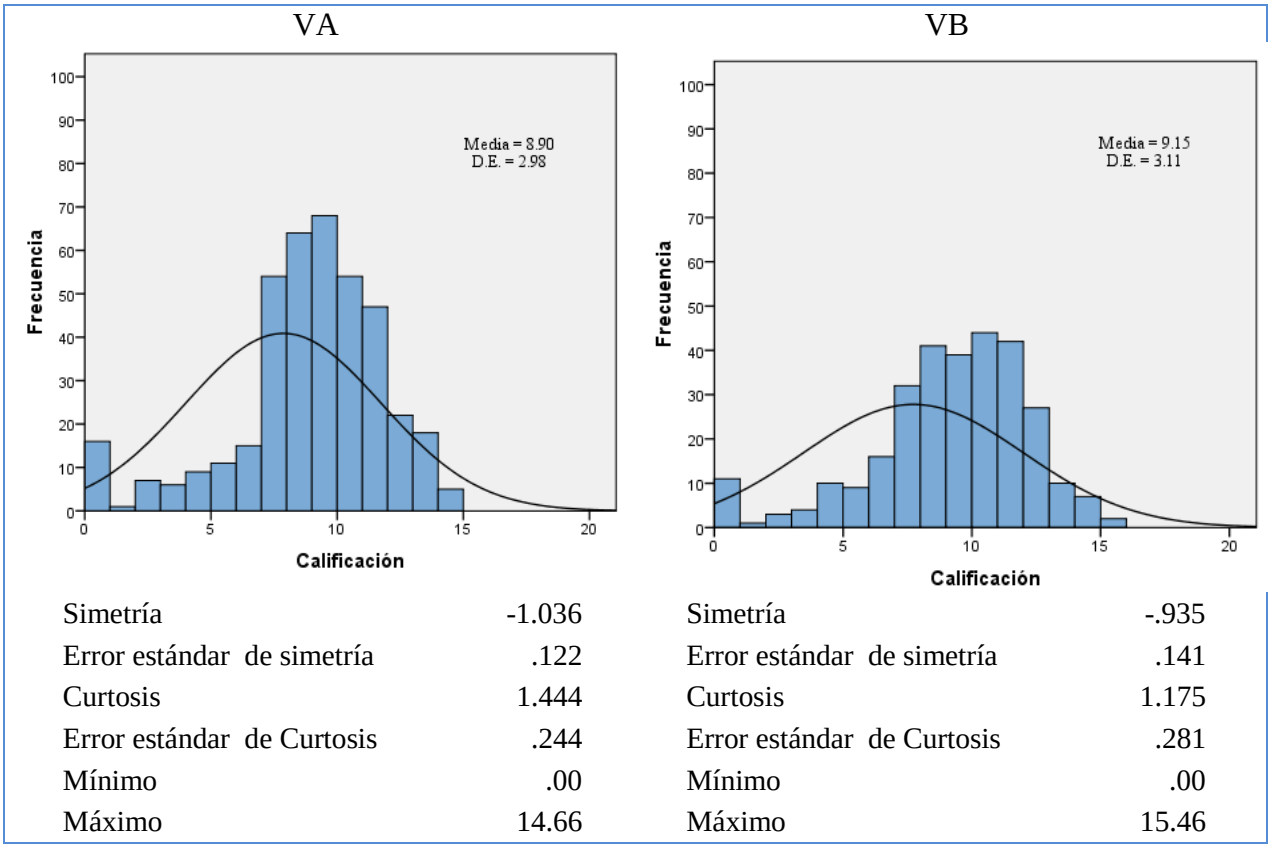


Figura D.43. Esquema del área de Ciencias sociales de VA y VB.



Simetría	-1.036	Simetría	-.935
Error estándar de simetría	.122	Error estándar de simetría	.141
Curtosis	1.444	Curtosis	1.175
Error estándar de Curtosis	.244	Error estándar de Curtosis	.281
Mínimo	.00	Mínimo	.00
Máximo	14.66	Máximo	15.46

Figura D.44. Distribución de las calificaciones del área de Ciencias sociales de VA y VB.

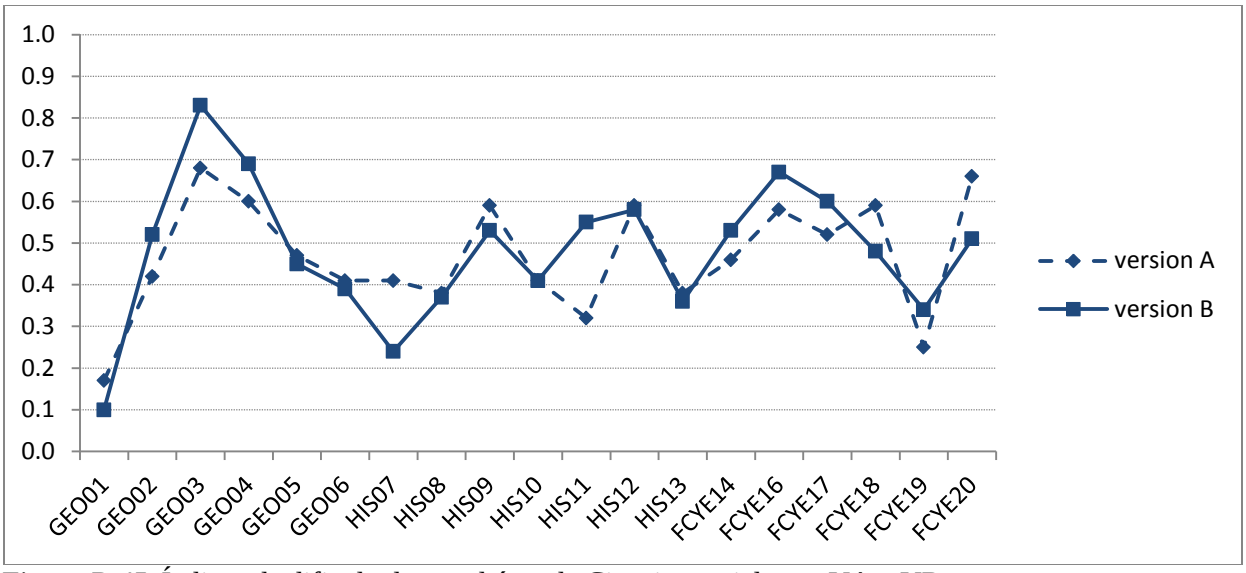


Figura D.45. Índices de dificultad para el área de Ciencias sociales en VA y VB.

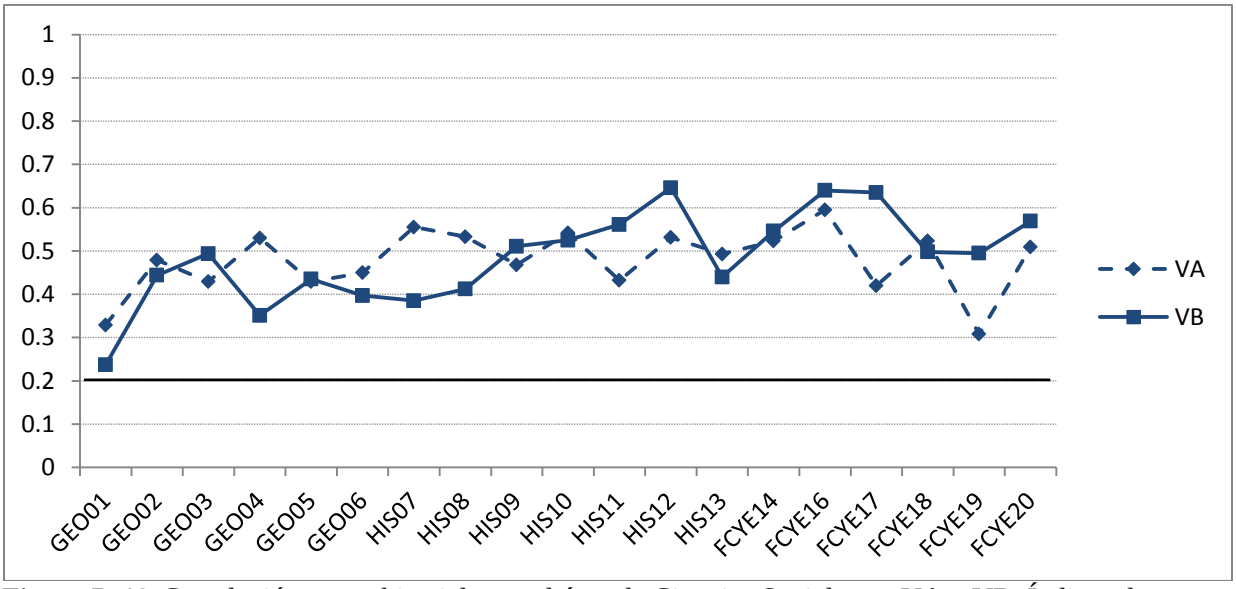


Figura D.46. Correlación punto biserial para el área de Ciencias Sociales en VA y VB. Índices de confiabilidad, alpha de Cronbach: VA = 0.869, VB = 0.877.

VA VB

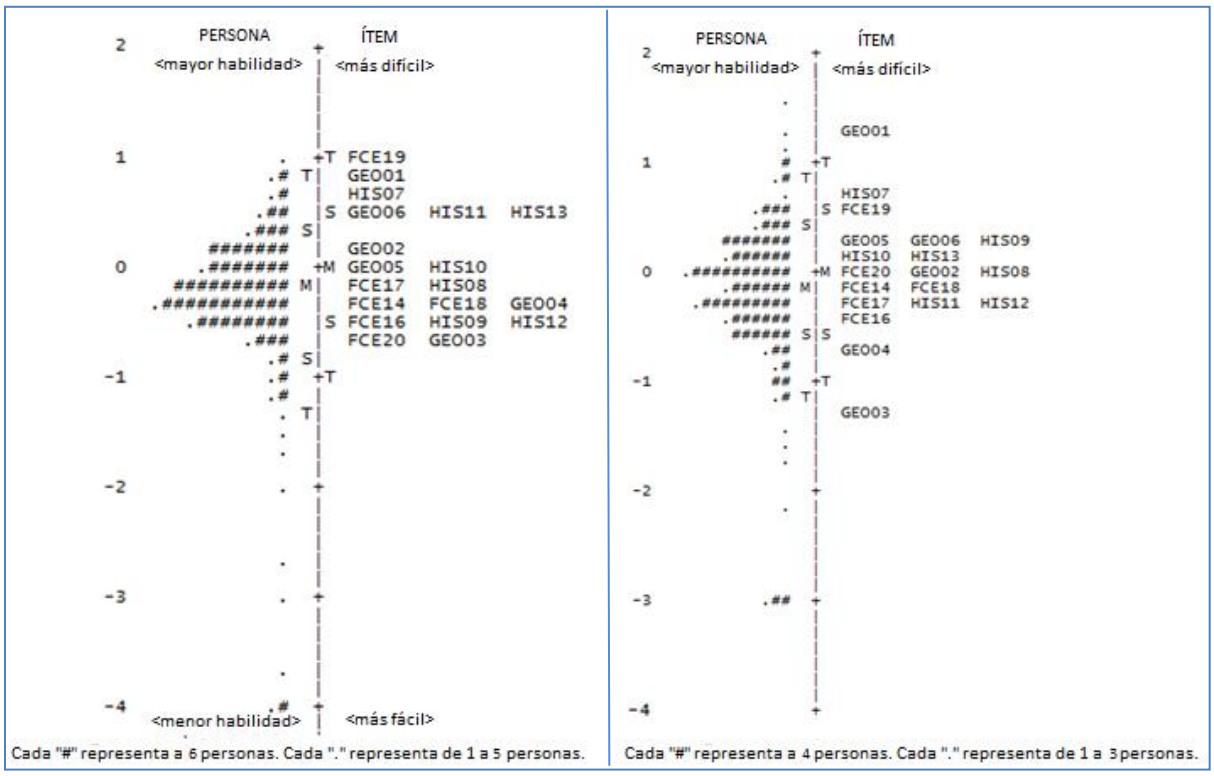


Figura D.47. Mapas de Wright de VA y VB para el área Ciencias sociales.

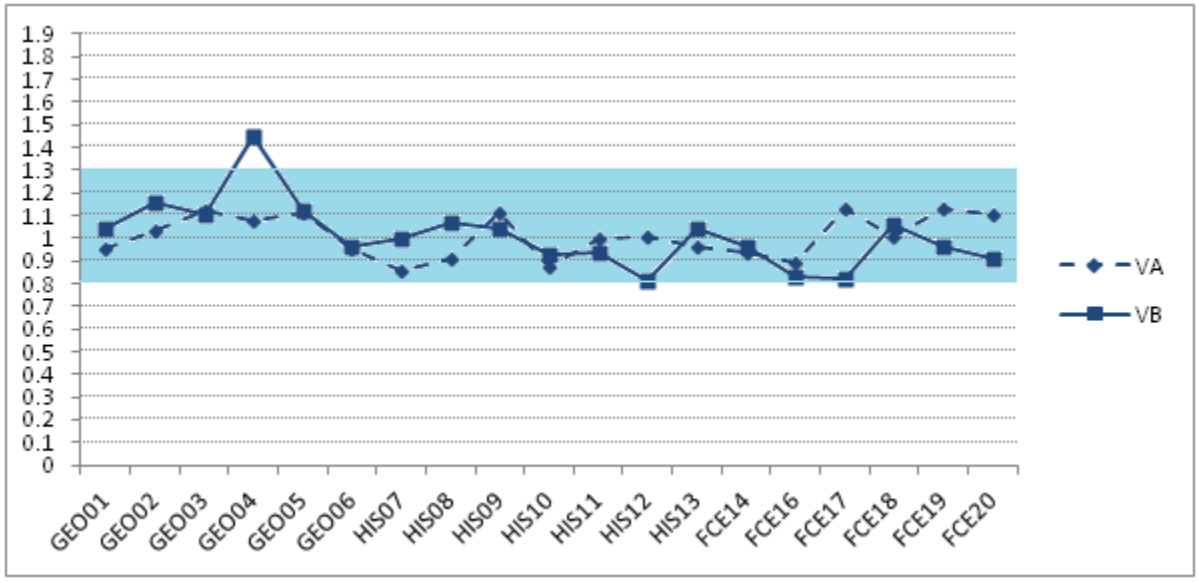


Figura D.48. Valores de *infit* para cada ítem del área de Ciencias Sociales en VA y VB.

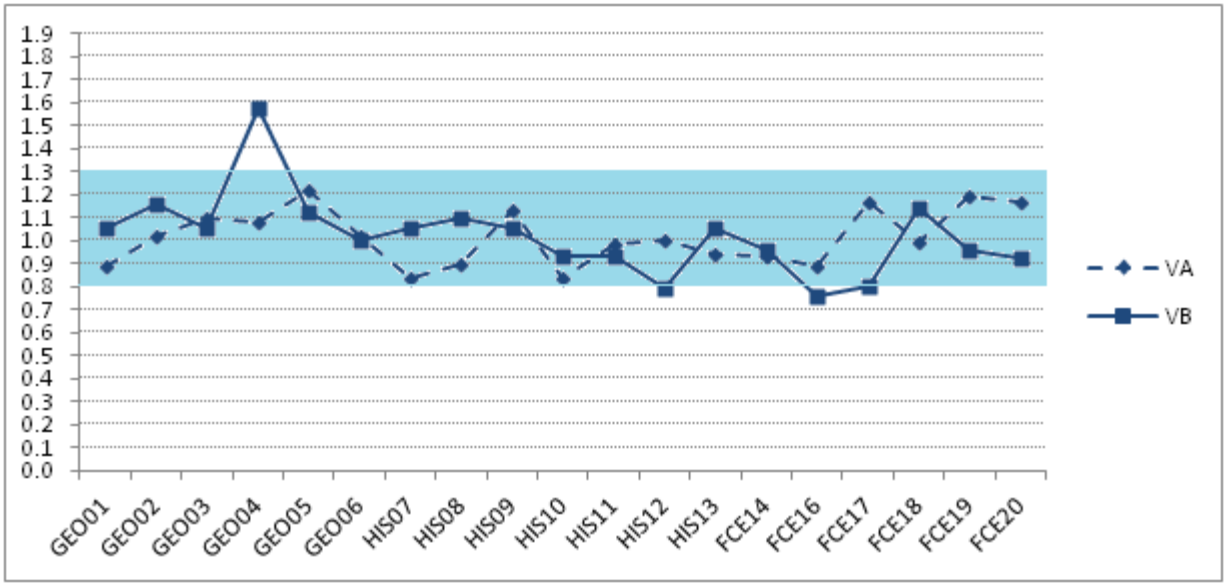


Figura D.49. Valores de *outfit* de cada ítem del área de Ciencias sociales de VA y VB.

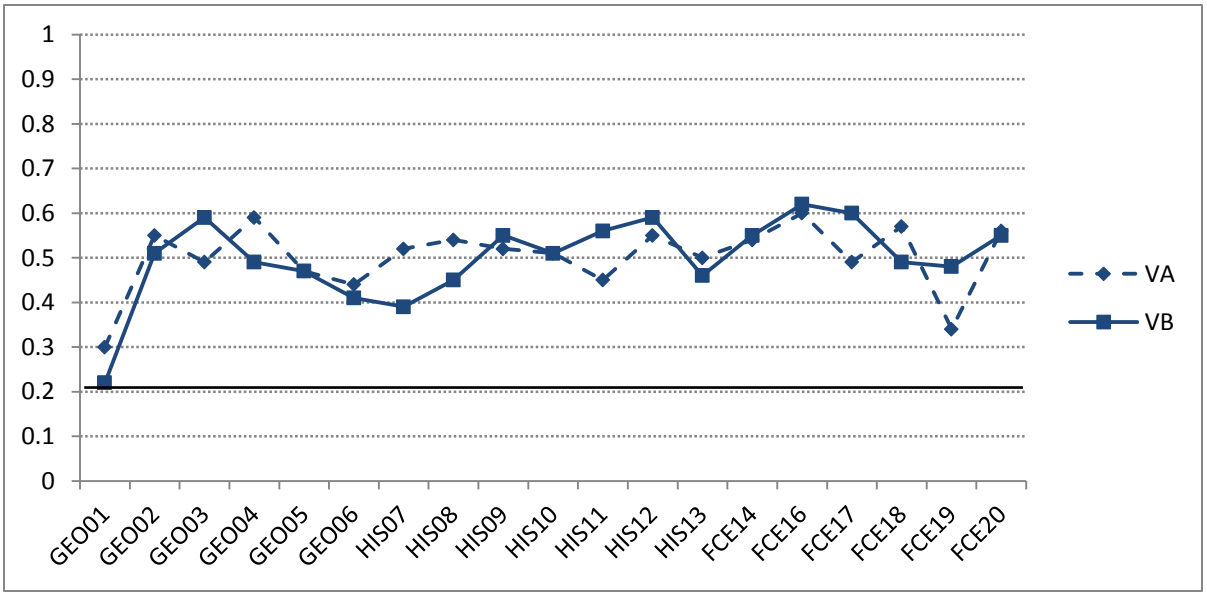


Figura 4.50. Correlación punto medida para el área de Ciencias Sociales en VA y VB.

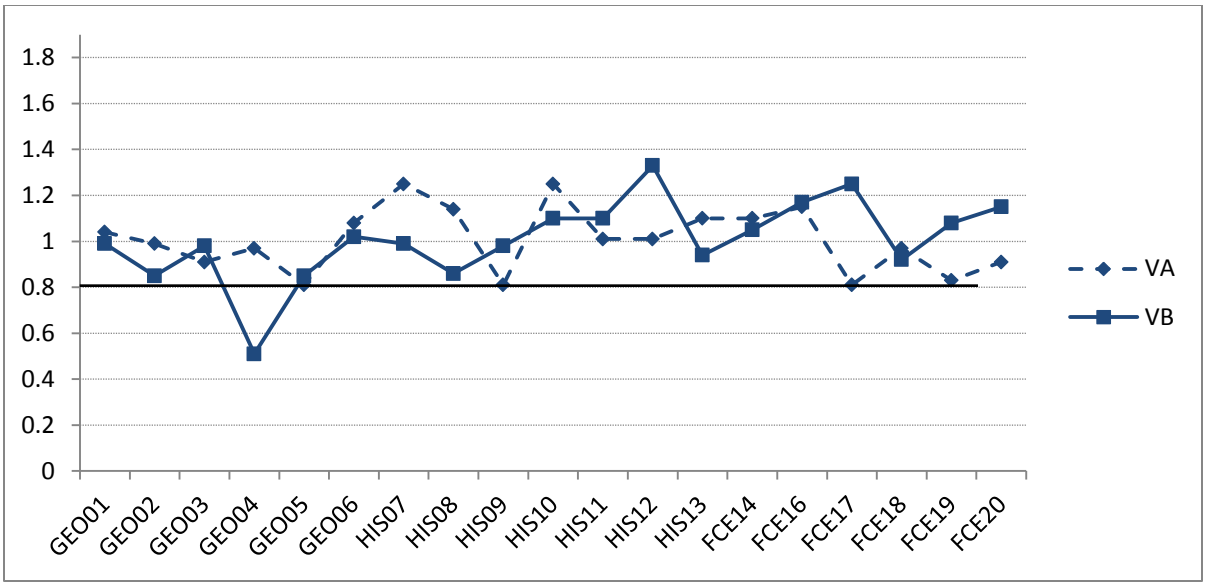


Figura 4.51. Discriminación para el área de Ciencias Sociales de VA y VB.

Tabla D.5.

Área de Ciencias Sociales. Índices de ajuste de los AFC para VA y VB, por modelo propuesto.

Índices	Versión A		Versión B	
	Modelo 1	Modelo 2	Modelo 1	Modelo 2
Chi cuadrado	132.262	185.590	144.458	209.144
Grados libertad	101	142	101	149
p	.020	**.001**	.003	**.001**
NNFI	0.970	0.970	0.948	0.951
CFI	.982	.975	.969	.957
RMSEA	.028	.028	.038	.037
Covarianzas				
F1-F2		.964		.792
F1-F3		.738		.643
F2-F3		.856		.873

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. Modelo 1: Un factor. Modelo 2: Tres factores que covarían.

Ítem	VA	VB
GEO01	.330	.237
GEO02	.510	.417
GEO03	.444	.462
GEO04	.551	.326
GEO05	.456	.446
GEO06	.488	.416
HIS07	.645	.448
HIS08	.586	.476
HIS09	.568	.589
HIS10	.648	.585
HIS11	.517	.655
HIS12	.613	.737
HIS13	.564	.515
FCYE14	.516	.539
FCYE16	.550	.619
FCYE17	.364	.596
FCYE18	.466	.482
FCYE19	.294	.508
FCYE20	.438	.551

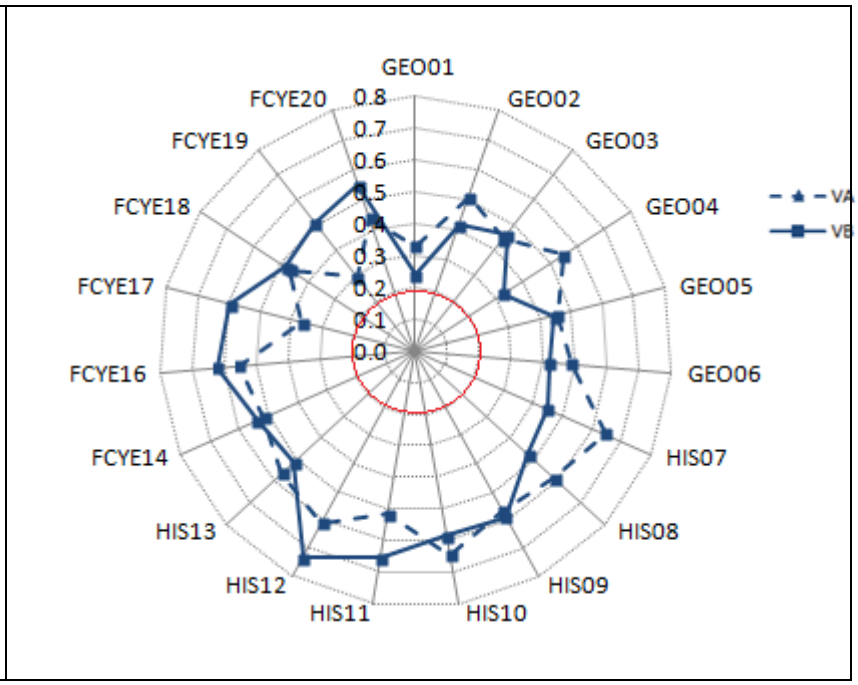


Figura 4.52. Cargas factoriales estandarizadas para Ciencias sociales de VA y VB. Modelo 1: unidimensional con errores que covarían.
Estadísticos significativos al nivel de 0.05.

Ítem	VA	VB
GEO01	.371	.272
GEO02	.568	.566
GEO03	.505	.651
GEO04	.619	.486
GEO05	.479	.537
GEO06	.519	.443
HIS07	.594	.417
HIS08	.602	.449
HIS09	.518	.544
HIS10	.579	.568
HIS11	.479	.627
HIS12	.580	.712
HIS13	.539	.488
FCYE14	.627	.596
FCYE16	.661	.748
FCYE17	.464	.762
FCYE18	.604	.577
FCYE19	.355	.548
FCYE20	.539	.667

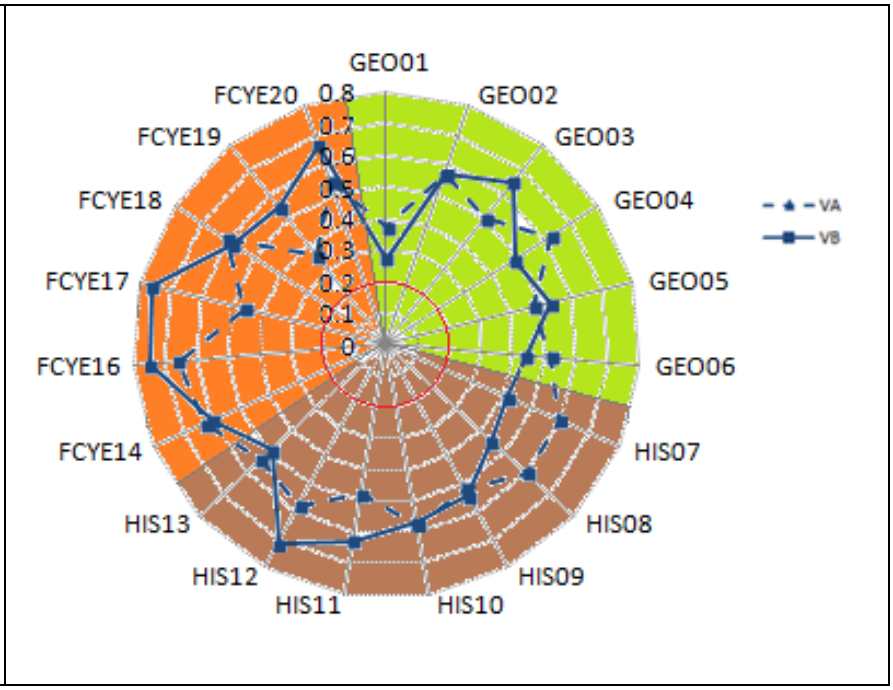


Figura 4.53. Cargas factoriales estandarizadas para Ciencias sociales de VA y VB. Modelo 2: tres factores que covarían, con errores que también covarían.
Estadísticos significativos al nivel de 0.05.

Anexo E

Resultados de los análisis psicométricos de las muestras HV, ESP, MAT, NAT y SOC

1. Habilidades del lenguaje (muestra HV)

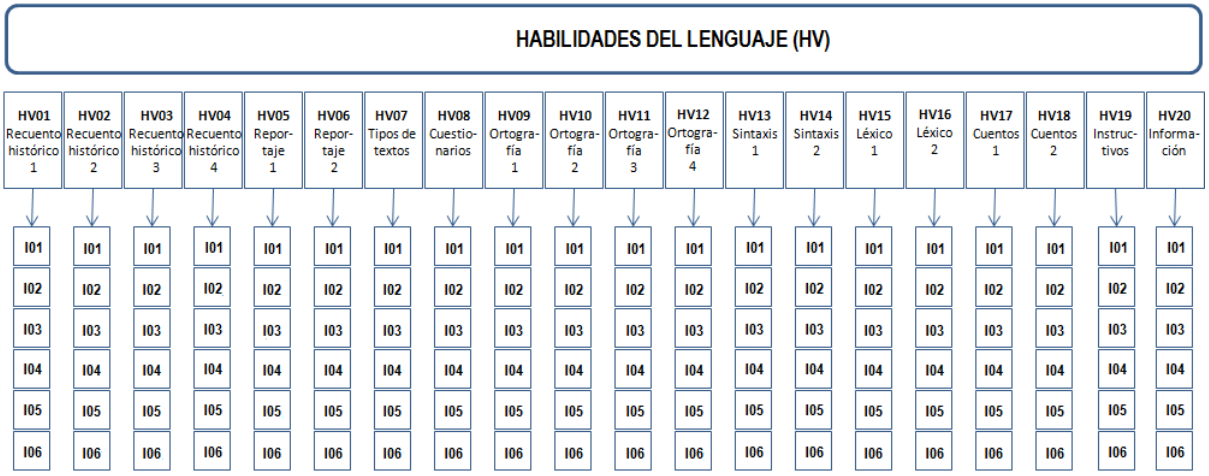
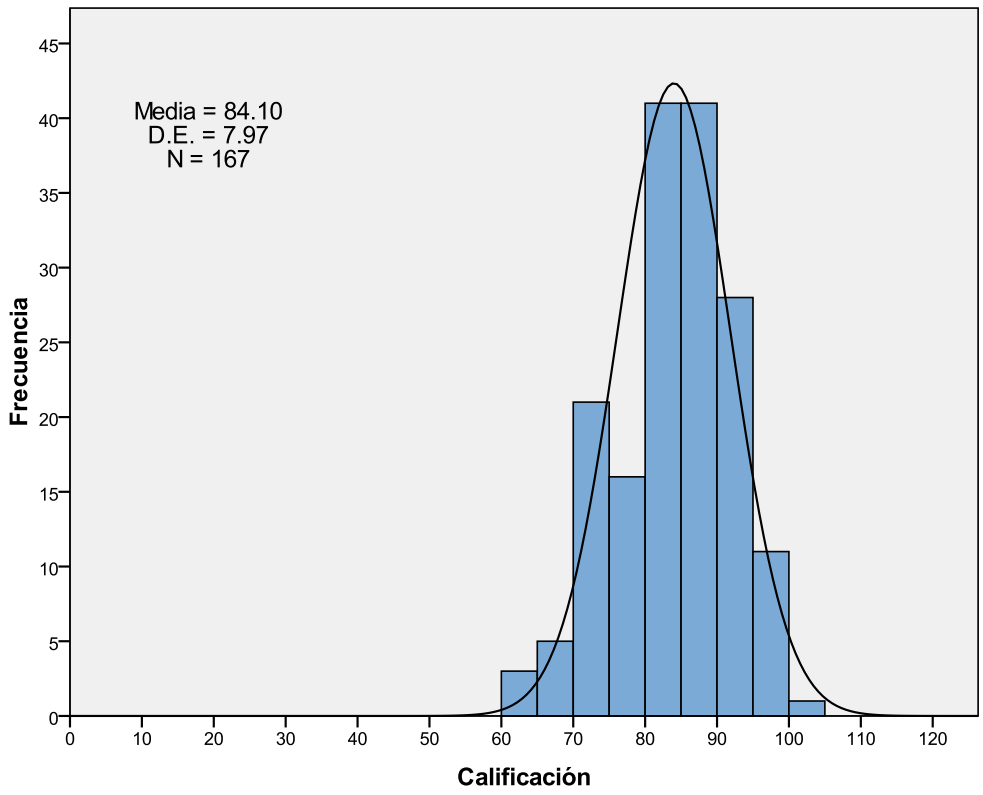


Figura E.1. Esquema de la muestra HV.



Simetría	-.406
Error estándar de simetría	.188
Curtosis	-.221
Error estándar de Curtosis	.374
Mínimo	61.56
Máximo	101.80

Figura E.2. Distribución de las calificaciones de la muestra HV.

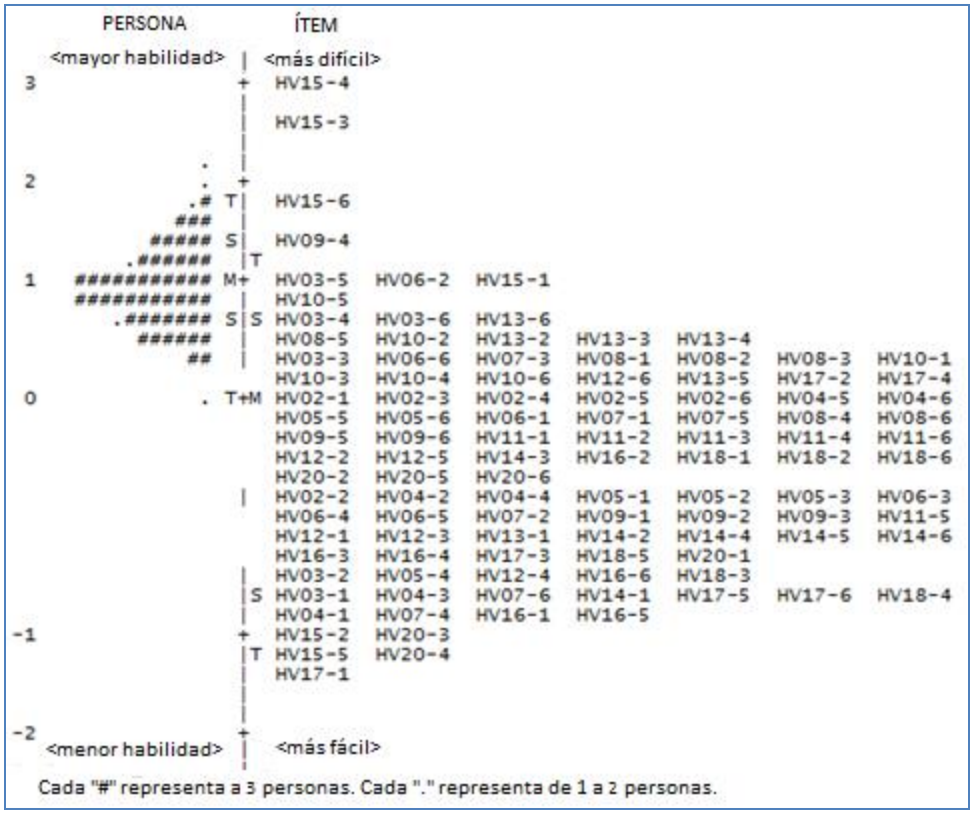


Figura E.3. Mapa de Wright de la muestra HV aplicada a 167 estudiantes de la universidad de Guanajuato.

Tabla E.1.
Infit, outfit, correlación punto medida y discriminación según el modelo de Rasch para la muestra HV aplicada en la UAG

IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS
V2-1	0.07	1.08	1.07	.35	0.94	V8-1	0.24	1.13	1.22	**.18**	0.90	V14-1	-0.60	0.98	0.92	.26	1.01
V2-2	-0.15	1.02	0.95	.35	0.99	V8-2	0.30	1.00	1.23	.27	0.97	V14-2	-0.24	0.98	1.02	.30	1.01
V2-3	-0.08	1.17	1.18	**.18**	0.84	V8-3	0.25	1.13	1.11	.27	0.87	V14-3	-0.07	0.96	0.94	.32	1.04
V2-4	0.10	1.10	1.10	.29	0.92	V8-4	-0.02	1.06	1.14	.25	0.95	V14-4	-0.10	0.95	0.93	.38	1.09
V2-5	0.07	1.11	1.24	.27	0.90	V8-5	0.39	1.01	1.02	.35	1.00	V14-5	-0.26	0.90	0.85	.40	1.10
V2-6	0.03	0.99	0.90	.38	1.02	V8-6	0.09	1.00	1.13	.28	0.98	V14-6	-0.11	0.89	0.85	.36	1.05
V3-1	-0.69	1.00	1.00	**.15**	0.99	V9-1	-0.12	1.09	1.12	.24	0.90	V15-1	1.10	1.05	1.05	**.10**	**0.52**
V3-2	-0.47	0.95	0.96	.33	1.10	V9-2	-0.18	1.03	0.99	.20	0.95	V15-2	-1.07	1.01	1.08	**.07**	0.98
V3-3	0.12	0.97	0.96	.32	1.08	V9-3	-0.20	1.13	1.17	**.07**	0.81	V15-3	2.52	1.01	1.06	**.12**	0.97
V3-4	0.70	0.89	0.86	.48	1.56	V9-4	1.45	0.99	0.99	.23	1.01	V15-4	2.95	1.01	1.08	**.10**	0.98
V3-5	1.08	0.92	0.92	.42	1.35	V9-5	0.00	0.90	0.85	.42	1.13	V15-5	-1.12	0.99	0.95	**.16**	1.01
V3-6	0.56	0.91	0.89	.44	1.36	V9-6	0.08	0.94	1.01	.34	1.02	V15-6	1.82	1.04	1.07	**.08**	0.87
V4-1	-0.85	1.01	1.07	.22	1.00	V10-1	0.28	1.14	1.14	.36	0.90	V16-1	-0.72	1.00	1.37	**.09**	0.99
V4-2	-0.21	1.12	1.38	**.11**	0.94	V10-2	0.41	0.98	0.98	.44	1.03	V16-2	0.05	1.10	1.12	**.07**	0.86
V4-3	-0.53	0.99	0.97	.20	1.01	V10-3	0.26	1.19	**1.34**	**.19**	0.81	V16-3	-0.17	1.02	1.08	.23	0.98
V4-4	-0.20	1.07	1.08	**.06**	0.93	V10-4	0.15	0.98	0.98	.43	1.04	V16-4	-0.10	1.10	1.28	**.17**	0.91
V4-5	-0.09	0.95	0.89	.41	1.07	V10-5	0.70	0.92	0.96	.49	1.10	V16-5	-0.73	0.97	0.65	.22	1.02
V4-6	0.00	1.05	1.29	**.19**	0.90	V10-6	0.30	0.88	0.87	.51	1.21	V16-6	-0.38	0.97	1.06	.21	1.00
V5-1	-0.29	1.03	1.03	.22	0.97	V11-1	-0.01	0.88	0.82	.47	1.09	V17-1	-1.33	1.02	1.20	**.08**	0.98
V5-2	-0.24	1.04	1.09	.27	0.98	V11-2	-0.01	0.86	0.85	.53	1.26	V17-2	0.25	0.98	0.97	.36	1.06
V5-3	-0.22	0.92	0.91	.42	1.13	V11-3	0.09	0.93	0.92	.46	1.09	V17-3	-0.19	0.92	0.84	.40	1.09
V5-4	-0.47	1.00	0.98	.28	1.02	V11-4	0.08	0.98	0.95	.33	1.02	V17-4	0.15	1.04	1.03	.29	0.96
V5-5	0.05	0.91	0.93	.42	1.13	V11-5	-0.20	0.86	0.78	.47	1.11	V17-5	-0.69	0.94	**0.67**	.28	1.03
V5-6	0.06	0.94	0.95	.35	1.04	V11-6	0.00	0.93	0.86	.39	1.07	V17-6	-0.70	0.91	**0.60**	.32	1.04
V6-1	0.07	1.09	1.07	**.03**	0.92	V12-1	-0.15	0.95	0.92	.36	1.07	V18-1	-0.03	1.23	**3.88**	.21	0.83
V6-2	1.02	0.99	0.95	**.14**	0.99	V12-2	0.07	1.06	1.20	.31	0.94	V18-2	-0.08	1.05	1.75	.22	0.98
V6-3	-0.23	1.12	1.20	**.17**	0.87	V12-3	-0.28	1.01	1.05	.24	1.00	V18-3	-0.39	1.03	0.96	**.18**	1.00
V6-4	-0.16	0.93	0.90	.37	1.04	V12-4	-0.30	0.93	0.75	.36	1.05	V18-4	-0.61	1.04	1.44	**.07**	0.98
V6-5	-0.14	1.08	1.27	**.19**	0.95	V12-5	-0.05	0.97	1.01	.32	1.02	V18-5	-0.16	1.01	2.94	**.12**	0.94
V6-6	0.17	1.04	0.97	**.14**	0.96	V12-6	0.25	0.86	0.83	.50	1.14	V18-6	0.00	1.03	**3.53**	.28	0.96
V7-1	0.01	0.95	0.89	.38	1.07	V13-1	-0.12	0.93	0.91	.39	1.09	V20-1	-0.24	1.06	1.01	**.15**	0.99
V7-2	-0.24	0.97	1.00	.34	1.04	V13-2	0.45	1.09	1.11	.21	**0.79**	V20-2	-0.08	1.01	0.98	.35	0.99
V7-3	0.22	1.10	1.12	**.19**	0.85	V13-3	0.40	1.01	1.01	.31	0.98	V20-3	-0.95	0.99	0.99	.23	1.02
V7-4	-0.72	0.94	0.51	.28	1.04	V13-4	0.35	0.89	0.87	.47	1.11	V20-4	-1.30	1.01	1.04	**.15**	0.99
V7-5	-0.02	1.06	1.18	.23	0.97	V13-5	0.23	0.91	1.09	.42	1.07	V20-5	-0.01	0.94	0.94	.32	1.04
V7-6	-0.50	0.83	0.75	.54	1.20	V13-6	0.56	1.00	1.00	.33	0.99	V20-6	0.00	1.00	1.00	.30	1.00

Nota: IT = ítem, M = medida, IN = infit, OU = outfit, PM = índice de correlación punto medida, V = HV.

2. Español (muestra ESP)

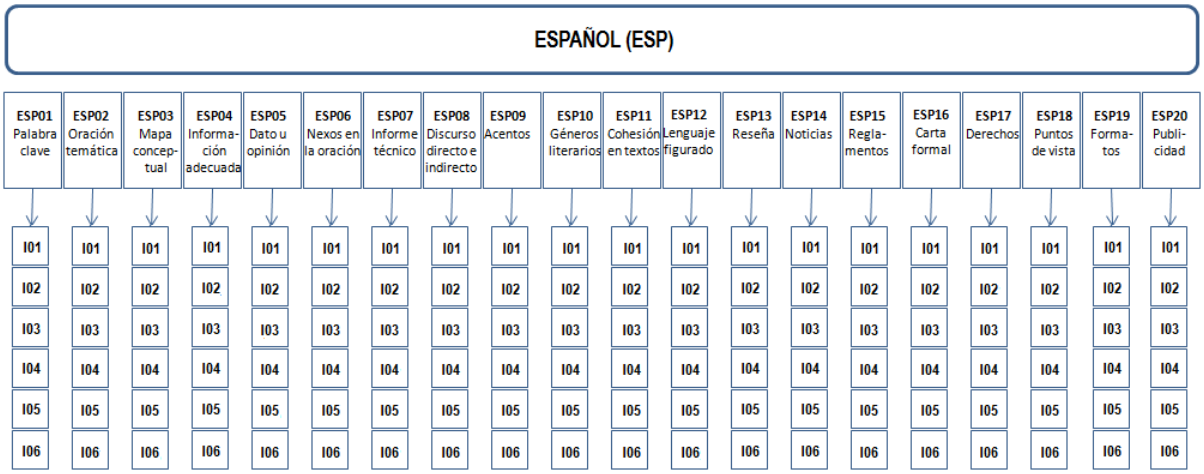
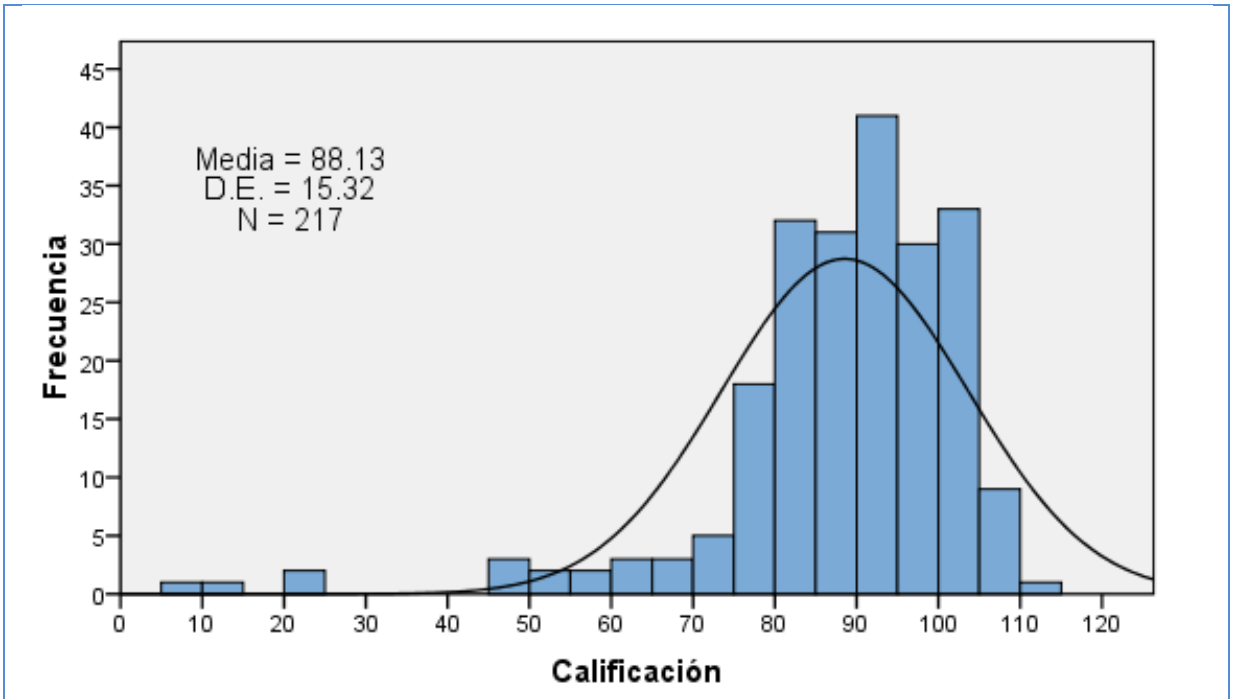


Figura E.4. Esquema de la muestra ESP



Simetría	-2.194
Error estándar de simetría	.165
Curtosis	7.407
Error estándar de Curtosis	.329
Mínimo	9.92
Máximo	110.51

Figura E.5. Distribución de las calificaciones de la muestra ESP.

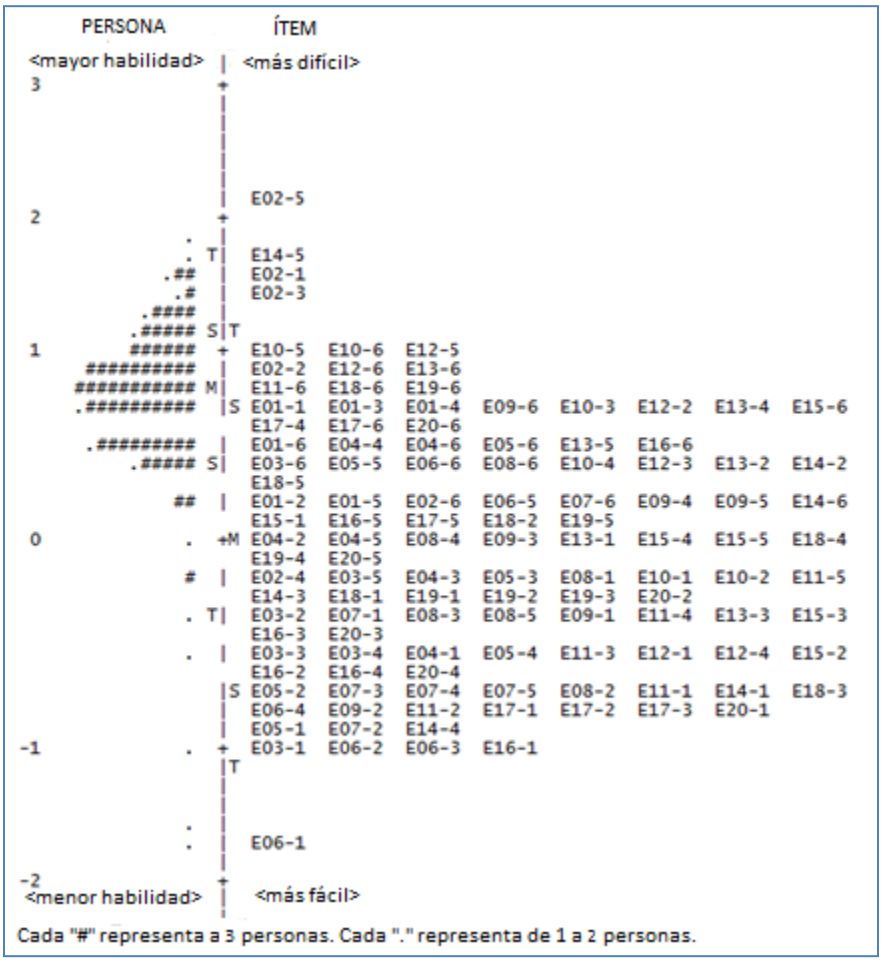


Figura E.6. Mapa de Wright de la muestra ESP aplicada a 217 estudiantes de la Universidad de Querétaro.

Tabla E.2.

Infit, outfit, correlación punto medida y discriminación según el modelo de Rasch para la muestra ESP aplicada a la Universidad de Querétaro

IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS
E1-1	0.54	1.25	1.27	**.14**	0.84	E08-1	-0.19	1.36	**1.81**	.30	0.92	E15-1	0.08	1.46	**1.51**	.27	0.94
E1-2	0.10	1.14	1.19	.38	0.92	E08-2	-0.53	1.16	1.00	.43	0.99	E15-2	-0.36	1.30	2.37	.40	0.93
E1-3	0.56	**1.56**	**2.46**	.28	**0.62**	E08-3	-0.25	1.09	1.31	.37	0.95	E15-3	-0.33	1.08	1.40	.41	0.95
E1-4	0.50	1.14	1.13	.34	0.83	E08-4	0.03	1.14	1.15	.30	0.85	E15-4	0.00	1.33	**1.96**	.36	0.89
E1-5	0.17	0.93	0.91	.51	1.02	E08-5	-0.34	0.69	0.88	.56	1.06	E15-5	-0.03	0.90	1.69	.50	1.00
E1-6	0.41	0.84	0.77	.58	1.10	E08-6	0.33	**0.77**	**0.58**	.61	1.13	E15-6	0.51	1.02	1.20	.52	0.97
E2-1	1.61	1.07	1.21	**-.01**	**0.74**	E09-1	-0.30	1.23	1.27	.29	**0.76**	E16-1	-0.94	1.09	1.08	**.11**	0.81
E2-2	0.84	1.02	1.02	**.16**	**0.73**	E09-2	-0.65	1.04	1.04	.39	0.96	E16-2	-0.50	0.96	1.02	.36	0.99
E2-3	1.38	0.96	0.96	.27	1.16	E09-3	0.03	0.97	1.00	.51	1.04	E16-3	-0.32	1.01	1.01	.34	0.98
E2-4	-0.10	1.00	0.98	.22	1.02	E09-4	0.15	1.02	1.00	.46	0.99	E16-4	-0.47	0.86	0.79	.47	1.03
E2-5	2.12	0.97	0.96	.22	1.04	E09-5	0.08	0.86	0.82	.57	1.12	E16-5	0.09	0.87	0.83	.49	1.12
E2-6	0.16	0.87	0.85	.49	1.62	E09-6	0.57	**0.74**	**0.66**	.65	1.25	E16-6	0.47	0.81	**0.74**	.57	1.13
E3-1	-1.02	1.16	0.96	.35	1.00	E10-1	-0.21	1.18	1.21	.25	0.90	E17-1	-0.78	1.34	1.72	.24	0.94
E3-2	-0.35	1.13	1.11	.34	0.97	E10-2	-0.15	1.27	1.23	**.19**	**0.69**	E17-2	-0.65	0.92	0.91	.42	1.02
E3-3	-0.43	0.94	0.83	.47	1.04	E10-3	0.63	1.01	1.01	.40	0.98	E17-3	-0.67	0.87	1.36	.43	1.02
E3-4	-0.43	0.89	0.80	.50	1.07	E10-4	0.34	1.09	1.08	.34	0.87	E17-4	0.56	0.97	0.95	.42	1.09
E3-5	-0.14	0.87	0.82	.50	1.07	E10-5	1.07	1.02	1.02	.38	0.98	E17-5	0.17	0.90	0.91	.49	1.13
E3-6	0.24	**0.77**	**0.64**	.60	1.17	E10-6	0.99	0.85	0.81	.57	1.14	E17-6	0.60	0.83	**0.78**	.55	1.17
E4-1	-0.46	**1.38**	**1.38**	.25	0.81	E11-1	-0.61	1.11	1.07	.24	0.90	E18-1	-0.18	1.08	1.04	.32	0.95
E4-2	0.02	**1.39**	**1.62**	.25	**0.60**	E11-2	-0.71	1.01	0.98	.38	1.00	E18-2	0.08	1.15	1.08	.33	0.94
E4-3	-0.13	1.20	**1.31**	.29	**0.77**	E11-3	-0.46	1.00	1.32	.36	0.98	E18-3	-0.55	0.97	0.97	.37	1.00
E4-4	0.36	1.11	1.15	.31	0.86	E11-4	-0.32	0.94	1.01	.41	1.02	E18-4	-0.01	0.93	0.90	.48	1.08
E4-5	0.07	0.94	0.97	.52	1.03	E11-5	-0.14	0.89	0.96	.47	1.08	E18-5	0.24	0.87	0.82	.52	1.11
E4-6	0.40	0.87	**0.76**	.57	1.09	E11-6	0.66	0.91	0.91	.48	1.23	E18-6	0.70	**0.78**	**0.75**	.58	1.27
E5-1	-0.79	**2.96**	**7.15**	**.09**	**0.78**	E12-1	-0.45	1.04	1.06	.25	0.96	E19-1	-0.09	1.02	1.03	.34	0.96
E5-2	-0.50	0.98	0.99	.42	1.01	E12-2	0.57	0.99	0.98	.32	1.06	E19-2	-0.20	1.03	0.98	.46	1.01
E5-3	-0.16	1.00	1.00	.35	0.98	E12-3	0.35	0.93	0.94	.39	1.12	E19-3	-0.19	0.99	1.17	.45	1.00
E5-4	-0.47	0.94	1.92	.44	0.99	E12-4	-0.42	0.83	**0.73**	.55	1.18	E19-4	0.04	0.97	0.98	.50	1.02
E5-5	0.28	1.21	**1.50**	.36	0.90	E12-5	0.93	0.95	0.94	.37	1.20	E19-5	0.16	0.85	0.79	.56	1.07
E5-6	0.49	0.98	1.26	.48	0.98	E12-6	0.84	0.87	0.86	.47	1.69	E19-6	0.66	0.88	0.84	.55	1.08
E6-1	-1.78	1.10	0.45	.39	1.04	E13-1	-0.03	1.07	1.08	.29	0.86	E20-1	-0.66	1.13	1.22	.32	0.97
E6-2	-0.93	0.99	1.17	.32	0.98	E13-2	0.30	1.12	1.16	.26	**0.79**	E20-2	-0.16	1.05	1.07	.27	0.92
E6-3	-1.02	0.76	0.68	.56	1.10	E13-3	-0.33	1.03	1.03	.37	0.99	E20-3	-0.29	1.00	1.02	.33	0.98
E6-4	-0.68	0.85	0.84	.48	1.04	E13-4	0.50	1.05	1.07	.32	0.92	E20-4	-0.46	0.81	0.73	.53	1.07
E6-5	0.08	0.88	0.87	.51	1.25	E13-5	0.47	1.12	1.13	.28	**0.66**	E20-5	-0.02	0.79	**0.61**	.58	1.12
E6-6	0.27	**0.69**	**0.57**	.67	1.26	E13-6	0.88	0.82	0.80	.55	1.43	E20-6	0.61	0.95	0.92	.47	1.05
E7-1	-0.27	1.20	1.23	**.15**	**0.77**	E14-1	-0.55	0.99	0.99	.33	1.03						
E7-2	-0.85	1.07	1.22	.32	0.98	E14-2	0.27	1.04	1.04	.25	0.95						
E7-3	-0.51	0.96	0.95	.37	1.04	E14-3	-0.10	0.96	0.95	.36	1.03						
E7-4	-0.61	0.85	0.80	.50	1.05	E14-4	-0.92	0.97	0.94	.39	1.04						
E7-5	-0.50	0.76	0.57	.56	1.07	E14-5	1.76	1.00	1.02	.21	0.99						
E7-6	0.18	**0.73**	**0.53**	.63	1.19	E14-6	0.16	0.90	0.89	.42	1.12						

Nota: IT = ítem, M = medida, IN = infit, OU = outfit, PM = índice de correlación punto medida, E = ESP.

3. Matemáticas (muestra MAT)

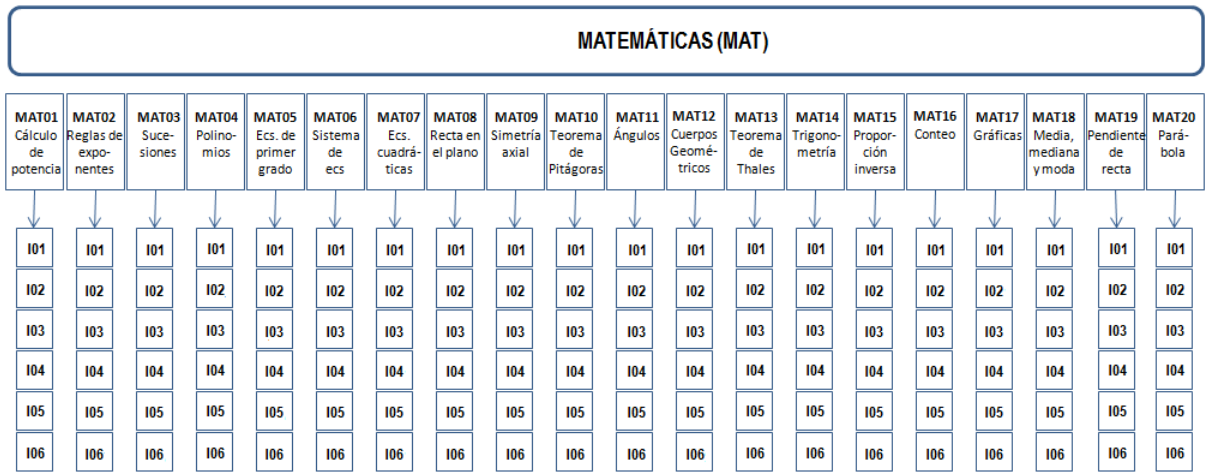
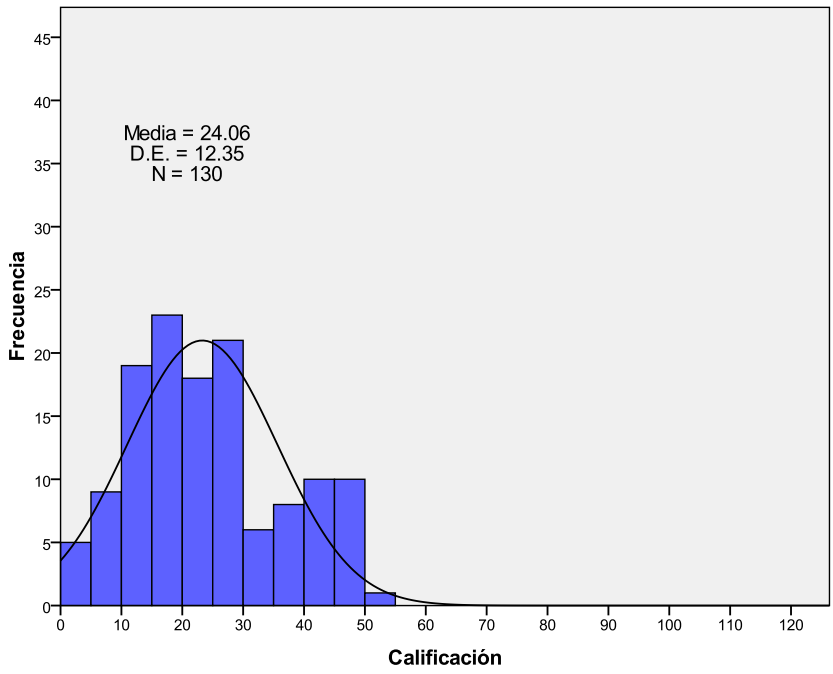


Figura 6.7. Esquema del la muestra MAT



Simetría	.392
Error estándar de simetría	.212
Curtosis	-.546
Error estándar de Curtosis	.422
Mínimo	.50
Máximo	50.97

Figura E.8. Distribución de las calificaciones de la muestra MAT de CESUES, Hermosillo.

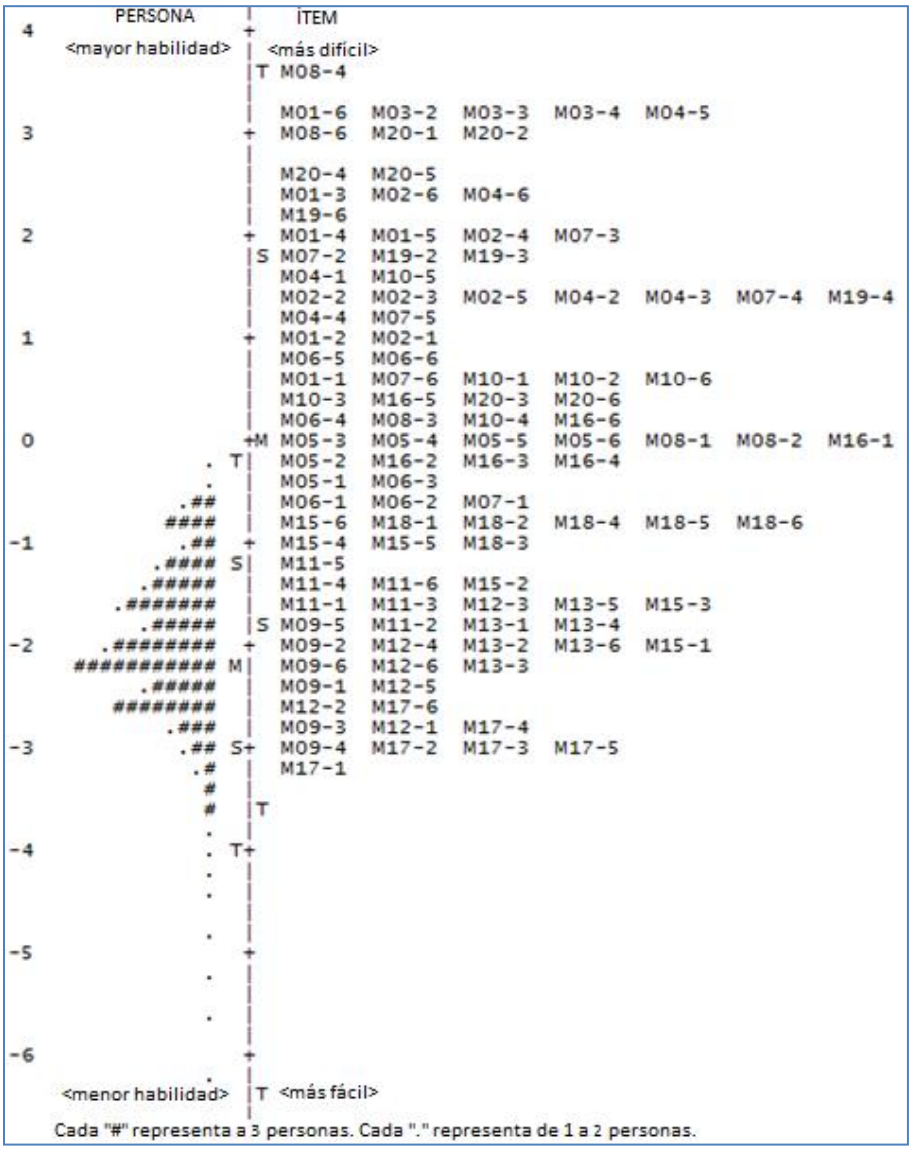


Figura E.9. Mapa de Wright de la muestra MAT aplicada a estudiantes de la Universidad Estatal De Sonora, sede Hermosillo.

Tabla E.3.

Medida, infit, outfit, correlación punto medida y discriminación de los ítems de la muestra MAT de CESUES, Hermosillo.

IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS
M1-1	0.68	1.04	0.98	**.14**	0.97	M07-1	-0.57	1.03	1.01	.24	0.96	M14-1					
M1-2	1.08	0.95	0.56	.25	1.06	M07-2	1.87	1.01	1.20	**.08**	0.99	M14-2					
M1-3	2.40	0.97	0.39	**.17**	1.04	M07-3	2.06	1.00	0.91	**.09**	1.00	M14-3			S/D		
M1-4	1.98	0.98	0.63	**.15**	1.03	M07-4	1.33	0.97	0.69	**.19**	1.03	M14-4					
M1-5	1.98	1.01	0.92	**.09**	1.00	M07-5	1.23	1.00	0.94	**.14**	1.00	M14-5					
M1-6	3.10	0.99	0.46	**.10**	1.02	M07-6	0.51	0.99	0.93	.20	1.01	M14-6					
M2-1	1.08	0.95	0.60	.24	1.06	M08-1	0.09	1.27	**1.40**	**.07**	**0.71**	M15-1	-2.08	1.05	1.05	.30	**0.77**
M2-2	1.44	0.96	0.59	.21	1.05	M08-2	0.08	1.09	1.07	.26	0.92	M15-2	-1.38	0.97	0.92	.36	1.11
M2-3	1.44	0.97	0.58	.20	1.04	M08-3	0.17	1.26	**3.00**	**.02**	**0.75**	M15-3	-1.52	0.95	0.93	.38	1.17
M2-4	1.98	1.00	0.78	**.12**	1.01	M08-4	3.70	1.00	0.56	**.06**	1.01	M15-4	-0.96	0.93	0.84	.38	1.16
M2-5	1.44	1.00	0.75	**.15**	1.01	M08-5	In.					M15-5	-0.96	0.91	0.81	.40	1.20
M2-6	2.40	0.99	0.54	**.13**	1.02	M08-6	3.00	1.01	0.72	**.07**	1.00	M15-6	-0.86	0.91	0.85	.39	1.17
M3-1	In.					M09-1	-2.46	1.30	**1.41**	.35	**0.71**	M16-1	-0.07	0.97	0.85	.28	1.05
M3-2	3.10	0.99	0.38	**.12**	1.03	M09-2	-2.00	1.01	1.00	.50	1.01	M16-2	-0.14	0.96	0.89	.28	1.04
M3-3	3.10	0.99	0.38	**.12**	1.03	M09-3	-2.75	1.00	0.98	.58	0.99	M16-3	-0.17	0.94	0.79	.32	1.08
M3-4	3.10	0.99	0.38	**.12**	1.03	M09-4	-2.99	1.17	1.36	.54	0.89	M16-4	-0.10	0.93	0.79	.32	1.08
M3-5	In.					M09-5	-1.71	**1.55**	**1.57**	.31	**0.10**	M16-5	0.41	1.07	1.31	**.08**	0.91
M3-6	In.					M09-6	-2.10	1.27	**1.40**	.42	**0.44**	M16-6	0.17	0.93	0.77	.30	1.07
M4-1	1.68	0.96	0.43	.22	1.06	M10-1	0.65	0.98	0.89	.22	1.03	M17-1	-3.10	0.97	0.94	.41	1.06
M4-2	1.44	0.95	0.46	.23	1.06	M10-2	0.59	0.97	**0.71**	.25	1.05	M17-2	-2.92	0.91	0.87	.47	1.20
M4-3	1.44	0.96	0.49	.22	1.05	M10-3	0.42	0.98	**0.79**	.24	1.04	M17-3	-3.01	0.89	0.84	.49	1.24
M4-4	1.25	1.00	0.71	**.17**	1.02	M10-4	0.22	0.97	0.87	.25	1.04	M17-4	-2.72	0.91	0.87	.46	1.25
M4-5	3.10	1.00	0.63	**.07**	1.01	M10-5	1.54	0.98	**0.57**	**.19**	1.04	M17-5	-2.92	0.86	0.79	.52	1.33
M4-6	2.40	0.99	0.51	**.13**	1.03	M10-6	0.53	1.00	0.95	**.18**	1.00	M17-6	-2.57	0.93	0.89	.45	1.23
M5-1	-.49	0.87	0.73	.41	1.20	M11-1	-1.56	0.81	**0.75**	.52	1.67	M18-1	-0.82	0.93	0.94	.47	1.05
M5-2	-.21	0.86	0.63	.41	1.19	M11-2	-1.81	0.85	0.81	.49	1.62	M18-2	-0.83	0.83	0.73	.51	1.12
M5-3	0.06	0.86	0.61	.39	1.15	M11-3	-1.64	0.80	**0.74**	.54	1.78	M18-3	-1.00	1.05	1.04	.45	0.98
M5-4	-.01	0.88	0.64	.38	1.14	M11-4	-1.37	**0.78**	**0.71**	.54	1.65	M18-4	-0.77	0.89	0.85	.47	1.07
M5-5	-.01	0.89	0.67	.36	1.13	M11-5	-1.28	0.91	0.95	.40	1.22	M18-5	-0.72	0.86	0.66	.49	1.10
M5-6	-.01	0.87	0.60	.39	1.16	M11-6	-1.30	0.87	0.89	.45	1.35	M18-6	-0.82	0.92	0.86	.47	1.04
M6-1	-.52	1.00	1.07	.25	0.98	M12-1	-2.74	**1.41**	**2.06**	.46	**0.74**	M19-1	In.				
M6-2	-.63	1.00	0.94	.28	1.01	M12-2	-2.63	1.26	**2.36**	.50	0.81	M19-2	1.71	1.01	0.70	**.13**	1.01
M6-3	-.43	0.98	0.94	.28	1.03	M12-3	-1.70	**1.36**	**1.40**	.25	**0.64**	M19-3	1.87	1.01	0.69	**.13**	1.01
M6-4	0.17	0.99	0.97	.23	1.01	M12-4	-1.99	1.10	1.14	.49	0.84	M19-4	1.33	0.99	0.91	**.15**	1.01
M6-5	0.82	0.98	0.75	.22	1.03	M12-5	-2.34	1.26	**1.78**	.45	**0.56**	M19-5	In.				
M6-6	0.82	0.96	0.59	.26	1.06	M12-6	-2.28	**1.39**	**1.72**	.42	**0.49**	M19-6	2.29	0.98	0.57	**.14**	1.03
						M13-1	-1.77	0.85	**0.79**	.50	1.62	M20-1	3.00	1.02	1.65	**-.02**	0.96
						M13-2	-2.05	0.80	**0.76**	.55	1.86	M20-2	3.00	1.02	1.65	**-.02**	0.96
						M13-3	-2.21	0.87	0.82	.50	1.56	M20-3	0.47	1.21	2.39	**.01**	**0.77**
						M13-4	-1.72	0.83	**0.79**	.51	1.67	M20-4	2.59	1.01	0.97	**.06**	0.99
						M13-5	-1.58	0.91	0.95	.41	1.27	M20-5	2.59	1.02	1.61	**.00**	0.96
						M13-6	-1.97	0.90	0.89	.45	1.42	M20-6	0.31	1.22	**4.16**	**-.01**	**0.75**

Nota: IT = ítem, M = medida, IN = infit, OU = outfit, PM = índice de correlación punto medida, M = MAT. "In" refiere a *inestimable*, no se pudieron obtener los parámetros solicitados debido a que ningún sustentante resolvió correctamente el ítem. S/D indica *sin datos*, no se tienen datos de esta familia de ítems.

4. Ciencias naturales (muestra NAT)

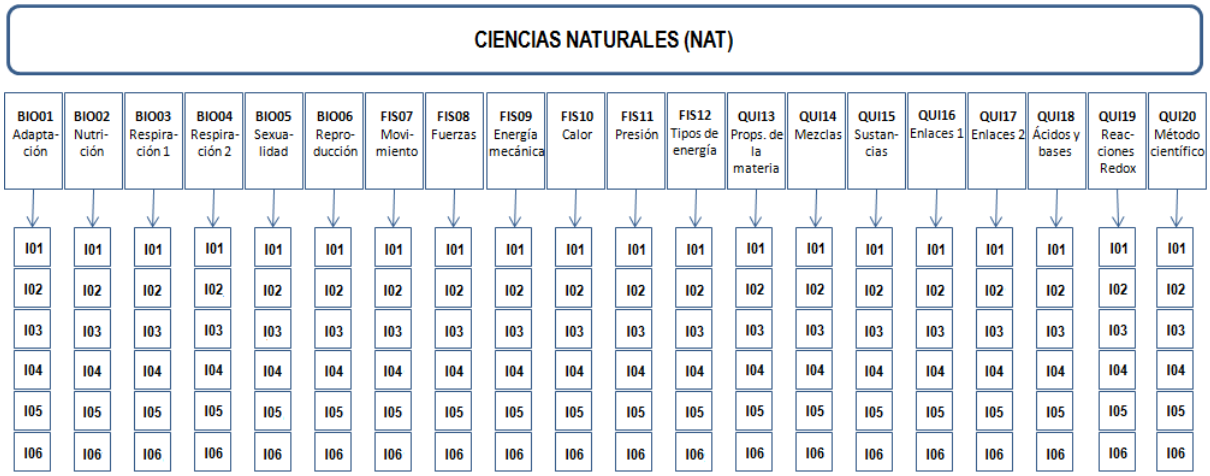
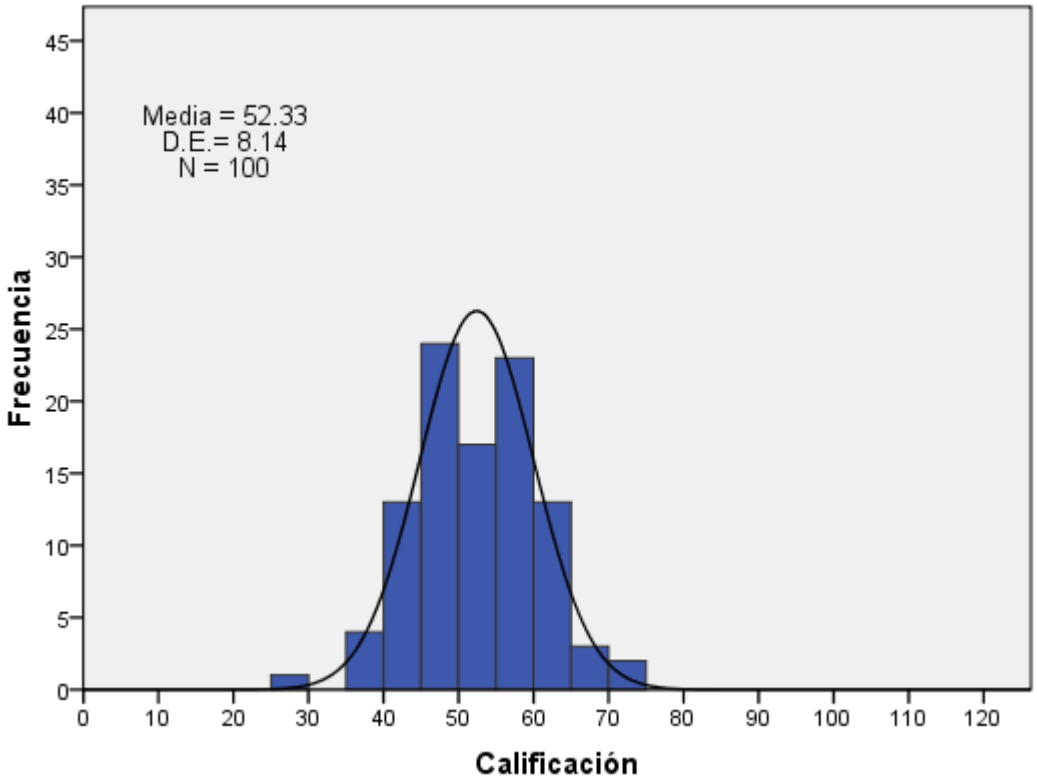


Figura E.10. Esquema del la muestra NAT



Simetría	-.193
Error estándar de simetría	.241
Curtosis	.200
Error estándar de Curtosis	.478
Mínimo	25.50
Máximo	70.20

Figura E.11. Distribución de las calificaciones de la muestra NAT de CESUES, San Luis Río Colorado.

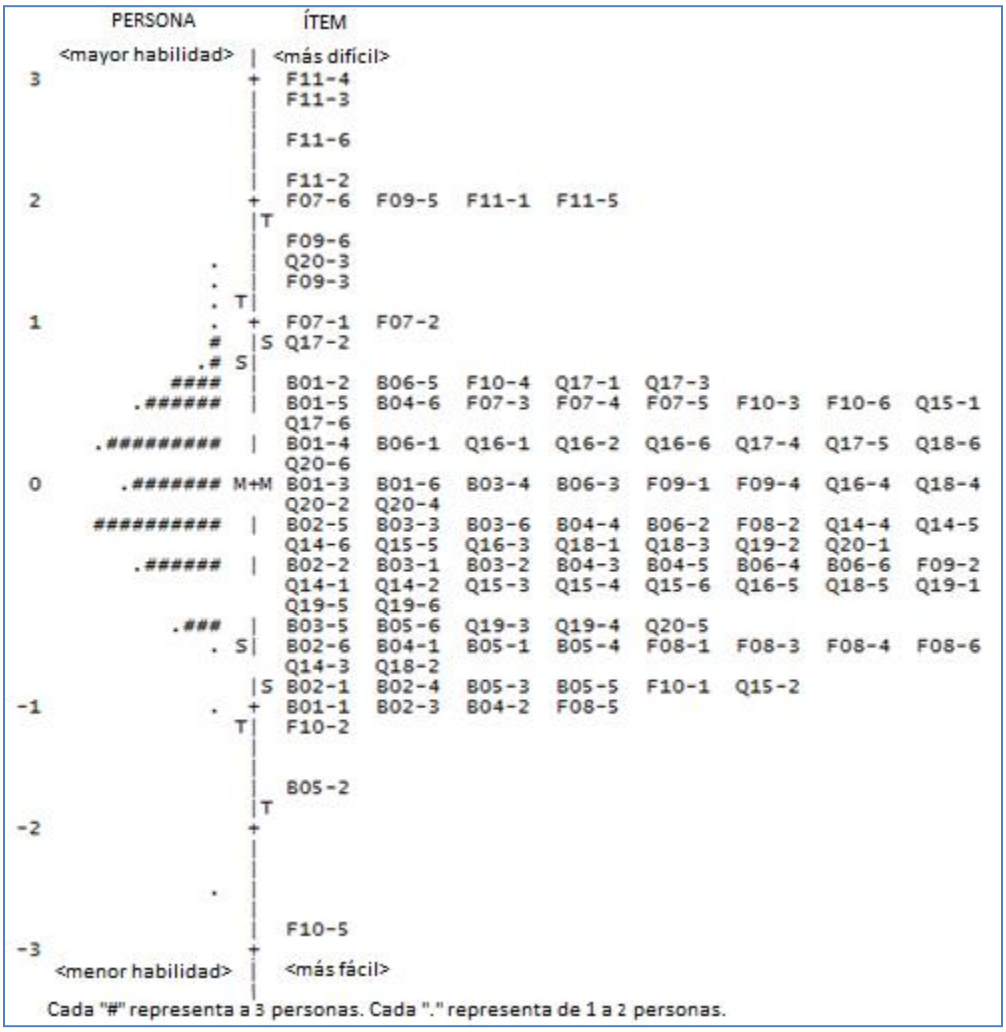


Figura E.12. Mapa de Wright de la muestra NAT de CESUES, San Luis Río Colorado, en caso de coincidir los ítems, se agregaron los datos de UACJ; también se agregó la familia FIS10, que se aplicó a UACJ.

Tabla E.4.

Medida, infit, outfit, correlación punto medida y discriminación de los ítems del área de Ciencias naturales (NAT)

IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS
B1-1	-0.95	1.24	1.27	.24	**0.69**	F07-1	1.00	1.00	0.98	**.19**	1.00	Q13-1					
B1-2	0.54	1.09	1.13	.26	0.93	F07-2	1.01	0.96	0.94	.29	1.08	Q13-2					
B1-3	0.08	1.08	1.08	.35	0.91	F07-3	0.35	0.87	0.86	.43	1.24	Q13-3		S/D			
B1-4	0.10	1.20	1.21	.27	**0.67**	F07-4	0.27	0.87	0.86	.43	1.28	Q13-4					
B1-5	0.35	1.09	1.08	.31	0.92	F07-5	0.37	0.82	0.81	.48	1.56	Q13-5					
B1-6	0.05	1.04	1.01	.40	0.96	F07-6	2.06	0.95	0.94	.29	1.09	Q13-6					
B2-1	-0.76	1.12	1.17	.37	0.82	F08-1	-0.69	1.07	1.13	.41	0.92	Q14-1	-0.37	1.14	1.17	.34	**0.72**
B2-2	-0.37	0.94	0.95	.48	1.10	F08-2	-0.19	0.99	1.00	.42	1.00	Q14-2	-0.39	0.98	1.01	.37	1.03
B2-3	-0.99	0.95	0.91	.55	1.07	F08-3	-0.64	0.87	0.86	.54	1.17	Q14-3	-0.59	1.09	1.14	.41	0.91
B2-4	-0.76	0.94	0.87	.51	1.06	F08-4	-0.69	0.95	0.96	.50	1.06	Q14-4	-0.18	1.07	1.08	.39	0.88
B2-5	-0.23	0.85	0.86	.51	1.31	F08-5	-0.97	0.78	0.71	.59	1.08	Q14-5	-0.12	0.97	0.96	.43	1.04
B2-6	-0.71	0.98	1.01	.49	1.02	F08-6	-0.63	0.98	1.00	.45	0.98	Q14-6	-0.10	1.00	1.05	.39	0.95
B3-1	-0.38	0.96	0.95	.45	1.05	F09-1	-0.05	0.96	0.96	.27	1.84	Q15-1	0.33	0.99	0.98	.39	1.01
B3-2	-0.41	1.07	1.09	.38	0.86	F09-2	-0.25	0.94	0.94	.34	2.09	Q15-2	-0.79	0.90	0.81	.55	1.14
B3-3	-0.09	1.04	0.99	.41	0.98	F09-3	1.40	0.93	0.91	.32	1.14	Q15-3	-0.28	0.89	0.94	.53	1.11
B3-4	0.03	1.10	1.08	.35	0.85	F09-4	-0.01	0.95	0.95	.31	2.06	Q15-4	-0.38	0.91	0.89	.51	1.10
B3-5	-0.47	0.91	0.91	.49	1.10	F09-5	1.92	0.95	0.95	.31	1.15	Q15-5	-0.12	0.85	0.87	.55	1.11
B3-6	-0.21	0.95	0.92	.47	1.11	F09-6	1.63	0.95	0.95	.29	1.09	Q15-6	-0.39	0.93	0.96	.49	1.08
B4-1	-0.72	1.25	1.3	.34	0.86	F10-1	-0.90	0.94	0.89	.38	1.11	Q16-1	0.13	0.99	0.95	.46	1.03
B4-2	-1.00	1.02	0.98	.50	0.99	F10-2	-1.10	1.26	**2.00**	**-.03**	**0.56**	Q16-2	0.21	1.23	1.28	**.07**	**0.56**
B4-3	-0.39	1.00	1.00	.45	1.00	F10-3	0.34	0.90	0.86	.35	2.15	Q16-3	-0.09	1.16	1.15	**.07**	0.82
B4-4	-0.20	1.08	1.11	.40	0.95	F10-4	0.47	0.95	0.92	.28	1.53	Q16-4	-0.07	1.07	1.07	.20	0.94
B4-5	-0.34	0.89	0.86	.53	1.07	F10-5	-2.78	1.70	**2.67**	**-.18**	**0.65**	Q16-5	-0.28	0.92	0.92	.40	1.07
B4-6	0.27	0.97	0.96	.41	1.05	F10-6	0.40	1.04	1.01	**.18**	**0.60**	Q16-6	0.23	1.09	1.08	.23	0.86
B5-1	-0.59	1.24	**1.99**	.33	**0.77**	F11-1	2.03	0.98	0.96	**.15**	1.02	Q17-1	0.44	0.88	0.84	.38	1.31
B5-2	-1.62	0.84	0.77	.67	1.09	F11-2	2.22	0.98	0.94	**.16**	1.02	Q17-2	0.84	1.02	0.95	.22	0.99
B5-3	-0.85	1.04	1.12	.45	0.97	F11-3	2.81	0.95	0.87	**.18**	1.03	Q17-3	0.55	0.89	0.84	.36	1.21
B5-4	-0.62	1.11	**1.49**	.40	0.91	F11-4	2.92	0.96	0.85	**.18**	1.03	Q17-4	0.20	0.93	0.91	.34	1.23
B5-5	-0.88	0.90	0.91	.51	1.04	F11-5	1.92	0.95	0.89	.22	1.05	Q17-5	0.12	0.86	0.84	.41	1.86
B5-6	-0.58	1.05	1.24	.40	0.92	F11-6	2.44	0.96	0.93	**.17**	1.03	Q17-6	0.34	0.86	0.84	.40	1.53
B6-1	0.22	1.04	1.04	.31	0.96	F12-1						Q18-1	-0.18	1.00	0.99	.41	1.00
B6-2	-0.21	1.10	1.13	.38	0.84	F12-2						Q18-2	-0.74	1.12	1.15	.31	0.88
B6-3	-0.07	1.01	1.06	.43	1.01	F12-3						Q18-3	-0.19	1.09	1.07	.33	0.93
B6-4	-0.40	1.21	**1.90**	.29	**0.70**	F12-4			S/D			Q18-4	-0.04	1.00	1.03	.42	0.98
B6-5	0.42	1.06	1.12	.33	0.93	F12-5						Q18-5	-0.34	0.93	0.93	.49	1.04
B6-6	-0.26	1.06	1.07	.39	0.91	F12-6						Q18-6	0.10	1.03	1.02	.38	0.97
												Q19-1	-0.38	1.04	1.06	.31	0.94
												Q19-2	-0.20	1.26	**1.40**	.31	**0.52**
												Q19-3	-0.55	0.89	0.90	.50	1.17
												Q19-4	-0.49	0.84	0.84	.56	1.25
												Q19-5	-0.30	0.80	0.80	.60	1.37
												Q19-6	-0.40	0.93	0.93	.49	1.12
												Q20-1	-0.18	0.94	0.94	.36	2.35
												Q20-2	-0.04	0.96	0.96	.27	1.84
												Q20-3	1.42	0.96	0.93	.27	1.05
												Q20-4	0.00	1.01	1.01	.12	0.85
												Q20-5	-0.55	0.92	0.90	.43	1.63
												Q20-6	0.18	0.97	0.96	.26	1.47

Nota: IT = ítem, M = medida, IN = infit, OU = outfit, PM = índice de correlación punto medida, B = Biología, F = Física, Q = Química.

Información obtenida de la base de datos NAT aplicada a CESUES, San Luis Río Colorado, en caso de coincidir los ítems, se agregaron los datos de UACJ; también se agregó la familia FIS10, que se aplicó a UACJ. S/D indica *sin datos*, no se tienen datos de esta familia de ítems. En las familias FIS10, FIS12 y QUI13 no se puso extraer la información de la base de datos original.

5. Ciencias sociales (muestra SOC)

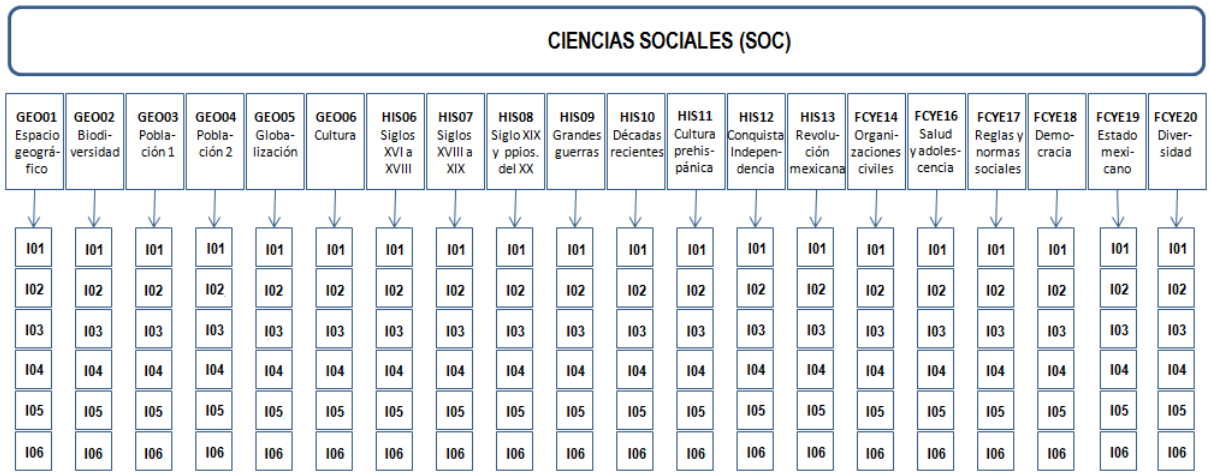
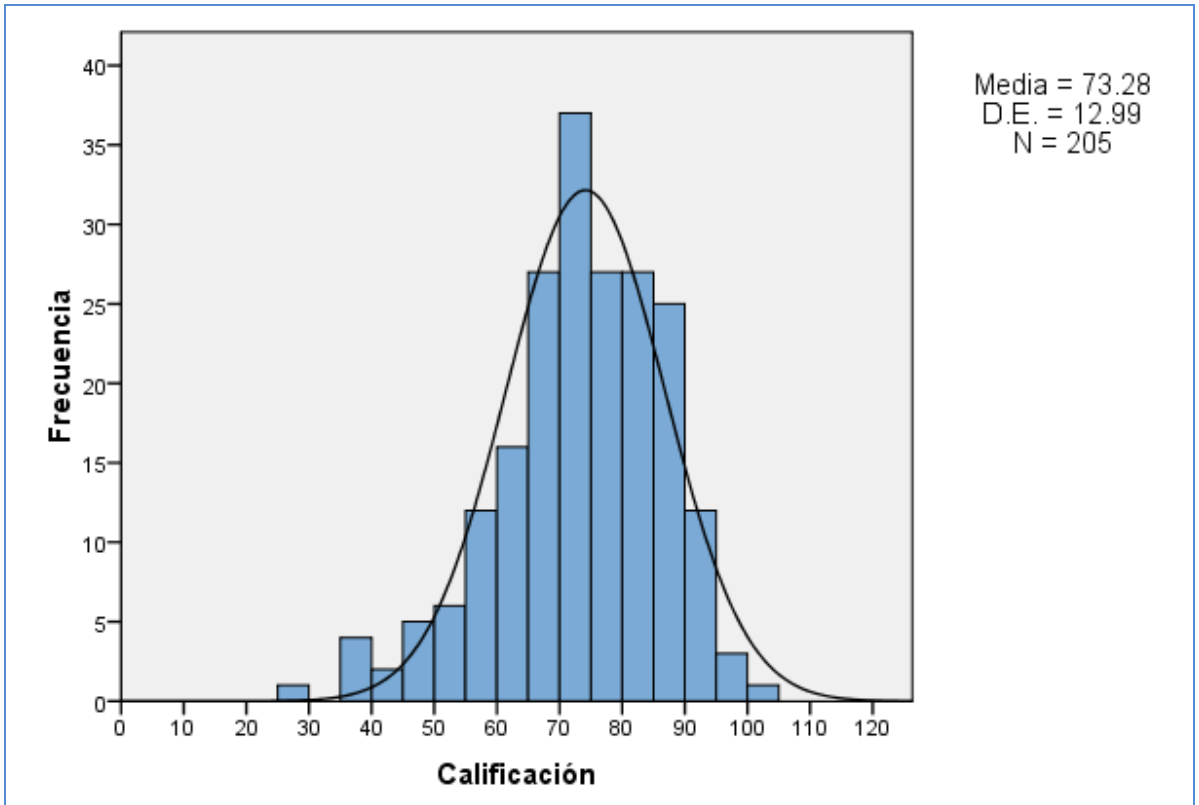


Figura E.13. Esquema del la muestra SOC.



Simetría	-.741
Error estándar de simetría	.170
Curtosis	.763
Error estándar de Curtosis	.338
Mínimo	25.57
Máximo	101.10

Figura E.14. Distribución de las calificaciones de la muestra SOC. Se eliminó un dato que correspondía a calificación total de ciencias sociales = 0, por considerarlo un caso atípico.

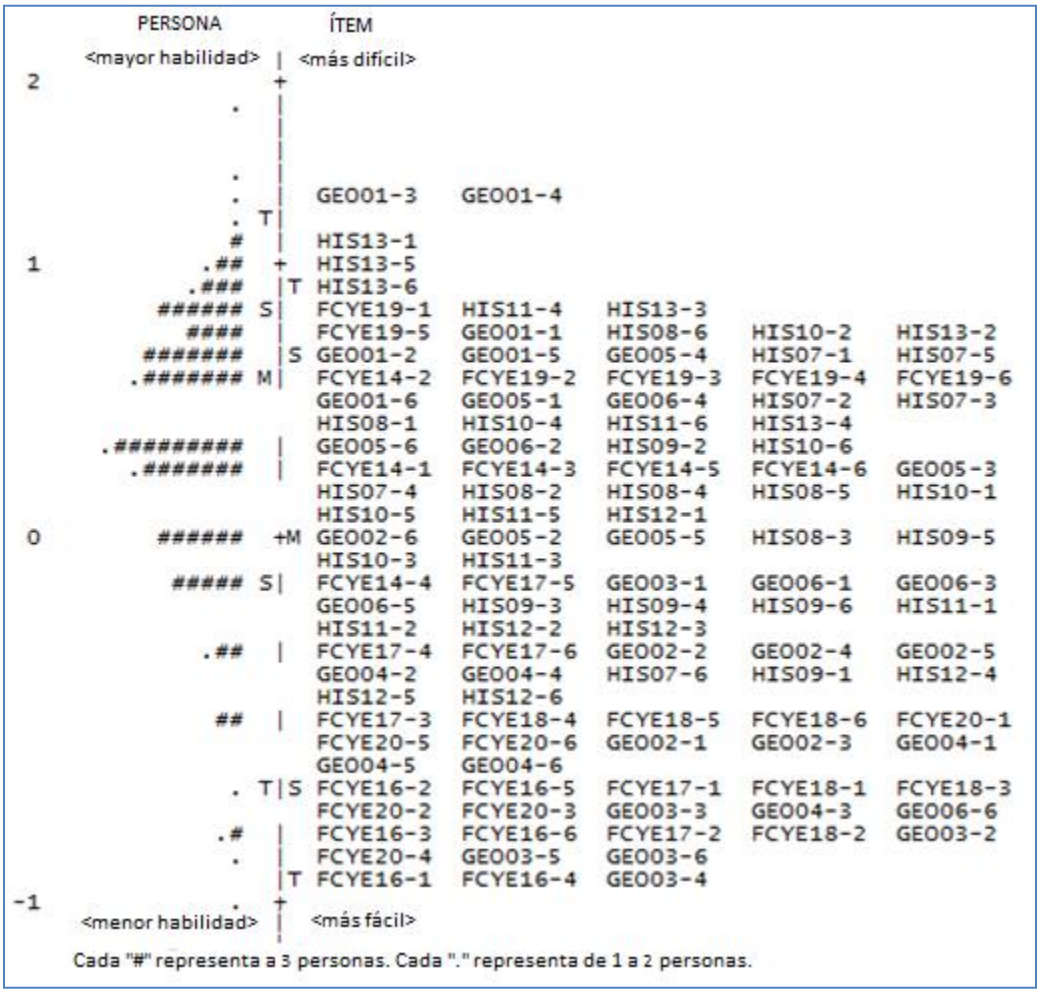


Figura E.15. Mapa de Wright de la muestra SOC aplicada a la Universidad Autónoma de Querétaro.

Tabla E.5.
Medida, Infit, Outfit, correlación punto medida y discriminación de los 114 ítems de Ciencias sociales

IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS	IT	M	IN	OU	PM	DIS
G1-1	0.59	1.10	1.12	.29	**0.70**	H07-1	0.47	0.98	1.07	.44	1.02	F14-1	0.14	1.27	1.27	.20	**0.77**
G1-2	0.55	1.07	1.11	.32	0.81	H07-2	0.42	1.02	1.02	.47	0.97	F14-2	0.42	1.27	**1.32**	**.17**	**0.65**
G1-3	1.38	0.99	1.00	.36	1.02	H07-3	0.40	0.81	0.80	.57	1.18	F14-3	0.17	1.04	1.06	.43	0.93
G1-4	1.41	1.08	1.09	.25	0.89	H07-4	0.18	0.92	0.94	.53	1.14	F14-4	-0.11	1.07	1.16	.43	0.93
G1-5	0.50	1.09	1.12	.32	**0.78**	H07-5	0.47	0.81	0.81	.60	1.20	F14-5	0.12	1.15	1.17	.32	0.80
G1-6	0.33	0.99	1.02	.42	0.99	H07-6	-0.19	**0.76**	**0.65**	.64	1.17	F14-6	0.17	1.08	1.11	.42	0.86
G2-1	-0.43	0.98	0.96	.43	1.03	H08-1	0.40	1.07	1.11	.33	0.92	F16-1	-0.83	1.25	1.15	.26	0.94
G2-2	-0.29	0.87	0.88	.56	1.18	H08-2	0.13	1.07	1.12	.40	0.87	F16-2	-0.53	1.21	1.28	.30	0.93
G2-3	-0.37	1.14	**1.71**	.32	0.85	H08-3	0.04	0.99	0.97	.45	1.01	F16-3	-0.60	1.10	1.22	.38	0.98
G2-4	-0.28	1.08	1.23	.41	0.94	H08-4	0.14	0.90	0.91	.50	1.14	F16-4	-0.82	0.99	1.34	.34	1.00
G2-5	-0.30	0.83	0.74	.60	1.17	H08-5	0.17	0.83	0.85	.58	1.17	F16-5	-0.50	1.01	1.17	.39	0.99
G2-6	0.02	0.86	0.85	.58	1.20	H08-6	0.65	0.92	0.92	.50	1.11	F16-6	-0.63	0.98	0.81	.46	1.03
G3-1	-0.07	1.10	1.08	.27	0.89	H09-1	-0.23	0.99	0.96	.47	1.05	F17-1	-0.44	**1.42**	**1.91**	**.17**	**0.76**
G3-2	-0.67	1.03	1.04	.31	1.00	H09-2	0.29	0.95	0.94	.42	1.05	F17-2	-0.59	1.16	1.17	.31	0.85
G3-3	-0.44	1.10	1.11	.21	0.89	H09-3	-0.07	1.01	1.01	.45	1.00	F17-3	-0.41	0.92	0.87	.52	1.11
G3-4	-0.82	1.00	1.04	.28	1.00	H09-4	-0.11	0.86	0.83	.54	1.13	F17-4	-0.27	1.09	1.19	.31	0.86
G3-5	-0.74	1.00	1.10	.30	1.00	H09-5	0.00	1.02	1.03	.44	0.98	F17-5	-0.18	0.94	0.96	.44	1.03
G3-6	-0.70	0.94	0.94	.35	1.03	H09-6	-0.14	0.97	0.93	.48	1.06	F17-6	-0.22	0.89	0.84	.50	1.10
G4-1	-0.31	1.16	1.29	.29	0.81	H10-1	0.12	0.95	1.00	.52	1.10	F18-1	-0.53	**1.32**	**1.71**	.20	0.85
G4-2	-0.28	1.18	**1.67**	.28	**0.78**	H10-2	0.60	0.88	0.90	.56	1.20	F18-2	-0.56	1.11	1.15	.35	0.97
G4-3	-0.55	1.11	1.30	.36	0.95	H10-3	0.00	**0.79**	**0.77**	.61	1.19	F18-3	-0.54	1.13	1.06	.38	0.97
G4-4	-0.27	0.98	1.11	.40	1.01	H10-4	0.40	0.82	0.82	.56	1.14	F18-4	-0.37	1.10	1.31	.39	0.95
G4-5	-0.39	0.98	1.10	.40	0.99	H10-5	0.15	**0.74**	**0.74**	.67	1.33	F18-5	-0.35	1.01	1.00	.45	1.01
G4-6	-0.34	1.10	1.24	.35	0.88	H10-6	0.30	**0.79**	**0.79**	.65	1.34	F18-6	-0.34	0.99	0.93	.46	1.04
G5-1	0.44	1.11	1.10	.28	0.91	H11-1	-0.08	1.04	1.03	.37	0.96	F19-1	0.70	1.16	1.16	.24	**0.71**
G5-2	0.06	1.06	1.06	.42	0.91	H11-2	-0.13	0.98	0.98	.48	1.08	F19-2	0.32	1.09	1.15	.39	0.80
G5-3	0.06	1.20	1.25	.25	**0.65**	H11-3	0.03	0.94	0.94	.47	1.06	F19-3	0.32	0.99	0.99	.45	1.02
G5-4	0.54	0.97	0.96	.46	1.04	H11-4	0.73	1.02	1.02	.42	0.97	F19-4	0.39	0.90	0.89	.57	1.17
G5-5	0.00	1.04	1.00	.46	0.95	H11-5	0.12	0.88	0.88	.56	1.18	F19-5	0.62	0.99	1.00	.46	1.02
G5-6	0.19	1.08	1.12	.38	0.88	H11-6	0.43	0.98	0.99	.45	1.03	F19-6	0.42	0.91	0.91	.54	1.15
G6-1	-0.15	0.93	0.93	.43	1.13	H12-1	0.08	1.05	1.03	.40	0.97	F20-1	-0.39	1.10	1.13	.33	0.90
G6-2	0.22	1.00	0.98	.33	1.01	H12-2	-0.09	1.22	1.22	.24	0.66	F20-2	-0.48	1.04	1.00	.43	0.99
G6-3	-0.15	1.00	0.99	.40	1.00	H12-3	-0.07	0.89	0.94	.51	1.10	F20-3	-0.46	1.03	1.09	.42	1.00
G6-4	0.32	1.00	1.02	.34	0.99	H12-4	-0.25	0.82	0.79	.60	1.21	F20-4	-0.81	0.81	**0.78**	.56	1.20
G6-5	-0.16	0.96	0.95	.42	1.06	H12-5	-0.20	0.96	1.06	.50	1.05	F20-5	-0.33	0.98	1.07	.45	1.03
G6-6	-0.52	0.86	0.89	.48	1.09	H12-6	-0.25	**0.79**	**0.74**	.62	1.25	F20-6	-0.41	0.96	0.95	.49	1.04
						H13-1	1.12	1.02	1.04	.40	0.96						
						H13-2	0.61	1.05	1.06	.41	0.92						
						H13-3	0.71	1.13	1.13	.31	0.83						
						H13-4	0.36	1.00	0.99	.45	0.99						
						H13-5	0.96	1.13	1.15	.29	0.81						
						H13-6	0.86	0.90	0.90	.51	1.12						

Nota: IT = ítem, M = medida, IN = infit, OU = outfit, PM = índice de correlación punto medida, G = Geografía, H = Historia, F = Formación Cívica y Ética.

Anexo F

Resultados de los análisis psicométricos de las muestras HV, ESP, MAT, NAT y SOC, por familia de ítems

1. Familias de HV

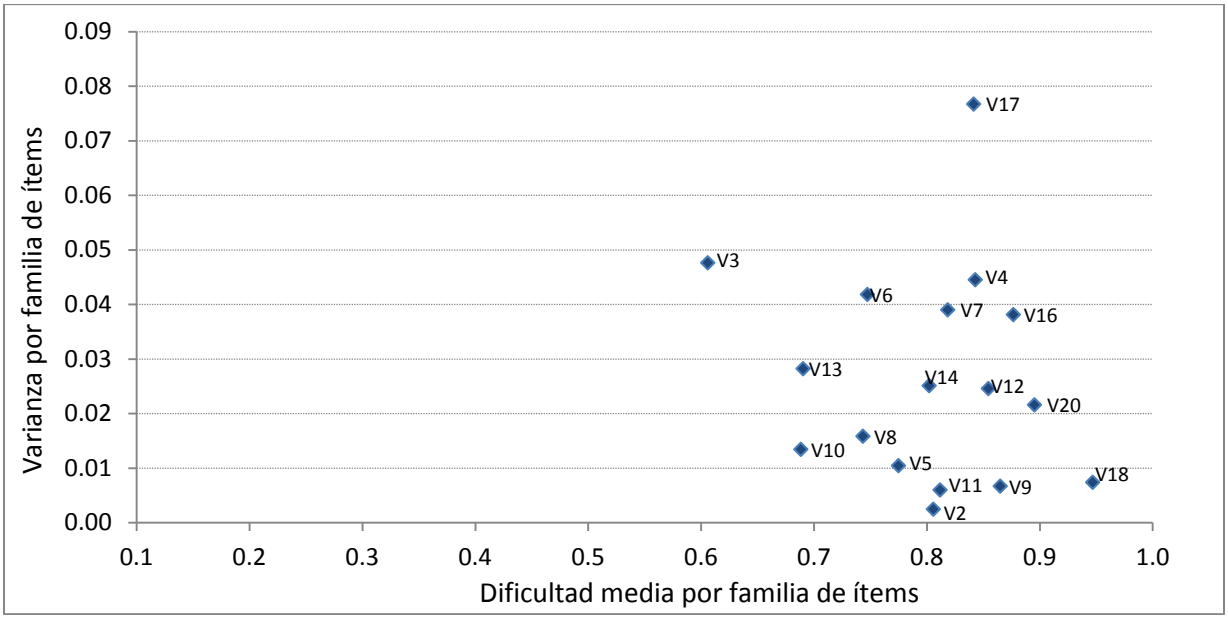


Figura F.1. Gráfica de la dificultad media por familia de 6 ítems de la muestra HV vs. varianza.

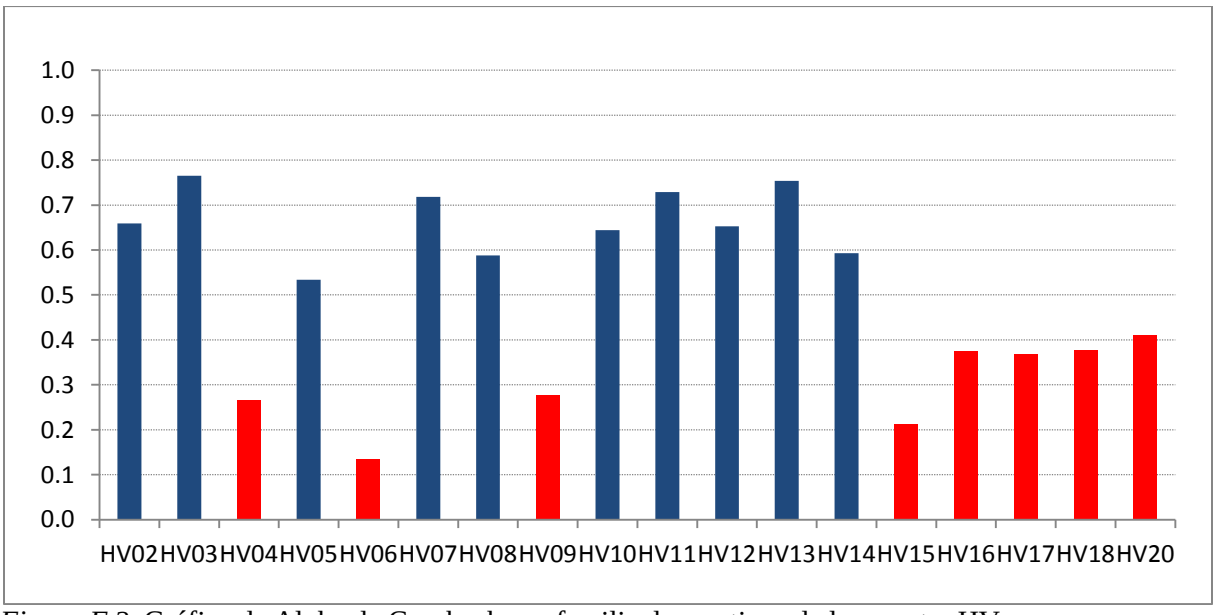


Figura F.2. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra HV.

Tabla F.1.
Infit y Outfit para cada ítem de cada familia de la muestra HV

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
HV02	1.14	1.06	0.99	0.93	0.88	0.99	1.13	1.10	1.03	0.83	0.84	0.99
HV03	1.31	**1.53**	1.04	**0.69**	**0.73**	**0.77**	**2.17**	**1.50**	0.95	**0.62**	**0.65**	0.75
HV04	1.15	0.76	0.78	1.26	0.89	1.07	0.89	0.22	0.70	1.25	0.75	0.97
HV05	1.13	1.10	0.95	0.93	0.92	0.97	1.24	1.19	0.93	0.90	0.89	1.07
HV06	1.10	1.01	0.89	0.96	0.87	1.18	1.02	0.88	0.82	0.79	0.55	1.05
HV07	0.98	0.84	1.09	1.04	0.85	1.21	0.98	0.83	1.07	0.91	0.87	1.17
HV08	1.18	0.95	1.01	0.98	0.90	1.00	1.25	0.92	1.04	0.97	0.91	1.22
HV09	1.15	0.91	1.03	1.11	1.01	0.75	1.00	0.76	0.85	1.12	1.04	0.57
HV10	1.01	0.88	1.22	1.09	0.85	0.97	0.97	0.84	**1.37**	1.10	0.84	0.92
HV11	0.86	0.87	1.03	0.99	0.89	1.25	0.90	0.81	0.97	0.95	0.82	1.26
HV12	0.87	1.16	1.25	0.71	1.07	0.89	0.87	1.10	1.23	**0.57**	1.18	0.88
HV13	1.07	1.29	1.13	0.78	**0.66**	0.96	1.03	1.27	1.15	0.72	**0.59**	0.96
HV14	0.90	1.09	1.12	0.95	0.90	1.03	0.90	1.05	1.16	0.93	0.87	0.95
HV15	0.86	1.03	1.03	1.10	1.03	0.89	0.84	1.27	0.92	1.40	1.32	1.11
HV16	0.98	1.18	1.09	0.85	0.85	1.04	0.66	1.13	1.08	0.74	0.31	0.68
HV17	1.07	1.00	0.94	0.97	1.04	1.00	1.25	0.98	0.96	0.91	1.01	1.40
HV18	1.00	0.82	0.97	1.12	0.93	1.06	0.98	0.82	1.01	1.66	1.00	1.01
HV20	1.36	1.07	0.89	1.04	0.94	0.86	1.71	1.09	0.90	1.09	0.91	0.78

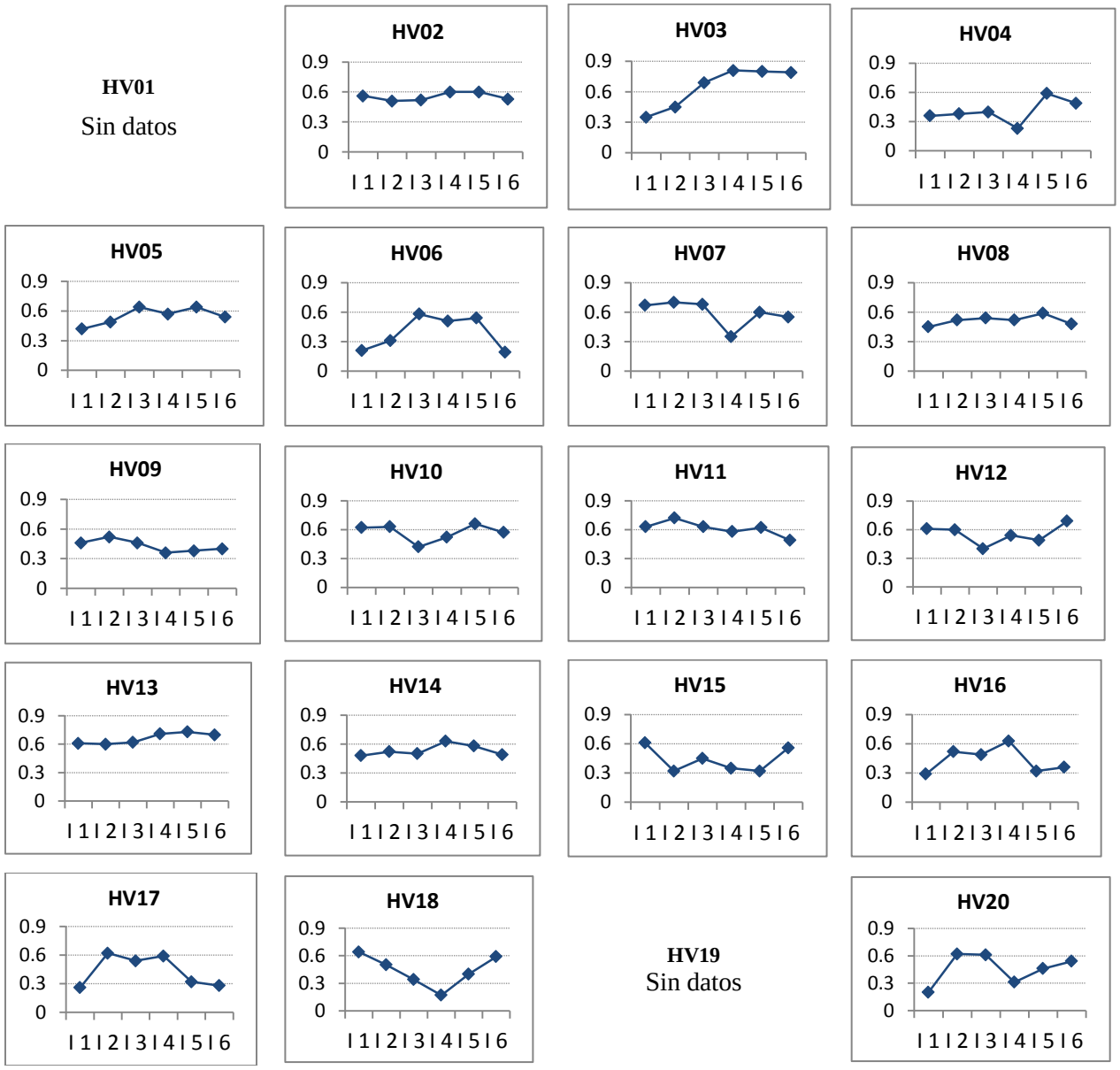


Figura F.3. Gráfica de correlaciones ítem medida por familia, la muestra HV.

Tabla F.2.
Discriminación de los ítems de cada una de las familias de la muestra HV.

	I1	I2	I3	I4	I5	I6
HV01	--	--	--	--	--	--
HV02	0.81	0.89	0.98	1.13	1.17	0.99
HV03	**0.70**	**0.33**	0.98	1.41	1.39	1.28
HV04	0.98	1.11	1.09	**0.74**	1.14	0.97
HV05	0.84	0.92	1.07	1.10	1.11	0.99
HV06	0.87	1.01	1.14	1.06	1.09	0.86
HV07	1.03	1.17	0.87	1.01	1.06	0.81
HV08	0.89	1.08	0.95	1.03	1.06	0.99
HV09	0.86	1.18	1.06	0.86	0.93	1.12
HV10	1.01	1.10	**0.78**	0.88	1.17	1.08
HV11	1.09	1.20	0.99	0.97	1.11	**0.66**
HV12	1.15	0.90	0.85	1.18	0.82	1.10
HV13	0.98	**0.63**	0.81	1.18	1.31	1.05
HV14	1.04	0.91	0.90	1.08	1.13	1.01
HV15	1.26	0.93	0.98	0.87	0.93	1.12
HV16	1.03	0.82	0.89	1.18	1.07	1.01
HV17	0.94	1.01	1.04	1.05	1.00	0.97
HV18	1.01	1.05	1.00	0.95	1.01	0.97
HV19	--	--	--	--	--	--
HV20	0.92	0.91	1.15	0.95	0.95	1.13

Tabla F.3.
Índices de ajuste de AFC por familia del área de Habilidades del lenguaje, con sus respectivas cargas factoriales

	Indices de ajuste						Cargas factoriales					
	Chi	Lib	p	NNFI	CFI	RMSEA	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6
HV01	--	--	--	--	--	--	--	--	--	--	--	--
HV02	3.744	5	0.586	1.031	1.000	0.000	.462	.488	.352	.458	.369	.442
HV03	0.376	3	0.945	1.046	1.000	0.000	.450	.616	.278	.276	.262	295
HV04	4.582	4	0.332	0.950	0.987	0.031	**.146**	.455	.844	.325	**.196**	.230
HV05 [a]	1.724	6	0.943	1.111	1.000	0.000	.328	.251	.418	.492	.388	.202
HV06	4.184	4	0.381	0.970	0.992	0.017	.667	**.198**	.209	.373	.502	**.009**
HV07	5.594	8	0.692	1.028	1.000	0.000	.505	.606	.591	.413	.683	.382
HV08	4.916	7	0.670	1.063	1.000	0.000	.319	.537	.399	.398	.357	.462
HV09	2.315	5	0.804	1.264	1.000	0.000	.202	**.128**	**.017**	.282	**.075**	1.000
HV10	3.502	8	0.899	1.078	1.000	0.000	.440	.587	.205	.505	.784	.462
HV11	8.652	6	0.194	0.964	0.986	0.052	.788	.375	.431	.487	.641	.385
HV12	2.936	4	0.568	1.029	1.000	0.000	.568	.291	**.152**	.721	.448	.755
HV13	2.164	4	0.705	1.032	1.000	0.000	.830	.388	.295	.487	.462	.793
HV14	4.748	8	0.784	1.090	1.000	0.000	.481	.436	.336	.475	.547	.477
HV15 [b]	1.406	5	0.923	2.118	1.000	0.000	**.083**	**.107**	.817	**.068**	**.001**	**-1.000**
HV16	6.193	6	0.401	0.984	0.994	0.014	.388	**.004**	.936	.245	.238	.558
HV17	0.714	4	0.949	1.318	1.000	0.000	.487	**.148**	1.000	**.137**	.205	**.003**
HV18	7.000	9	0.637	1.142	1.000	0.000	.324	.662	.270	**-.052**	.340	.294
HV19	--	--	--	--	--	--	--	--	--	--	--	--
HV20	8.548	9	0.480	1.018	1.000	0.000	**.003**	.311	.636	**.182**	.296	.608

[a] Item2,F1; Item3,F1; Item4,F1; Item5,F1; Item6,F1: linealmente dependientes de otros parámetros.
[b] Como se trata de datos dicotómicos, se utilizó la matriz de correlaciones de Pearson para ejecutar los análisis.

2. Familias de ESP

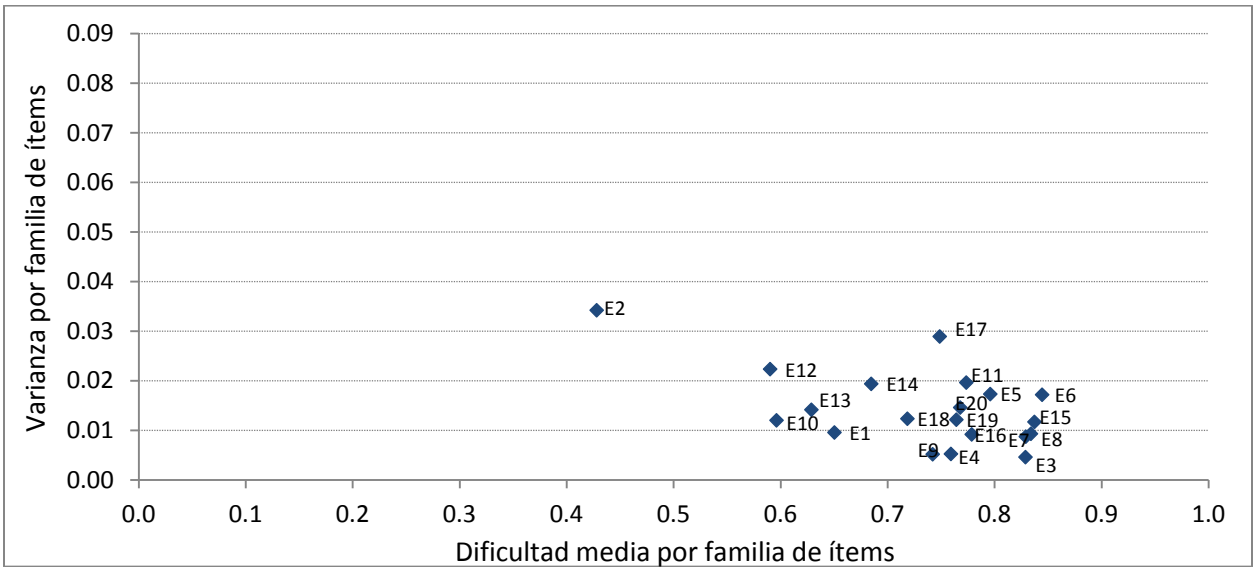


Figura F.4. Gráfica de la dificultad media por familia de 6 ítems de la muestra ESP vs. Varianza.

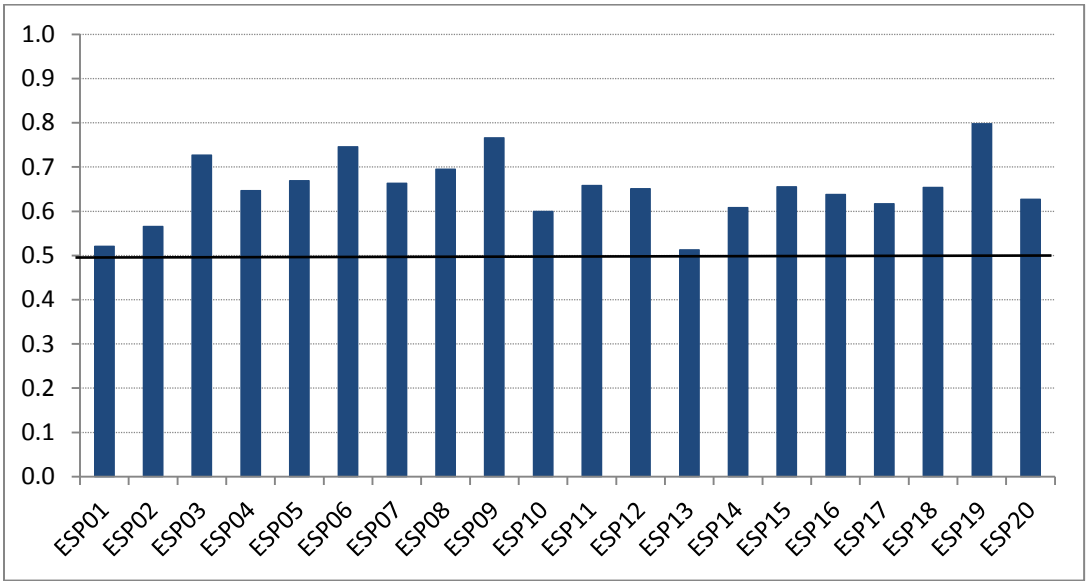


Figura F.5. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra ESP.

Tabla F.4.
Infit y Outfit para cada ítem de cada familia de la muestra ESP

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
ESP01	1.13	1.02	1.01	0.91	0.97	0.94	1.12	1.03	1.03	0.88	0.98	0.87
ESP02	0.92	0.87	0.82	0.85	0.81	**1.53**	0.95	0.85	0.79	0.82	**0.53**	**2.49**
ESP03	1.08	1.04	1.02	0.88	0.88	1.19	1.07	1.09	1.02	0.87	0.85	1.03
ESP04	1.30	1.11	1.04	0.97	0.83	0.83	**1.33**	1.12	1.02	0.94	0.75	0.74
ESP05	1.79	1.03	1.10	1.00	0.80	0.84	1.90	1.04	1.10	1.37	0.65	0.55
ESP06	1.07	**1.45**	0.89	0.91	0.94	0.92	0.69	1.43	0.97	1.31	0.89	0.85
ESP07	1.29	1.02	1.03	0.79	0.81	0.90	1.21	0.73	1.01	0.64	0.70	0.66
ESP08	1.06	1.02	1.09	1.16	0.79	0.86	1.02	0.79	1.09	1.12	0.37	0.66
ESP09	1.04	0.94	0.99	0.96	0.92	1.10	0.91	0.88	0.99	0.95	1.07	1.09
ESP10	1.09	1.23	0.94	0.90	0.95	0.92	1.03	1.21	0.94	0.87	0.94	0.90
ESP11	1.05	0.96	1.05	0.98	0.82	1.14	1.01	0.71	1.26	1.10	0.82	1.13
ESP12	1.07	1.05	0.87	0.89	1.05	1.05	**1.34**	1.09	0.86	0.74	1.01	1.01
ESP13	0.95	1.02	0.99	1.00	0.99	1.08	0.93	1.01	0.94	1.02	0.99	1.07
ESP14	1.13	0.88	0.97	0.93	1.14	0.90	1.17	0.88	0.98	0.99	**1.63**	0.91
ESP15	1.25	1.15	0.98	1.12	0.76	**0.78**	1.13	0.92	1.12	0.77	0.44	**0.46**
ESP16	1.17	1.00	1.10	0.88	0.88	0.84	1.15	1.01	1.11	0.76	0.85	0.70
ESP17	1.55	0.95	0.86	1.06	0.88	1.03	1.52	0.99	1.00	1.02	0.83	1.06
ESP18	1.17	0.94	1.05	0.85	0.85	1.12	1.19	0.85	1.28	0.80	0.75	1.00
ESP19	1.15	0.95	0.80	0.94	0.93	1.19	1.14	0.96	0.79	0.99	0.82	**2.04**
ESP20	1.40	1.12	1.05	0.82	**0.74**	0.93	1.58	1.07	1.03	0.75	**0.56**	0.90

Nota: los *infit* y *outfit* superiores a 1.30, o inferiores a 0.8 se encuentran marcados en celeste. I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente.

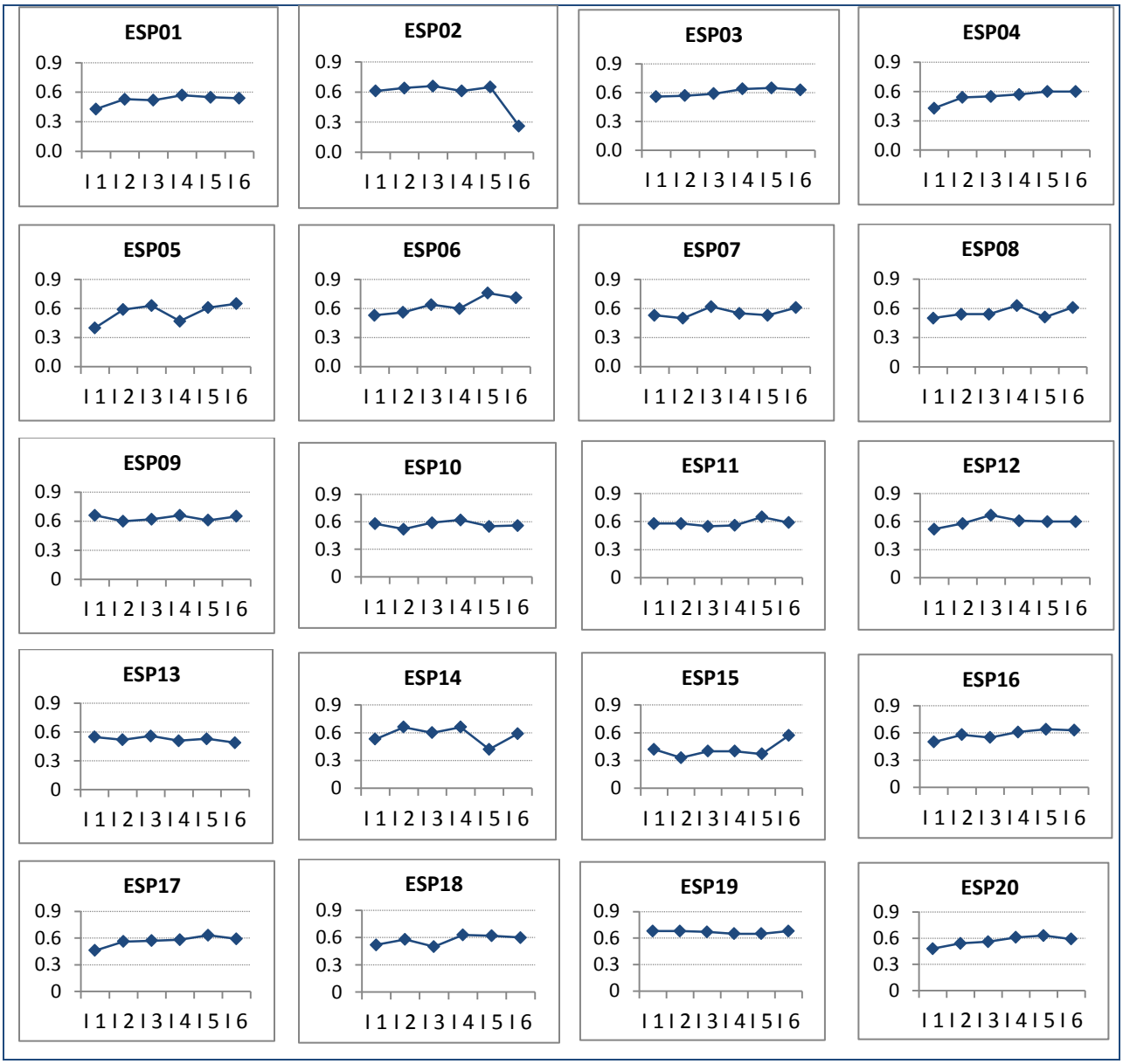


Figura F.6. Gráfica de correlaciones punto medida por familia, de la muestra ESP.

Tabla F.5.
Discriminación de los ítems de cada una de las familias de la muestra ESP

FLIA	I 1	I 2	I 3	I 4	I 5	I 6
ESP01	0.92	0.97	0.99	1.13	1.00	1.05
ESP02	1.10	1.27	1.30	1.23	1.23	**-0.26**
ESP03	0.99	0.98	1.01	1.08	1.08	0.87
ESP04	**0.79**	0.97	0.95	1.00	1.12	1.15
ESP05	0.89	0.98	0.88	0.97	1.08	1.09
ESP06	1.02	**0.78**	1.02	0.96	1.11	1.08
ESP07	**0.65**	1.01	0.95	1.08	1.05	1.10
ESP08	0.98	1.03	0.94	0.84	1.06	1.09
ESP09	1.05	1.03	0.94	1.00	0.95	1.02
ESP10	0.97	**0.76**	1.05	1.08	1.05	1.08
ESP11	0.96	1.05	0.98	0.97	1.16	**0.78**
ESP12	0.86	0.87	1.17	1.15	0.94	0.92
ESP13	1.07	0.97	1.03	0.98	1.01	0.91
ESP14	0.80	1.14	1.01	1.07	0.86	1.09
ESP15	0.90	0.98	0.95	0.97	1.07	1.15
ESP16	**0.73**	0.96	0.88	1.05	1.12	1.08
ESP17	0.91	1.02	1.04	0.89	1.17	0.97
ESP18	0.80	1.04	0.90	1.16	1.14	0.92
ESP19	**0.79**	1.04	1.12	1.04	1.07	0.85
ESP20	0.90	0.84	0.92	1.07	1.15	1.05

Tabla F.6.
Índices de ajuste de AFC por familia de la muestra ESP

	Indices de ajuste						Cargas factoriales					
	Chi	Lib	p	NNFI	CFI	RMSEA	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6
ESP01	4.396	7	0.733	1.059	1.000	0.000	.296	.538	.161	.394	.491	.406
ESP02[a]	9.225	7	0.236	0.976	0.989	0.041	.418	.704	.640	.365	.473	**-.138**
ESP03	6.262	8	0.617	1.013	1.000	0.000	.626	.596	.618	.620	.587	.369
ESP04	7.733	7	0.356	0.990	0.995	0.022	.310	.450	.568	.657	.433	.341
ESP05	8.328	6	0.215	0.975	0.990	0.042	.721	.743	.619	.630	.380	.292
ESP06	13.037	9	0.160	0.982	0.989	0.406	.746	.628	.787	.764	.497	.403
ESP07	3.534	7	0.831	1.033	1.000	0.000	.253	.642	.548	.765	.636	.325
ESP08	3.644	5	0.601	1.018	1.000	0.000	.635	.710	.547	.410	.515	.372
ESP09	6.063	5	0.300	0.990	0.997	0.031	.905	.644	.573	.704	.674	.498
ESP10	8.903	8	0.350	0.986	0.992	0.023	.485	.362	.598	.478	.549	.247
ESP11	5.572	7	0.590	1.015	1.000	0.000	.431	.561	.506	.594	.644	.248
ESP12	5.610	8	0.690	1.029	1.000	0.000	.417	.441	.509	.619	.553	.308
ESP13	8.807	8	0.358	0.979	0.989	0.022	.599	.533	.555	.311	.314	**.181**
ESP14	5.626	7	0.584	1.022	1.000	0.000	.428	.325	.606	.520	.325	.268
ESP15	4.546	5	0.473	1.006	1.000	0.000	.883	.420	.410	.339	.972	.298
ESP16	6.115	4	0.190	0.952	0.987	0.049	.460	.596	.520	.655	.482	**.145**
ESP17	11.232	9	0.260	0.978	0.987	0.034	.512	.698	.724	.373	.474	.232
ESP18	12.386	9	0.192	0.961	0.977	0.042	.402	.522	.429	.657	.616	.366
ESP19	2.135	1	0.144	0.963	0.998	**0.072**	.922	.581	.499	.476	.350	.545
ESP20	7.404	6	0.285	0.982	0.993	0.033	.462	.526	.680	.605	.427	.212

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. [a] Como se trata de datos dicotómicos, se utilizó la matriz de correlaciones de Pearson para ejecutar los análisis

3. Familias de MAT

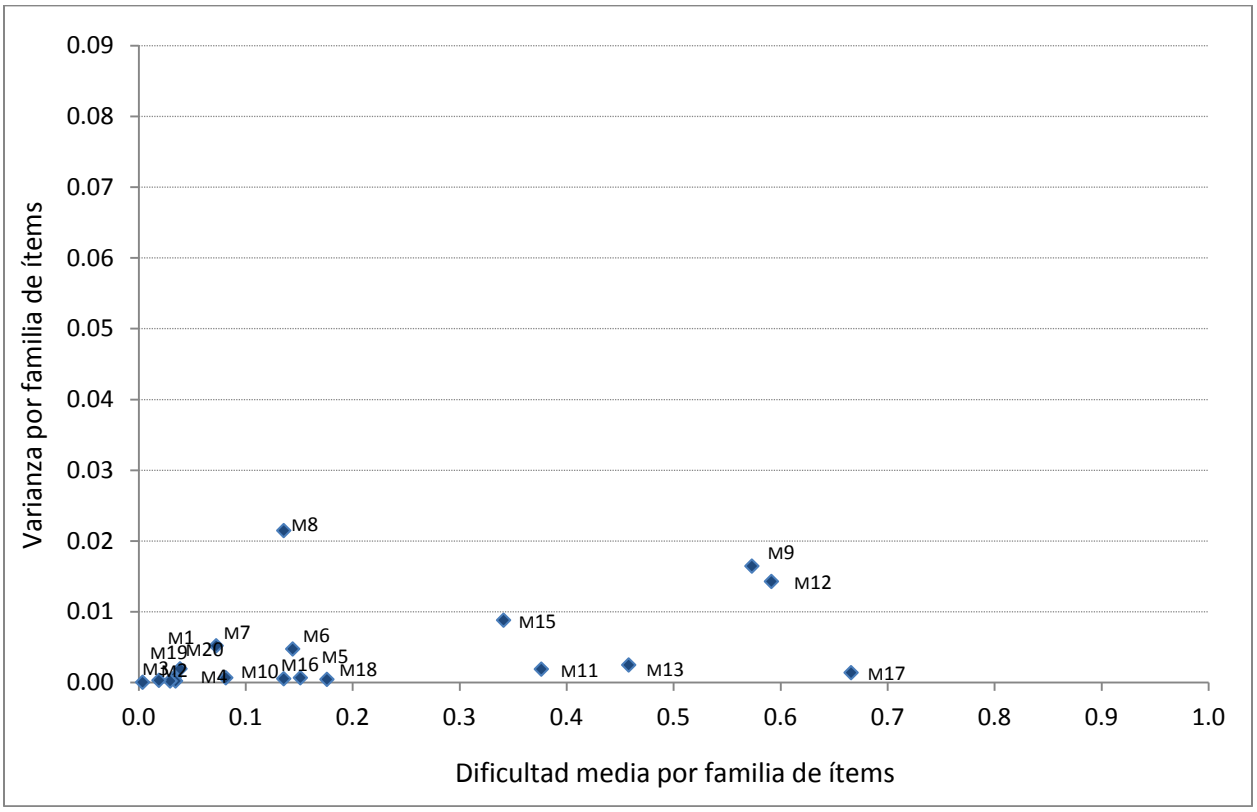


Figura F.7. Gráfica de la dificultad media vs. Varianza, por familia de 6 ítems de la muestra MAT

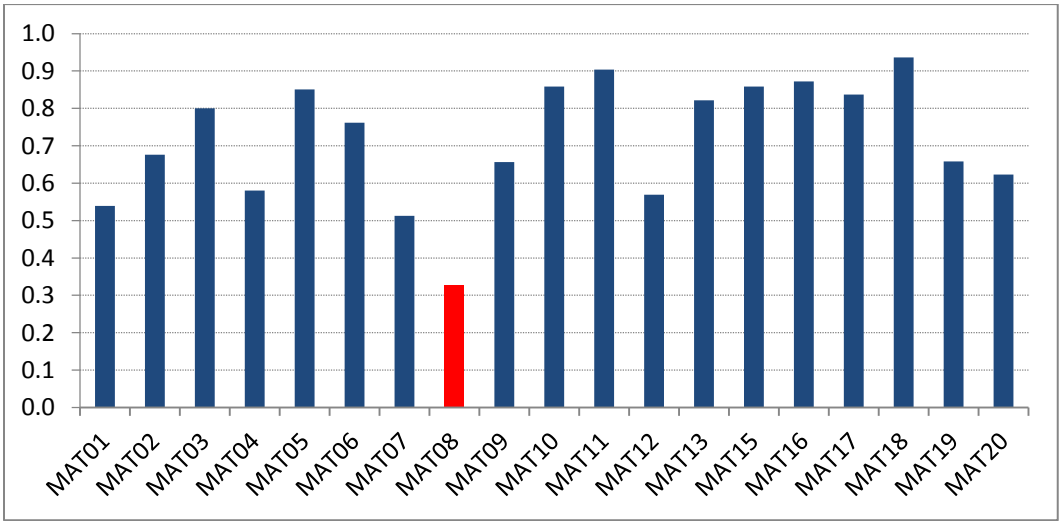


Figura F.8. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra MAT.

Tabla F.7.
Infit y Outfit para cada ítem de cada familia de la muestra MAT

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
MAT01	1.19	0.86	0.69	0.73	1.39	1.13	1.16	0.83	0.46	0.53	1.60	0.79
MAT02	1.12	1.13	0.74	0.91	0.87	1.17	1.11	1.12	0.65	1.09	0.87	1.60
MAT03	---	---	---	---	---	---	---	---	---	---	---	---
MAT04	1.03	0.73	0.88	1.14	1.04	1.18	1.05	0.68	0.88	1.15	0.91	1.55
MAT05	1.31	0.90	0.92	0.82	0.95	1.10	1.37	0.87	0.98	0.84	0.85	1.07
MAT06	0.96	1.08	1.06	1.00	0.81	1.01	0.95	1.25	1.02	0.87	0.73	1.13
MAT07	1.03	1.09	1.00	0.83	1.10	0.97	0.98	1.01	0.99	0.81	1.07	0.96
MAT08	1.14	0.85	0.95	1.09	---	1.14	1.13	0.85	0.95	1.23	---	0.78
MAT09	1.22	0.92	0.84	0.94	1.04	1.02	1.22	0.92	0.81	0.79	1.02	1.12
MAT10	1.02	0.73	**0.59**	0.93	1.21	**1.46**	1.04	0.69	**0.52**	0.82	**3.19**	**1.49**
MAT11	0.94	**1.31**	**0.75**	**0.63**	1.18	1.24	0.89	**1.46**	**0.67**	**0.57**	1.13	1.17
MAT12	1.12	0.88	1.21	1.00	0.93	0.91	1.02	**0.78**	1.25	1.01	0.87	0.84
MAT13	1.11	0.94	0.93	0.88	1.06	1.06	1.16	0.95	0.87	0.84	1.14	1.06
MAT14	S/D											
MAT15	1.27	0.85	1.00	0.95	0.83	0.96	**2.59**	0.81	1.14	0.95	0.86	1.01
MAT16	0.92	**0.64**	**0.55**	**0.58**	**2.43**	0.83	0.82	**0.58**	**0.49**	**0.52**	**3.75**	0.78
MAT17	1.06	1.06	0.95	0.97	0.98	0.98	1.06	1.10	0.97	0.95	0.96	0.96
MAT18	1.14	0.81	0.84	**0.56**	1.20	1.29	1.15	0.89	0.87	**0.52**	1.11	**1.31**
MAT19	---	0.78	0.63	1.42	---	1.08	---	0.74	0.55	**1.83**	---	1.20
MAT20	0.96	0.96	1.03	0.61	0.65	1.14	0.16	0.16	1.04	0.28	1.33	1.16

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente.

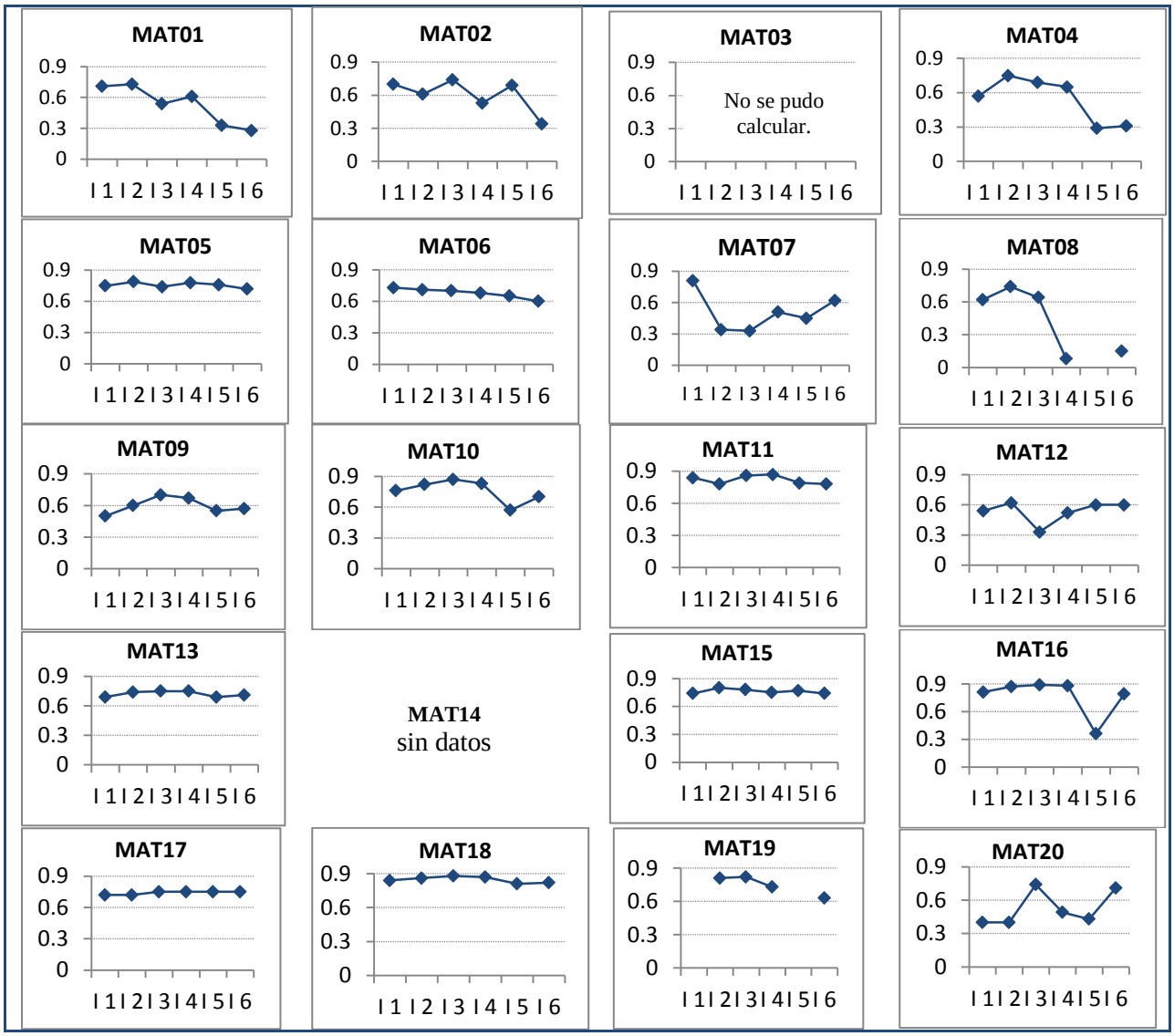


Figura F.9. Gráfica de correlaciones ítem medida por familia, de la muestra MAT.

Tabla F.8.

Discriminación de los ítems de cada una de las familias de la muestra MAT

	I1	I2	I3	I4	I5	I6
MAT01	0.14	1.51	1.24	1.29	**0.60**	0.96
MAT02	**0.56**	**0.70**	1.70	1.05	1.31	**0.79**
MAT03	--	--	--	--	--	--
MAT04	0.90	2.09	1.45	**0.25**	0.98	**0.76**
MAT05	**0.37**	1.20	1.10	1.28	1.13	0.85
MAT06	1.08	**0.74**	0.89	1.04	1.15	0.98
MAT07	0.95	0.95	1.00	1.14	0.92	1.06
MAT08	**0.75**	1.30	1.05	0.92	--	0.94
MAT09	0.82	1.09	1.18	1.04	0.93	0.91
MAT10	0.95	1.56	1.97	1.22	**0.62**	**0.02**
MAT11	1.11	**0.47**	1.41	1.59	**0.74**	**0.66**
MAT12	0.95	1.08	**0.78**	0.99	1.13	1.12
MAT13	**0.73**	1.12	1.16	1.27	0.84	0.87
MAT14	S/D					
MAT15	**0.46**	1.25	0.95	1.07	1.24	1.04
MAT16	1.16	1.57	1.72	1.65	**-1.32**	1.26
MAT17	0.91	0.87	1.07	1.08	1.05	1.06
MAT18	0.84	1.15	1.19	1.39	**0.78**	**0.66**
MAT19	--	1.64	1.84	**-0.45**	--	0.87
MAT20	1.09	1.09	0.93	1.15	1.06	**0.71**

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente.

Tabla F.9.

Índices de ajuste de AFC por familia de la muestra MAT, con sus respectivas cargas factoriales

	Indices de ajuste						Cargas factoriales[a]					
	Chi	Lib	p	NNFI	CFI	RMSEA	IT1	IT2	IT3	IT4	IT5	IT6
MAT01	0.153	7	0.999	1.180	1.000	0.000	.347	.910	.576	.459	**-.033**	**-.024**
MAT02	0.017	2	0.991	1.094	1.000	0.000	.330	.418	.413	.564	.413	.708
MAT03[b]	--	--	--	--	--	--	--	--	--	--	--	--
MAT04[c]	--	--	--	--	--	--	--	--	--	--	--	--
MAT05	2.049	4	0.726	1.021	1.000	0.000	.603	.763	.639	.972	.565	.586
MAT06	8.306	9	0.503	1.004	1.000	0.000	.626	.562	.594	.635	.656	.566
MAT07	7.077	7	0.420	0.998	0.999	0.007	.503	.297	**.170**	.559	**.102**	.557
MAT08	2.465	5	0.781	1.118	1.000	0.000	**.171**	1.000	.324	**.092**	--	**.038**
MAT09	4.128	5	0.531	1.011	1.000	0.000	.603	.684	.547	.389	.331	.735
MAT10	45.673	7	0.000	0.890	0.949	0.154	.678	.681	.853	.851	.401	.596
MAT11	5.207	4	0.266	0.995	0.999	0.036	.838	.754	.946	.872	.605	.602
MAT12	7.750	7	0.355	0.987	0.994	0.021	.505	.608	.215	.431	.344	.377
MAT13	3.510	4	0.476	1.004	1.000	0.000	.641	.731	.593	.674	.607	.492
MAT14	--	--	--	--	--	--	--	--	--	--	--	--
MAT15	2.577	6	0.859	1.015	1.000	0.000	.648	.837	.775	.615	.615	.578
MAT16	5.224	7	0.632	1.004	1.000	0.000	.773	.847	.869	.919	**.135**	.771
MAT17	6.440	3	0.090	0.964	0.993	0.069	.743	.654	.578	.659	.571	.631
MAT18	3.930	2	0.140	0.988	0.998	0.064	1.000	.872	.894	.928	.799	.752
MAT19[d]	5.147	2	0.076	0.967	0.989	0.081	--	.841	.907	.413	--	.444
MAT20[e]	2.537	2	0.281	0.991	0.997	0.030	--	--	.338	.898	.737	.249

Nota: NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. IT = item.

[a] Para datos dicotómicos se utilizó la matriz de correlaciones de Pearson para ejecutar los análisis.

[b] Se cancela la estimación porque la matriz no está definida positiva.

[c] Item2,F1; ítem3,F1; ítem4,F1; ítem5,F1; ítem6,F1: linealmente dependiente de otros parámetros.

[d] No se pudo calcular para ítems con dificultad cero.

[e] Se eliminaron ítems con dificultad tendiente a cero.

4. Familias de NAT

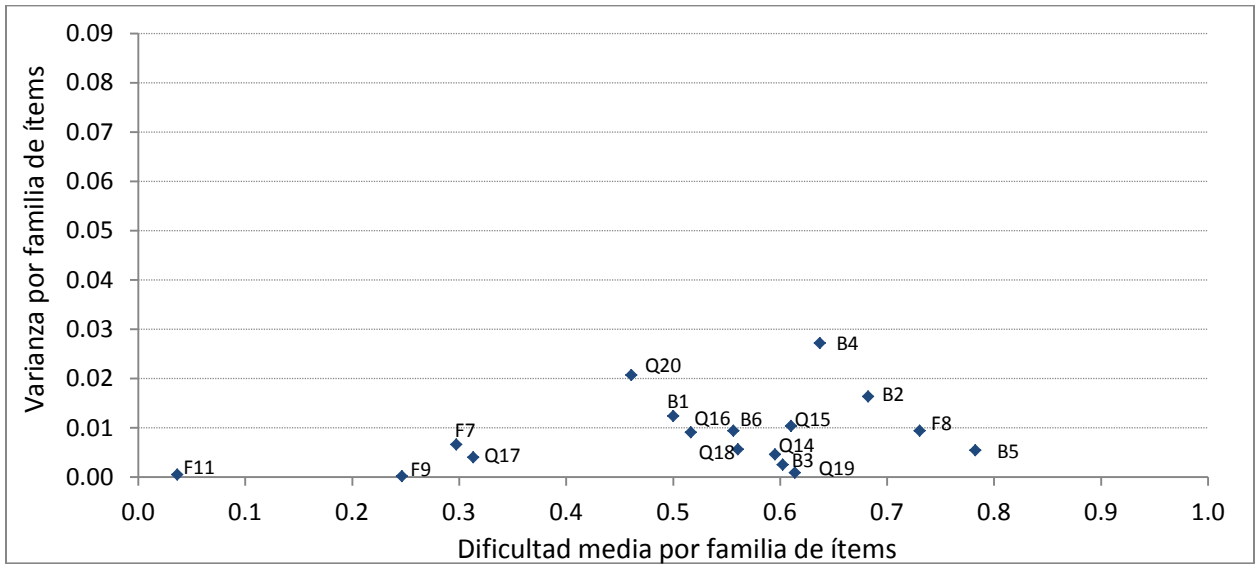


Figura F.10. Gráfica de la dificultad media vs. Varianza, por familia de 6 ítems de la muestra NAT de CESUES, San Luis Río Colorado.

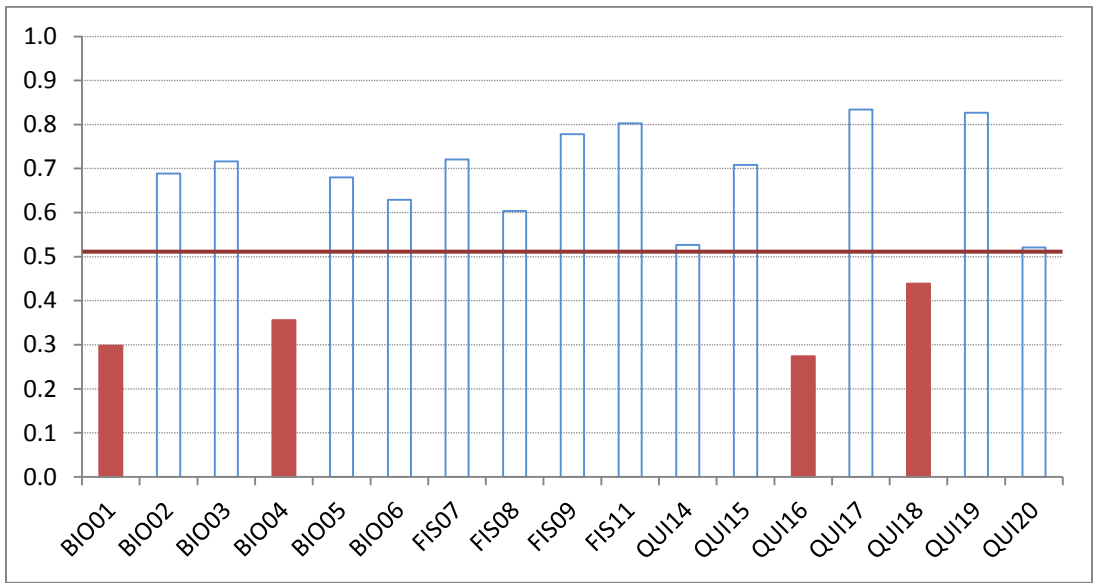


Figura F.11. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra NAT de CESUES, San Luis Río Colorado.

Tabla F.10.
Infit y Outfit para cada ítem de cada familia del área de la muestra NAT

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
BIO01[a]	0.91	0.94	1.39	0.89	0.91	0.92	1.00	1.01	**1.48**	0.89	0.89	0.94
BIO02[a]	1.05	0.96	0.92	0.93	0.92	1.12	1.04	0.91	1.01	0.95	0.91	**1.59**
BIO03[a]	0.89	1.16	0.94	1.12	1.02	0.88	0.91	1.18	0.93	1.12	1.00	0.85
BIO04[a]	1.15	0.95	0.95	0.97	0.91	1.02	1.11	0.85	0.96	0.97	0.94	1.05
BIO05[a]	**1.32**	0.83	0.93	1.22	0.76	0.82	**1.33**	0.73	0.90	1.27	**0.76**	0.81
BIO06[a]	1.23	0.91	**0.70**	1.24	1.00	0.89	1.17	0.93	**0.72**	1.52	1.00	0.88
FIS07[b]	**1.34**	1.24	**0.64**	**0.69**	0.88	1.11	**2.03**	1.30	**0.62**	**0.67**	0.88	1.11
FIS08[a]	1.18	1.23	0.89	0.96	0.71	0.92	1.20	1.17	0.84	0.84	0.64	1.09
FIS09[b]	1.04	0.98	0.95	**0.74**	0.83	**1.38**	1.14	0.83	0.89	0.82	0.77	**1.56**
FIS10[c]	0.71	1.36	0.95	0.87	1.25	0.97	0.58	1.52	0.89	0.72	3.73	0.83
FIS11[a]	1.19	**1.45**	0.75	0.65	**0.64**	1.16	1.16	1.55	0.89	0.39	0.55	1.24
FIS12[d]	1.16	1.23	0.94	0.68	0.90	0.89	1.17	1.21	1.20	**0.62**	0.86	0.85
QUI13	S/D											
QUI14[b]	1.14	1.03	0.91	0.98	0.89	1.06	1.12	1.03	0.91	0.97	0.93	1.03
QUI15[a]	1.24	0.83	0.98	0.95	1.00	0.93	1.26	0.78	1.03	0.97	1.10	0.94
QUI16[b]	1.04	0.92	1.10	0.94	1.14	0.86	1.00	0.92	1.12	0.94	1.12	0.84
QUI17[a]	0.82	**1.34**	1.07	0.94	0.84	1.02	0.88	1.33	1.13	0.91	0.89	0.98
QUI18[a]	0.87	1.10	1.11	0.96	0.86	0.98	0.88	1.16	1.10	1.03	0.87	1.00
QUI19[b]	**1.55**	1.13	0.89	**0.65**	0.95	0.84	**1.70**	1.20	0.93	0.64	0.90	0.72
QUI20[b]	0.82	1.09	1.18	0.91	1.11	0.95	0.74	1.12	1.36	0.86	1.13	0.88

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente. S/D: sin datos disponibles.

[a] Muestra de 160 estudiantes (CESUES de SLRC y UACJ). [b] Muestra de 100 estudiantes (CESUES de SLRC). [c] Muestra de 60 estudiantes (UACJ). [d] Muestra de 239 estudiantes (CESUES, Hermosillo).

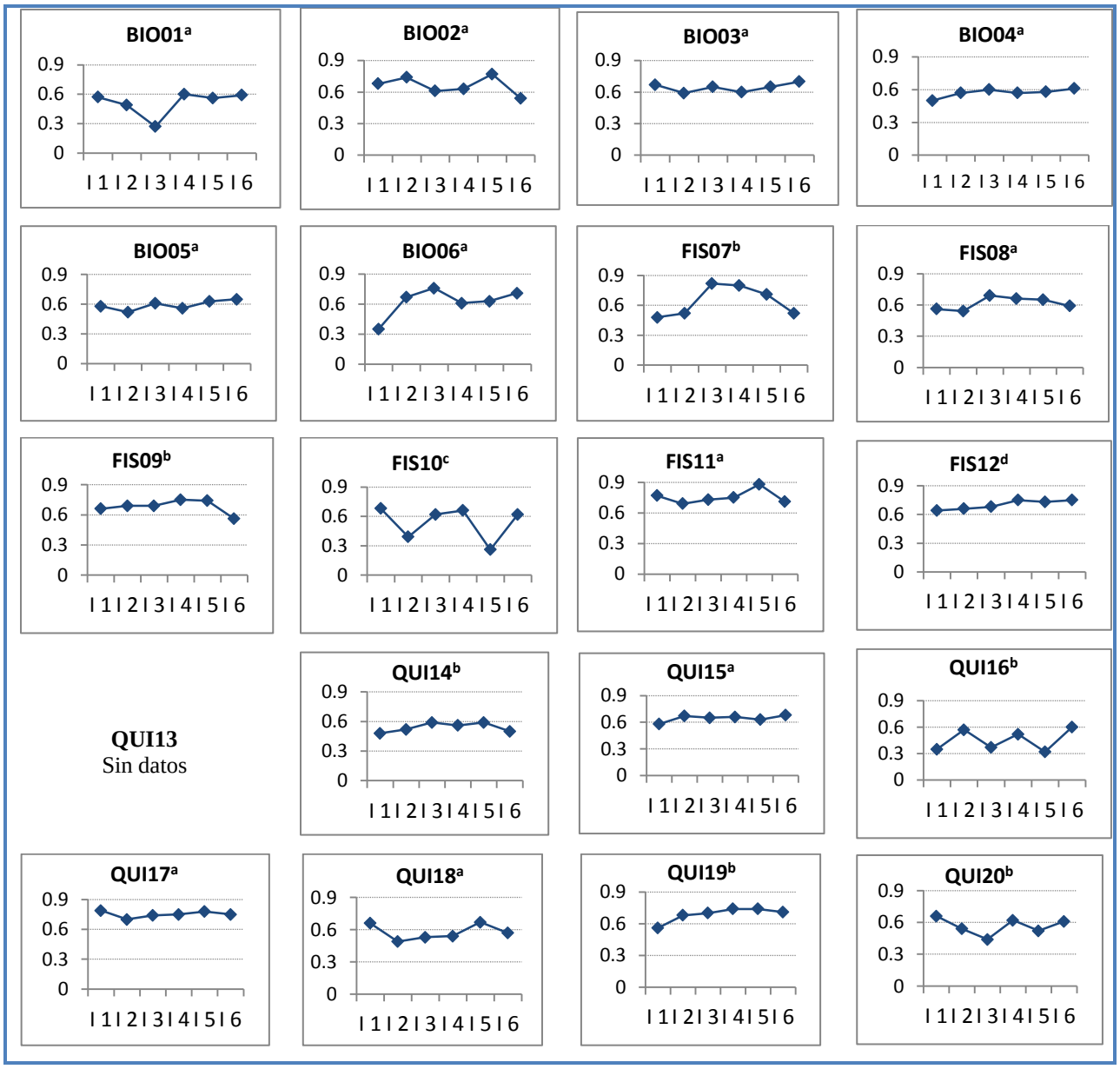


Figura F.12. Gráfica de correlaciones ítem medida por familia, de la muestra NAT.
[a] Muestra de 160 estudiantes (CESUES de SLRC y UACJ). [b] Muestra de 100 estudiantes (CESUES de SLRC). [c] Muestra de 60 estudiantes (UACJ). [d] Muestra de 239 estudiantes (CESUES, Hermosillo).

Tabla F.11.
Discriminación de los ítems de cada una de las familias de la muestra NAT

	I1	I2	I3	I4	I5	I6
BIO01[a]	1.09	1.05	**0.56**	1.15	1.10	1.08
BIO02[a]	0.96	1.09	1.08	0.98	1.10	0.85
BIO03[a]	1.12	0.80	1.06	0.85	1.00	1.17
BIO04[a]	0.90	1.07	1.03	1.01	1.04	0.98
BIO05[a]	0.77	1.10	1.03	0.87	1.15	1.08
BIO06[a]	0.83	1.06	1.32	**0.75**	1.01	1.15
FIS07[b]	**0.45**	**0.63**	1.37	1.31	1.15	0.82
FIS08[a]	0.84	**0.79**	1.16	1.12	1.14	0.94
FIS09[b]	0.92	1.05	1.01	1.39	1.22	**0.57**
FIS10[c]	1.41	**0.52**	1.12	1.33	**0.72**	1.13
FIS11[a]	**0.65**	**0.21**	1.25	1.44	1.71	**0.73**
FIS12[d]	0.80	**0.77**	0.94	1.17	1.08	1.12
QUI13	S/D					
QUI14[b]	**0.79**	0.94	1.12	1.02	1.12	0.97
QUI15[a]	**0.70**	1.19	1.05	1.01	1.00	1.06
QUI16[b]	0.95	1.14	0.88	1.05	0.84	1.20
QUI17[a]	1.21	**0.68**	0.88	1.11	1.26	1.00
QUI18[a]	1.15	0.88	0.90	0.97	1.13	1.01
QUI19[b]	**0.32**	0.95	1.08	1.41	1.06	1.23
QUI20[b]	1.43	**0.78**	0.80	1.24	**0.77**	1.14

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente.
[a] Muestra de 160 estudiantes (CESUES de SLRC y UACJ). [b] Muestra de 100 estudiantes (CESUES de SLRC). [c] Muestra de 60 estudiantes (UACJ). [d] Muestra de 239 estudiantes (CESUES, Hermosillo).

Tabla F.12.
Índices de ajuste de AFC por familia del área de Ciencias naturales (NAT), con sus respectivas cargas factoriales

	Indices de ajuste							Cargas factoriales[a]					
	α	Chi	Lib	p	NNFI	CFI	RMSEA	I1	I2	I3	I4	I5	I6
BIO01[b]	**.500**	5.695	7	.575	1.038	1.000	.000	.275	.368	**-.155**	.386	.464	.655
BIO02[b]	.804	9.847	8	.275	0.986	.993	.038	.647	.719	.631	.601	.736	.464
BIO03[b]	.747	6.396	9	.699	1.026	1.000	.000	.660	.454	.629	.486	.595	.657
BIO04[b]	.621	8.175	8	.416	0.996	.998	.012	.400	.529	.560	.417	.683	.323
BIO05[b]	.746	8.189	6	.224	0.977	.991	.048	.424	.716	.545	.522	.581	.494
BIO06[b]	.734	12.934	6	.044	0.922	.969	**.085**	.314	.447	.619	.472	.587	.758
FIS07[b]	.721	3.947	4	.413	1.001	1.000	.000	.995	.394	.862	.837	.739	.355
FIS08[b]	.736	4.392	4	.355	0.993	.998	.025	.697	.377	.641	.691	.657	.481
FIS09[c]	.778	4.589	7	.710	1.035	1.000	.000	.678	.530	.550	.806	.682	.407
FIS10[d]	.554	6.267	9	.712	1.161	1.000	.000	.913	**-.034**	.461	.430	**-.008**	.533
FIS11[b]	.848	4.414	6	.620	1.009	1.000	.000	.659	.617	.673	.665	.929	.690
FIS12[e]	.839	9.594	4	.047	0.963	.990	**.077**	.565	.599	.568	.759	.587	.733
QUI13	S/D												
QUI14[c]	.527	5.886	6	.436	1.009	1.000	.000	.249	.485	.308	.933	.568	.411
QUI15[b]	.750	9.934	9	.355	.991	.995	.026	.370	.698	.615	.607	.562	.621
QUI16[c]	**.274**	6.158	7	.521	1.163	1.000	.000	**.196**	.479	**.166**	.417	**-.102**	.423
QUI17[b]	.860	11.372	9	.251	0.989	.993	.041	.794	.610	.688	.716	.749	.706
QUI18[b]	.616	1.756	4	.780	1.096	1.000	.000	.580	.327	.306	.324	.712	.324
QUI19[c]	.827	5.326	4	.255	0.976	.994	.058	.395	.721	.601	1.00	.793	.626
QUI20[c]	.591	2.818	5	.727	1.100	1.000	.000	.358	.523	.430	**.082**	.597	.211

Nota: α: coeficiente Alpha de Cronbach, NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation*. I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente. S/D: sin datos disponibles.

[a] En los casos de datos dicotómicos se utilizó la matriz de correlaciones de Pearson para ejecutar los análisis.

[b] Muestra de 160 estudiantes (CESUES de SLRC y UACJ).

[c] Muestra de 100 estudiantes (CESUES de SLRC).

[d] Muestra de 60 estudiantes (UACJ).

[e] Muestra de 239 estudiantes (CESUES, Hermosillo).

5. Familias de SOC

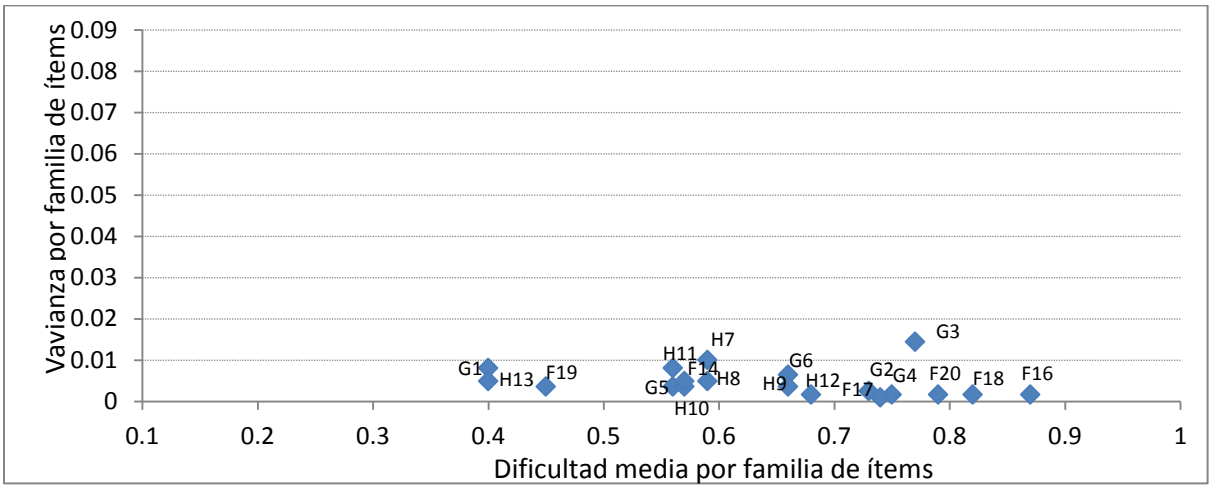


Figura F.13. Gráfica de la dificultad media por familia de 6 ítems de la muestra SOC vs. Varianza.

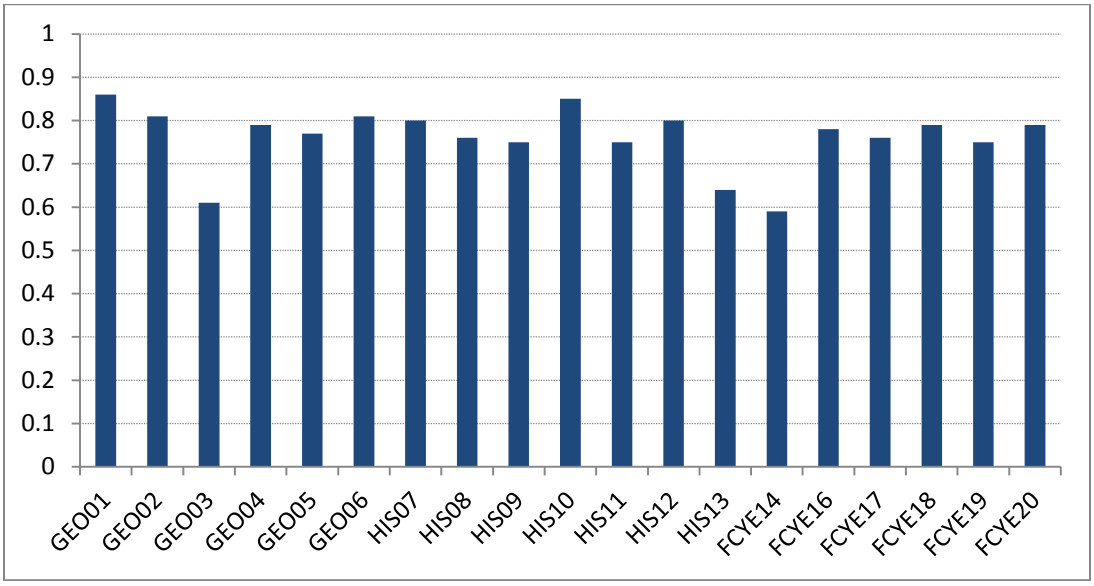


Figura F.14. Gráfica de Alpha de Cronbach por familia de reactivos de la muestra SOC.

Tabla F.13.
Infit y Outfit para cada ítem de cada familia de la muestra SOC.

FLIA	*INFIT*						*OUTFIT*					
	I1	I2	I3	I4	I5	I6	I1	I2	I3	I4	I5	I6
GEO01	0.96	1.18	0.85	0.95	0.86	1.19	0.98	1.20	0.84	0.88	0.81	1.17
GEO02	1.06	1.03	1.11	0.96	0.85	1.01	1.05	1.04	1.13	0.91	**0.70**	0.96
GEO03	1.10	1.08	1.27	0.80	0.85	0.82	1.18	0.96	**1.48**	0.62	0.77	0.73
GEO04	0.97	1.01	0.99	1.08	0.98	0.96	0.96	1.01	1.14	1.10	1.00	0.85
GEO05	1.18	0.90	1.11	**0.75**	1.03	1.03	1.18	0.89	1.06	**0.72**	1.08	1.04
GEO06	1.01	1.13	0.98	1.04	0.97	0.84	0.98	0.99	1.04	1.10	0.82	**0.65**
HIS07	1.06	1.20	0.85	1.09	1.06	0.68	1.03	1.16	0.84	1.12	1.05	0.67
HIS08	1.08	1.04	1.03	1.02	0.82	0.96	1.12	1.01	0.99	1.03	0.82	0.97
HIS09	0.89	1.02	1.01	1.01	1.09	0.91	0.87	1.02	1.01	0.99	1.09	0.88
HIS10	0.97	1.18	0.85	0.95	0.86	1.19	0.98	1.20	0.84	0.89	0.81	1.17
HIS11	1.28	0.99	0.85	0.92	0.89	1.06	**1.32**	0.95	0.86	0.92	0.88	1.04
HIS12	1.13	**1.35**	0.94	0.90	0.96	**0.78**	1.15	**1.32**	0.95	0.83	0.89	**0.77**
HIS13	1.01	0.91	1.02	1.08	1.15	**0.79**	1.02	0.91	1.01	1.08	1.16	0.82
FCYE14	1.04	1.18	0.98	0.92	0.94	0.91	1.02	1.22	1.01	1.03	0.94	0.90
FCYE16	1.08	1.15	1.03	0.84	1.08	0.84	0.92	1.03	0.93	0.75	1.05	**0.73**
FCYE17	**1.31**	1.01	0.92	0.86	0.96	0.91	1.28	1.12	1.22	0.83	1.01	0.94
FCYE18	1.17	0.94	0.81	1.14	0.93	0.98	1.15	0.97	1.10	1.16	0.85	1.01
FCYE19	1.25	1.11	0.98	**0.79**	0.92	0.91	1.28	1.21	0.97	0.80	0.92	0.90
FCYE20	1.23	1.24	0.80	0.80	0.97	0.99	1.14	1.29	1.01	0.77	0.90	0.94

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente.

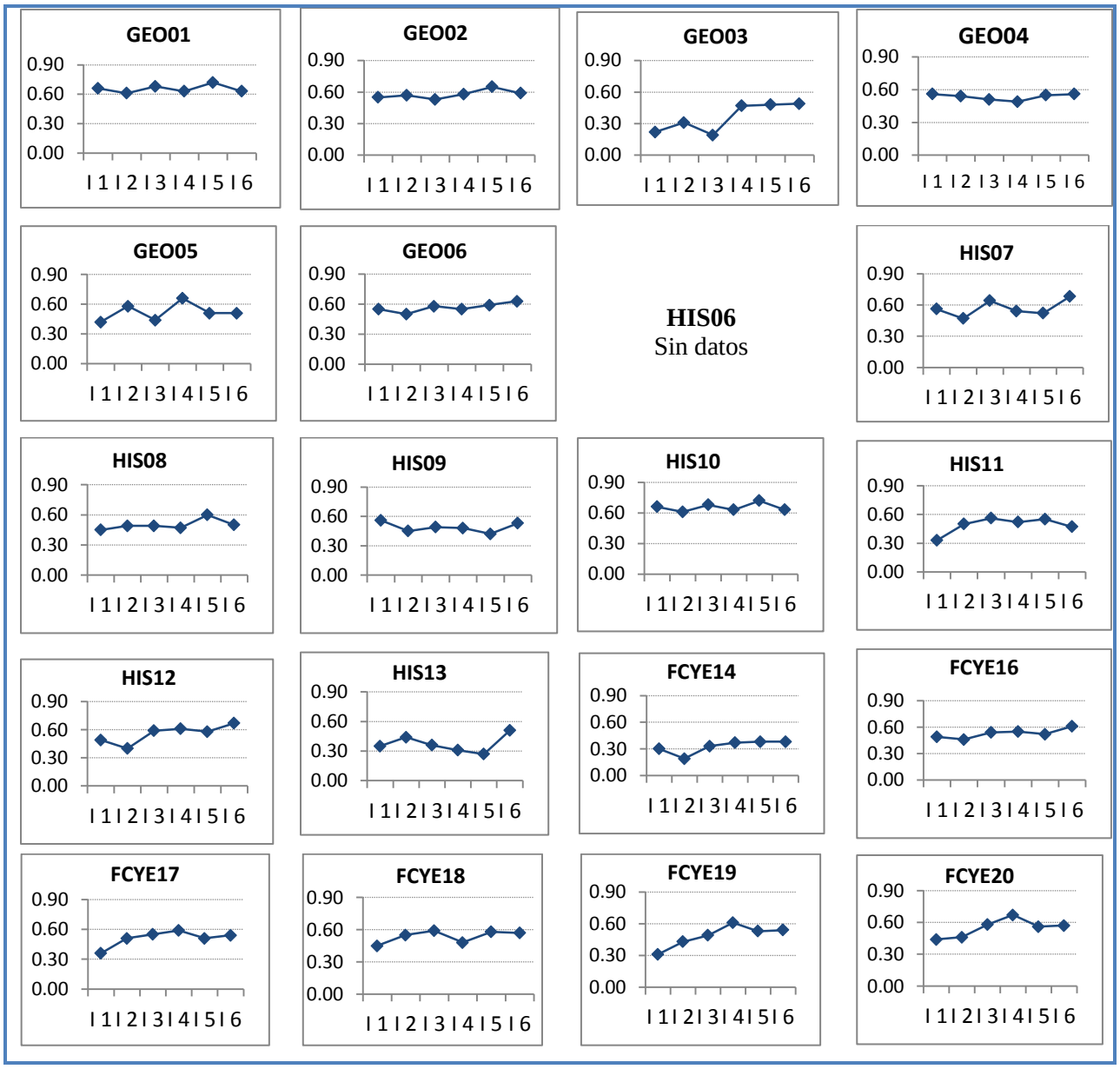


Figura F.15. Gráfica de correlación punto medida por familia de la muestra SOC.

Tabla F.14.
Discriminación de los ítems de cada una de las familias de la muestra SOC

	I1	I2	I3	I4	I5	I6
GEO01	1.01	**0.75**	1.16	1.09	1.20	**0.78**
GEO02	0.90	0.94	0.91	1.02	1.20	0.98
GEO03	0.88	0.97	**0.60**	1.11	1.18	1.20
GEO04	1.00	1.04	0.97	0.91	1.02	1.09
GEO05	0.87	1.12	0.88	1.28	0.92	0.98
GEO06	1.01	0.91	1.04	0.90	1.09	1.14
HIS06			S/D			
HIS07	1.02	**0.75**	1.14	0.86	0.95	1.20
HIS08	0.87	0.94	0.97	0.99	1.15	1.08
HIS09	1.15	0.96	0.99	0.98	0.92	1.06
HIS10	1.01	**0.75**	1.16	1.09	1.20	**0.78**
HIS11	**0.68**	1.09	1.14	1.13	1.10	0.92
HIS12	0.89	**0.61**	1.06	1.14	1.09	1.24
HIS13	0.97	1.15	0.98	0.88	0.82	1.21
FCYE14	0.93	0.80	1.02	1.05	1.10	1.15
FCYE16	0.99	0.93	0.99	1.07	0.86	1.08
FCYE17	0.84	0.97	0.98	1.16	1.00	1.09
FCYE18	0.87	1.03	1.04	0.92	1.14	0.99
FCYE19	**0.61**	0.83	1.03	1.25	1.12	1.14
FCYE20	**0.78**	**0.79**	1.08	1.21	1.11	1.00

Nota: I1, I2, I3, I4, I5, I6 refieren a ítem1, ítem2, ítem3, ítem4, ítem 5 e ítem 6, respectivamente. S/D: sin datos disponibles.

Tabla F.15.
Índices de ajuste de AFC por familia de la muestra SOC, con sus respectivas cargas factoriales

	Indices de ajuste						Cargas factoriales					
	Chi	Lib	p	NNFI	CFI	RMSEA	ITEM 1	ITEM 2	ITEM 3	ITEM 4	ITEM 5	ITEM 6
GEO01	6.029	5	.303	0.994	.998	.032	.728	.644	.756	.680	.789	.663
GEO02	1.891	4	.756	1.023	1.000	.000	.733	.566	.652	.695	.562	.558
GEO03	0.064	4	.999	1.099	1.000	.000	.508	.352	.365	.590	.733	.574
GEO04	2.502	6	.868	1.032	1.000	.000	.725	.573	.493	.509	.722	.654
GEO05	0.011	3	.999	1.051	1.000	.000	.664	.665	.612	.737	.386	.465
GEO06	11.234	7	.128	0.977	.989	.054	.686	.633	.535	.577	.560	.684
HIS06	S/D											
HIS07	6.567	8	.583	1.008	1.000	.000	.625	.496	.786	.602	.565	.839
HIS08	4.079	9	.906	1.037	1.000	.000	.518	561	.581	.542	.720	.590
HIS09	12.366	9	.193	0.975	.985	.043	.667	.544	.572	.577	.487	.632
HIS10	8.238	7	.312	0.994	.997	.029	.666	.731	.617	.611	.799	.693
HIS11	9.716	9	.373	0.995	.997	.020	.378	.596	.678	.585	.652	.571
HIS12	9.648	9	.379	0.997	.998	.019	.529	.439	.663	.717	.671	.777
HIS13	7.933	8	.440	1.001	1.000	.000	.443	.664	.444	.369	.423	.638
FCYE14	3.928	4	.415	1.003	1.000	.000	.320	.336	.334	.438	.530	.556
FCYE16	11.744	8	.162	0.974	.986	.048	.569	.529	.627	.615	.603	.696
FCYE17	6.447	9	.694	1.017	1.000	.000	.406	.582	.645	.683	.620	.650
FCYE18	10.090	9	.343	0.993	.996	.024	.522	.621	.694	.550	.667	.650
FCYE19	4.965	3	.195	0.964	.993	.052	.997	.567	.562	.643	.700	.661
FCYE20	11.477	9	.244	0.987	.992	.037	.479	.502	.672	.789	.681	.633

Nota: α: coeficiente Alpha de Cronbach, NNFI: *Non-Normed Fit Index*. CFI: *Comparative Fit Index*. RMSEA: *Root mean-square error of approximation.* S/D: sin datos disponibles.