Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Doctorado en Ciencias Educativas

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

"Análisis del aspecto sustantivo de la validez de constructo de una prueba de habilidades cuantitativas"

T E S I S
Que para obtener el grado de
DOCTOR EN CIENCIAS EDUCATIVAS
Presenta:

Juan Carlos Pérez Morán

DIRECTORA DE TESIS:

Dra. Norma Larrazolo Reyna

Ensenada B. C. México; Febrero de 2014

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Doctorado en Ciencias Educativas

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

"Análisis del aspecto sustantivo de la validez de constructo de una prueba de habilidades cuantitativas"

T E S I S
QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS EDUCATIVAS
PRESENTA:

Juan Carlos Pérez Morán

APROBADO POR:

Dra. Norma Larrazolo Reyna
Directora de tesis

Dr. Eduardo Backhoff Escudero
Sinodal

Dr. Luis Ángel Contreras Niño
Sinodal

Dr. Jesús Miguel Jornet Meliá
Sinodal

Dr. Felipe de Jesús Tirado Segura
Sinodal

Dra. Virginia Velasco Ariza
Sinodal

Ensenada B. C. México, febrero de 2014

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

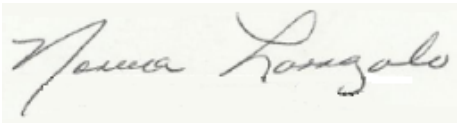
Dr. Lewis McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS*".*

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente



_____

Dra. Norma Larrazolo Reyna

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

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

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS".

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

_____
Dr. Eduardo Backhoff Escudero

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

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

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS*".*

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

_____
Dr. Luis Ángel Contreras Niño

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

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

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS".

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

_____
Dr. Jesús Miguel Jornet Meliá

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

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

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS".

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

_____
Dr. Felipe de Jesús Tirado Segura

Ensenada, B.C. a 12 de Enero de 2014

ASUNTO: Voto aprobatorio sobre trabajo de tesis de grado de Doctor.

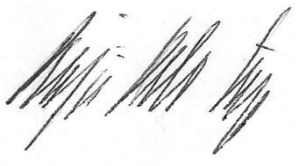
Dr. Lewis McAnally Salas
Coordinador del Doctorado en Ciencias Educativas
Presente.

Después de haber efectuado una revisión minuciosa sobre el trabajo de tesis presentado por el **C. Juan Carlos Pérez Morán** para poder presentar la defensa de su examen y obtener el grado de Doctorado en Ciencias Educativas, me permito comunicarle que he dado mi VOTO APROBATORIO, sobre su trabajo intitulado:

"ANÁLISIS DEL ASPECTO SUSTANTIVO DE LA VALIDEZ DE CONSTRUCTO DE UNA PRUEBA DE HABILIDADES CUANTITATIVAS*".*

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente



_____

Dra. Virginia Velasco Ariza

Dedicatoria

A Dios padre por todas sus bendiciones derramadas en mi vida

A mis padres y familiares por su cariño

A mi esposa e hijos por su amor y su apoyo incondicional

Agradecimientos

Agradezco a Dios por la existencia y a mis padres por apoyarme en cada momento de mi vida, inculcándome que con trabajo arduo se alcanzan los sueños.

A mi esposa por su paciencia, su comprensión y su ayuda, pero principalmente por mostrarme que el amor es siempre un motor para vivir y para salir adelante. A mis hijos por ser inspiración y esperanza viva en mi existencia.

A mis hermanos Daniel, Guillermo y Omar, quienes siempre están conmigo acompañándome y compartiendo sus experiencias.

A mi directora de tesis, la Dra. Norma Larrazolo, a quien admiro, quiero y respeto mucho; agradezco su amistad y su apoyo incondicional en cada una de las etapas de mi formación a lo largo de mis estudios de posgrado.

A cada uno de mis profesores y sinodales del doctorado por su invaluable apoyo, especialmente al Dr. Eduardo Backhoff, al Dr. Luis Ángel Contreras, al Dr. Juan Carlos Rodríguez, al Dr. Joaquín Caso, a la Dra. Paola Ovalle, a la Dra. Edna Luna, al Dr. Javier Organista, al Dr. Lewis McAnally, a la Dra. Virginia Velasco, al Dr. Felipe Tirado y al Dr. Jesús Jornet, quienes compartieron conmigo su amplio conocimiento durante mis estudios de posgrado.

Al Dr. Vicente Ponsoda, a la Dra. Carmen García y, en especial a la Dra. Sonia Romero y al Mtro. Guarner Rojas, quienes durante mis dos estancias en la Universidad Autónoma de Madrid me apoyaron, me ofrecieron su amistad y sus conocimientos sobre los modelos componenciales.

A todas las personas del equipo administrativo y de intendencia del IIDE, en especial a la Mtra. Estrella Velasco, quien con su trabajo profesional y su entrega a la mejora de la calidad educativa hizo posible que me beneficiara durante mi doctorado con estancias, becas, congresos, excelentes espacios y materiales de estudio.

A todos los que estuvieron conmigo siempre, a mis amigos y a mis compañeros del doctorado, gracias totales.

Juan Carlos P. M.

Índice

Índice de tablas

Índice de figuras

Siglas y acrónimos

ACT	*American College Testing*
AERA	*American Educational Research Association*
AFC	Análisis Factorial Confirmatorio
AHM	*Attribute Hierarchy Method*
APA	*American Psychological Association*
CIA	Criterio de Información de Akaike
CRESST	*National Center for Research on Evaluation, Standards and Student Testing*
DINA	*Deterministic Input, Noisy And Gate*
EDC	Evaluación Diagnóstica Cognitiva
EDU	Evaluación del Diseño Universal
ESDC	Enfoque Sistémico de Diseño Cognitivo
ETS	*Educational Testing Service*
EXHCOBA	Examen de Habilidades y Conocimientos Básicos
GAÍ	Generación Automática de Ítems
GAÍRC	Generación Automática de Ítems de Respuesta Compleja
LLTM	Modelo Logístico Lineal de Rasgo Latente
LSDM	Método de las distancias mínimo-cuadráticas
MDC	Modelos de Diagnóstico Cognitivo
MDE	Modelos para el Diseño de las Evaluaciones
MPC	Modelos Psicométricos Componenciales
NCME	*National Council on Measurement in Education*
PFLC	Preparatoria Federal Lázaro Cárdenas

REESCO	Reactivos Estructurales Constructivos
RMSD	Raíz de la Diferencia Media al Cuadrado
RSM	*Rule Space Method*
SAT	*Scholastic Aptitud Test*
TAI	Test Adaptativos Informatizados
TCT	Teoría Clásica de los Tests
TRI	Teoría de Respuesta al Ítem
UNAM	Universidad Nacional Autónoma de México
UABC	Universidad Autónoma de Baja California
UG	Universidad de Guanajuato
UNISON	Universidad de Sonora
UAQ	Universidad Autónoma de Querétaro
V-ÍOM	Versión de Ítems de Opción Múltiple
V-ÍRC	Versión de Ítems de Respuesta Compleja

RESUMEN

El conocimiento de las características de los procesos cognitivos que utilizan los estudiantes para responder a los ítems de pruebas psicológicas y educativas puede ser útil en diferentes contextos y procesos educativos. Uno de los procesos dentro del ámbito educativo que se beneficia en forma directa —y casi inmediata— es la evaluación educativa, principalmente el desarrollo y la validación de los instrumentos de medición (Borsboom & Mellenbergh, 2007; Embretson, 1998; Messick, 1989b; *National Research Council*, 2001; Snow & Lohman, 1989).

El EXHCOBA es un examen computarizado de selección para ingreso a la educación media superior y superior. Este examen se aplica desde 1992. Actualmente está desarrollándose una nueva versión del EXHCOBA, basada en la Generación Automática de Ítems de Respuesta Compleja (GAÍRC).

La fundamentación teórica de la nueva versión está basada en cuatro directrices importantes: (a) evaluar competencias estructurales de inclusión del conocimiento (competencias básicas) que den soporte al aprendizaje subsecuente y que se encuentran presentes en el currículum nacional de la educación básica; (b) evaluar primordialmente procesos cognitivos complejos de comprensión, aplicación y evaluación; (c) evaluar de forma auténtica, o lo más cercano posible, cómo aprenden naturalmente los estudiantes de los diferentes niveles educativos; y (d) evaluar los dominios comprometidos en el EXHCOBA utilizando la GAÍRC (Backhoff, 2012). De esta manera, el EXHCOBA se caracteriza por ser una de las pruebas con mayor nivel de innovación en México.

Asimismo, la nueva versión del EXHCOBA se desarrolló bajo un modelo innovador de Generación Automática de Ítems (GAÍ). La GAÍ se desarrolla por modelos de tareas con

los cuales se producen *ítems base* (también conocidos como *ítems padres),* que a su vez generan ítems isomorfos (también conocidos como *ítems hijos*) con propiedades psicométricas similares. Para esto, los desarrolladores del EXHCOBA elaboraron templetes de ítems, los cuales funcionan como estructuras para sostener, mantener y resguardar la fidelidad de las interpretaciones del dominio especificado en modelos de tareas. Dichos templetes tienen la función de asegurar que los ítems de una familia (conjunto de ítems hijos generados del mismo ítem padre) se comporten de manera similar en términos de sus características operativas (ver Embretson, 2001; Luecht, 2008; Gierl & Lai, 2013).

Es por lo anterior que se consideró relevante la realización de un estudio con el propósito de analizar el aspecto sustantivo de la validez de constructo del área de Habilidades Cuantitativas (HC) del Examen de Habilidades y Conocimientos Básicos (EXHCOBA) en dos de sus versiones, una con 30 Ítems de opción múltiple y otra con 20 Ítems de respuesta compleja. Para alcanzar el propósito de la investigación fue adaptado y aplicado un modelo teórico–metodológico con enfoque *top-down* (ver Bejar, 2002, 2010; Gorin y Embretson, 2013; Messick, 1989b), basado en los principios del Enfoque Sistémico de Diseño Cognitivo (ESDC) propuesto por Embretson (1994). Para la aplicación de dicho modelo se desarrollaron cinco fases.

Las dos primeras fases abarcan el diseño, el pilotaje y la aplicación de distintos estudios cognitivos para obtener y para analizar las primeras evidencias de validez basadas en el proceso de respuesta (AERA, APA & NCME, 1999). Los estudios cognitivos utilizados fueron el método de *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983) con el apoyo del *análisis de expertos* en el área de dominio de la prueba, recomendado por Rupp, Templin, & Henson (2010). Para evaluar el

diseño del interfaz de los ítems y para verificar el modelo del proceso de respuesta elaborado por los expertos se utilizó la *técnica de pensamiento de voz alta* con el *análisis de protocolos* concurrentes y retrospectivos (Ericsson & Simon, 1984, 1993; Leighton, 2009; Leighton & Gierl, 2007b). También se siguieron las recomendaciones hechas por algunos autores (por ejemplo, Snow & Lohman, 1989 y Sternberg, 1977) en el campo de la psicología cognitiva, referentes al *análisis de protocolos* acompañado del *análisis del sendero de la vista* (Newell & Simon, 1972) y al *análisis cronométrico* o de tiempo de latencia de respuesta (Fredericksen, 1980; Posner, 1978; Posner & Rogers, 1978).

Durante la tercera fase se modeló el proceso de respuesta subyacente a los ítems del área de HC del EXHCOBA con ayuda de los expertos en el área de matemáticas. También: se definieron y describieron cada uno de los atributos u operaciones subyacentes a los ítems, se estructuraron cada uno de los modelos cognitivos de las dos versiones estudiadas, se verificó la similitud entre los modelos del proceso de respuesta definidos por los expertos y los procesos de respuesta utilizados por los estudiantes de tercero de secundaria ante los ítems, y se evaluó el diseñó de la interfaz de la prueba.

En la cuarta fase se aplicó el análisis psicométrico básico y el componencial para obtener y analizar las evidencias de validez basadas en la estructura del modelo cognitivo del área de HC del EXHCOBA. Para el análisis psicométrico básico se calibraron cada una de las versiones de la prueba con la aplicación del modelo de la Teoría Clásica de los Tests (TCT), se analizó su unidimensionalidad con la aplicación del modelo de Análisis Factorial Confirmatorio (AFC) de Fraser (1988), y se analizaron los parámetros de dificultad de los ítems, su ajuste a los datos con el modelo de RASCH unidimesional y el estadístico de la Razón de Verosimilitudes Condicional (CLR, por sus siglas en inglés) (Fischer & Ponocny-Seliger, 1998).

Para los análisis psicométricos componenciales se aplicó el Modelo Logístico Lineal de Rasgo Latente (LLTM, por sus siglas en inglés) desarrollado por Fischer (1973; 1985) y el Método de la Distancias de Mínimo-Cuadráticas (LSDM, por sus siglas en inglés) desarrollado por Dimitrov (2007). Se realizaron distintas pruebas de ajuste entre los parámetros del modelo RASCH y los parámetros de los distintos modelos psicométricos componenciales aplicados. Además, se realizó un proceso reiterativo de reconfiguración de la matriz Q y un análisis de validez cruzada entre los modelos LSDM, LLTM y LLTM con matriz Q reconfigurada. Cabe mencionar que la tercera etapa (Interpretación de los resultados de los examinados) de la Fase IV no se desarrolló al estar fuera de los objetivos de la presente tesis.

Con la aplicación de los diferentes análisis y procedimientos prescritos en el modelo teórico-metodológico que fue adaptado se obtuvieron diversas evidencias de validez. Estas con relación al aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA. De las evidencias de validez del proceso de respuesta obtenidas durante las tres primeras fases del modelo, cinco resultaron básicas: (1) los modelos definidos por expertos del proceso de respuesta subyacente a los ítems, (2) la lista con atributos que explican la dificultad de los ítems, (3) la estructura de los modelos cognitivos del área de HC para cada una de las versiones estudiadas, (4) los reportes verbales con los procesos de respuesta de los estudiantes de tercero de secundaria ante los ítems y (5) la evaluación del diseño del interfaz de los ítems. De la cuarta fase del modelo resultaron seis evidencias de validez relacionadas con el análisis de la estructura del modelo cognitivo de la prueba: (1) la calibración de la prueba con la aplicación del modelo de la TCT, (2) el análisis de la estructura interna bajo el modelo de *redes nomológicas* y la dimensionalidad de la prueba, (3) el ajuste del modelo de RASCH a los datos de la prueba, (4) el ajuste entre los distintos

modelos psicométricos aplicados, (5) el análisis componencial de la estructura del modelo cognitivo de la prueba y (6) el análisis de la validez cruzada entre los resultados de los modelos LLTM y el LSDM.

La aplicación del método de *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983), el cual se apoyó con el *análisis de expertos* en el área de dominio de la prueba, permitió definir los modelos generales del proceso de respuesta de cada uno de los ítems de las dos versiones estudiadas. Los modelos generales del proceso de respuesta fueron descritos como pasos básicos requeridos para contestar correctamente los ítems. Por ejemplo, el proceso cognitivo requerido por los participantes para contestar correctamente el ítem dos de la versión con ítems de opción múltiple con el que se mide la *obtención del valor faltante en secuencias numéricas* es: (a) *leer detalladamente las indicaciones del ítem*, (b) *comprender el objetivo de la tarea evaluativa*, (c) *obtener la diferencia entre cada una de las cantidades*, (d) *identificar el patrón que rige en la secuencia*, (e) *aplicar el patrón identificado para obtener la última cantidad de la secuencia* y (f) *reconocer el valor faltante en la secuencia numérica dentro las opciones de respuesta*.

Como resultado del análisis por expertos se definieron y describieron un total de 14 operaciones o atributos cognitivos para la V-ÍOM, y 9 operaciones cognitivas para la V-ÍRC. Según los expertos, los atributos cognitivos subyacentes a la V-ÍOM son: *conocer el nombre y ubicación de los dígitos, fracción y visualización de figuras geométricas; comprensión de problemas matemáticos contextualizados; aplicación de operaciones aritméticas básicas; adición de ángulos de un triángulo; aplicación de fracciones; identificación de patrones de secuencias numéricas; representación de modelos exponenciales; representación del modelo matemático del perímetro, el área y el volumen;*

uniformidad de unidades de medida diferentes; aplicación de la regla de tres simple; suma de fracciones; aplicación del modelo matemático de probabilidades; y aplicación de expresiones algebraicas.

Los atributos cognitivos definidos para la V-ÍRC por los expertos son: *representación de lugares geométricos; posicionamiento y ubicación de valores; aplicación de operaciones aritméticas básicas; aplicación de escalas gráficas en mapas; identificación de patrones de secuencias numéricas; interpretación de la información del problema; representación de modelos matemático-aritméticos; cálculo de equivalencias de unidades de medida; y representación de modelos matemático-geométricos.* Una vez definidos estos atributos requeridos para responder los ítems del área HC del EXHCOBA se obtuvieron: una matriz Q de 30_iX14_k para la V-ÍOM y una matriz Q de 20_iX9_k para la V-ÍRC.

Esta aplicación de las *técnicas de pensamiento en voz alta* con *análisis de protocolos concurrentes y retrospectivos* a 16 estudiantes de tercero de secundaria consiguió tres tipos de resultados: (a) los procesos de respuesta que utilizan los estudiantes de tercero de secundaria ante los ítems de la prueba, (b) la verificación de la similitud entre los modelos del proceso de respuesta definidos por los expertos y los procesos de respuesta utilizados por los estudiantes de tercero de secundaria ante los ítems de la prueba, y (c) la evaluación del diseño del interfaz de los ítems. Los resultados, con respecto a la verificación de las similitudes entre el modelo del proceso cognitivo definido por los expertos y el proceso de respuesta de los examinados ante los ítems de la prueba, permitieron analizar una serie de problemas asociados a diferentes factores, tanto del interfaz como del propio estudiante. Estas problemáticas deben ser tomadas en cuenta para reducir la varianza irrelevante en las respuestas de los examinados.

Los tipos de problema encontrados con mayor frecuencia en el interfaz de los ítems de la V-ÍOM estuvieron relacionados con la *Comprensión y/o legibilidad de la base del ítem* y las *Respuestas al azar o por adivinación,* donde la frecuencia fue de 67 (60%) y 27 (24%), respectivamente. Los tipos de problemas en el procesos de respuesta de los examinados con menor frecuencia fueron aquellos relacionados con la *Comprensión y/o legibilidad de las indicaciones del ítem* y la *Comprensión de las opciones de respuesta*, cuya frecuencia fue de 13 (12%) y 4 (4%), respectivamente.

Los tipos de problema presentes con mayor frecuencia en los procesos de respuesta de los examinados ante los ítems de la V-ÍRC fueron los relacionados con la *Usabilidad de las operaciones de respuesta del ítem* y la *Comprensión y/o legibilidad de la base del ítem,* siendo la frecuencia 23 (36%) y 21 (33%), respectivamente. En contra parte, los tipos de problemas con menor frecuencia en los procesos de respuesta de los estudiantes fueron: *Respuestas al azar o por adivinación* por parte de los examinados, donde la frecuencia fue del 13 (20%); *Comprensión y/o legibilidad de las indicaciones del ítem,* cuya frecuencia fue de 5 (8%); y *Estructura y/o formato del ítem* con 2 (3%) frecuencias.

Como resultado de la aplicación de la Teoría Clásica de los Test (TCT), el promedio de dificultad obtenido fue de 0.54 para la V-ÍOM y 0.36 para la V-ÍRC. La proporción de aciertos de los ítems de V-ÍOM oscila entre 0.27 (alta dificultad) y 0.87 (baja dificultad), mientras que la proporción de aciertos de los ítems de V-ÍRC oscila entre 0.07 (elevada dificultad) y 0.64 (media dificultad). El promedio del índice de discriminación *bajos-altos* fue de 0.54 para la V-ÍOM y 0.47 para la V-ÍRC, respectivamente. Con respecto al promedio del coeficiente de correlación puntual-biserial (*Rpbis*), el valor promedio fue de 0.47 para la V-ÍOM y 0.43 para la V-ÍRC. Según el estándar técnico de calidad establecido (*Rpbis* =>

0.20), los coeficientes obtenidos para las dos versiones del área de habilidades cuantitativas del EXHCOBA fueron adecuados. Por último, el **α** de confiabilidad de la V-ÍOM fue de 0.880 y el de la V-ÍRC fue de 0.776.

En la aplicación de los modelos componenciales LLTM y LSDM se realizaron pruebas de ajuste al modelo de RASCH, las cuales permitieron descartar los ítems que no cumplieron con los *estándares* básicos para los subsecuentes procedimientos relacionados con el análisis de la estructura del modelo cognitivo. Respecto a los resultados de la aplicación del modelo AFC de Fraser (1988), se obtuvo un Índice de Bondad de Ajuste (GFI, por sus siglas en inglés) (Tanaka & Huba, 1985) para la versión con 30 ítems de opción múltiple de 0.993 con un Residuo Cuadrático Medio (RMSR, por sus siglas en inglés) de 0.01, mientras que para la versión con ítems de respuesta compleja reducida a 17 ítems se obtuvo un GFI de 0.986 con un RMSR de 0.01, lo cual confirma que los ítems de las dos versiones del área HC del EXHCOBA miden una sola dimensión, respectivamente. En cuanto a la aplicación del modelo unidimensional de RASCH, los rangos de los parámetros de dificultad ($\boldsymbol{\delta_i}$) de cada uno de los ítems de la V-ÍOM fueron desde -2.021 (fácil) hasta 1.550 (difícil), mientras que para la V-ÍRC fueron desde -1.461 (fácil) hasta 2.246 (difícil). Tomando en cuenta el criterio de calidad establecido de -1.8 a 1.8 para los valores $\boldsymbol{\delta_i}$ de la prueba, el ítem 6 de la V-ÍOM y los ítems 11 y 12 de la V-ÍRC no cumplen con dicho criterio. Por su parte, el estadístico de la Razón de Verosimilitudes Condicional (CLR, por sus siglas en inglés) resultó significativo para la V-ÍOM y la V-ÍRC; por lo tanto, no se confirma el ajuste y el supuesto de unidimensionalidad. Esto no es sorprendente, dado el rigor divulgado de la prueba CLR en la literatura (Fischer, 1995).

En los resultados de la aplicación del modelo LLTM, cada uno de los parámetros básicos de las dos versiones analizadas fue significativo, lo cual indica que las operaciones correspondientes contribuyen a la dificultad de los ítems. En cuanto a las operaciones que contribuyen a la dificultad de los ítems (parámetros básicos negativos), fueron 3 tanto para la V-ÍOM como para la V-ÍRC; por su parte, las operaciones que contribuyen a la "facilidad" (parámetro básico positivo) fueron 8 para la V-ÍOM y 3 para la V-ÍRC. Un análisis a profundidad permite observar las operaciones cognitivas que introducen mayor dificultad a los ítems del área Habilidades cuantitativas de la V-ÍOM, las cuales son: la *comprensión de problemas matemáticos contextualizados*, la *aplicación de operaciones aritméticas básicas* y la *aplicación de fracciones*; con respecto a las operaciones cognitivas que introducen mayor dificultad de la V-ÍRC, estas son: la *aplicación de operaciones aritméticas*, la *representación de valores en sistemas posicionales* y la *representación de modelos matemáticos-geométricos*.

Con respecto a los resultados de la aplicación del LSDM, y una vez analizadas las Curvas de Probabilidad de los Atributos (CPA) de la matriz Q obtenidas con el LSDM para la V-ÍOM y la V-ÍRC, se encontró una pobre estimación de varias operaciones cognitivas con valores MAD (en inglés *Mean Absolute Difference*) superiores a 0.1 en las dos versiones analizadas. Por lo anterior, se decidió reconfigurar reiteradamente la matriz Q hasta alcanzar una mejor estimación de tales valores. Del proceso de reconfiguración de la matriz Q para la V-ÍOM resultaron un total de 10 operaciones cognitivas para 21 ítems (Matriz Q, 21_i X10_k) y 5 operaciones cognitivas para 14 ítems (Matriz Q, 14_i X5_k) de la V-ÍRC. De esta manera, pudo obtenerse una estimación de todas las operaciones cognitivas en las dos versiones.

Para el análisis de la validación cruzada entre los modelos LLTM y LSDM se realizó la prueba de ajuste del LLTM con el modelo de RASCH mediante la matriz Q reconfigurada. Como resultado, se encontró que para la V-ÍOM mejoró mucho la correlación (de r_{xy}=0.829 a r_{xy}=0.95; *p*<.05) entre los parámetros de dificultad del LLTM y del modelo de RASCH unidimensional. Por su parte, la V-ÍRC mejoró su ajuste con una correlación significativa r_{xy}=0.91 con *p* <.05 en comparación al valor anterior de r_{xy}=0.78, *p* < .05. También disminuyó el Criterio de Información de Akaike (CIA), mostrándose un mejor ajuste del modelo LLTM reconfigurado al modelo de RASCH, en las dos versiones estudiadas. Ahora bien, se puede decir que los valores de los parámetros básicos, estimados con el LLTM y ordenados de fácil a difícil, coinciden perfectamente con el orden de las CPA estimadas con el LSDM para la validación cruzada de los resultados de los dos modelos utilizados. Además, todos los parámetros básicos del LLTM en las dos versiones estudiadas son significativamente diferentes de 0 (*p* < 0.01), lo que indica que todas las operaciones cognitivas contribuyen a la explicación de la dificultad de los ítems.

En cuanto a las aportaciones del presente estudio, la adaptación de un modelo teórico-metodológico con enfoque *top-down* para obtener evidencias de validez del proceso y de la estructura del modelo cognitivo del EXHCOBA es una de las aportaciones de mayor impacto y alcance de esta investigación. En primer lugar porque el modelo teórico-metodológico adaptado favorece a los desarrolladores de pruebas e investigadores en la medición y la evaluación, pues aporta una plataforma operativa para el análisis del aspecto sustantivo de la validez de constructo de pruebas educativas y psicológicas (Messick, 1989b; AERA, APA & NCME, 1999). En segundo lugar porque dicho modelo puede ser aplicado en una gran diversidad de contextos evaluativos para guiar el fortalecimiento de la validez de instrumentos que no iniciaron su desarrollo desde un

modelo cognitivo o teoría *fuerte* (Griel & Lai, 2013), como en el caso de la mayoría de las pruebas nacionales utilizadas para evaluar el aprendizaje, y desarrolladas bajo el modelo de *redes nomológicas*. En tercer lugar porque la aplicación de dicho modelo permitió conocer a profundidad diferentes evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo del área de HC del EXHCOBA, en las dos versiones estudiadas.

Así, se estructuró el modelo cognitivo subyacente al área de HC y se fortaleció el modelo de GAÍ de la prueba. Lo anterior representa un gran impacto, ya que se obtuvieron fuertes evidencias de validez del EXHCOBA y se potencializó su modelo de GAÍ, siendo el fundamento una teoría cognitiva *fuerte* (Gorin & Embretson, 2013; Griel & Lai, 2013). De esta manera, la investigación aquí expuesta es una base sólida para que los resultados e interpretaciones del EXHCOBA presenten un alto grado de validez e innovación.

Palabras clave: Validez de constructo; evidencias de validez; aspecto sustantivo; proceso cognitivo; modelo cognitivo; reportes verbales; análisis psicométricos componenciales.

I. INTRODUCCIÓN

El conocimiento de las características de los procesos cognitivos que utilizan los estudiantes para responder a los ítems de pruebas psicológicas y educativas puede ser útil en diferentes contextos y procesos educativos. Uno de los procesos dentro del ámbito educativo que se beneficia en forma directa —y casi inmediata— es la evaluación educativa, principalmente el desarrollo y la validación de los instrumentos de medición (Borsboom & Mellenbergh, 2007; Embretson, 1998; Messick, 1989b; *National Research Council*, 2001; Snow & Lohman, 1989). Por otra parte, algunos de los procesos educativos que también se benefician al conocer las características de los atributos cognitivos que subyacen a las puntuaciones de las pruebas son: el diseño y la operación del currículum, el diagnóstico de las estrategias de enseñanza y la mejor comprensión del proceso de enseñanza-aprendizaje (Leighton & Gierl, 2007a; National Research Council, 2001).

Con relación al desarrollo y a la validación de los instrumentos de medición, se han presentado diversas propuestas que integran una variedad de innovaciones de la psicología cognitiva, la psicometría y la teoría de la validez, en las últimas décadas. Modelos de medición como los propuestos por Embretson (*The cognitive design system approach*, 1998) y Mislevy (*The Model-Based Reasoning*, 2009) son algunos ejemplos. En el caso específico del modelo propuesto por Embretson (1998), además de incorporar las innovaciones de la psicología cognitiva y la psicometría, se toman en cuenta las innovaciones referentes a la Generación Automática de Ítems (GAÍ) (Bejar, 1990;

Fredericksen, Mislevy & Bejar, 1993; Hornke & Habon, 1986; Shye, Elizur & Hoffman, 1994) y al uso de ítems de respuesta compleja, que integran diferentes herramientas computarizadas. Así, los modelos de medición señalados constituyen un nuevo enfoque en la teoría de los tests, basado en la integración sustancial de la psicología cognitiva y la psicometría. Dicha integración es comúnmente conocida en el campo de la medición como Evaluación Diagnóstica Cognitiva (EDC) (Embretson, 1994; Leighton & Gierl, 2007a; Messick, 1989b, 1995; Mislevy, 2009; Nichols, Chipman & Brenan, 1995; Rupp & Mislevy, 2007). La EDC ha permitido que las fortalezas de una área compensen las deficiencias de otra, y viceversa (Romero, Ponsoda & Ximenez, 2008).

Algunos desarrolladores internacionales de pruebas educativas (por ejemplo, Bejar & Yocom,1986; Leighton & Gierl, 2007a; Li, s.f.), al ver los beneficios que resultan de la incorporación de los aportes de la psicología cognitiva en el desarrollo y en la validación de las pruebas, se han interesado en la aplicación de diversos modelos de medición que toman en cuenta los procesos y atributos cognitivos utilizados para resolver los ítems. De manera que, los modelos de EDC como los mencionados en el párrafo anterior, son actualmente utilizados por algunos desarrolladores de pruebas educativas reconocidos a nivel internacional (por ejemplo, Ayala, Ayala & Shavelson, 2001; Gierl, Leighton, Wang, Zhou & Gokiert, 2009; Li, s.f.; Gierl, Leighton, Changjiang, Jiawen, Rebecca & Tan, 2009; Tatsuoka, 2009) para el establecimiento de los objetivos de la evaluación, así como para guiar los procesos de desarrollo y validación de sus instrumentos de medición. Con ello, se enfatiza la importancia de la relación entre la

psicología cognitiva y la psicometría; se resalta su utilidad para que las pruebas estén al servicio de la comprensión de los procesos educativos.

En México, la historia en cuanto a la aplicación de modelos de medición que integren las innovaciones de la EDC es diferente. Los desarrolladores nacionales de pruebas psicológicas y educativas no han incorporado aún modelos de medición como los propuestos por Embretson (1983, 1994, 1998) o Mislevy (2007, 2009). Sin embargo, en la actualidad algunos investigadores y desarrolladores de pruebas a gran escala se encuentran interesados en la aplicación de dichos modelos (por ejemplo, Backhoff & Larrazolo, 2012; Pérez, Larrazolo, Backhoff & Rojas, 2013). Tal es el caso del grupo de investigadores que se encargan del desarrollo del Examen de Habilidades y Conocimientos Básicos (EXHCOBA), el cual se aplica cada año en el proceso de selección de aspirantes en distintas instituciones de prestigio de nivel medio superior y superior en México (por ejemplo: la preparatoria Federal Lázaro Cárdenas (PFLC), la Universidad de Guanajuato (UG), la Universidad de Sonora (UNISON) y la Universidad Autónoma de Querétaro (UAQ), entre otras). De manera especial, del grupo EXHCOBA se encuentra desarrollando e implementando un Generador Automático de Ítems (GAÍ) y Reactivos Estructurales Constructivos (REESCO). Dichos reactivos presentan un formato de respuesta compleja y, según Tirado (2010), representan una estrategia socio-cognitiva para la evaluación de conocimientos y habilidades básicas estructurales en donde se hace uso de diferentes elementos y herramientas informáticas (ver Apéndice 1).

Dado lo anterior —y con el interés de realizar una aportación al campo del desarrollo e innovación en la aplicación de modelos que integren los avances de la psicología cognitiva, la psicometría, la teoría de la validez y, en particular, las innovaciones de la EDC—, la presente tesis muestra un trabajo de investigación centrado en el análisis del aspecto sustantivo de la validez de constructo del área de Habilidades Cuantitativas (HC) del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. Aunado a ello, con el fin de guiar el trabajo analítico de la tesis, se realizó la adaptación y la aplicación de un modelo teórico-metodológico con enfoque *top-down* (ver Bejar, 2002, 2010; Gorin y Embretson, 2013; Messick, 1989b), basado en el Enfoque Sistémico de Diseño Cognitivo (ESDC) propuesto por Embretson (1994). También, de manera especial se pretende contribuir en la generación de las condiciones necesarias para aplicar eventualmente un programa integral de validez en las distintas áreas del EXHCOBA, tanto en sus versiones con ítems de opción múltiple como en las versiones con ítems de respuesta compleja. Cabe señalar que el EXHCOBA es una prueba de gran escala y de alto impacto para la vida académica de los estudiantes que desean ingresar a las distintas instituciones usuarias. Es por eso que la prueba tiene una alta relevancia para los tomadores de decisiones y desarrolladores, ya que permite contar con evidencias de validez basadas en el aspecto sustantivo.

1.1. Antecedentes del estudio

La tercera edición del libro de *Educational Measurement* editado por Linn (1989) Messick presenta un capítulo relacionado con la validez. En el mismo libro, Snow y Lohman, presentan un capitulo relacionado con las Implicaciones de la psicología cognitiva para la evaluación educativa. Ambos capítulos señalan el creciente interés y la necesidad por la EDC, debido a que esta solidifica la integración entre la psicología cognitiva, la psicometría y la evaluación educativa (Leighton & Gierl, 2007a). Desde dicha publicación, otros investigadores y académicos, influenciados por los beneficios de la EDC en el campo de la medición, escribieron artículos, capítulos y libros sobre el tema. Entre las producciones bibliográficas más notables se encuentran los libros coeditados por Nichols, Chipman, y Brennan (1995) y el de Leighton y Gierl (2007). En dichas publicaciones se muestra una amplia gama de modelos y técnicas relacionadas con la EDC. También se presentan distintas perspectivas y formas de implementar la EDC en la medición educativa; en especial, Leighton y Gierl (2007a) mencionan que la EDC —con su capacidad de integrar la psicología cognitiva y la psicometría— representa actualmente uno de los caminos más convincentes para establecer la validez de un test.

1.1.1. Discusión del concepto de validez en el campo de la medición

A la par con las innovaciones de la EDC y de su integración en el desarrollo de las pruebas educativas, se presentaron cambios significativos en los temas relacionados con el concepto de validez, consignados en los *estándares* internacionales y nacionales de pruebas psicológicas y educativas. Uno de los cambios más importantes fue la

redefinición del concepto mismo de validez. Dicho concepto, entre los años de 1920 y 1950 presentaba un modelo de validez de criterio (Angoff, 1998; Cronbach, 1971; Kane, 2006; Moss, 1992; Shepard, 1993). Después, a principios de 1950 surge el modelo de la validez de contenido en donde se incorpora la necesidad de contar con una muestra representativa del dominio a medir para establecer un vínculo racional entre los procedimientos utilizados que generan los puntajes del criterio medido y la interpretación propuesta o el uso de los puntajes (Cureton, 1951; Ebel, 1961; Kane, 2006). Para 1955, Cronbach y Meehl (1955) se presenta el modelo de validez de constructo como una alternativa a los anteriores modelos de validez. Dicho modelo, en 1980, presentaba una gran aceptación por parte de los teóricos y de los investigadores en el campo de la validez (Anastasi, 1986; Embretson, 1983, Guion, 1977; Messick, 1980, 1988, 1989b).

Con todo ello, el concepto de validez durante las últimas décadas ha pasado de un concepto que se distinguía por su criterio a una concepción unitaria con varias fuentes de evidencia (Martínez, 2001; Mislevy, 2009). En especial, la teoría de validez propuesta por Messick (1980, 1989b, 1995) y mayormente extendida por Kane (1992, 2001, 2006), puede verse fielmente aplicada en la propuesta de Embretson (1983, 1994, 1998), donde se relaciona con el uso de la teoría cognitiva para guiar el diseño, el desarrollo y la validación de las tareas evaluativas o ítems. Sin embargo, en la actualidad hay un nuevo debate sobre la teoría de la validez que cuestiona sus límites conceptuales, teóricos y metodológicos (ver Borsboom, 2006; Borsboom, Mellenbergh, & van Heerden, 2004; Kane, 2006; Messick, 1989b; Mislevy, 2009; Yang & Embretson, 2007).

Por su parte, en la última versión de los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999) se ve reflejada la Ley Federal de Educación de los Estados Unidos, las nuevas tendencias del concepto de validez, lo concerniente a las pruebas para personas con discapacidad, las diferencias entre personal de diferente ámbito lingüístico, los nuevos tipos de pruebas y los usos de las pruebas ya existentes. Estos *estándares* retoman especialmente la visión teórica de validez desde un constructivismo-realista (Embretson, 1998; Kane, 2001; Messick, 1989b).

Una de las novedades de la última versión de los *estándares* (AERA, APA & NCME, 1999) es el establecimiento del estándar 1.8, relacionado con la obtención de evidencias en el proceso de respuesta ante los ítems de las pruebas. En dicho estándar se puede ver claramente reflejado el aspecto sustantivo de la validez de constructo que Messick (1989b) coloca como uno de los más importantes en su teoría de la validez de constructo. Además, se establece en el mismo estándar que la razón fundamental de una prueba o de las interpretaciones de las puntuaciones depende de las premisas sobre los procesos psicológicos o las operaciones cognitivas usadas por los examinados para responder a la prueba; de esta manera, las evidencias teóricas y empíricas soportan aquellas premisas que deben ser provistas para el argumento de validez.

1.1.2. Integración de la psicología cognitiva y los modelos de medición

Con la creciente influencia de la psicología cognitiva en la construcción y en la validación de los test, se ha provocado un mayor interés en el análisis de las operaciones mentales y de los procesos de respuesta requeridos ante los ítems de una prueba, con el objetivo

de lograr una mejor comprensión de los constructos o atributos medidos (Embretson, 1994; Snow & Lohman, 1989; Yang & Embretson, 2007). Por ejemplo, los análisis cognitivos —como una de las aportaciones de la psicología cognitiva al campo de la medición— pueden utilizarse para la creación de bancos de ítems con características psicométricas conocidas (Bejar, 2007; Embretson & Gorin, 2001), sin necesidad de calibrar los ítems en una muestra de sujetos reales. También, con el resultado de dichos análisis se pueden establecer reglas o modelos cognitivos que fundamenten el diseño de modelos de tareas para la Generación Automática de Ítems (GAÍ). Asimismo, un microanálisis sociocognitivo de los procesos de respuesta de los examinados, ante los ítems de las pruebas, puede ayudar a los desarrolladores y usuarios a conocer con mayor profundidad las diferentes explicaciones del conocimiento o de la habilidad de los examinados (Snow & Lohman, 1989).

Como ya se mencionó, hay diferentes modelos basados en la EDC, que en la actualidad pueden ayudar a medir con mayor profundidad estructuras especificas del conocimiento, habilidades y procesos de respuesta de examinados, con el fin de proveer información acerca de sus fortalezas y debilidades cognitivas. Los modelos propuestos por Mislevy (*The Model-Based Reasoning*, 2009) y Embretson, (*The cognitive design system approach*, 1998), anteriormente mencionados, son un ejemplo de ello. Cada uno de estos modelos propone la integración sustancial de la psicología cognitiva en el proceso de medición; asimismo, hacen énfasis a partir del diseño de las pruebas desde una teoría *fuerte* que guíe el desarrollo y la validación de éstas (Griel & Lai, 2012, 2013). Es por ello que en la actualidad varios académicos, investigadores y desarrolladores de

pruebas (por ejemplo, Borsboom & Mellenbergh, 2007; Embretson, 1998; Leighton & Gierl, 2007a; Yang & Embretson, 2007; entre otros) consideran que la EDC presenta un gran potencial para el desarrollo y el análisis de la validez de los test.

Una de las implementaciones a nivel técnico y más significativas de la psicología cognitiva en la medición durante los últimos años son los estudios y métodos cognitivos (por ejemplo, Gierl, Leighton, Changjiang, Jiawen, Rebecca & Tan, 2009; Gierl, Wang & Zhou, 2008; Hoppmann, 2007; Johnstone, Bottsford-Miller & Thompson, 2006; Rupp, Templin & Henson, 2010), en especial las *técnicas de pensamiento en voz alta* (Ericsson & Simon, 1984, 1993; Leighton, 2009; Leighton & Gierl, 2007b). Sin embargo, otros tipos de estudios cognitivos comúnmente utilizados son el *modelado computacional de simulación de problemas* (Anderson, 1976; Dehn & Shank, 1982; Newell & Simon, 1972; Tomkins & Messick, 1963), el *análisis cronométrico* o de latencia de respuesta (Fredericksen, 1980; Posner, 1978; Posner & Rogers, 1978), el modelado matemático (Embretson, 1983), el *método de correlaciones cognitivas* (Pellegrino & Glaser, 1979), el *método del análisis del seguimiento del sendero de la vista* (Newell & Simon, 1972), el *método de análisis de los errores sistemáticos en la ejecución de la tarea* (Brown & Burton, 1978) y el *método de análisis de las estrategias y estilos de resolución de problemas* (Dunker, 1945; van Lehn, 1989), por mencionar algunos.

Aunado a ello, otro de los grandes resultados de la integración de la psicología cognitiva con los modelos de medición son los modelos psicométricos componenciales. Dichos modelos se han denominado de diferentes maneras (Romero, 2010): *modelos componenciales* (van der linden & Hambleton, 1997), *modelos de clase latente*

restringida (Haertel, 1989), *modelos de la Teoría de Respuesta al Ítem* (TRI) estructurados (Rupp & Mislevy, 2007) y *Modelos de Diagnóstico Cognitivo* (MDC) (Nichols, Chipman & Brennan, 1995).

1.1.3. Innovaciones y desarrollos del EXHCOBA

El EXHCOBA es un examen computarizado de selección para ingreso a la educación media superior y superior, el cual se aplica desde 1992. Se caracteriza por ser una de las pruebas con mayor nivel de innovación en México (Backhoff, 2001). En la actualidad, el EXHCOBA presenta una versión innovadora relacionada con la Generación Automática de Ítems de Respuesta Compleja (GAÍRC). Tirado (2010) denomina Reactivos Estructurales Constructivos (REESCO) a los ítems de respuesta compleja de la nueva versión mencionada, los cuales —según el autor— consisten en una estrategia socio-cognitiva para evaluar conocimientos y habilidades básicas mediante el uso de diferentes elementos y herramientas informáticas (ver Apéndice 1).

Cabe señalar que además del desarrollo de las diferentes versiones del EXHCOBA, los desarrolladores e investigadores de la prueba han realizado una gran cantidad de esfuerzos y trabajos de investigación, encaminados a la mejora de los procesos de evaluación y de sus indicadores técnicos. El equipo de desarrolladores buscan constantemente el logro de los criterios y *estándares* (AERA, APA & NCME, 1999) para el desarrollo de pruebas psicológicas y educativas (por ejemplo, Antillón, Larrazolo & Backhoff, 2008; Backhoff & Larrazolo, 2012; Backhoff & Tirado, 1992, 1994; Backhoff, Aguilar & Larrazolo, 2006; Backhoff, Tirado & Larrazolo, 2001; Ferreyra,

Larrazolo & Backhoff 2010; González-Montesinos, 2004; Pérez, Backhoff, Larrazolo & Rojas, 2013; Tirado, 2010; Tirado, Backhoff, Larrazolo & Rosas, 1997). Así, el EXHCOBA es una de las pruebas en México que cuenta con la más amplia gama de trabajos de investigación relacionados con la calidad técnica de todos sus procesos y desarrollos.

Aunque el EXHCOBA tiene una amplia variedad de evidencias de calidad técnica, recabadas a lo largo de su consolidación, las innovaciones mencionadas vuelven fundamental al hecho de incorporar nuevos modelos analíticos capaces de poder explicar y dar cuenta de sus indicadores de validez. Por ello, la incorporación de modelos analísticos y de desarrollo, que tomen en cuenta las innovaciones relacionadas con la GAÍRC, es de suma importancia para la mejora de los procesos relacionados con el desarrollo y la validación del EXHCOBA. Además, en la actualidad es fundamental para cualquier desarrollador de pruebas a gran escala y de alto impacto estar al pendiente en la incorporación de los nuevos avances en teorías de la validez y en lo referente a los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999).

1.2. Planteamiento del problema

Si se voltea la mirada a la historia del desarrollo de las pruebas psicológicas y educativas, se puede reconocer fácilmente que dicho campo representa una de las más grandes contribuciones de las ciencias del comportamiento a nuestra sociedad. Sin embargo, no todas las pruebas y las evaluaciones desarrolladas son adecuadas o benéficas para los usuarios y los destinatarios. En algunas ocasiones, tanto el desarrollo como el uso de las pruebas, pueden no resultar en la mejora de la toma de decisiones

para la vida académica, laboral y clínica de las personas o para la mejora de los diferentes procesos de las instituciones usuarias en donde son aplicados.

Debe tomarse en cuenta que los beneficios de las pruebas y de las evaluaciones dependen en gran medida del buen o mal uso de las mismas. Por un lado, si el uso e interpretación de los puntajes de las pruebas es adecuado, estas pueden proporcionar una ruta más amplia y un acceso más equitativo a la educación, al empleo y a la salud. Por otro lado, la interpretación y el uso inadecuado de las pruebas y de las evaluaciones pueden causar un daño considerable a los examinados, y a los usuarios en general. De modo que, las decisiones basadas en interpretaciones de los resultados de evaluaciones mal fundamentadas o resultados no sustentados en evidencias científicas pueden llevar directamente a la ineficacia en la toma de decisiones en cualquiera de los contextos de aplicación. Al respecto, hay una vasta documentación sobre *estándares* rigurosos de calidad para pruebas psicológicas y educativas que pueden ayudar a guiar el buen uso de las mismas (ver AERA, APA & NCME, 1999; Conbrach, 1971, 1990; Conbrach & Meehl, 1955; Bejar, 2010; Kane, 2006; Martínez, Backhoff, Castañeda, De la Orden, Schmelkes, Solano-Flores, Tristán & Vidal, 2000; Messick, 1989a, 1989b, 1995; Mislevy, 2009; National Research Council, 2001). Con todo esto, se puede decir que realizar estudios de validez, así como mejorar y evaluar la calidad de las pruebas, ayuda al uso justo y ético de sus resultados en los distintos contextos y procesos de aplicación (AERA, APA & NCME, 1999).

En el caso del EXHCOBA —una prueba de gran escala que se aplica en el proceso de admisión de más de una docena de instituciones de nivel medio superior y

superior—, toma mayor importancia el seguimiento de *estándares* de calidad rigurosos que guíen el desarrollo de la misma. También es importante agregar que en la actualidad se evalúa a más de 100 mil estudiantes por año, de los cuales sólo serán seleccionados aquellos que cumplan con el puntaje establecido por la institución educativa a ingresar. Es por esto que para el uso e interpretación adecuada de las puntuaciones de la prueba, resulta de gran relevancia que todas sus versiones y aplicaciones alcancen los parámetros y *estándares* de calidad más rigurosos (Kane, 2006; Messick, 1989a, 1989b, 1995).

Por otro lado, según Backhoff y Larrazolo (2012) y Tirado (2010), el EXHCOBA se presenta como una de las pruebas más vanguardista a nivel nacional por sus innovaciones relacionadas con la GAÍ, la construcción de ítems respuesta compleja, la integración de aspectos sustanciales de la psicología cognitiva y la psicometría, la evaluación de competencias básicas estructurales presentes en el currículum nacional y el desarrollo de tareas evaluativas con un enfoque más auténtico. Junto a esto aparece la necesidad de mejorar sus procesos de desarrollo y de validación, así como de incorporar modelos de medición que integren de forma sustancial las innovaciones mencionadas (Backhoff & Larrazolo, 2012; Backhoff & Tirado, 1992, 1994; Tirado, 1986). Es también fundamental en la actualidad para cualquier desarrollador de pruebas a gran escala y de alto impacto estar al pendiente de la incorporación de las nuevas propuestas en teorías de la validez y de los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999).

Al respecto, en las propuestas basadas en la EDC (ver Embretson, 1998; Embretson & Gorin, 2001; Huff, Alves, Pellegrino & Kaliski, 2013; Mislevy, 2009; Pellegrino, Baxter & Glaser, 1999) se hace mención de la importancia que tiene utilizar modelos de medición, los cuales ayuden a proporcionar una base más sólida para organizar las observaciones y orientar la toma de decisiones. En la actualidad, distintos teóricos en el tema de desarrollo y validación de pruebas (ver Bejar, 2010; Borsboom & Mellenbergh, 2007; Embretson, 1994; Kane, 2006; Messick, 1995; Mislevy, 2009; Pellegrino, Baxter & Glaser, 1999), hacen hincapié en disminuir la sobre-simplificación de los argumentos fundamentales de las evaluaciones educativas, de sus variables internas y de los procesos cognitivos subyacentes a ellas. Dichos teóricos también hacen hincapié en integrar a los modelos de medición las aportaciones de la psicología cognitiva con el fin de obtener mayor comprensión de las variables internas desde un nivel más fino de análisis socio-cognitivo (por ejemplo: Embretson, 1998; Mislevy, 2009; Pellegrino, Baxter & Glaser, 1999).

Por su parte, el EXHCOBA se encuentra en un momento de grandes transformaciones. Como ya se mencionó, los desarrolladores de la prueba, junto con sus asesores e investigadores, trabajan en la transición de un modelo de medición de *redes nomológicas* (ver Embretson, 1994; Kane, 2006; Messick, 1989b; Mislevy, 2009) a un modelo de medición con GAÍ basada en un enfoque *top-down* (ver Bejar, 2002, 2010; Gorin y Embretson, 2013; Messick, 1989b). Con ello, se pretende fortalecer y colocar en un plano superior el modelo de desarrollo y validación del EXHCOBA, acercándolo a los modelos emergentes para el desarrollo de instrumentos basados en una teoría *fuerte*

(por ejemplo: *The cognitive design system approach*, Embretson, 1998; *The Model-Based Reasoning*, Mislevy, 2009; *The Assessment and Reasoning from Evidence*, Pellegrino, Baxter & Glaser, 1999) y en la integración de la psicología cognitiva y la psicometría (por ejemplo: DiBello, Stout, Roussos, 1995; Dimitrov, 2007; Romero, 2010; Embretson, 1984; 1995; Fischer, 1973; Rupp, Templin & Henson, 2010; Tatsuoka, 1985). También se resuelve una necesidad para el EXHCOBA el hecho de integrar modelos teóricos de medición y de validez más acordes con las innovaciones y desarrollos implementados en su versión de GAÍRC.

Con todo lo anterior, realizar estudios sobre el aspecto sustantivo de la validez de constructo, tomando en cuenta tanto las características de los procesos cognitivos requeridos para responder a los ítems del EXHCOBA como la estructura del modelo cognitivo subyacente a los ítems de una prueba, representa grandes beneficios para su proceso de desarrollo y validación (Messick, 1989b). A su vez, la mejora en dichos procesos asegura el uso y la interpretación más adecuados y éticos de sus resultados para la toma de decisiones. Asimismo, otros contextos como la realimentación y operación del currículum, el diagnóstico de las estrategias de enseñanza y la mejor comprensión del proceso de aprendizaje se verán directamente beneficiados (Cortada de Kohan, 2000).

En síntesis, realizar el análisis del aspecto sustantivo de la validez de constructo del EXHCOBA es de alta relevancia social y de gran avance teórico-metodológico en el campo de las pruebas estandarizadas a gran escala en México. Sin embargo, es necesario recordar que las áreas y los nodos de la prueba estudiada presentan una gran

diversidad y amplitud, por lo que se decidió que era suficiente para la presente tesis doctoral delimitar el análisis en una sola de las áreas de la prueba, en este caso, el área de Habilidades Cuantitativas (HC) en dos de sus versiones (una con ítems de opción múltiple y otra con ítems de respuesta compleja). Para dicha decisión, se procuró que el área elegida tuviera más posibilidad de ser contrastada con otros estudios y experiencias parecidas en otros países del mundo (por ejemplo, Chen & Macdonald, 2011; Gierl *et al*, 2009; Revuelta & Ponsoda, 1998; Romero, Ponsoda & Ximénez, 2008) y, así, poder llevar a cabo aportaciones a la discusión y al debate en el ámbito disciplinar de pertinencia. También, se procuró que el universo de reactivos elegidos fuera suficiente y asequible para ilustrar la aplicación del modelo teórico-metodológico adaptado en el presente trabajo. Además, se buscó establecer las bases y las condiciones para que en futuros estudios se desarrolle un programa de validación para obtener evidencias basadas en el proceso de respuesta y de la estructura del modelo cognitivo a todas las áreas y versiones de la prueba. Tomando en cuenta lo anterior, se considera de gran relevancia analizar las evidencias de validez obtenidas bajo el modelo teórico-metodológico adaptado, el cual integra las innovaciones y las características que actualmente presenta el EXHCOBA.

1.3. Objetivos del estudio

Con el propósito de guiar el trabajo analítico de la presente tesis, se estableció como objetivo general analizar el aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. Para alcanzar dicho propósito fueron precisados diferentes objetivos específicos:

- Documentar los fundamentos teóricos del análisis de evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas psicológicas y educativas.
- Adaptar y aplicar un modelo teórico-metodológico con enfoque *top-down* para obtener y analizar las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas psicológicas y educativas computarizadas.
- Obtener y analizar evidencias de validez basadas en los procesos de respuesta subyacentes a los ítems del área de HC del EXHCOBA.
- Evaluar el diseño del interfaz de los ítems del área de HC del EXHCOBA.
- Definir y estructurar el modelo cognitivo del área de HC del EXHCOBA.
- Obtener y analizar evidencias de validez basadas en la estructura del modelo cognitivo del área de HC del EXHCOBA.

1.4. Justificación

Hay diversas razones que justifican el presente análisis del aspecto sustantivo de la validez de constructo del área HC del EXHCOBA. Una de las principales razones es que se aportan diversas evidencias de validez y de calidad técnica de la prueba, lo que a su vez da pie a un uso más adecuado y ético de sus resultados, así como a una mejor toma de decisiones en sus contextos de desarrollo y de aplicación (AERA, APA & NCME, 1999).

En cuanto al desarrollo de la prueba, los investigadores y los desarrolladores pueden obtener del presente estudio mayor información para la mejora del diseño y de las propiedades técnicas del área de HC del EXHCOBA. Además, el modelo teórico-metodológico para el análisis del aspecto sustantivo de la validez de constructo planteado en esta tesis puede ser utilizado por los desarrolladores de la prueba para continuar con un programa de validación, con el fin de obtener evidencias del proceso de respuesta y de la estructura del modelo cognitivo en el resto de las áreas. Aunado a ello, el presente estudio de validez fortalece y coloca en un plano superior el modelo de desarrollo y validación del EXHCOBA, acercándolo a los modelos emergentes para el desarrollo de instrumentos basados en una teoría *fuerte* (por ejemplo, *The cognitive design system approach*, Embretson, 1998; *The Model-Based Reasoning*, Mislevy, 2009; *The Assessment and Reasoning from Evidence*, Pellegrino, Baxter & Glaser, 1999), en el uso novedoso de la GAÍ (por ejemplo, Bejar, 1993; Hornke & Habon, 1986; Collis, Tapsfield, Irvine, Dann & Wright, 1995; Gierl & Lai, 2012) y en la integración de la psicología cognitiva y la psicometría (por ejemplo, DiBello, Stout, Roussos, 1995;

Dimitrov, 2007; Romero, 2010; Embretson, 1984; 1995; Fischer, 1973; Rupp, Templin & Henson, 2010; Tatsuoka, 1985).

En cuanto al contexto de aplicación, con la información aquí presentada sobre las evidencias de validez del área de HC del EXHCOBA, los usuarios podrán —con mayor certidumbre y seguridad— tomar decisiones fundamentadas para sus procesos de admisión. Hay que recordar la importancia de que el EXHCOBA —al ser una prueba de alto impacto y de gran escala que se aplica en distintas instituciones de educación media superior y superior en México y que se utiliza para evaluar a más de 100,000 aspirantes al año (Backhoff & Larrazolo, 2012; Rosas, Ramírez & Larrazolo, 2009) — presenta los mejores parámetros de calidad en todos sus procesos. Lo anterior da cuenta del impacto que tiene la prueba en la vida académica de los estudiantes y de las instituciones usuarias; por lo que es evidente el gran beneficio para los tomadores de decisiones de las instituciones usuarias el saber que cuentan con un instrumento que tiene excelentes condiciones, evidencias e indicadores de su calidad. De modo agregado, con los resultados del presente estudio se puede proporcionar a diferentes actores de las instituciones usuarias, a los padres de familia y a los mismos examinados, mayor información y realimentación diagnóstica sobre los procesos y las estrategias cognitivas utilizadas para responder a los ítems del área de HC del EXHCOBA. Dado lo anterior, se mencionan de forma resumida diversos beneficios y alcances que justificaron el desarrollo del presente estudio:

- **Teóricos.** Se presenta la aplicación de un modelo teórico-metodológico adaptado con el fin de analizar el aspecto sustantivo de la validez de constructo de pruebas computarizadas basadas en la GAÍ y que utilizan ítems de respuesta compleja.

Dicho modelo puede ser útil para investigadores y para desarrolladores de pruebas psicológicas y educativas en diversos contextos de aplicación. También se muestra un modelo teórico de los procesos cognitivos que utilizan los examinados para responder a los ítems de respuesta compleja del área de HC del EXHCOBA. Lo anterior presenta una alta relevancia social e institucional, pues el área de HC del EXHCOBA, y en general todas las áreas de la prueba, evalúan competencias básicas presentes en el currículum nacional. De tal forma que los resultados de la evaluación, enriquecidos con el reporte del dominio de los atributos cognitivos subyacentes al área de HC de la prueba por parte de los examinados, abona a la comprensión de la operación del currículum de matemáticas a nivel primaria.

- **Prácticos.** Se obtienen evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo del área de HC del EXHCOBA. De forma puntual, se obtiene información sobre la calidad del diseño de la prueba, así como de los procesos y las estrategias utilizadas por los examinados para resolver los ítems del área de HC. Además, se obtiene información sobre las propiedades técnicas de los ítems y de la estructura del modelo cognitivo de la prueba con la aplicación del modelo RASCH, y de los modelos componenciales.

- **Institucionales.** Se aportan evidencias de validez y de calidad técnica del EXHCOBA, lo que a su vez da pie a un uso más adecuado y ético de sus resultados, así como a una mejor toma de decisiones en los diversos contextos de aplicación. Por su parte, los tomadores de decisiones de las instituciones usuarias podrán tomar decisiones para sus procesos de admisión con mayor certidumbre y seguridad.

- **Sociales.** Se le puede proporcionar mayor información sobre la validez y calidad técnica de la prueba a los más de 100,000 examinados que contestan el EXHCOBA anualmente. A su vez, esto ayuda a generar en ellos mayor seguridad y confianza al saber que son evaluados bajo condiciones adecuadas, justas y éticas. De forma agregada, se puede proporcionar información a los diferentes actores de las instituciones usuarias (por ejemplo: directores, profesores, orientadores

educativos y pedagógicos, tutores, entre otros), a los padres de familia y a los mismos examinados; también, mayor información y realimentación diagnóstica sobre los procesos y estrategias cognitivas utilizadas para responder los ítems del área de HC del EXHCOBA.

- **Técnicos.** Se presenta una variedad de técnicas analíticas, y se ilustra su aplicación, lo cual resulta valioso para los investigadores y los desarrolladores del EXHCOBA, ya que así se ayuda a la identificación de posibles mejoras e implementaciones del diseño y del desarrollo de la prueba y, en específico, del área de HC. También se ilustra un programa integral de validez que puede ser aplicado al resto de las áreas del EXHCOBA, abonando a la generación de las condiciones necesarias para mejorar la obtención de las evidencias de validez del proceso de respuesta y de la estructura del modelo cognitivo, lo que a su vez ayuda al análisis del aspecto sustantivo de validez de constructo de la prueba. De forma agregada, se proporciona e ilustra la aplicación de diversas técnicas que pueden ser útiles para otros investigadores y desarrolladores en otros procesos y contextos de medición que deseen realizar estudios para obtener, y analizar, las evidencias de validez del proceso de respuesta, y de la estructura del modelo cognitivo.

1.5. Estructura de la tesis

La tesis consta de cinco capítulos y dos secciones al final (referencias y apéndices). En el primer capítulo, como ya se observó, se describe la introducción al tema central de la tesis. En ella se mencionaron algunos antecedentes sobre la discusión de las teóricas del concepto de validez y de los *estándares* internacionales para el desarrollo de pruebas psicológicas y educativas. A su vez, se mencionaron algunos modelos de medición que incorporan diferentes aportaciones de la psicología cognitiva y la psicometría. De igual manera, se presentaron algunos antecedentes e innovaciones recientes del EXHCOBA. Con ello, se dio pauta a la problematización del tema de investigación. Después, se describieron los objetivos del estudio que proporcionaron una guía para realizar el trabajo analítico de la presente tesis y, consecutivamente, se presentó la justificación del estudio y algunos de sus beneficios y sus alcances.

El segundo capítulo, presentado en las siguientes páginas, estructura los fundamentos teóricos que soportan el trabajo analítico de la tesis. Para ello, se abordan primero los fundamentos e innovaciones del EXHCOBA y el tema relacionado con las evidencias de validez del proceso de respuesta y de la estructura del modelo cognitivo de pruebas psicológicas y educativas. También se abordan algunos modelos para el desarrollo de pruebas basados en el constructo. Aunado a ello, se muestran algunas técnicas para analizar los procesos que utilizan los examinados para responder a los ítems de una prueba y se comenta lo referente a la clasificación y a las características de aplicación de los modelos psicométricos componenciales.

En el tercer capítulo se presenta el modelo teórico-metodológico adaptado y el conjunto de procedimientos realizados en el presente estudio en cada una de las fases, etapas y actividades establecidas en dicho modelo. Consecutivamente, en el capítulo cuatro se presentan las últimas dos fases relacionadas con el análisis de los resultados de la aplicación del modelo teórico-metodológico adaptado para obtener evidencias de validez basadas en el proceso de respuesta y de la estructura del modelo cognitivo del área de HC del EXHCOBA.

En el quinto y último capítulo de la tesis se describen las conclusiones de la investigación, entre las cuales destacan las aportaciones del modelo teórico-metodológico adaptado al análisis de las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de la prueba. Sin embargo, también se reconocen limitaciones, como el problema relacionado al descarte de los datos de algunos reportes verbales aplicados a estudiantes de secundaria que no aportaron la información necesaria sobre los procesos de respuesta ante los ítems de la prueba. Finalmente, se formulan perspectivas para futuras investigaciones que permitan la toma de decisiones sobre bases más objetivas en el campo joven en nuestro país de los estudios para obtener evidencias basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas educativas y psicológicas.

II. MARCO TEÓRICO

A lo largo de este capítulo se muestran los fundamentos teóricos requeridos para el trabajo analítico desarrollado en la tesis. Para ello, se muestran cinco grandes temáticas relacionadas con los fundamentos del EXHCOBA y con el análisis del aspecto sustantivo de la validez de constructo de pruebas psicológicas y educativas. La primera temática presentada al inicio de la sección del capítulo es sobre los fundamentos teóricos de las innovaciones, incorporados actualmente en el EXHCOBA. Después, en la segunda sección se discute sobre el debate teórico del concepto de validez y, en particular, el aspecto sustantivo de la validez de constructo. En la tercera sección se abordan algunos de los modelos de medición que integran diferentes aportaciones de las nuevas teorías de la validez, así como de la integración de la psicología cognitiva y la psicometría. Para la cuarta sección del capítulo se revisan diferentes estudios cognitivos que comúnmente son utilizados para el análisis del proceso de respuesta de los examinados ante las tareas evaluativas. En la quinta y última sección del capítulo del marco teórico se muestra la temática relacionada con los modelos de la perspectiva psicométrica-cognitiva, conocidos también como modelos componenciales.

2.1. Innovaciones y desarrollos del EXHCOBA

El EXHCOBA es un examen computarizado de selección para ingreso a la educación media superior y superior, el cual se caracteriza por ser una de las pruebas con mayor nivel de innovación en México. Actualmente cuenta con diferentes versiones,

desarrolladas y aplicadas a lo largo de varios años. De 1992 a 1993 se dispuso de una primera versión en formato lápiz y papel, de opción múltiple, de gran escala y basada en la Teoría Clásica de la Medida, la cual se adaptó para evaluar el aprendizaje escolar de los estudiantes que deseaban ingresar a la UABC; en 1994 se presentó esta primera versión en formato computarizado (con el Sistema Computarizado de Exámenes: SICODEX) para administrar anualmente el EXHCOBA a los 14'000 aspirantes de la UABC (Backhoff y Tirado, 1992; Backhoff, Ibarra y Rosas, 1994; 1995). A finales de 1994 se crearon dos versiones más (versiones dos y tres); de 1995 a 1996 se modificaron las versiones 2 y 3 y se crearon las versiones 4 y 5; de 1997 al 2007 se dispone de las versiones 2, 3, y 4 modificadas (Backhoff, 2001) y, posteriormente, se crearon las versiones 6, 7 y 8, que actualmente operaran. Todas estas versiones, evalúan contenidos curriculares básicos de los tres niveles educativos (primaria, secundaria y medio superior), quince áreas o nodos relacionados con las asignaturas de matemáticas, español, ciencias naturales y ciencias sociales a lo largo de los tres niveles educativos mencionados y, en total, cada versión contiene 310 ítems (ver Tabla 2.1).

Tabla 2.1. Secciones y áreas de las versiones 6, 7 y 8 con ítems de opción múltiple del EXHCOBA

Nivel	Sección	Área	No. de ítems
Primaria	Habilidades Básicas (60 contenidos curriculares representados)	Habilidades Verbales	30
		Habilidades Cuantitativas	30
Secundaria	Conocimientos Básicos (70 contenidos curriculares representados)	Español	15
		Matemáticas	15
		Ciencias Naturales (C.N.)	20
		Ciencias Sociales (C.S.)	20
Total de ítems que deberá responder un aspirante para ingresar a la educación media superior			130
Media superior	Conocimientos Básicos para Especialidad * (60 contenidos curriculares representados según el área disciplinar de la carrera a la que se desea ingresar)	Matemáticas para el cálculo	20
		Matemáticas para la estadística	20
		Física	20
		Química	20
		Biología	20
		Ciencias Sociales	20
		Humanidades	20
		Lenguaje	20
		Cs. Económico-Administrativas	20
Total de ítems que deberá responder un aspirante para ingresar a la educación superior			190
Total de ítems del EXHCOBA V-ÍOM con base en la totalidad de sus áreas			310
Total de ítems del EXHCOBA V-ÍOM, tomando en cuenta las últimas tres versiones actualizadas			930

En la actualidad se encuentra en desarrollo una nueva versión del EXHCOBA basada en la Generación Automática de Ítems de Respuesta Compleja (GAÍRC). La fundamentación teórica de dicha versión tiene cuatro directrices importantes: (a) evaluar competencias estructurales de inclusión del conocimiento (competencias básicas) que

dan soporte al aprendizaje subsecuente y que se encuentran presentes en el currículum nacional de educación básica; (b) evaluar primordialmente procesos cognitivos complejos de comprensión, de aplicación y de evaluación; (c) evaluar de forma auténtica, o lo más cercano posible, cómo aprenden naturalmente los estudiantes de los diferentes niveles educativos; y (d) evaluar los dominios comprometidos en el EXHCOBA mediante la utilización de la GAÍRC (Backhoff, 2012). Asimismo, la nueva versión del EXHCOBA basada en la GAÍRC se desarrolló bajo un modelo de cinco fases generales (ver Tabla 2.2).

Tabla 2.2. Modelo general de evaluación del EXHCOBA

Fases	Etapas
I. Planeación general	1. Diseño del plan general de evaluación
	2. Diseño y elaboración de cuestionarios de contexto
	3. Diseño y desarrollo del sistema informático
II. Diseño de la prueba	4. Selección y justificación de contenidos por especialistas en el currículum
III. Construcción	5. Elaboración de las especificaciones y de los modelos de ítems
	6. Validación del contenido de los reactivos
	7. Piloteo de las pruebas y de los cuestionarios de contexto
IV. Administración	8. Aplicación de las pruebas y administración de los resultados
V. Análisis e interpretación de los resultados	9. Análisis psicométrico inicial
	10. Estudios de validez
	11. Análisis y reporte de resultados

Asimismo, la Versión con Ítems de Respuesta Compleja (V-ÍRC) —como todas las Versiones con Ítems de Opción Múltiple (V-ÍOM) del EXHCOBA— fue desarrollada con apoyo de diferentes teorías cognitivas (Backhoff & Larrazolo, 2012; Backhoff & Tirado, 1992, 1994; Tirado, 1986). Sin embargo, es importante señalar que la estructura interna de cada una de estas versiones se encuentra construida bajo un modelo de medición de *redes nomológicas* (ver Conbrach & Meehl, 1955; Kane, 2006; Messick, 1989b). Según el modelo general de evaluación del EXHCOBA, los ítems se desarrollan con base en competencias básicas estructurales ubicadas en diferentes áreas del currículum nacional, las cuales son seleccionadas para su evaluación por un comité de especialistas del currículum de la educación primaria, secundaria y media superior en México (Backhoff & Tirado, 1992). Después, durante la etapa cinco del modelo de evaluación se desarrollan las especificaciones y las tareas evaluativas de la prueba (ver Apéndice 1). Entonces, con base en la definición de dichas tareas, se desarrollan cada uno de los templetes en el editor informatizado de la prueba. Se puede observar que los procedimientos para estructurar la prueba no están basados en reglas o en atributos determinados por la estructura de un modelo cognitivo sustantivo que tome en cuenta los procesos naturales de los examinados. En la Tabla 2.3 aparecen las secciones y las áreas de la versión del EXHCOBA basada en la Generación Automática de Ítems de Respuesta Compleja (GAÍRC).

Tabla 2.3. Secciones y áreas de la versión del EXHCOBA basada en la GAÍRC

Nivel	Sección	Área	No. de ítems base (padre)
Primaria	Habilidades básicas (40 contenidos curriculares representados)	Habilidades Verbales	20
		Habilidades Cuantitativas	20
Secundaria	Conocimientos básicos (70 contenidos curriculares representados)	Español	20
		Matemáticas	20
		C.N. (Física, química y biología)	20
		C.S. (Formación cívica y ética, Geografía e Historia)	20
	Total de ítems para ingreso a la educación media superior		120
Media superior	Conocimientos básicos para la especialidad * (200 contenidos curriculares representados)	Matemáticas	20
		Lenguaje	20
		Lengua extranjera	20
		Comunicación e informática	20
		Física	20
		Química	20
		Biología	20
		Ciencias Sociales	20
		Humanidades	20
		Cs. Económico-Administrativas	20
	Total de ítems para ingreso a la educación superior		180
	Total de ítems base tomando en cuenta la totalidad de sus áreas		320
		Total de ítems isomorfos (hijos)	no calculado

Cabe puntualizar que el modelo ingenieril de la GAÍRC de la nueva versión de EXHCOBA se fundamenta en modelos de tareas con los cuales se producen *ítems base* (también conocidos como *ítems padres)* y, con los que a su vez, se generan ítems isomorfos (también conocidos como *ítems hijos*), los cuales presumiblemente presentan

propiedades psicométricas similares. Para ello, cualquier modelo ingenieril de GAÍ requiere de templetes de ítems que funcionan como estructuras para sostener, mantener y resguardar la fidelidad de las interpretaciones del dominio especificado en modelos de tareas, es decir, un templete ayuda a asegurar que los ítems de una familia (conjunto de ítems hijos generados del mismo ítem padre) se comporten de manera similar en términos de sus características operativas (ver Embretson, 2001; Luecht, 2008; Gierl & Lai, 2013).

Diferentes autores denominan este tipo de GAÍ como basada en una teoría *débil* (por ejemplo, Bejar, 2007; Haladyna, 2007; Gierl & Lai, 2013). Esta definición se debe a que el diseño de los ítems no está basado en componentes definidos por una teoría cognitiva o teoría *fuerte* (Bejar, 2010; Griel & Lai, 2013; Lohman, 1989; Messick, 1989b), sino en modelos de tareas estructuradas.

Tanto las V-ÍOM como las V-ÍRC del EXHCOBA presentan diversas áreas, elementos y operaciones en su interfaz: los controles de administración, las tareas operativas de respuesta, los controles de ayuda, el texto y los elementos (imágenes y dibujos) de las indicaciones, así como el texto y los elementos de la base del ítem (ver Figura 1). En cuanto a las tareas operativas de respuesta, los ítems de respuesta compleja presentan tres tipos: (a) selección de elementos (ver Figura 2), (b) arrastre de elementos (ver Figura 3), y (c) escritura numérica y algebraica (ver Figura 4). En cambio, los ítems de opción múltiple sólo requieren de la tarea operativa de selección (ver Figura 5).

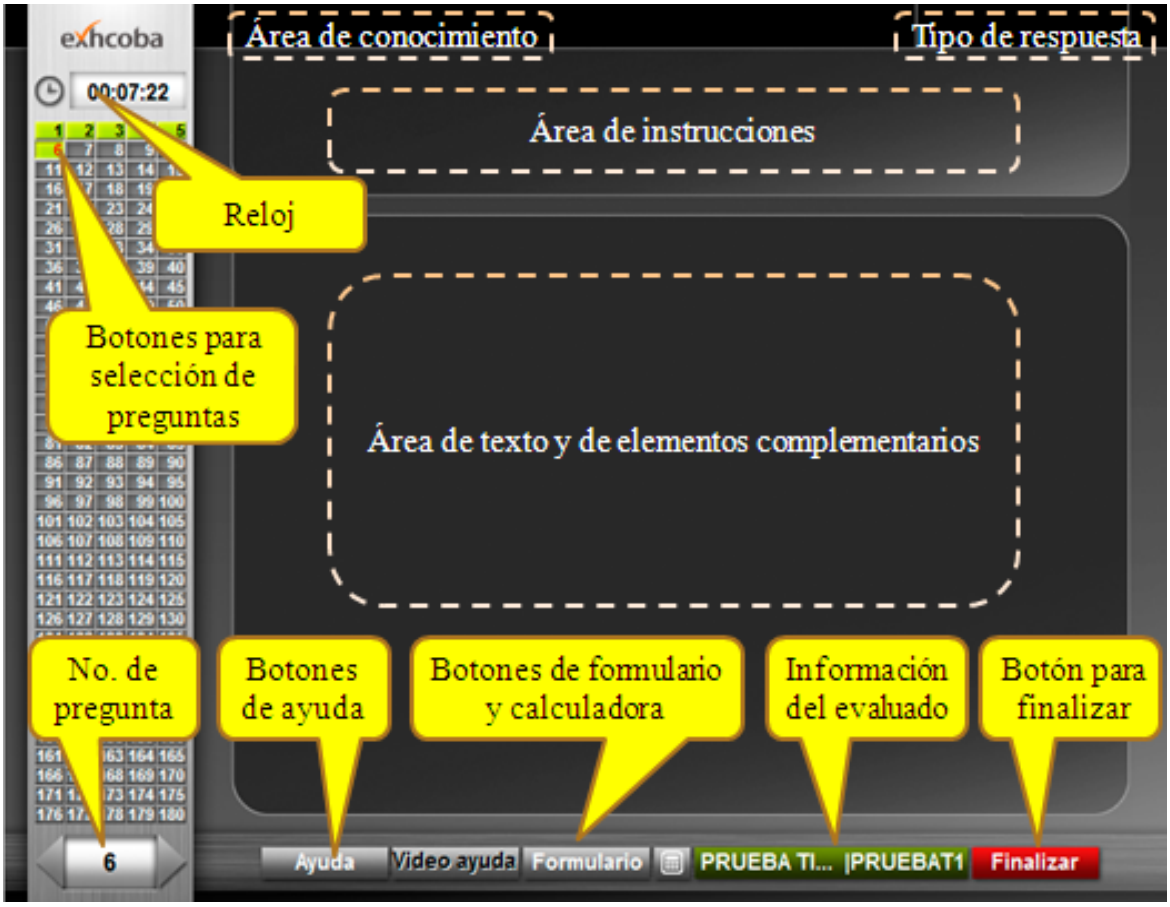


Figura 1. Áreas y elementos del interfaz del EXHCOBA

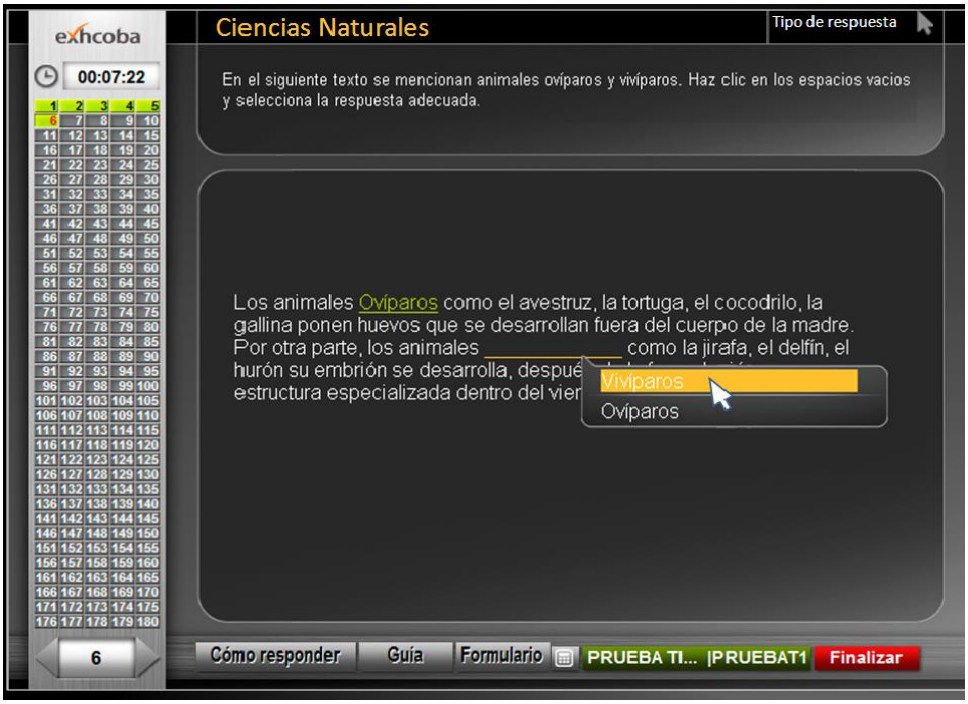


Figura 2. Ítems de respuesta compleja con tarea operativa de selección de elementos

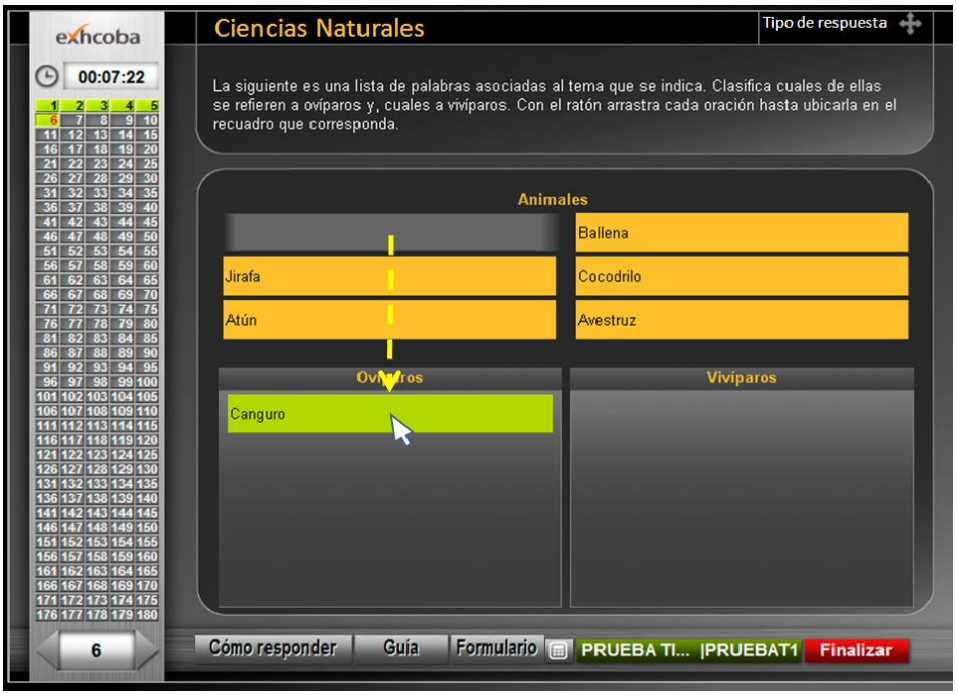


Figura 3. Ítems de respuesta compleja con tarea operativa de arrastre de elementos

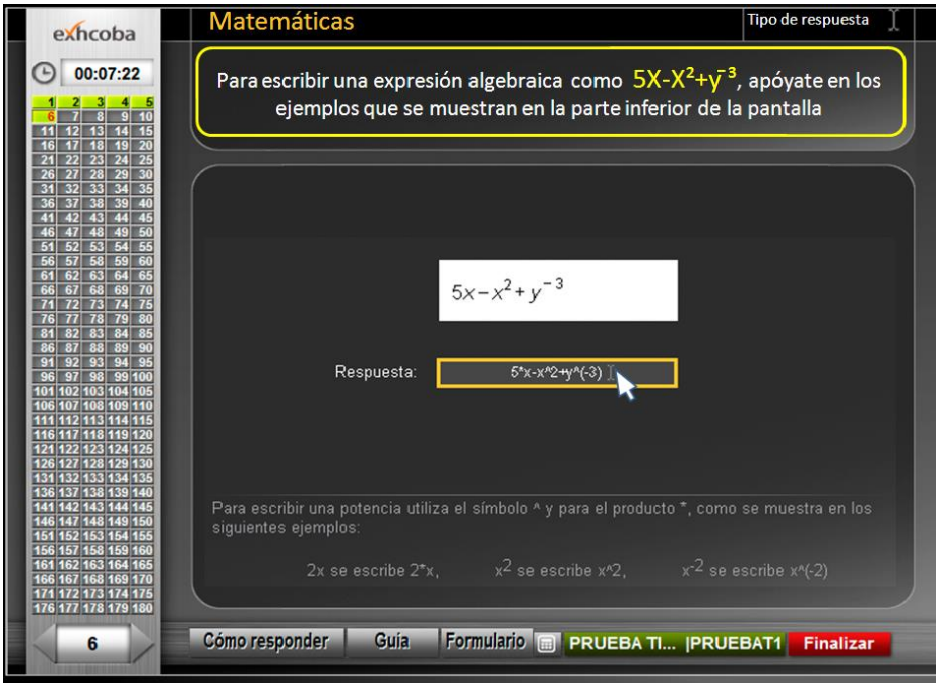


Figura 4. Ítems de respuesta compleja con tarea operativa de escritura numérica y algebraica

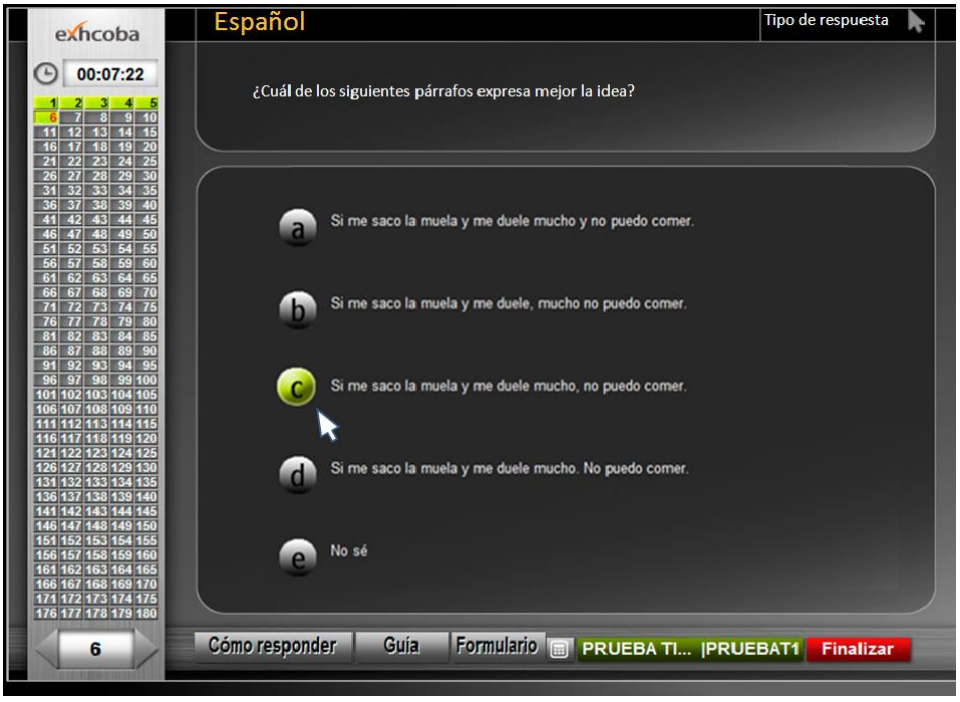


Figura 5. Ítems de opción múltiple con tarea operativa de selección

2.2. Evidencias de validez del aspecto sustantivo de pruebas psicológicas y educativas

En el campo de la medición, el tema de la validez merece un profundo análisis para poder identificar sus distintas aproximaciones, características y aplicaciones. Para la presente tesis se decidió revisar dos de las posturas más representativas sobre el concepto de validez, debido a su relevancia e impacto a nivel internacional en el desarrollo de las pruebas educativas y psicológicas. También se consideró relevante el análisis de los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999) en relación al tema de la validez y, en especial, al estándar 1.8 que aborda el tema de las evidencias de validez basadas en el proceso de respuesta.

2.2.1. Posturas teóricas sobre el concepto de validez

El concepto de validez ha sufrido cambios importantes durante las últimas dos décadas en cuanto a sus límites y a sus alcances teóricos, metodológicos y sociales. Como ya se mencionó durante la introducción de la presente tesis, dicho concepto pasó de distinguirse por su criterio predictivo a una concepción unitaria e integrada por diversos aspectos sustantivos (Mislevy, 2009; Sireci, 2008). Sin embargo, en la actualidad hay un nuevo debate sobre la teoría de la validez que cuestiona sus límites y alcances (Borsboom, 2006; Borsboom & Mellenberg, 2007; Kane, 2006; Mislevy, 2009; Yang & Embretson, 2007). Hay que puntualizar que —aunque la validez es un aspecto crucial relacionado a la calidad técnica de las evaluaciones educativas y psicológicas, dependiendo del enfoque o marco de referencia— la validez puede referirse, o hacer

énfasis, en aspectos psicométricos, en el uso e interpretación de resultados o en las consecuencias y las repercusiones socio-políticas de sus usos.

Una de las primeras conceptualizaciones de la validez fue la que propusieron Conbrach y Meehl (1955). Dichos autores denominaban la validez como *validez de constructo* y, además, la definieron como un indicador que describe el significado de las puntuaciones de un instrumento de medición. Además, mencionan que hay cuatro tipos de validez referidos por distintos tipos de investigación y que requieren diferentes tipos de interpretación: (a) validez predictiva, (b) validez concurrente, (c) validez de contenido y (d) validez de constructo o de estructura interna (Conbrach, 1971, 1990).

Para fundamentar su propuesta, Conbrach y Meehl (1955) definen al constructo como un postulado sobre algún atributo de los individuos, el cual —se asume— está reflejado en el rendimiento de una prueba. Estos autores mencionan también que, en las pruebas de validez, el atributo del cual se hacen declaraciones para las interpretaciones de una prueba es un constructo. Desde dicha postura de la validez, se espera que una persona ostente o no (cualitativamente) un atributo, o que tenga en cierto grado (cuantitativamente) un atributo. Con ello, se implica la lógica de la validación del constructo cuando este es fuerte o débilmente sistematizado, ya sea a través de la ramificación teórica y de proposiciones simples o del uso de las proposiciones absolutas o las afirmaciones probabilísticas.

En la actualidad, según Rupp, Templin y Henson (2010), el debate teórico de la validez se enmarca en la propuesta de dos perspectivas teóricas: la messickiana y la holandesa. La primera propone la interrelación de los temas clásicos de la validez como

el aspecto fundamental de una teoría más comprensiva que aborda el significado de las puntuaciones, los valores sociales de sus interpretaciones y el uso de las pruebas (Kane, 2001; Messick, 1989a y 1989b). Con ello, se hace énfasis en el aspecto sustantivo y en el aspecto consecuencial de la validez de constructo. En especial, como ya también se mencionó en la introducción, la teoría de validez —propuesta por Messick (1989b), extendida por Kane (2001) y fielmente representada por Embretson (1998) — se basa en una concepción *constructivista-realista* del concepto de validez. En cambio, la teoría de validez propuesta por Borsboom, Mellenbergh y van Heerden (2004) se basa en un *constructivismo radical*. Dicha escuela Holandesa pone énfasis la validez como un atributo de la prueba relacionado con la exactitud de ésta y con la noción de causalidad de las puntuaciones. Borsboom (2006) menciona que una prueba es válida para la medición de un atributo si, y sólo si, la variación en el atributo causa variación en los resultados de la medición a través de los procesos de respuesta que suscita la prueba (Borsboom & Mellenberg, 2007).

Para Messick (1989a y 1989b), las decisiones y las consecuencias resultantes de los participantes son claves en el diseño de las evaluaciones. Estas deben ser explícitamente evaluadas y comprometidas en la noción de validez de constructo. Cuando Messick habla de la validación consecuencial se refiere a las consecuencias positivas y negativas asociadas a las fuentes de invalidación provenientes del diseño y de la aplicación de las evaluaciones. Dicha noción consecuencial de la validez es vista por los profesionales en el campo de la medición como una de las contribuciones más importantes de Messick al campo de la medición. Otra contribución importante es la

integración de varios aspectos de la teoría de validez en un paraguas conceptual coherente y completo bajo el término de validez de constructo (Yang & Embretson, 2007). Sin embargo, según Messick (1995), la validez como un concepto unificado implica también distinguir entre aspectos de esta, con el fin de enfatizar temas importantes y matices que, de no tomarse en cuenta, se les restaría importancia o se pasarían por alto, como en el caso de las consecuencias sociales de las evaluaciones del desempeño o del papel del significado de las puntuaciones en el contexto de aplicación. La intención de dichas temáticas y matices es proporcionar un medio para abordar los aspectos funcionales de la validez con el fin de desentrañar las complejidades inherentes a la conveniencia, a la significación y a la utilidad de las puntuaciones y de las inferencias de los resultados de las evaluaciones. De forma puntual, se distinguen siete aspectos de la validez de constructo, los cuales pueden ser brevemente resumidos de la siguiente forma (adaptado de Messick, 1995, p.745):

- **Aspecto de contenido.** Se busca si el contenido de la evaluación representa el dominio en estudio.
- **Aspecto sustantivo.** Se busca si los participantes comprometen los procesos cognitivos apropiados al momento de responder las tareas evaluativas o ítems de las pruebas.
- **Aspecto estructural.** Se busca que los puntajes de respuesta de los examinados reflejen la interacción de las variables o atributos latentes del dominio.
- **Aspecto predictivo.** Se busca que los puntajes como resultado de las evaluaciones puedan usarse para predecir una variable de interés.
- **Aspecto consecuencial.** Se busca que las interpretaciones presenten consecuencias justas y defendibles para los examinados.

- **Aspecto externo.** Se busca que los participantes respondan de forma similar en las evaluaciones que miden constructos similares y de forma diferente en las evaluaciones con diferentes constructos.
- **Generalizabilidad.** Se busca que los resultados de la evaluación puedan ser generalizados a través de diferentes condiciones de tiempo, contexto, administración y muestras de los examinados.

Por su parte, la perspectiva propuesta por Borsboom, Mellenbergh y van Heerden (2004) difiere en buen grado de la teoría de validez de constructo de Messick (1995). Estos tres autores argumentan que el concepto de validez de constructo en la perspectiva messickiana se encuentra innecesariamente sobresaturado. Es por ello que sugieren distinguir el lugar en donde debería ubicarse la validez de una prueba en tres familias conceptuales (adaptado de Rupp, Templi & Henson, 2010):

- conceptos de medición (ej. validez, unidimensionalidad y precisión);
- conceptos de decisión (ej. exactitud predictiva, optimización); y
- conceptos de impacto (ej. aceptabilidad y justicia).

En especial, Borsboom y colaboradores consideran que la validez debería ser investigada únicamente con métodos psicométricos. Así, todas aquellas propiedades y facetas del concepto de medición deben ser investigadas directamente con métodos psicométricos cuantitativos. Por otra parte, los conceptos de decisión pueden ser investigados igualmente con métodos psicométricos cuantitativos, pero pueden utilizarse evaluaciones adicionales o criterios, donde las puntuaciones del diagnóstico de los atributos pueden ser relacionados empíricamente (Borsboom, Mellenbergh & van

Heerden, 2004). Para los conceptos relacionados con el impacto de las evaluaciones se deben manejar cualitativamente, tomando en cuenta las decisiones basadas en las evaluaciones. Entonces, no pueden trazarse o analizarse directamente todos los conceptos mencionados con el estudio de las propiedades psicométricas de las evaluaciones. Por ello, Borsboom y colaboradores sugieren la asociación del término validación con el concepto de medición, y verlo como un problema de causalidad.

De forma específica, los métodos de validez, desde la perspectiva holandesa, comprenden que en cada tipo de evaluación se hace énfasis en distintas fases de su desarrollo. En el caso de la Evaluación Diagnóstica Cognitiva (EDC), el efecto causal capturado por el puntaje implica que la validación sea centralizada por la misma evaluación, dado que su objetivo explícito es capturar el proceso de respuesta de los examinados. Puesto de otra manera, la evaluación diagnóstica a menudo pone un gran énfasis en el proceso cognitivo de respuesta en contraste con otros tipos de evaluación.

Además, la evaluación diagnóstica continúa requiriendo evidencia más comprensiva a lo largo de las distintas facetas de su desarrollo. Por su parte, las evaluaciones no diagnósticas ponen un gran énfasis en otras facetas, pero continúan requiriendo alguna evidencia para dar cuenta de la fidelidad de los procesos de respuesta con el constructo medido. De hecho, la perspectiva holandesa concibe que los métodos utilizados en la EDC son los que presentan mayor poder para capturar y para modelar el proceso cognitivo de respuesta, ya que estos analizan la validez de constructo de forma más directa.

Un punto de convergencia entre las dos propuestas teóricas de validez mencionadas tiene relación con el énfasis en una mejor representación del constructo. Es por ello que la integración de la teoría cognitiva y del análisis de los procesos de respuesta de los examinados ante los ítems de las pruebas proporciona un camino más seguro y apropiado para obtener fuertes evidencias de validez, evitando de esta forma la sobre-simplificación y la sobre-estimación en las mediciones educativas y psicológicas (Kane, 2001; Mislevy, 2009; Snow & Lohman, 1989).

2.2.2. Cambios en los estándares internacionales para las pruebas educativas y psicológicas

Antes de iniciar con el análisis de los cambios sufridos por los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999), es importante recordar que fueron escritos especialmente para profesionales y para expertos en el área del desarrollo de pruebas. Dichas normas se ocupan de aspectos profesionales y técnicos relacionados con el desarrollo de las pruebas y su uso en la educación, la psicología y el empleo. Con ello, los *estándares* son una referencia importante para los elaboradores de pruebas profesionales, patrocinadores, editores, usuarios, políticos, empresarios y estudiantes de educación y psicología. Entre los puntos más relevantes tratados en los *estándares* de 1999, se encuentran los relacionados con la construcción de las pruebas y su documentación; la equidad de las pruebas y la aplicación de las pruebas. En esta tesis, un apartado importante de los *estándares* es el relacionado al tema de validez, el cual se encuentra considerado en los *estándares* relacionados con la construcción de las pruebas.

Para iniciar con la discusión de los cambios sufridos en la versión de los *estándares* de 1999, es necesario señalar que estos son el resultado más importante del aún actual debate del concepto de validez (ver Borsboom, 2006; Borsboom & Mellenberg, 2007; Kane, 2006; Mislevy, 2009; Yang & Embretson, 2007). En dicha versión, la validez se conceptualiza de forma diferente que en la pasada versión (AERA, APA & NCME, 1985), en donde se consideraban diferentes tipos de validez (validez de constructo, validez de contenido, validez de criterio-concurrente), dependiendo del tipo de evaluación a desarrollar. Básicamente, en los *estándares* de 1999 se retoma la visión teórica de la validez de constructo de Messick (1995), pero procurando dejar clara la importancia de la aplicación técnico-legal para el desarrollo de los test educativos y psicológicos. De tal manera que, a la par con el cambio en la noción del concepto de validez dentro de los *estándares* cambia también la conceptualización de los tipos de evidencias de validez.

Ahora bien, es importante recordar que un evento previo y próximo al desarrollo de los *estándares* de 1999, fue el establecimiento de criterios para revisar la calidad técnica de las evaluaciones por el *National Center for Research on Evaluation, Standards and Student Testing* (CRESST, 1994). Los criterios establecidos por dicho organismo, apelan a tomar en cuenta la complejidad cognitiva, la calidad del contenido, la significatividad, la propiedad del lenguaje, la trasferencia y la generalizabilidad, la justicia, la confiabilidad y las consecuencias pretendidas en el desarrollo de una evaluación. Con ello, la versión de 1985 de los *estándares* fue revisada y reestructurada, dando como resultado la versión mencionada de 1999.

Como cambios visibles, la versión del manual 1999 de los *estándares* tiene más material de referencia en cada uno de sus apartados y una mayor cantidad de normas especializadas, además de un glosario e índices ampliados. También, se ve especialmente reflejada la Ley Federal de Educación actual en los Estados Unidos, la medición de diversas tendencias que afectan a la validez, lo concerniente a las pruebas para personas con discapacidad o diferente ámbito lingüístico, los nuevos tipos de pruebas y, los usos de las pruebas ya existentes. También se agregan otros tipos de evidencias considerados de igual o mayor relevancia que los ya existentes.

En la actualidad se establecen cinco los tipos de evidencias de validez en la versión de los *estándares* de 1999. El primer tipo de evidencia está basada en el contenido de la prueba; la segunda tiene que ver con el proceso de respuesta; la tercera evidencia de validez es la referida a la estructura interna; la cuarta evidencia se refiere a la relación con otras variables; y la quinta evidencia es la relacionada con las consecuencias de la prueba. Sin embargo, como ya se mencionó, no sólo se agregaron o cambiaron los tipos de evidencias de validez, sino que también cambió su conceptualización, y métodos para obtenerlas.

Por ejemplo, un cambio importante relacionado con las evidencias basadas en el contenido tiene que ver con la relevancia y la representatividad que tomó del contenido del test para definir y para clarificar el significado del constructo a medir. Con ello, las evidencias de contenido ahora retoman importancia para evaluar la propiedad de las inferencias hechas en una evaluación. Asimismo, las evidencias de validez basadas en el contenido deben dar cuenta de la superficialidad o profundidad de las características

estructurales del contenido de una prueba con respecto al constructo a medir (Yang & Embretson, 2007).

Otro de los cambios tiene que ver con la redefinición del aspecto estructural de una prueba. Básicamente, ahora las evidencias basadas en la estructura interna de la prueba tienen que ver con la relación del sistema de puntaciones, dada la estructura interna del constructo o dominio a medir, la cual debe estar basada tanto en un modelo nomológico como en un modelo de la teoría sustantiva. Con ello, para los constructos basados en un modelo nomológico los análisis factoriales toman relevancia en la obtención de evidencias de validez basadas en la estructura interna. Si los puntajes entre ítems y factores se relacionan, la evidencia empírica da cuenta de la plausibilidad de dicha estructura relacional.

Para la medición de constructos, a partir de un modelo de la teoría sustantiva, la aplicación de la EDC proporciona un camino más convincente y completo para la obtención de evidencias de validez basadas en la estructura interna, y de otros tipos de evidencias como las basadas en el proceso de respuesta y en el contenido. Dado lo anterior, diversos autores (Borsboom & Mellenbergh, 2007; Embretson, 1998; Leighton & Gierl, 2007a; Yang & Embretson, 2007) consideran que la EDC presenta un gran potencial para el desarrollo y el análisis de la validez de los test, pues las evidencias obtenidas bajo dicho modelo profundizan en cuanto a que las evidencias que se aportan con los modelos componenciales sobre la estructura de un modelo cognitivo de una prueba se relacionan con los componentes o atributos cognitivos que subyacen y se encuentran presentes en los procesos de respuesta de los examinados.

Otro de los cambios en los *estándares* tiene que ver con la incorporación del análisis de las evidencias de validez basadas en el proceso de respuesta. Dicho tipo de evidencias tratan de la racionalidad teórica y de las evidencias acerca del proceso cognitivo subyacente a los ítems de un test. Para el caso particular de los test de habilidad, en los *estándares* (AERA, APA & NCME, 1999) se establece que el proceso utilizado por los examinados para resolver los ítems debe ser evaluado con base en las características del constructo *a-priori*. Con ello, el análisis de las evidencias basadas en el proceso de respuesta, aporta información sobre el grado de ajuste entre el constructo y la práctica natural de las respuestas, o las respuestas reales comprometidas por los examinados ante los ítems. Por ejemplo, si la solución de ítems de opción múltiple de razonamiento cuantitativo depende principalmente del uso de la información presente en los distractores y no del dominio de interés, la lógica de la medición de dicho atributo se presentaría de forma inversa presentándose un proceso de respuesta inadecuado por parte de los examinados ante los ítems de la prueba.

Cada uno de los cambios en la última versión de los *estándares* (AERA, APA & NCME, 1999) relacionados con la redefinición de las evidencias basadas en el contenido y en la estructura interna, y con la incorporación de las evidencias basadas en el proceso de respuesta, tienen su explicación en los fundamentos teóricos de la concepción unitaria de la validez propuesta por Messick (1989b). En dicha concepción de validez, como ya se mencionó, se toma como punto de partida el enfoque constructivista-realista (ver Figura 6).

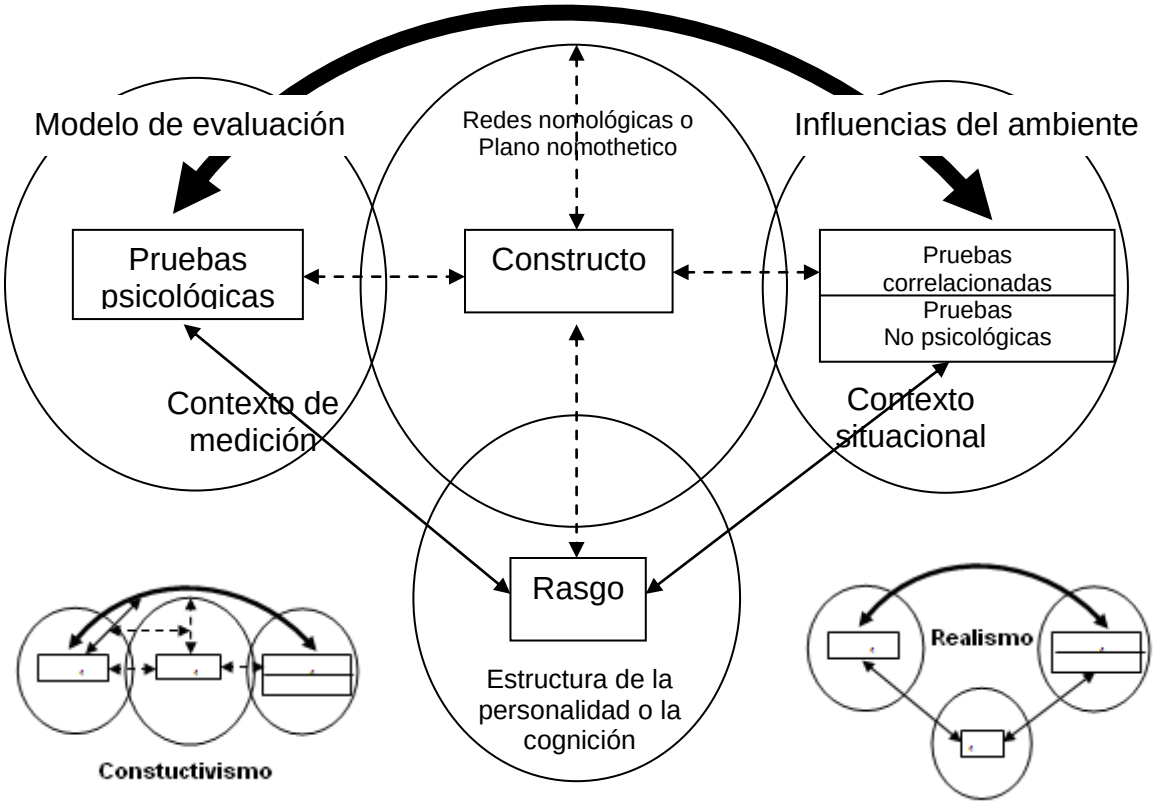


Figura 6. Enfoque constructivista-realista (adaptado de Messick, 1989b, p. 30)

La Figura 6 muestra un diagrama que representa la integración del enfoque *constructivista* y *realista* de la validación de constructo. Los tres círculos superiores del diagrama, junto con sus relaciones y términos, pertenecen al modelo del enfoque *constructivista*. Por su parte, los tres círculos externos del diagrama junto con sus relaciones y términos pertenecen al modelo del enfoque *realista*. En especial, el modelo integrador aporta un fuerte soporte teórico a la validez de constructo y por ende una base para la interpretación de las evidencias de validez. Cabe recordar, que en la propuesta de Messick (1989b) la validez de constructo, en esencia, comprende las evidencias y los

fundamentos que apoyan la exactitud de las interpretaciones de los resultados en términos de conceptos explicativos, que revelan tanto el desempeño de una prueba como las relaciones con otras variables.

Asimismo, Messick (1989b) menciona que el enfoque *constructivista-realista* para la interpretación de pruebas psicológicas y no psicológicas es un enfoque que aporta a la validez de constructo un fuerte argumento predictivo (Constructivismo) y explicativo (Realismo). Con ello, se retoman diversas cuestiones de la validación de constructos propuesta por Loeveringer (1957): (a) ¿En qué medida la prueba mide lo que se especifica mediante un constructo determinado? En términos generales, esto corresponde a lo que Campbell (1960) llamaba *validez de rasgo*; (b) ¿En qué medida el constructo incorpora una hipótesis válida? Por lo general, esto corresponde a lo que Campbell (1960) denominaba *validez nomológica* y Embretson (1983) etiqueta como *plano nomothetico*; (c) ¿En qué medida la prueba mide “algo” que “realmente” existe? Esta es una pregunta básica y previa a la consistencia conductual en el rendimiento de una prueba, que indica que "algo" de hecho se está midiendo; y (d) ¿Qué tan bien la propuesta de interpretación corresponde a lo que se mide en la prueba? Esencialmente, esto equivale a lo que llama Loevering (1957) como *validez sustantiva*, que evalúa qué tan bien en la interpretación de una prueba se capta la naturaleza de ese "algo" que se está midiendo.

Con respecto al cuestionamiento propuesto por Loeveringer (1957), relacionado con la medida en que una prueba mide “algo” que “realmente” existe, se puede decir que dicho cuestionamiento examina la magnitud y la consistencia de las relaciones, ambas

internas a la prueba y externas a otras medidas, es decir, ambas pertenecen al *aspecto estructural* de las relaciones internas y externas de la validez de constructo. Por otro lado, para el cuestionamiento relacionado con la exactitud en que la propuesta de interpretación del constructo corresponde a lo que se mide en la prueba, este examina el contenido y la naturaleza de las relaciones internas y externas de la validez de constructo, es decir, el *aspecto sustantivo* de la validez de constructo. Dado lo anterior, los dos aspectos internos mencionados, tanto el *estructural* como el *sustantivo*, son incorporados por Embretson (1983) en el concepto de *representación del constructo*, y los aspectos externos son incorporados en su noción del *plano nomothetico*.

A su vez, Messick (1989) menciona que para todos los aspectos de la validez de constructo hay dos amenazas principales a tomar en cuenta: (a) *insuficiente representación del constructo* y (b) *varianza irrelevante del constructo*. La *insuficiente representación del constructo* ocurre cuando un importante aspecto de la validez o alguna faceta de lo que se está midiendo se omiten. Por ejemplo, si un test de razonamiento cuantitativo incluye sólo problemas de algebra sería demasiado corto y, en consecuencia, sería representado insuficientemente el constructo de interés. Con respecto a la *varianza irrelevante del constructo*, esta ocurre cuando el desempeño depende de cualidades que no son consideradas parte del constructo a medir. Por ejemplo, si en el caso del mismo test de razonamiento cuantitativo se presenta una excesiva dependencia del lenguaje, la *varianza irrelevante del constructo* es introducida por tal hecho; es decir, para los individuos que presenten menor dominio de la

competencia lingüística, dicho test será más una medida del dominio del idioma, que del razonamiento cuantitativo.

Retomando la discusión sobre el enfoque *constructivista-realista*, Messick (1989b), se puede decir que dicho enfoque es *realista* porque asume que los rasgos y otras entidades causales existen fuera de la mente del teórico y es *constructivista-realista* porque se supone que dichas entidades causales no pueden ser comprendidas directamente, sino que deben ser estudiadas a través de construcciones de la mente o modelos representacionales. Para atribuir realidad a entidades causales, que simultáneamente requieren una construcción teórica de las relaciones observadas, con el enfoque *constructivista-realista* se aspira a alcanzar la riqueza explicativa de la posición puramente *realista* y, a su vez, se trata limitar los excesos metafísicos a través del análisis racional. Al mismo tiempo, el enfoque *constructivista-realista* espera retener las ventajas y el poder de predicción del enfoque *constructivista*.

Por su parte Loevinger (1957) argumenta a favor de la aplicación del enfoque *constructivista-realista* a la teoría de la validez. Dicho autor menciona que el concepto *constructo* connota tanto construcción como artificialidad. Con ello, lo que queda en cuestión es la validez respecto a lo que exactamente el psicólogo no construye. Asimismo, la validez de una prueba es una medida de los rasgos que existen antes e independientemente del acto de medición. De tal manera que, aun cuando algunos psicólogos sólo conocen los rasgos estudiados indirectamente a través del cristal de su construcción, es importante recordar que los datos a ser juzgados son manifestaciones de los rasgos, y no manifestaciones del constructo.

Hay otros cambios importantes de mencionar de la última versión de los *estándares* relacionados con el aspecto externo, consecuencial y de la generalizabilidad de la validez de constructo. Algunos de los aspectos mencionados, han sufrido fuertes cambios y, en algunos de los casos, se presentan algunos completamente innovadores como en el caso del aspecto sustantivo de la validez. Cada uno de los cambios presentes en la actual versión de los *estándares* ha impactado, en menor o mayor medida, en el desarrollo y en la aplicación de las pruebas, siendo esencial conocerlos a profundidad por los desarrolladores de pruebas y, en general, por todo actor en el campo de la evaluación.

En resumen, la última versión de los *Estándares para las pruebas educativas y psicológicas* (AERA, APA & NCME, 1999) presenta fuertes cambios en sus límites teóricos, metodológicos y sociales. En especial, autores como Embretson y Gorin (2001), Mislevy (2007), Yang y Embretson (2007), Leighton y Gierl (2007a), Bejar (2010) y Chen y Macdonald (2011) mencionan que los *estándares* presentan un nuevo enfoque en donde se hace mayor énfasis en las interpretaciones y en las inferencias de los resultados de las evaluaciones, y en sus consecuencias.

Por su parte, las evidencias más importantes a tomar en cuenta para la presente tesis son las relacionadas con el proceso de respuesta (estándar 1.8) y la estructura interna de la prueba (estándar 1.11). Según Loevinger (1957, en Messick, 1989b) las evidencias de validez basadas en el proceso de respuesta representan la habilidad de construir teoría que explique los resultados de una prueba. Con ello, la aplicación de estudios cognitivos puede ser de gran utilidad tanto para aportar información valiosa de

insumo durante el modelado del proceso cognitivo requerido para responder a los ítems y para aportar evidencias sobre el ajuste entre el modelo cognitivo de los ítems de una prueba y los procesos de respuesta naturales de los examinados ante dichos ítems (Leighton & Gierl, 2007b; Snow & Lohman, 1989).

También, evidencias recabadas con los métodos cognitivos aportan información para una mejor definición del constructo. Con ello, además de las evidencias de validez basadas en el proceso de respuesta, se pueden obtener, en cierto nivel, evidencias de validez relacionadas con el aspecto de contenido (Leighton y Gierl, 2007a). En especial, con la aplicación de la EDC y de los estudios cognitivos se aporta información valiosa sobre la adecuación de la definición y la descripción del contenido de la prueba. A través de la respuesta de los examinados ante los ítems o tareas evaluativas se puede obtener información que ayude a identificar errores de adecuación y de propiedad relacionados con la sobre-simplificación o sobre-estimación, tanto de la descripción de los contenidos como del desarrollo de los ítems. Con dicha información se puede enriquecer el propio desarrollo de las especificaciones de los reactivos y en general la homogeneidad de la estructura de la prueba, a la luz de los datos empíricos de los procesos de respuesta de los examinados.

Por ejemplo, si con una prueba se pretende evaluar el razonamiento matemático, es importante determinar si los examinados realizan el razonamiento esperado ante los ítems de dicha prueba o si simplemente siguen un algoritmo de respuesta estándar o un proceso de respuesta inadecuado, debido a problemas en el diseño de los ítems. Otra contribución importante a mencionar con la aplicación de los estudios cognitivos para el

análisis del aspecto sustantivo de validez de una prueba es la posibilidad de obtener diferentes significados válidos para la interpretación de los resultados (Ericsson & Simon, 1984, 1993, 1998; Messick, 1989b; Snow & Lohman, 1989).

Por otra parte, para las evidencias de validez basadas en la estructura interna de la prueba, la aplicación de los modelos componenciales puede arrojar evidencias sobre la propiedad de la estructura del modelo cognitivo que subyace a la prueba. Para dicho propósito se toman en cuenta los atributos relevantes, ya sean predefinidos por un modelo teórico u obtenidos de forma inductiva con ayuda de estudios, tales como los *análisis verbales* y los *análisis de protocolos*. Aunado a ello, algunos de los Modelos de Diagnóstico Cognitivo (MDC), como el de Embretson (1983), incorporan el aspecto interno estructural y sustantivo del enfoque *constructivista-realista* a través de la noción de la *representación del constructo*, y el aspecto externo de las relaciones de criterio a través de la noción del *plano nomothetico*.

Es por esto que para algunos investigadores en el campo de la medición, como Leighton y Gierl (2007a), Borsboom y Mellenbergh (2007) y Yang y Embretson (2007), la EDC toma un lugar privilegiado al representar uno de los caminos más convincentes para establecer la validez. Un ejemplo de que la EDC representa uno de los modelos de medición con mayor potencial para el desarrollo y para la validación de los nuevos tipos de test —y que además puede incorporar las innovaciones de la tecnología informática como la GAÍ (Bejar, 2002, 2010; Gorin & Embretson, 2013; Leighton & Gierl, 2007a) — es el Enfoque Sistémico del Diseño Cognitivo (ESDC) propuesto por Embretson (1983, 1998). El objetivo del ESDC es colocar a la teoría cognitiva como base en el desarrollo

de los ítems de una prueba y así mejorar el significado y el uso de sus puntuaciones y, con ello, su validez.

A pesar del panorama prometedor en el campo del desarrollo y de la validación de pruebas que presumen los modelos basados en la EDC, hay temas relacionados con la validez de constructo que todavía necesitan discutirse: (a) la idoneidad, exhaustividad y granularidad de la representación del constructo; (b) el diseño y selección de los indicadores observables para una medida del constructo finamente granulada; (c) la posibilidad de medición del constructo con respecto a los formatos de los ítems o los procedimientos de administración de la prueba; y (d) la adecuación de los fundamentos teóricos de la medición que son relevantes para el propósito específico de la evaluación diagnóstica. Esta lista de requerimientos queda lejos de estar completa, debido a que hay pocos estudios respecto a la aplicación de la EDC, especialmente desde la perspectiva de la validación de constructo.

2.2.3. El aspecto sustantivo de la validez de constructo y la evaluación diagnóstica cognitiva

El estudio del proceso de respuesta de los examinados ante los ítems de una prueba es cada vez más frecuente para obtener fuertes evidencias de validez de las evaluaciones psicológicas y educativas (ver Leignton & Gierl, 2007a; Nichols, Chipman & Brenan, 1995). En especial, la última versión de los *estándares* (AERA, APA & NCME, 1999) retoma el análisis del aspecto sustantivo de la validez de constructo en el estándar 1.8 relacionado con obtención de evidencias basadas en el proceso de respuesta. Como ya se mencionó, dicho tipo de evidencias se refieren a la recolección y análisis teórico-

empírico de los procesos cognitivos que los examinados pueden proveer como evidencia concerniente al ajuste entre el constructo y la práctica natural de sus respuestas implicadas ante los ítems. Con ello, la razón fundamental de una prueba o de las interpretaciones de sus puntuaciones depende de las premisas acerca de los procesos psicológicos o de las operaciones cognitivas usadas por los examinados. En otras palabras, las evidencias teóricas y empíricas dadas en los procesos de respuesta de los examinados sustentan las premisas que deben ser provistas para el argumento de validez. Asimismo, el modelo cognitivo subyacente a la prueba es previamente establecido por los evaluadores y, por lo tanto, los procesos de respuesta de los examinados deberán proveer información similar para dar cuenta del ajuste entre los procesos de respuestas evocados por los examinados ante los ítems de la prueba y el constructo a medir.

Por su parte, Messick (1989b) anticipó la importancia de proveer información acerca de los procesos de respuesta de los examinados en contraposición del análisis tradicional de la validez de contenido. Por su parte, la noción básica de la validez de contenido trata de que los ítems de una prueba sean una muestra de un universo teórico articulado (constructo), donde las inferencias se puedan representar. Sin embargo, dichas inferencias se pueden obtener —incluso de manera tácita— a través del análisis de los procesos psicológicos de los examinados ante los ítems, lo cual no se puede lograr a través del análisis del contenido con técnicas de acuerdo entre jueces.

De acuerdo con Messick (1989b), comprender el aspecto sustantivo de una prueba en términos de los procesos mentales utilizados por los examinados para

contestar o resolver los ítems es una característica nuclear de la teoría de la validez de constructo que representa un papel definitivo en la especificación del dominio de una prueba:

> En el enfoque sustantivo, los ítems están incluidos en el grupo original básico juzgado como relevante ante un dominio ampliamente definido, pero, son seleccionados para incorporarse a la prueba sobre la base de las consistencias de las respuestas empíricas. El componente sustantivo de la validez de constructo, tiene que ver con la capacidad de la teoría del constructo para tomar en cuenta el contenido resultante de la prueba... la estructura interna y la sustancia de una prueba, pueden ser abordadas de forma más directa a través del modelado causal del rendimiento de un ítem o tarea evaluativa. En este enfoque, para construir la representación del constructo se realizan intentos importantes para identificar los mecanismos teóricos del desempeño subyacentes a las tareas evaluativas, principalmente, obtenidos a través de la descomposición de la tarea en componentes en términos de procesos requeridos para responderla (Embretson, 1983). En la psicología cognitiva del procesamiento de la información, está firmemente arraigada la construcción de las representaciones relativas de dependencia entre los procesos cognitivos, las estrategias, los conocimientos (incluyendo el auto-conocimiento) y los procesos de respuesta involucrados en el rendimiento de una prueba (Traducido de Messick, 1989; pp. 42-45).

Además de Messick (1989b), Snow y Lohman (1989) también consideran que la psicología cognitiva es una fuerte aliada para complementar los análisis y los resultados de las evaluaciones educativas. Dichos autores mencionan que la psicología cognitiva, al tener como uno de sus campos de estudio predilectos la solución de problemas, se convierte en una fuerte aliada de la medición educativa debido a que todo test de inteligencia, en cierto sentido, es una tarea de solución de problemas. Así, para estos

autores, la psicología cognitiva es altamente requerida para informar sobre el diseño de las pruebas y, en especial, sobre los resultados de las evaluaciones educativas.

Un aporte que considera Messick (1989b) significativo por parte de la psicología cognitiva a la medición educativa y a los modelos psicométricos es la definición del término *fidelidad estructural*. Dicho término se refiere a la medida en que las relaciones estructurales entre los ítems de una prueba son paralelas a las relaciones entre las características de los atributos del dominio a medir. El aspecto estructural de la teoría de la validez de constructo incluye tanto el grado de *fidelidad* entre los puntajes del modelo estructural con respecto a las características y manifestaciones del constructo medido, como el grado de relación *estructural* entre los ítems de una prueba (Embretson & Gori, 2001; Loevinger, 1957 en Leighton & Gierl, 2007a).

Dado lo anterior, no es raro imaginar que la psicología cognitiva pueda incidir de forma directa en los modelos de medición educativa y, en el caso específico de algunos modelos psicométricos, puede incluso generar un cambio sustancial (Embretson, 1998). Desafortunadamente, dichos cambios no son tan fáciles de incorporar a los modelos de medición puesto que hay una gran cantidad de modelos cognitivos que pueden explicar la estructura de una prueba y el dominio o constructo a medir. Aún con ello, autores como Snow y Lohman (1989) consideran que una de las aportaciones más significativas de la psicología cognitiva al campo de la medición, es el desarrollo de teorías sustantivas sobre las aptitudes, el aprendizaje y el logro educativo que ayuden a guiar el diseño de las evaluaciones educativas.

Sería ideal que los psicólogos cognitivos desarrollaran tales teorías, exclusivamente para el desarrollo de tareas de medición educativa. Sin embargo, los desarrolladores de pruebas necesitan a su vez incorporar y adaptar en sus procesos de medición métodos, técnicas, y herramientas que ayuden a integrar las aportaciones de la psicología cognitiva y las teorías de la cognición. Además de la aportación mencionada, Snow y Lohman (1989) identifican otras aportaciones significativas de la psicología cognitiva al campo de la medición como: (a) informar sobre los análisis psicométricos ya existentes para aclarar la teoría subyacente, (b) aclarar los objetivos de los test en términos de conocimiento y de habilidades que sean genuinos indicadores del dominio y la comprensión y (c) mejorar las teorías sobre las aptitudes, el aprendizaje y el logro educativo en diferentes tipos de dominio.

Por su parte, Pellegrino (2010) hace énfasis en la integración de las ciencias cognitivas y del aprendizaje a la evaluación, argumentando el gran progreso que estas han generado en la comprensión de la naturaleza, del aprendizaje y del conocimiento (National Research Council, 2001). Dicho autor considera que la psicología cognitiva puede y debe informar acerca del desarrollo de los sistemas de evaluación, además de ayudar en la medición del estado y el avance de los logros académicos en las áreas del currículum.

Por otra parte, Pellegrino, Baxter y Glaser (1999) proponen que la evaluación debe integrarse con el currículum y con la instrucción, así también, que las tres juntas deben ser guiadas por las teorías e investigaciones sobre la naturaleza del aprendizaje y los conocimientos expresados en los contenidos curriculares. También, dichos autores

señalan que una de las problemáticas en la evaluación educativa es que los evaluadores deben aceptar con humildad que nunca se podrá saber con total certeza lo que un estudiante sabe. Con ello, la mejor evaluación es aquella en donde se aplican procesos rigurosos y cuidadosamente estructurados, a partir de pruebas impulsadas por datos y teorías sobre la cognición del estudiante.

Básicamente, el modelo propuesto por Pellegrino, Baxter y Glaser (1999), prioriza la relación de la evaluación con el currículum y la instrucción conformando una triada guiada por las teorías de la cognición y el aprendizaje (ver Figura 7) (Pellegrino, 2010). Los tres elementos de esta tríada están vinculados, aunque la naturaleza de sus vínculos e influencia recíproca no es tan clara como se quisiera.

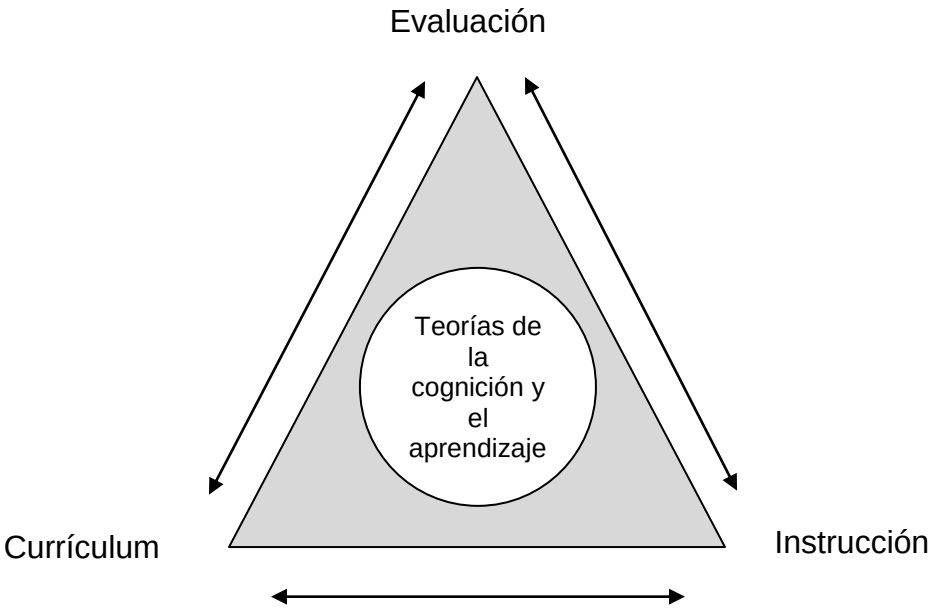


Figura 7. Representación de las interconexiones entre currículum, instrucción y evaluación, guiadas por las teorías de la cognición y del aprendizaje.

Tomando en cuenta lo mencionado en los párrafos anteriores, se puede decir que los evaluadores necesitan poner las evaluaciones educativas y psicológicas bajo el "microscopio cognitivo". Con dicho microscopio se puede analizar de forma puntual el aspecto sustantivo de la validez de dichas evaluaciones (Leighton & Gierl, 2007b; Messick, 1989b; Snow & Lohman, 1989). Asimismo, los desarrolladores de pruebas deben considerar y priorizar el aspecto de la *fidelidad estructural* de las evaluaciones, antes que los contenidos de aprendizaje como objetivos de la medición (National Research Council, 2001; Snow & Lohman, 1989). Para tal fin, la Evaluación Diagnóstica Cognitiva (EDC) puede ayudar a hacer explícitos los procesos y estructuras de conocimiento de los examinados ante los ítems de un test y contrastarlos con las hipótesis sustantivas del dominio a evaluar propuestas al inicio por los desarrolladores de la evaluación. Además, la EDC puede ayudar a conocer cómo se desarrollan las estructuras de los procesos y conocimientos, y cómo los individuos con mayor dominio del atributo difieren de los individuos con menor dominio (Nichols, 1994; Nichols, Chipman & Brenan, 1995).

De forma resumida, Yang y Embretson (2007) mencionan que las pruebas basadas en la EDC aplicadas en un entorno psicológico o educativo se enfocan principalmente en al menos tres aspectos de las características cognitivas:

- Perfiles de habilidad o listas de conocimiento, que son esenciales en un dominio cognitivo. Dichos grupos de habilidades y conocimientos representan los más importantes conceptos del dominio y sirven como los bloques básicos del constructo para el desarrollo de cualquier otro nivel de competencia.

- Estructuras de procesos cognitivos o redes de conocimientos, teóricos y prácticos, representados en nuestra mente de manera altamente estructurada (Collins Loftus, 1975; Rumelhart, 1980). Los conocimientos en un dominio están representados no sólo por el número de las competencias básicas, o trozos de conocimiento poseídos en el dominio, sino también por la estructura y la organización de tales conocimientos y habilidades (Chi, Glaser, Farr, 1988; Ericsson & Charness, 1994).
- Procesos cognitivos, componentes o capacidades. El paradigma de la investigación cognitiva de procesamiento de la información proporciona métodos para aprovechar los procesos internos de la cognición y desarrollar modelos cognitivos específicos para un determinado tipo de tareas cognitivas. Procesos de respuestas presentados por examinados ante alguna tarea evaluativa pueden explicarse con ayuda de métodos cognitivos.

A partir de la teoría de la validez de constructo propuesta por Messick (1989b), las evidencias de validez basadas en el proceso de respuesta de los examinados pueden ser analizadas desde el aspecto sustantivo de la validez de constructo, destacando el papel de las teorías que dan cuenta de los atributos del dominio de las pruebas y los métodos que ayudan a asegurar el ajuste de dichas teorías sustantivas con los procesos de respuesta, manifestados por los examinados ante los ítems. Con ello, el aspecto sustantivo de la validez de constructo se puede visualizar en dos puntos importantes: (a) la necesidad de que una prueba contenga un conjunto de tareas evaluativas que proporcionen un muestreo adecuado de los atributos del dominio a medir, además de la cobertura tradicional de sus contenidos, y (b) la necesidad de ir más allá del juicio tradicional de expertos en el contenido con la acumulación de evidencias empíricas relacionadas con una muestra de procesos realmente comprometidos por los

examinados ante las tareas evaluativas (Borsboom & Mellenbergh, 2007; Messick, 1989a y 1989b).

Por su parte, desde la propuesta de validez de Borsboom y Mellenbergh (2007), el análisis del proceso de respuesta subyacente a los ítems de una prueba puede realizarse de dos formas diferentes. Una de las formas es la obtención de información introspectiva o retrospectiva (Ericsson & Simon, 1984, 1993; Leigntong & Gierl, 2007a) del proceso de respuesta de los examinados ante las tareas evaluativas. La segunda forma es el uso de modelos psicométricos (por ejemplo, el Modelo Logístico Lineal de Rasgo Latente de Fisher, 1973) para el análisis de la estructura del modelo cognitivo construida a través de los procesos de respuesta de los examinados. Dichas estructuras de los modelos cognitivos deben ser probadas con ayuda de los modelos psicométricos, utilizando datos empíricos. El uso de los métodos cognitivos (introspectivos y retrospectivos), y de los modelos psicométricos que incorporan el análisis de la estructura del modelo cognitivo para obtener fuertes evidencias de validez, son discutidos y promovidos ampliamente por autores como Embretson y Reise (2000), Maris (1995, 1999), Snow y Lohman (1989), Rupp y Mislevy (2007), Roussos, DiBello, Stout, Hartz, Henson y Templin (2007) y Gierl, Wang y Zhou (2008).

Los procesos cognitivos obtenidos en el *análisis verbal* y en los *protocolos cognitivos* son usados para construir modelos cognitivos que sirven de insumo a los modelos psicométricos, que integran aspectos de la psicología cognitiva y que subsecuentemente ayudan al ajuste ante los datos empíricos de un test (Borsboom & Mellenbergh, 2007). Las dos estrategias mencionadas presentan un excelente método

combinado para la validación de un test. Un ejemplo exitoso de la aplicación de ambos métodos para obtener evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de una prueba es el presentado por Gierl, Wang y Zhou (2008) ante los ítems del SAT.

Algunos procedimientos básicos sugeridos en la última versión de los e*stándares* (AERA, APA & NCME, 1999) para el análisis de los procesos de respuesta son los *análisis verbales* y los *análisis de protocolos cognitivos*. En ambos métodos se explora a los examinados con respecto a sus estrategias de respuesta o de cómo responden a ciertos ítems en particular. También, en los *estándares* mencionados se proponen técnicas de registro en donde se monitoreen electrónicamente (videograbaciones) las respuestas de los examinados ante las tareas evaluativas. Además, se propone incorporar documentación de otros aspectos de los procesos de respuestas como el *seguimiento visual* o los *estudios de tiempos de respuesta* que para la revisión de algunos constructos toman una alta relevancia. Otro tipo de análisis sugerido en los *estándares* es el estudio del diseño y del formato de los ítems de una prueba, así como el análisis entre la prueba y otras variables, por ejemplo, las diferencias individuales entre expertos y novatos en el contenido que pueden ser reveladoras y, a su vez, guiar a la reconsideración del diseño de las pruebas (Messick 1989b; Snow & Lohman, 1989).

Por otra parte, en los mismos *estándares* mencionados en el párrafo anterior, se aclara que los estudios de los procesos de respuesta no están limitados al examinado. En algunos estudios a menudo se confía en jueces o expertos para definir los atributos sustantivos y modelar los procesos de respuesta ambos subyacentes a los ítems de la

prueba. En estos casos, la evidencia de validación relevante incluye el grado donde los procesos de respuesta, propuestos por los expertos o jueces, son consistentes para interpretar los puntajes. Sin embargo, se sugiere tener cuidado en dichos procedimientos, pues es necesario comprobar que el criterio de los jueces no se encuentre influenciado por factores irrelevantes que afecten la propiedad de las interpretaciones correspondientes. Así, la validación del proceso de respuesta debe incluir estudios empíricos sobre cómo los expertos o los jueces evalúan la información y realizan el análisis de los procesos de respuestas hacia la interpretación que se pretende, o hacia la definición del constructo.

Las evidencias que se basan en el proceso de respuesta sólo toman relevancia en evaluaciones referidas a un criterio, dado el aporte de estas en el diagnóstico de un dominio determinado. Sin embargo, en las evaluaciones de selección que son de tipo normativas —como en el caso de la prueba analizada en el presente proyecto—, se pone un gran énfasis en otras facetas del proceso, distintas al diagnóstico (por ejemplo, en la predicción del éxito escolar). Ello no significa que la validación deba encontrarse inherente al tipo de evaluación, sino que el significado deberá analizarse con respecto al efecto causal de los atributos en el puntaje de la prueba (Borsboom, Mellenbergh & van Heerden, 2004); es decir, dichos tipos de evaluaciones, aun definiéndose previamente como normativas, continúan requiriendo evidencias relacionadas con la fidelidad de los procesos de respuesta ante la prueba y el constructo supuestamente medido.

Es por ello que una prueba de selección para ingreso a la universidad o para ingreso a cualquier nivel educativo debe desarrollarse con base en los constructos que

mejor predigan el rendimiento y el éxito escolar de los examinados. Asimismo, un examen que mide las variables con mayor poder de predicción del éxito escolar, y que tiene fuertes evidencias de validez basadas en el proceso de respuesta, presenta mayor certidumbre para su uso en la selección de aspirantes a ingresar al nivel educativo de interés y, por lo tanto, del éxito en la vida escolar (Tirado, et al. 1997). Visto de otra manera, cuando intentamos obtener evidencias del poder predictivo de una prueba, lo que estamos buscando son evidencias de criterio. Sin embargo, en la misma línea, toma mayor relevancia el aspecto sustantivo de la validez de constructo cuando el constructo de interés es demostrar mediante métodos o mediciones cognitivas, si la justificación para el uso de la prueba (sea para selección o de diagnóstico) o la interpretación de sus puntuaciones (criterial o normativa) depende de las premisas teóricas de los procesos psicológicos o de las operaciones cognitivas utilizadas por los examinados (AERA, APA & NCME, 1999).

Por otro lado, la EDC presenta una novedosa metodología que trasciende distintas problemáticas en los actuales modelos de medición. Como un antecedente, Anastasi (1967) advirtió que algunos psicólogos se especializaron en el análisis psicométrico haciéndose cada vez más y más devotos de los esfuerzos para refinar las técnicas de medición y las técnicas para la construcción de los tests, perdiendo así de vista la *fidelidad estructural* del constructo a medir. En palabras de Leignton y Gierl (2007b, p.5): “los psicómetras a la fecha se enfocan más en la métrica que en la psicología”. También, autores como Snow y Lohman (1989), encuentran serias limitaciones en los modelos psicométricos tradicionales. Tales autores, mencionan que dichos modelos fallan en

incorporar: (a) teorías sustantivas que expliquen las respuestas ante los ítems; (b) supuestos realistas acerca de las variables psicológicas que influyen en las respuestas y el desempeño ante los ítems de una prueba; y (c) delimitaciones claras de los procesos psicológicos que en conjunto reflejan el constructo a medir. Además, los modelos cognitivos, implícitos en la mayoría de las pruebas educativas, son todavía un reflejo de las expectativas particulares de los evaluadores sobre cómo razonan los estudiantes y resuelven los ítems de una prueba en el dominio de interés. Dichos razonamientos no están basados en evidencias empíricas sobre cómo razonan los examinados y responden ante dichos ítems (Leignton & Gierl, en prensa; Nichols, 1994). Aunado a ello, en el enfoque tradicional de medición se confía únicamente en una taxonomía lógica y en contenidos específicos para describir los objetivos de medición, pero no se toma en cuenta que dichos contenidos (principalmente los presentes en el currículum o planes de estudio) son ambiguos y vagamente revelan los mecanismos cognitivos naturales directamente utilizados por los estudiantes que se examinan en su aprendizaje (Leighton & Gierl, 2007a).

Sin embargo, dichas limitaciones de los modelos psicométricos de medición tradicionales son claramente disminuidas y, finalmente, superadas con el uso de la EDC. Por supuesto, ante teorías sustantivas sofisticadas probablemente sería de poca utilidad la aplicación de la EDC, si dichas teorías no pueden ser incorporadas en los análisis psicométricos. Pero, en la actualidad los modelos psicométricos se están adaptando para asimilar y acomodar las diferentes características (por ejemplo complejidad, continuidad, jerarquía, peso, isomorfismo, entre otras) de los componentes presentes en los procesos

de respuesta de los examinados ante los ítems de una prueba y en los mecanismos cognitivos naturales directamente utilizados por los examinados en su aprendizaje.

De forma especial, el análisis de las evidencias de validez, basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas educativas, encuentra un gran cobijo metodológico a través de la EDC. Con la integración de los principios de la psicología cognitiva al campo de la medición, dicho tipo de modelo evaluativo toma un lugar privilegiado en el desarrollo y validación de evaluaciones educativas y psicológicas (Yang & Embretson, 2007). En lugar de inferir una tendencia general de la respuesta de un examinado ante un ítem, los resultados de la EDC proveen un cálculo más detallado de los procesos cognitivos básicos utilizados por los examinados ante las evaluaciones. Sin embargo, asegurar que la base teórica sustantiva de la EDC esté bien investigada y articulada (además de comprobar su *fidelidad estructural*) requiere un compromiso nada trivial para la comprensión de la estructura que está siendo medida. Es por ello que se han desarrollado sofisticados análisis psicométricos como el RSM (Rule-Space Method en inglés; Tatsuoka, 1995; 2009), el LLTM (Linear Logistic Latent Trait Model en inglés; Fischer, 1973), el AHM (Attribute Hierarchy Method; Gierl, Wang, & Zhou, 2008) y, el LSDM (Least Squares Distance Model en inglés, Dimitrov, 2007; Dimitrov, Romero, Ponsoda & Ximénez, 2006), entre otros, con el fin de analizar la información de los procesos de respuesta y los mecanismos cognitivos naturales directamente utilizados por los examinados en su aprendizaje, además de aportar información y evidencias basadas en la validez de la estructura del modelo cognitivo e información diagnóstica puntual

sobre los diferentes procesos de respuesta utilizados por los examinados ante la evaluaciones.

Con todo lo que se ha mencionado a lo largo de este apartado sobre la EDC, se puede afirmar que dicho modelo de evaluación es irreconciliable con programas débiles de validez debido a que se fundamenta en el análisis detallado del proceso de respuesta de los examinados ante los ítems de un prueba y en la *fidelidad estructural* del constructo a medir (Nichols, 1994). En resumen, la EDC obliga a los evaluadores y a los desarrolladores de pruebas a ejecutar un programa riguroso de validación centrado en la medición de los procesos mentales de los estudiantes ante una evaluación, y a utilizar la información resultante de dicha evaluación para mejorar las oportunidades de los estudiantes en relación a lo que aprenden y al éxito en sus estudios.

2.3. Modelos de medición que integran los principios de la psicología cognitiva

Tradicionalmente, el diseño de los ítems es visto por los desarrolladores de pruebas como un arte (Embretson & Gorin, 2001). Dicho reconocimiento fue ganado porque se requieren de labores especializadas y, en ocasiones extremadamente creativas para planear y diseñar una prueba. Un ejemplo de ello, es la selección de contenidos del currículum durante la planeación de evaluaciones educativas y la elaboración de las especificaciones de los ítems. En especial, para el diseño de los ítems de pruebas de logro educativo es necesario seleccionar contenidos generales del currículum representados en términos de temáticas generales u objetivos de aprendizaje que contienen un alto grado de ambigüedad (Contreras, 2000; Embretson & Gorin, 2001).

Con ello, los desarrolladores de las pruebas se encuentran obligados a realizar estrategias encaminadas a clarificar, a delimitar y, en ocasiones, a traducir de manera artesanal los contenidos del currículum, con el fin de que dichos contenidos sean asequibles para medirse. Una vez producidos los ítems, un comité de especialistas los examina con el fin de asegurar su representatividad. Tal actividad, es comúnmente conocida como análisis o juicio por expertos y es comúnmente utilizada para obtener evidencias de validez basadas en el contenido de una prueba.

Por su parte, los métodos y los análisis psicométricos se utilizan después de obtener resultados de su piloteo en una población dada. La aplicación de dichos análisis psicométricos, es importante para determinar la calidad técnica de las pruebas y, en especial, para evaluar la propiedad de los procesos de medición con respecto al constructo a medir. Siguiendo la tradición, con el resultado de los análisis psicométricos, los ítems que no presentan una alta correlación con los otros ítems diseñados no son seleccionados para formar parte de la prueba.

Dado lo anterior, es evidente que el modelo para el desarrollo y para la validación de los ítems presentado en los dos párrafos anteriores se basa en el enfoque clásico de la validación de constructo propuesto por Conbrach y Meehl (1955). En dicho enfoque, la validación del constructo se define como el significado empírico obtenido después de la construcción de la prueba en base al análisis de la correlación entre los ítems y algún criterio externo. En otras palabras, el significado del constructo de un test se caracteriza a través de estudios empíricos de correlación de las puntuaciones con otros criterios externos.

Por otra parte, el diseño de los ítems con base en la teoría clásica de la validación de constructo presenta algunas limitaciones para incorporar las aportaciones de la psicología cognitiva, principalmente, porque la noción conceptual de las *redes nomológicas* en el enfoque clásico de la validación de constructo implica desarrollar una prueba antes de elaborar su significado teórico. Dado lo anterior, hay señalar que una implicación importante de sentido común en cualquier medición es que los contenidos (por ejemplo los objetivos de aprendizaje y los contenidos establecidos en el currículum) con los que se pretende representar al constructo (por ejemplo los procesos cognitivos naturales en los individuos) a medir no deben modificarse con base en los resultados de una prueba (Embretson & Gorin, 2001; Leighton & Gierl, 2007a). En otras palabras, el proceso de medición debe de representar *fielmente* el constructo a medir, y no a la lógica inversa. Resultando así ilógico pensar que el proceso de medición debe ser representado por el constructo a medir, debido a que la naturaleza teórica de un constructo existe *a priori* al proceso de medición y, por lo tanto, antes del diseño de los ítems.

Al respecto, Pellegrino (2010) considera que la psicología cognitiva puede complementar el diseño de las pruebas como otro aspecto de la validación de constructo clásica. En el mismo sentido, Messick (1989b) menciona que la psicología cognitiva puede igualmente aportar al aspecto sustantivo de la validez de constructo de pruebas consolidadas diseñadas sin la guía de una teoría sustantiva. Para ello, propone utilizar un enfoque *top-down,* justificando que es también válido partir en busca del significado de una prueba consolidada cuando ésta presenta fuertes evidencias basadas en los aspectos de contenido, de estructura interna y predictivo. Sin embargo, promueve el uso

de la EDC como uno de los caminos más convincentes para la obtención de evidencias de validez de constructo, debido a que dicho modelo de medición ofrece uno de los programas de validez más fuertes y acordes a la teoría sustantiva de la validez (Leighton & Gierl, 2007a).

En resumen, los principios de la psicología cognitiva pueden beneficiar en gran medida al diseño y a la validación de las evaluaciones educativas en diferentes sentidos (Embretson & Gorin, 2001). Primero, con la ayuda de métodos cognitivos se puede hacer una mejor definición de los contenidos que se seleccionan para representar fielmente el constructo a medir. También, se pueden seleccionar tareas educativas más acordes a las características del constructo de interés. Segundo, con los resultados de los estudios cognitivos se puede proveer de una base teórica *fuerte* (Griel & Lai, 2013) para las interpretaciones de los resultados de las pruebas. Tercero, con el análisis de los procesos cognitivos naturales que los individuos utilizan ante una tarea evaluativa, se pueden definir principios que ayuden a que los resultados se presenten de forma automática. Además, de proveer estructuras y atributos para elaborar algoritmos de generación de ítems con posibilidades de automatizarse con la ayuda de los avances de la informática.

Dado lo anterior, es importante conocer diferentes modelos de la EDC que integran las aportaciones de la psicología cognitiva para el diseño, desarrollo y validación de las pruebas. Al respecto, durante las últimas dos décadas se han desarrollado distintas aproximaciones y métodos de medición con el fin de incorporar y aprovechar de la mejor forma los principios de la psicología cognitiva en el campo de la medición. Estos

modelos generalmente pueden dividirse en dos categorías: (a) modelos estadísticos que incorporan parámetros relacionados con los componentes de procesamiento cognitivo o atributos y (b) Modelos para el Diseño de las Evaluaciones (MDE) que integran los principios de la teoría cognitiva a lo largo de su desarrollo.

Los modelos estadísticos que incorporan parámetros relacionados con los componentes de procesamiento cognitivo o atributos son también conocidos como Modelos Psicométricos Componenciales (MPC). Dichos modelos son esencialmente técnicas de análisis estadístico diseñadas para incorporar la teoría cognitiva con las propiedades psicométricas de los ítems. En dichos análisis, los atributos cognitivos necesarios para responder correctamente los ítems de una prueba son especificados y su impacto es estimado. Además, se estima y reporta el dominio de las habilidades de los examinados requeridas para responder a cada uno de los ítems.

A su vez, los MPC se pueden clasificar en dos tipos: los modelos derivados de la Teoría de Respuesta al Ítem (TRI) y los Modelos de Diagnóstico Cognitivo (MDC). Los modelos componenciales derivados de la TRI buscan descomponer los parámetros de los ítems en atributos subyacentes. Un ejemplo de estos modelos es el Modelo Logístico Lineal de Rasgo Latente (LLTM por sus siglas en inglés) de Fischer (1973, 1985). Por su parte, los MDC clasifican a los examinados en *estados de conocimiento* (Tatsuoka, 1995 en Romero, 2010) con respecto al dominio que presentan los sujetos en cada uno de los atributos. Tres ejemplos representativos de los MDC son el RSM (en inglés *Rule Space Model*; Tatsuoka, 1983), los modelos estadísticos de redes Bayesianas (Mislevy, 1995) y

los modelos DINA (en Inglés *Deterministic Input, Noisy And Gate*; Junker & Sijtsma, 2001), entre otros.

Ahora bien, se pueden mencionar tres importantes modelos para el diseño de las evaluaciones, los cuales se destacan porque integran los principios de la teoría cognitiva a lo largo de su desarrollo: (1) el modelo de cinco pasos basados en principios psicológicos para el desarrollo de pruebas, propuesto por Nichols (*Model of five steps for psychology-driven test development*, 1994); (2) el modelo del diseño de evaluación basado en evidencias de Mislevy (*Evidence-centered Assessment Design*, 2007); y (3) el Enfoque sistémico del diseño cognitivo de Embretson (*The cognitive design system approach*, 1998). Cada uno de los modelos aquí señalados presenta un gran vínculo con los MPC, pues ayudan a direccionar el uso de las teorías cognitivas sustantivas en las diferentes etapas durante el desarrollo de los ítems de una prueba y del análisis psicométrico componencial. Al inicio del desarrollo de un test, el uso tanto de los MDE como de los MPC se relaciona con la definición del constructo. Más adelante, su uso apoya la elaboración de los ítems y, al final, a los procedimientos de validación de la prueba. Dado lo anterior, es obvio el extenso uso y aplicación de los MDE y los MPC a lo largo del desarrollo de una prueba. Es por ello que para varios autores reconocidos en el campo de la medición (por ejemplo, Borsboom & Mellenbergh, 2007; Embretson, 1998; Leighton & Gierl, 2007a; Yang & Embretson, 2007, entre otros) la EDC, en la actualidad, representa el más prometedor y convincente camino para el diseño, desarrollo y validación de las evaluaciones educativas y psicológicas.

2.3.1. Modelo de cinco pasos basados en principios psicológicos para el desarrollo de pruebas

Uno de los primeros modelos de medición basados en los principios de la psicología cognitiva, es el que desarrolló Nichols (1994). Este modelo presenta cinco pasos generales a realizar: (a) construcción de la teoría sustantiva, (b) selección del diseño de la evaluación, (c) administración y aplicación del test, (d) presentación de los resultados de la evaluación, y (e) revisión del diseño del test. Como se puede observar, Nichols (1994) propone dos pasos generales para el proceso del diseño y del desarrollo de los ítems. Para el primer paso, relacionado con la construcción de la teoría sustantiva, se requiere del desarrollo de un modelo o teoría que caracterice la estructura hipotética del conocimiento y de los procesos necesarios para responder a los ítems de un test. A la par, es importante en el contexto del modelo identificar las variables subyacentes al ítem que evocan procesos cognitivos y estructuras del conocimiento.

Para el segundo paso relacionado con la selección del diseño de la evaluación, los desarrolladores guiados por el modelo o la teoría sustantiva desarrollada en el paso uno, seleccionan el diseño general de la medición. Para seleccionar o elaborar el tipo de tareas evaluativas que conformarán la evaluación, los desarrolladores deberán tomar en cuenta las posibles respuestas o procesos naturales de los examinados en relación con el modelo o la estructura teórica de los conocimientos y/o habilidades definidos en el paso uno.

En el tercer paso, relacionado con la administración y aplicación del test, los investigadores deben tomar en cuenta los detalles del medio y del contexto en el cual los

examinados realizan su evaluación, como el formato (por ejemplo de opción múltiple o de respuesta construida) y el medio (por ejemplo un examen a papel y lápiz o computarizado) de la presentación del ítem. En este tercer paso, Nichols (1994) recomienda que las decisiones sobre la administración de la prueba estén basadas en investigación sobre cómo diferentes variables presentes en la administración del test influyen en la ejecución de los examinados.

Durante el cuarto paso, concerniente a la presentación de los resultados de la evaluación, los evaluadores deberán presentar los puntajes resultantes de la evaluación cognitiva, de tal forma que resulten informativos sobre el nivel del dominio que tienen los examinados del constructo medido. Por último, para el quinto paso del modelo de Nichols se reexamina el diseño de la prueba con el fin de observar si apoya el modelo teórico en el que se fundamentó la medición. Con lo que, una vez obtenidos los resultados de la evaluación, deberá revisarse la base substantiva de la prueba. Si los desarrolladores de evaluaciones basadas en los principios de la psicología cognitiva siguen fielmente los cinco pasos modelo aquí descrito, se puede lograr el apego al enfoque sustantivo y la *fidelidad estructural* durante el desarrollo de la evaluación cognitiva.

Por su parte, Nichols (1994) enfatiza en su propuesta que la base substantiva no sólo se desarrolla de la revisión bibliográfica o de la investigación básica. Según el autor, también es posible desde el establecimiento de supuestos acerca de cuál es la mejor representación de los procesos naturales del aprendizaje y de las diferencias individuales. Sin embargo, debe tenerse mucho cuidado con los supuestos que se asumen, debido a que el aspecto sustantivo es la base de la evaluación cognitiva. De tal

forma que los desarrolladores de la prueba deben tener la seguridad y buen nivel de confianza sobre el modelo de los procesos cognitivos de los examinados en el que se basará la medición. Es por ello que las suposiciones sobre el aspecto sustantivo de las estructuras del conocimiento y de los procesos naturales de los examinados deben someterse al escrutinio empírico.

2.3.2. Diseño de evaluación centrado en evidencias

El origen del Diseño de Evaluaciones Centrado en Evidencias (ECD, por sus siglas en inglés) inició en 1990 en el ETS (en inglés Educational Testing Service). Desde ese entonces, el objetivo del ECD fue la acumulación de evidencias que sustentaran las inferencias del rendimiento de un examinado. Por supuesto, antes del ECD ya había otros modelos para el diseño de pruebas que trabajaban en la misma línea de investigación (por ejemplo el *modelo de argumentos de representación del constructo*, propuesto por Embretson en 1983). El modelo para el diseño de pruebas propuesto por Mislevy (1994), a diferencia de las evaluaciones tradicionales que se limitan a la medición de una sola competencia, se puede aplicar para la medición de dominios complejos con más de un atributo subyacente. Es por ello que es un modelo adecuado para realizar inferencias diagnósticas sobre las fortalezas y las debilidades de los examinados.

Principalmente, el ECD incide en el desarrollo de un test en tres modelos y sus componentes: (a) el *modelo del estudiante* o el tipo de inferencias que se desean obtener de un examinado, (b) el *modelo de las inferencias* o el tipo de evidencias que

fundamentan las inferencias deseadas y (c) el *modelo de las tareas* o el tipo de tareas que pueden producir las evidencias necesarias (Mislevy, 1994, 2009; Mislevy, Steinberg & Almond, 2002).

Por su parte, el *modelo del estudiante* especifica las habilidades, conocimientos o estrategias como el foco de las inferencias. La formulación de este modelo está vinculada naturalmente con el propósito de la prueba. En lo que respecta al *modelo de las evidencias*, su aporte es describir los comportamientos observables o los procesos de respuestas que proveen las evidencias de los componentes del *modelo del estudiante*. Con ello, las estimaciones del estatus de los examinados en las variables del *modelo del estudiante* son renovadas, de acuerdo a las reglas del *modelo de las evidencias*. Finalmente, el *modelo de las tareas* define la naturaleza específica de un ítem, incluyendo las condiciones en las cuales dichas tareas son realizadas, los materiales presentados y la naturaleza del producto generado por los examinados. Además, cuando se selecciona un formato particular del *modelo de la tarea*, el objetivo es crear un ambiente en el cual un examinado genere comportamientos observables que correspondan, de la mejor manera, al *modelo de las evidencias*. Una vez que las tareas evaluativas son generadas y relacionadas con las evidencias y el *modelo del estudiante*, el desarrollador tiene la base para la generación de una gran cantidad de ítems para el montaje de una prueba.

La propuesta de los tres modelos mencionados, se encuentra basada en los aspectos del modelo basado en los razonamientos de Stewart y Hafner (1994) y, Gobert y Buckleys's (2000) (ver Tabla 2.4). Cabe mencionar que un modelo es la representación

simplificada que se focaliza en ciertos aspectos de un sistema (Ingham & Gilbert, citado en Gobert & Buckley, 2000). Con lo que los componentes, las relaciones entre ellos y los procesos de un modelo son la base de su estructura. Lo anterior provee un fundamento para reflexionar acerca de los patrones de cruce de cualquier cantidad de situaciones únicas en el mundo real (por ejemplo, en la evaluación educativa).

Tabla 2.4. Aspectos del modelo basado en razonamientos (Stewart & Hafner, 1994; Gobert & Buckleys's, 2000)

Etapas	Descripción
Diseño del modelo	Establecer una correspondencia entre un fenómeno del mundo real y un modelo o estructura abstracta, en términos de entidades, relaciones, procesos, comportamientos, etc. Incluye el alcance y el tamaño del modelo, determina qué aspectos del modelo deben establecerse y cuáles dejar de lado.
Elaboración del modelo	Combinar, extender y añadir detalles al modelo estableciendo correspondencias a través de la superposición con otros modelos. A menudo hay que ensamblar modelos pequeños en otros más grandes o reestructurar modelos generales en modelos más detallados.
Uso del modelo	Reflexionar posibles predicciones, explicaciones o conjeturas con base en la estructura del modelo.
Evaluación del modelo	Evaluar la correspondencia entre los componentes del modelo y sus homólogos del mundo real, haciendo énfasis en las anomalías y características importantes no estipuladas en el modelo.
Revisión del modelo	Modificar o elaborar un modelo para representar un fenómeno y establecer una mejor correspondencia. A menudo dicha correspondencia se inicia en los procedimientos de la evaluación del modelo.
Investigación fundamentada del modelo	Trabajar en forma interactiva entre los fenómenos y en los modelos, utilizando todos los aspectos anteriores. Se debe hacer énfasis en el monitoreo y en la toma de acciones con respecto a las inferencias basadas en el modelo y en la retroalimentación del mundo real.

En otro punto, Messick (1989b) define la validez como "un juicio evaluativo integral del grado en que las evidencias empíricas y los razonamientos teóricos fundamentan la adecuación y la propiedad de las inferencias y las acciones basadas en los puntajes o las formas de evaluar". Sin embargo para Mislevy (2009), la validez tiene que enfocarse en la "adecuación y conveniencia de las inferencias y los usos". Es así como el criterio relacionado con "el grado en que la evidencia empírica y los razonamientos teóricos",

apoya dicho razonamiento. Por lo tanto, Mislevy (2009) se encuentra en desacuerdo con Messick (1989b) y compagina en cierto grado con la propuesta de validez de Borsboom, Mellenbergh y van Heerden (2004) en lo referente a que la validez no es un juicio absoluto, sino la propiedad a ser juzgada.

Para el caso específico de la validación de un test, y en especial para el caso de la medición educativa y psicológica (Kane, 2006), la ECD aborda la propiedad a ser juzgada desde la relación *modelo-sistema* en cuatro vías: (a) la teoría y la experiencia que fundamentan el nivel narrativo o científico del modelo; (b) la teoría y la base empírica de las tareas evaluativas; (c) la teoría y la base empírica de los procedimientos de calificación de las tareas evaluativas; y (d) la evaluación empírica del ajuste interno y otros criterios.

Para la vía analítica de la teoría y la experiencia que fundamenta el nivel narrativo o científico del modelo, Mislevy menciona que se debe partir de que todos los modelos, en cierto grado, están equivocados y que, en general, somos más propensos a elegir e inferir modelos que son más consistentes con la investigación cognitiva y que han demostrado prácticamente su utilidad en distintas experiencias evaluativas (por ejemplo, Gierl, Wang, & Zhou, 2008). Como en la física, no es que un modelo debe ser una representación fiel de un sistema, pero sí debe poder capturar los patrones más importantes, de tal forma que se adapten a las deducciones previstas (Mislevy 2009). En la misma línea, algo interesante que menciona el autor es el hecho de que las aplicaciones de los modelos de la TRÍ y los MDC tienen más probabilidades de tener éxito si los psicómetras no creyeran que el modelo es correcto, pues en general los

investigadores serían más aptos y conscientes de las explicaciones alternativas, además de ser más diligentes en la crítica del modelo.

En cuanto a la vía analítica de la teoría y la base empírica de las tareas evaluativas, es necesario tomar en cuenta que la formación del modelo no sólo es un asunto de construcción y elección, sino también un asunto de diseño de situaciones en que se observa el rendimiento (Mislevy, 2009). Para iniciar con el análisis de dicha vía es importante preguntarse: ¿Qué pueden decir las teorías cognitivas y del aprendizaje sobre las características de las tareas que se necesitan para incitar un proceso de respuesta específico? ¿Cómo alinear las características y las situaciones de la tarea evaluativa con las características y las situaciones futuras de lo que se pretende inferir? Embretson (1983) llama a esta la línea de argumentación de validez como *representación del constructo*.

Para el análisis de la teoría y la base empírica de los procedimientos de calificación de la tarea evaluativa, Mislevy, (2009) propone el argumento en evidencias, basadas en los procesos de repuesta de los examinados para fundamentar los valores de las variables observables requeridas en los análisis psicométricos componenciales. En sí, esta línea de investigación se refiere a poner mayor atención dentro de la medición psicológica y educativa a la identificación de los procesos de respuesta, con el fin de conocer de forma más detallada la naturaleza cognitiva de la competencia y el rendimiento. Al respecto, en esta misma línea de investigación se han desarrollado en los últimos años diferentes innovaciones relacionadas con GAÍ basada en una teoría sustantiva, la cual según Williamson, Mislevy y Béjar (2006) representa una mejor vía

para tomar en cuenta los procesos de respuesta de los examinados y presentar puntuaciones automatizadas con formatos más completos y a profundidad. Para ello, los análisis cognitivos pueden utilizarse para crear bancos de ítems de características psicométricas similares y conocidas, sin necesidad de calibrar los ítems en una muestra de sujetos reales (Embretson, 1983, 1995). A su vez estos bancos pueden utilizarse para crear ítems informatizados generados automáticamente como en el caso de los Test Adaptativos Informatizados (TAIs) (ver van der Linden & Glass, 2000; Wainer, 1990), basados en la GAÍ. Algunos ejemplos de estas técnicas pueden verse en Collis, Tapsfield, Irvine, Dann y Wright (1995), Bejar (1993), Hornke y Habon (1986) y Gierl, Lai, y Turner (2012).

Para la última vía analítica propuesta por Mislevy (2009), relacionada con la evaluación empírica del ajuste interno del modelo y otros criterios, se estudia si el razonamiento es compatible con la estructura de un modelo de medición, más allá de lo que nos pueden decir los datos. En modelos como la TRI y el MDC a un nivel probabilístico, así como a un nivel semántico y metafórico, se debe analizar cómo la representación del modelo concuerda con los datos observados. Lo más importante para la ECD es la correspondencia entre el modelo, el sistema y las fallas en formas que podrían ser predichas por las explicaciones alternativas (por ejemplo, con la aplicación de investigaciones externas predictivas).

2.3.3. El enfoque sistémico del diseño cognitivo

El Enfoque Sistémico del Diseño Cognitivo (ESDC) de Embretson tiene sus orígenes aproximadamente en 1983. En ese año, dicha autora presentó una propuesta diferente al modelo de validación de constructo de Combrach y Meehl (1955) que llamó *argumentos de representación del constructo* y, en 1985, presentó un caso en donde ilustró un diseño para integrar los principios de la psicología cognitiva, el diseño de tareas evaluativas y los modelos psicométricos. Con ello, la propuesta de Embretson representa una fuerte crítica a los enfoques del diseño de ítems tradicionales que recolectan *débiles* evidencias de validez de constructo.

Para Embretson, la generación de los ítems con base en los principios de la psicología cognitiva es probablemente la mejor opción por muchas razones. Primero, la validez de constructo proporciona una fuerte base para los test con una amplia variedad de principios teóricos para generar ítems. Segundo, surge un nuevo conjunto de *estándares* de calidad para el desarrollo de ítems que se basa principalmente en los principios de la psicología cognitiva. Los ítems, además de los criterios tradicionales (por ejemplo: dificultad adecuada y el alto índice de discriminación), también presentan evidencias acerca del aspecto sustantivo de la validez de constructo (Messick, 1995). Tercero, se presenta un nuevo tipo de información sobre el nivel de dominio de los examinados en términos micro-moleculares del proceso de respuesta. En este punto, cabe recordar que las especificaciones de los ítems presentan tradicionalmente la definición de contenidos bajo la noción de *redes nomológicas,* lo que contribuye poco a la comprensión del nivel y complejidad del dominio con base en las características de la

teoría sustantiva. Cuatro, una gran cantidad de ítems puede ser rápidamente desarrollada, como en el caso de los test adaptativos computarizados basados en la GAÍ. En cambio, en el desarrollo tradicional de ítems estos se producen de forma muy lenta. Cinco, los test del área intelectual basados en la generación automática de ítems pueden ser factibles de aplicarse vía internet de forma eficiente, y sin problemas de envejecimiento.

Ahora bien, el objetivo del ESDC es colocar a la teoría cognitiva como base en el desarrollo de los ítems de una prueba con el fin de mejorar el significado y el uso de sus puntuaciones. Para ello, Embretson (1983) toma como ejemplo los estudios típicos de la psicología cognitiva en donde el diseño de las tareas evaluativas se basa explícitamente en las hipótesis acerca de las características específicas del constructo. Con ello, se incorpora al diseño de las tareas evaluativas la noción de que sus características pueden variar sistemáticamente para producir niveles diferenciales de dificultad en diferentes procesos.

Asimismo, el ESDC se estructura en un *marco conceptual* y un *marco procedimental* con el fin de focalizar el desarrollo de un test bajo los principios de la psicología cognitiva. En lo que respecta al *marco conceptual*, uno de sus aspectos se refiere a la representación del constructo, que es en donde la teoría cognitiva tiene un rol central al momento de desarrollar un test y al momento de realizar las interpretaciones de los resultados. Por su parte, el *marco procedimental* es una serie de etapas que guían la incorporación de la teoría cognitiva en el diseño del test (Embretson, 1983, 1998; Gorin & Embretson, 2013).

Para el *marco conceptual*, Embretson (1983, 1995) menciona que la *representación del constructo* y el *plano nomológico* son dos aspectos de la validez que corresponden al significado y a la significancia, respectivamente. Dichos aspectos de la validez del constructo tienen diferentes funciones y, por lo tanto, estos requieren diferente tipo de evidencias.

Por su parte, la *construcción del significado* concierne a los procesos, estrategias, y estructuras del conocimiento subyacentes a las respuestas de un ítem. Dichas evidencias pueden recolectarse gracias a los métodos y tipos de estudios usados y desarrollados por la psicología cognitiva. Por otra parte, el *plano nomológico* corresponde al análisis de la correlación entre los puntajes y los ítems del test con otra medición. Este tipo de análisis es definido por Cronbach y Meehl (1955) como un aspecto relacionado con las *redes nomológicas* debido a dos razones: primero, porque el *plano nomothetico* al igual que las *redes nomológicas* se refieren a la significancia, pero no al significado; segundo, porque un fuerte sistema de hipótesis generado desde investigaciones sobre la *representación del constructo* pude guiar los estudios de validez de una forma mejor que el análisis desde el *plano nomothetico*. Es por ello que distinguir la *representación del constructo* (relacionada al significado) del *plano nomothetico* (relacionado a la significancia) ayuda a focalizar el desarrollo de un test en los principios de la teoría cognitiva (ver Messick, 1989b).

Desde el marco procedimental, Embretson y Gorin (2002) aclaran que para generar ítems que midan constructos específicos es necesaria una explicación integral y valida de cómo influyen los procesos de respuesta de los examinados en su rendimiento

ante los ítems de una prueba. Hay que tomar en cuenta que los ítems de habilidad y logro requieren de tareas complejas que implican múltiples procesos, estrategias u operaciones para su solución. Asimismo, los examinados varían sus competencias, o utilizan diferentes procesos o estrategias de respuesta, para responder correctamente a los ítems, lo cual habla de procesos complejos multidimensionales. De tal manera que, para Embretson (1983) los procesos sustantivos son más importantes que el conjunto de ítems y, por lo tanto, esos procesos sustantivos deben determinar qué dimensión se mide.

El marco procedimental del ESDC no sólo enmarca las etapas involucradas en el desarrollo de los modelos para generar ítems, sino que también se relaciona con diferentes procesos para probar la validez de la prueba. Son siete las etapas que integran el ESDC y van desde especificar los objetivos de la medición hasta la evaluación del modelo de las pruebas generadas (ver Tabla 2.5). Embretson (1983) menciona que aunque dichas etapas se presentan en un orden sugerido, se debe tener en cuenta que todo el proceso es iterativo y que se puede requerir regresar a etapas anteriores del marco procedimental con el fin de mejorar continuamente los ítems generados. Por ejemplo, si el modelo cognitivo desarrollado al inicio no es lo suficientemente amplio para fundamentar la generación de ítems, entonces, sería necesario volver a la etapa del desarrollo del modelo teórico del dominio, incluso, después de haber generado los ítems. La secuencia del marco pretende destacar la importancia de las primeras etapas del desarrollo de la evaluación, que deben abordarse con igual consideración durante el análisis psicométrico (Embretson & Gorin, 2001).

Tabla 2.5. Etapas del Enfoque Sistémico del Diseño Cognitivo (Embretson, 1998; Embretson & Gorin, 2001)

Etapas	Objetivos específicos
Especificar los objetivos de la medición	Definir los objetivos tanto para la representación del constructo como para el *plano nomothetico*
Identificar las características del diseño en el dominio de la tarea	Identificar el tipo de estrategias o tareas evaluativas más idóneas con los principios psicológicos
Desarrollar el modelo cognitivo	Crear el modelo cognitivo para el diseño de los modelos y los tipos de ítems
Generación de ítems	Definir las reglas y las operaciones para generar los ítems
Evaluar el modelo de las pruebas generadas	Evaluar empíricamente el modelo de los ítems generados
Generar un banco de ítems según su complejidad cognitiva	Buscar evidencias predictivas de las propiedades psicométricas de los ítems
Validación bajo el *plano nomothetico*	Buscar la correlación con otros criterios externos

En cada una de las etapas hay ciertas especificaciones que se deben tomar en cuenta para comprender de mejor forma el ESDC. Para la primera etapa, relacionada a la especificación de los objetivos de la medición, se debe especificar un camino a la significancia desde el *plano nomothetico* y otro camino a la construcción del significado desde el *plano sustantivo* o en específico del modelo cognitivo. Para la segunda etapa, asociada a la especificación del diseño de las tareas evaluativas, es importante determinar las características específicas del diseño de las mismas. Con ello, se ayuda a sistematizar y a focalizar el diseño cognitivo antes que el diseño por *redes nomológicas,* tradicionalmente usado para el desarrollo de las pruebas. De esta forma, las características de los ítems son manipuladas con base en la representación del

constructo, tomando en cuenta los procesos cognitivos, las estrategias de respuesta naturales y las estructuras del conocimiento de los examinados.

Para la etapa tres, relacionada con el desarrollo del modelo cognitivo, se debe considerar que es esencial para el ESDC focalizar los esfuerzos en el aspecto sustantivo para el propio desarrollo del modelo con que se generaran los ítems de una prueba. Embretson y Gorin (2001) mencionan que en esta etapa se deben resolver tres aspectos del desarrollo del modelo cognitivo. Primero, se debe reconocer la relevancia de los procesos cognitivos, las estrategias y las estructuras del conocimiento necesarias para identificar y organizar el modelo sustantivo unificado. Dado esto, es importante el papel que juega la revisión de la literatura para integrar modelos teóricos sobre el dominio a medir, que ayuden a guiar el diseño de los modelos y tipos de ítems. Segundo, se deben de igual forma operacionalizar las características de los procesos cognitivos. Para crear un modelo cognitivo que fundamente el diseño de los modelos y los tipos de ítems se deben cuantificar sus características. Si el objetivo es generar modelos bases de ítems, sus características deben ser manipulables al igual que sus puntajes. Tercero, se debe estudiar empíricamente el impacto de las funciones cognitivas con ayuda en las propiedades psicométricas de los ítems ya existentes. El impacto relativo de las características (dificultad y discriminación entre otras) de los ítems, debe ser evaluado de forma reiterativa con varios modelos cognitivos en la generación de ítems. Esta etapa es fundamental para asegurar la *fidelidad estructural* de la prueba.

En la cuarta etapa, se trata la generación de ítems. De manera puntual, en dicha etapa se desarrollan y se establecen las reglas y las operaciones cognitivas en las que se basará el desarrollo de los ítems con el fin de operacionalizar sus características. Si la

etapa anterior se realiza con éxito, las variaciones en los procesos mentales de los examinados, previstas en el modelo cognitivo de los ítems, estarán representadas por las variaciones en los procesos de medición. Así, se seleccionan los ítems característicos (también conocidos como *ítems hijos* o *isomorfos*) que cumplen con las reglas (atributos, operaciones o componentes) y la estructura del modelo del ítem (también conocidos como *ítem padre* o *base*). Inmediatamente después, los ítems generados desde los modelos base son ensamblados y montados para realizar una prueba empírica.

En la etapa cinco, los modelos para la generación de los ítems deben ser evaluados. El modelo subyacente a la generación de ítems debe ser evaluado de forma empírica. Para el ESDC en su conjunto, el éxito de la quinta etapa es esencial para fundamentar tanto el aspecto de la *representación de constructo* de la validez como la GAÍ de la prueba. Embretson y Gorin (2001) realzan la importancia de que el sistema de la generación de ítems se evalué y se confirme a través del análisis cognitivo y psicométrico con el fin de obtener evidencia predictivas o experimentales de la ejecución de los ítems. En este esquema las variables dependientes son el tiempo promedio de las respuestas y la dificultad de los ítems mientras que las variables independientes son los atributos estructurados en los modelos cognitivos de los ítems y las características de los estímulos en los ítems que operacionalizan los procesos cognitivos representativos del dominio de interés. Con ello, el modelo psicométrico debe ser evaluado con base en el ajuste a los datos de ejecución ante los ítems.

Para la sexta etapa se genera un banco de ítems tomando en cuenta su complejidad. Si el sistema de generación es efectivo en predecir o en explicar las propiedades de los ítems, estos podrán ser utilizados para evaluar los diferentes niveles

de complejidad cognitiva. De tal forma, que si el modelo estadístico provee suficientes evidencias predictivas de la dificultad de los ítems de una prueba, esta podrá ser descompuesta en los diferentes parámetros de dificultad atribuidos a las diferentes operaciones cognitivas previstas por el modelo cognitivo. Con ello, los ítems se pueden clasificar según su complejidad cognitiva y su dificultad empírica. En la séptima y última etapa del ESDC se lleva a cabo la validación de la prueba bajo el *plano nomothetico*. Los ítems generados deben ser evaluados por cada uno de los objetivos específicos del análisis predictivo con criterios externos definidos desde los resultados de la representación del constructo y desde criterios con dominios similares.

2.4. Métodos para la construcción y definición de modelos cognitivos de pruebas psicológicas y educativas

A lo largo de los temas discutidos en la presente tesis se ha enfatizado la importancia de fundamentar o basar el desarrollo y la validación de test psicológicos y educativos en el aspecto sustantivo del constructo. Sin embargo, para lograr dicho cometido es necesario algún método que ayude a desarrollar las teorías sustantivas o los modelos cognitivos, y a relacionar estos con los ítems de las pruebas. Por ejemplo, los MPC —brevemente mencionados al inicio del apartado anterior—, a pesar de su sofisticación matemática, requieren de un modelo cognitivo tanto para su desarrollo como para el análisis de los datos obtenidos tras su aplicación. En otras palabras, el modelo cognitivo debe ser construido antes de la aplicación del modelo psicométrico. Con ello, contar con una teoría sustantiva siempre provee una buena base de inicio para modelar los procesos de respuesta ante los ítems de una prueba (Gorin, 2013). Desafortunadamente, la mayoría

de los constructos de las pruebas de logro educativo no han sido examinados por los profesionales de la psicología cognitiva.

Irónicamente, una de las implementaciones de la psicología cognitiva más significativas al campo de la medición son los estudios y métodos cognitivos. En especial, los métodos cognitivos más utilizados por investigadores en el campo de la medición (por ejemplo, Gierl, Leighton, Changjiang, Jiawen, Rebecca & Tan, 2009; Gierl, Wang & Zhou, 2008; Hoppmann, 2007; Johnstone, Bottsford-Miller & Thompson, 2006; Rupp, Templin & Henson, 2010) son el *análisis verbal* (también conocidos como *reportes verbales*) y los *análisis de protocolos* (también conocidos como *protocolos cognitivos*) (Ericsson & Simon, 1984, 1993; Leighton, 2009; Leighton & Gierl, 2007b).

2.4.1. Construcción y definición de modelos cognitivos

Como se ha comentado a lo largo del capítulo, desde el enfoque del aspecto sustantivo del constructo, es importante que las pruebas cuenten con un modelo cognitivo bien estructurado y definido con el fin de aportar las evidencias requeridas para el argumento de validez (Embretson, 1983, 1998; Messick, 1989b). A su vez, el adherirse a una teoría *fuerte* del constructo a lo largo del diseño de los ítems, puede maximizar la validez de las interpretaciones de los resultados de una evaluación (Embretson & Gorin, 2001; Mislevy, 2007). Por su parte, la mayoría de los modelos para el desarrollo de test presentan como primer paso la selección y la generación de una definición clara del constructo a medir (DeVellis, 1991; Wilson, 2005). Así, contar con definiciones claras y comprensibles de los constructos ayuda a los desarrolladores de pruebas durante el proceso de medición a mantenerse focalizados en el rasgo de interés. Además, la vinculación de los ítems de

una prueba con el rasgo a medir es un punto crítico para la representación del constructo y, por lo tanto, un componente importante para la validez (Gorin, 2007). Para ello, el uso de la psicología cognitiva en la evaluación y en la psicometría ha incorporado una nueva dimensión para la definición de los constructos subyacentes a las pruebas.

La definición del constructo puede ser reportada de diversas maneras. Usualmente, son determinadas y reportadas según los propósitos de la evaluación. Por ejemplo, para interpretar las puntuaciones de una prueba de selección o de admisión a la universidad sería suficiente con una descripción general de la teoría sustantiva del constructo a medir (Gorin, 2007). Por otro lado, para la interpretación de pruebas que presentan puntuaciones más complejas (por ejemplo, evaluación de estrategias de solución de problemas) es necesaria una definición del constructo más detallada.

Los desarrolladores de pruebas comúnmente elaboran especificaciones de ítems que proveen con gran detalle la definición del constructo a medir (ver Contreras, 2000; Backhoff & Tirado, 1994; Leighton, 2004). Dichas especificaciones comúnmente presentan una definición del constructo en términos de dimensiones (por ejemplo, contenidos curriculares, competencias, habilidades y conocimientos esperados) y ayudan a asegurar la cantidad suficiente de ítems para el desarrollo de la prueba. Sin embargo, las descripciones del constructo en las especificaciones de los ítems pueden carecer de información sustantiva necesaria para las inferencias de las interpretaciones de la prueba y el argumento de validez. Actualmente, teóricos como Embretson y Gorin (2001), Gorin (2007) y Messick (1989b, 1995) defienden una definición del constructo más detallada, recomendando a los desarrolladores incluir en las especificaciones: (a) una declaración verbal de la descripción del constructo, (b) un modelo del proceso de respuesta, (c) una

lista de habilidades, operaciones o atributos y (d) las relaciones teóricas entre las dimensiones del rasgo.

La definición clara del constructo a medir toma una alta importancia en el diseño de pruebas automatizadas. Gorin (2006) sugiere que la calidad de los ítems se relacione directamente con la especificidad de la definición del constructo. Sin embargo, una fundamentación teórica *fuerte* dada por una definición del constructo clara es sólo el primer paso para el diseño de los ítems de una prueba (Embretsom, 1983; Embretson & Gorin, 2001). Por su parte Bejar (1993) menciona que el éxito obtenido en diferentes proyectos de medición basados en la GAÍ se debe tanto a la fuerza de la fundamentación teórica como a los avances en los modelos algorítmicos de los procesos de generación automática de los ítems.

Idealmente, un modelo teórico del constructo debe preceder y guiar el desarrollo de los ítems. Dicho modelo teórico *a-priori* no debe solamente describir la naturaleza del constructo, sino también conectar las características de los ítems con el modelo que define los procesos cognitivos. Bejar (2002) llama a este enfoque *top-down* y lo recomienda como el más prometedor método de GAÍ en contraparte con el enfoque *buttom-up,* que comienza con los datos resultantes de la aplicación de la prueba para después derivar estadísticamente el modelo del proceso de respuesta de los ítems. En la Figura 8 puede observarse el modelo de desarrollo de una prueba basada en la GAÍ desde un enfoque *top-down*. Dicho modelo comienza con un paso critico relacionado con la definición del modelo cognitivo para el desarrollo de los ítems.

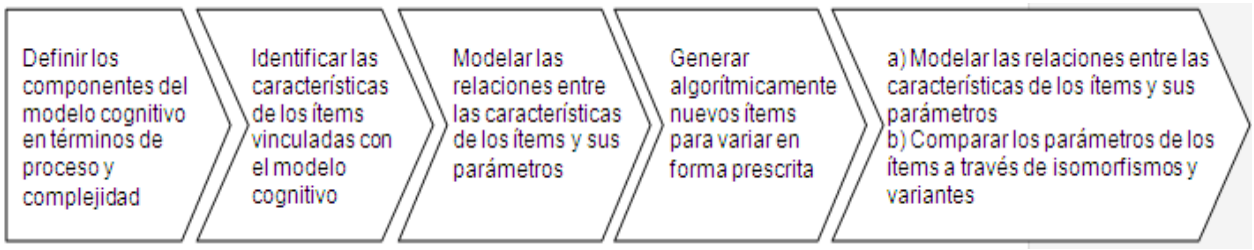


Figura 8. Modelo de desarrollo de una prueba basada en la GAÍ desde un enfoque *top-down*

Según Gorin (2007), los primeros requerimientos para la implementación de la GAÍ giran en torno al modelo psicométrico-cognitivo (matriz Q) de un sistema estático de ítems con el fin de exhibir los componentes generativos para la estructura de la GAÍ. Dado que en el enfoque *top-down* no hay ítems que requieran ser generados durante las primeras etapas del modelo prescrito por Bejar (2002), el desarrollo del modelo cognitivo, la identificación de la estructura y el modelado de la dificultad del ítem pueden ser considerados un prerrequisito para la GAÍ.

2.4.2. Técnicas de pensamiento en voz alta

La *técnicas de pensamiento en voz alta* son muy utilizadas por psicólogos cognitivos para observar, definir y medir los procesos cognitivos que los examinados utilizan para responder los ítems de una prueba de aptitudes o de logro educativo. En dichas técnicas, expertos en el campo (Ericsson & Simon, 1984, 1993; Leighton, 2009; Leighton & Gierl, 2007b) proponen una serie de procedimientos formales sobre cómo llevar a cabo y registrar los datos producidos a partir de la evocación del pensamiento en voz alta para que el contenido de la mente del evaluado se mida con precisión. Tales procedimientos formalizados y datos producidos de los mismos variarán dependiendo de si a los

examinados se les pide (a) verbalizar los procesos de resolución de problemas (es decir, los procesos de respuesta, las cogniciones, las estrategias y planes utilizados) que utilizan ante una tarea o (b) verbalizar sus conocimientos, creencias y actitudes acerca de la tarea.

Aunque actualmente no existe ningún método único disponible con el que de forma directa se identifique el contenido de la mente de una persona, la *técnica de pensamiento en voz alta* y los datos resultantes (reportes verbales) de ella son considerados por muchos psicólogos educativos y cognitivos como una buena alternativa para el estudio de estos. Cabe recordar que son dos los tipos de *técnicas de pensamiento en voz alta* que pueden utilizarse para desarrollar modelos cognitivos del desempeño subyacente a las tareas de medición: (a) El a*nálisis de protocolos* (Ericsson Simon, 1993; Ericsson, 2006) y (b) el *análisis verbal* (Chi, 1997). Ambos métodos pueden utilizarse para identificar y para medir los conocimientos y las habilidades que los examinados usan para resolver tareas evaluativas. Sin embargo, hay diferencias importantes entre estos dos métodos que deben tomarse en cuenta para su aplicación.

Según Ericsson y Simon (1993), el *análisis de protocolos* es usado para identificar y medir el proceso de solución de problemas. Durante la aplicación de dicho método es requerido el reporte de las verbalizaciones sobre los pensamientos dados de forma simultánea a la solución de la tarea. Tal requerimiento es con el fin de identificar, lo más directamente posible, los contenidos de la memoria de trabajo de los examinados. Cabe señalar que los contenidos de la memoria de trabajo son primordiales para el *análisis de protocolos* dado que se encuentran asociados con las funciones ejecutivas (sistema de la memoria donde se asignan los recursos atencionales cuando las personas tratan de

resolver problemas) (Baddeley, 2006) y, por lo tanto, son de vital interés para la adquisición de evidencias que apoyen las inferencias sobre cómo un examinado soluciona algún problema.

Por otro lado, el *análisis verbal* es usado para identificar y medir estructuras del conocimiento, incluyendo creencias y actitudes (Chi, 1997). En este tipo de técnica, es menos específico el momento en el que se deben articular los pensamientos de los examinados, puesto que el foco de interés son los contenidos de la memoria a largo plazo, dado que allí residen las estructuras del conocimiento (Ericsson & Simon, 1984). De tal manera que para el *análisis verbal* no es necesario articular de forma estrictamente simultánea los pensamientos de los individuos con la solución del problema. Lo importante es poder articular el pensamiento del examinado *durante (concurrente) o después* (retrospectivo) de haber solucionado el problema. Lo anterior se justifica debido a que parte de las estructuras del conocimiento pueden ser utilizadas por la memoria de trabajo para resolver problemas.

Sin embargo, la cohesión entre las estructuras del conocimiento y el proceso de solución de problemas solo puede ser plenamente identificada mediante el análisis de la memoria a largo plazo. Además las estructuras del conocimiento ubicadas en la memoria de largo plazo, son considerablemente más estables y poco susceptibles a corromperse por distracciones momentáneas durante las técnicas de pensamiento en voz alta. Con ello, el *análisis verbal* presenta grandes fortalezas en el desarrollo y en la validación de modelos cognitivos de pruebas educativas. Es por ello también que los psicómetras se encuentran en la actualidad más interesados en utilizar el *análisis verbal* que el *análisis de protocolos* (Leighton, 2009). Además, según Chi (1997), se ha asociado más el uso

de los *análisis verbales* a los estudios estadísticos para cumplir con las normas tradicionales relacionadas con la validez que el uso del *análisis de protocolos*.

Asimismo, el contar o no en las especificaciones de los ítems con la matriz de conocimientos y habilidades puede ser un punto a tomar en cuenta por el psicómetra para la selección del tipo específico de *técnica de pensamiento en voz alta*. De ello dependerá la elección más acertada para lograr, de la mejor manera, el desarrollo o la validación del modelo cognitivo de los procesos de respuesta del conjunto de ítems de una prueba. Por ejemplo, para un ítem, del cual solo se tiene una vaga descripción del dominio a medir, es preferible utilizar el análisis de verbalizaciones para poder obtener información con mayor profundidad sobre cómo el examinado utiliza las estructuras del conocimiento para resolver un problema. Con ello, se pueden articular tanto contenidos de la memoria a largo plazo (estructuras semánticas del conocimiento) como contenidos de la memoria de trabajo (habilidades, estrategias y procesos de solución de problemas) y, a su vez, se pueden estudiar la cohesión y la estructura de los procesos cognitivos de los examinados con diferente nivel de dominio del constructo a medir (Leighton, 2009).

Tanto el *análisis de protocolos* como el *análisis verbal* pueden utilizarse para explorar y confirmar hipótesis sobre la cognición de los examinados. Las hipótesis sobre los modelos cognitivos del desempeño de la tarea (Leighton & Gierl, 2007b) ilustran los conocimientos específicos y habilidades que los examinados deben tener y utilizar para resolver una tarea. Sin embargo, dichas hipótesis requieren de evidencias o pruebas para ser aceptadas como verdaderas. Por lo tanto, los modelos de los procesos de respuesta ante los ítems de pruebas psicológicas y educativas requieren apoyarse de evidencias empíricas. Tanto el *análisis de protocolos* como el *análisis verbal* pueden

utilizarse para producir las evidencias requeridas para las pruebas de hipótesis. De tal manera que para la inferencia respecto a si los examinados utilizan ante un ítem ciertos procesos de resolución de problemas, el *análisis de protocolos* puede ser de utilidad con la generación de reportes verbales que den cuenta de la secuencia de los procesos utilizados para responder dichos ítems. Por su parte, para la inferencia sobre si los examinados requieren de la comprensión para resolver un ítem, el *análisis verbal* puede ser útil en la generación de reportes verbales que muestren la estructura del conocimiento requerido para responder dicho ítem.

Es importante señalar que cuando se realizan *análisis de protocolos* la mayor parte del trabajo se hace antes de recoger los reportes verbales. El análisis de las tareas evaluativas y la generación del modelo cognitivo constituyen una parte importante del trabajo que se utilizará de manera descendente para muestrear y codificar los reportes verbales. En contraste, la mayor parte del trabajo en el *análisis verbal* se realiza después de obtener los reportes verbales. En el *análisis verbal* primero se obtienen los reportes verbales y después se identifican las estructuras de conocimiento relevantes. Es en ese momento que los investigadores deben decidir cuál es la mejor estrategia para segmentar y estructurar los informes, y codificar las estructuras del conocimiento. De tal manera, la estructuración, la segmentación y la codificación de los informes se realiza de forma exploratoria, de abajo hacia arriba, sin un modelo para guiar el proceso (Chi, 1997). Según Leighton (2009), son ocho pasos realizados generalmente en el análisis verbal:

- reducir o muestrear los reportes de las verbalizaciones;
- segmentar los informes de las verbalizaciones reducidas o muestreadas (opcional);

- desarrollar o elegir un esquema o formula de codificación;
- operacionalizar las evidencias de los protocolos de codificación que constituyen una asignación a un formalismo elegido;
- representar el formalismo asignado;
- interpretar los patrones; y
- repetir todo el proceso, quizás, en un nivel de granularidad diferente (opcional).

2.4.3. Otras técnicas para el análisis del proceso cognitivo

Además de las *técnicas de pensamiento en voz alta*, hay otros tipos de técnicas para el análisis del proceso cognitivo comúnmente utilizadas en el campo de la medición, algunas de las cuales ya se han mencionado en el texto (ver Snow & Lohman, 1989; Messick, 1989b). Por ejemplo: el *análisis del seguimiento del sendero de la vista* (Newell & Simon, 1972), el *análisis cronométrico* o de latencia de respuesta (Fredericksen, 1980; Posner, 1978; Posner & Rogers, 1978), el *modelado matemático* (Embretson, 1983), el *modelado computacional de simulación de problemas* (Anderson, 1976; Dehn & Shank, 1982; Newell & Simon, 1972; Tomkins & Messick, 1963), el *método de correlaciones cognitivas* (Pellegrino & Glaser, 1979), el *método de análisis de los errores sistemáticos en la ejecución de la tarea* (Brown & Burton, 1978) y el *método de análisis de las estrategias y estilos de resolución de problemas* (Dunker, 1945; van Lehn, 1989).

En especial, para la presente tesis, otra de las técnicas que resulta importante revisar con mayor detenimiento es el *modelado matemático de sub-tareas de respuesta* (Embretson, 1983). Dicha técnica es comúnmente aplicada por investigadores junto a las *técnicas de pensamiento en voz alta* con el fin contar con una buena representación del

constructo de interés (Snow & Lohman, 1989). En particular, el *modelado matemático de sub-tareas de respuesta* junto con el *análisis cronométrico* y el *análisis del sendero de la vista* pueden ayudar al análisis de ciertos procesos que se suscitan en tan sólo algunos segundos y que, por lo tanto, no es posible su introspección (Sternberg, 1977). La obtención de información sobre las diferencias en los tiempos de respuesta puede proveer una base para inferir etapas, pasos o ciclos en el proceso mismo de respuesta, así como para conocer la duración de algunos procesos aislados (Messick, 1989b).

Para la representación del constructo en el *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983), se parte de la base de que los ítems están compuestos por sub-tareas y la dificultad del ítem se puede modelar matemáticamente en términos de procesos identificados a partir de las respuestas para completar una serie de sub-tareas (Sternberg, 1977). Usualmente en dicha técnica se emplean medidas de los procesos subyacentes, dadas por algún método cognitivo, como las *técnicas en pensamiento en voz alta* y el *análisis de expertos* (Rupp, Templin, & Henson, 2010) para dar cuenta de la dificultad de la tarea. Sin embargo es todavía más común para el *modelado matemático* el apoyo de expertos en el dominio de interés. También sucede lo mismo con el desarrollo de la teoría sustantiva y, en particular, con la elaboración de matrices que representen las relaciones entre los ítems y los componentes cognitivos subyacentes a la prueba.

Sin embargo, es importante que los expertos cuenten con un profundo conocimiento en el dominio de interés acerca de los procesos de respuesta que utilizan los individuos. También es importante que cuenten con el conocimiento de diferentes

caminos para el desarrollo de los componentes o atributos, y de los contextos en los que los examinados adquieren y utilizan dichos atributos (Rupp, Templin & Henson, 2010).

Además, en el enfoque de diseño cognitivo propuesto por Embretson (1983) —descrito anteriormente en el presente capitulo— se presenta el *modelado matemático* como uno de los métodos más acordes para la representación del constructo. Sin embargo, también se menciona que es necesario probar la validez del modelo cognitivo obtenido bajo dicha técnica. Para ello, Embretson (1983) propone estudiar empíricamente el impacto de los componentes o los atributos explícitos en el modelo cognitivo de la prueba con el análisis de las propiedades psicométricas de los ítems. De tal manera que, con la aplicación de un modelado psicométrico se represente la probabilidad de obtener una respuesta correcta en función de la dificultad del ítem y el nivel de dominio por parte de los examinados sobre el constructo o rasgo latente subyacente a los ítems (Messick, 1989b; Embretson, 1983, 1998). En otras palabras, el modelo psicométrico ayuda a explicar la relación de la probabilidad de responder correctamente al ítem asociada a los resultados de las sub-tareas representadas en el modelo matemático con las diferentes sub-tares o componentes cognitivos utilizados por los examinados para responder a dicho ítem (Embretson, 1984).

2.5. Modelos psicométricos componenciales

Como ya se mencionó en el apartado de antecedentes de la presente tesis, uno de los grandes resultados de la integración entre la psicología cognitiva y la psicometría es el desarrollo de los modelos componenciales (Van der Linden & Hambleton, 1997 en Romero, 2010). Dichos modelos se han denominado de diferentes maneras (Rupp,

Templin & Henson, 2010): *modelos de clasificación diagnóstica* (Rupp, Temlin & Henson), *modelos psicométricos cognitivos* (Rupp, 2007), *Modelos de Diagnóstico Cognitivo* (MDC) (Nichols, 1994; Nichols, Chipman & Brennan, 1995; Templin & Henson, 2006), *modelos de respuesta latente* (ver Maris, 1995), *modelos de clase latente restringida* (Haertel, 1989; Nacready & Dayton, 1976), *modelos de clase latente de clasificación múltiple* (Maris, 1995, 1999), *modelos de clase latente de localización estructurada* (Xu & von Davier, 2008a, 2008b) y *modelos de la TRI estructurados* (Rupp & Mislevy, 2007).

También es importante señalar que los modelos componenciales comparten la mayoría de sus elementos claves con otros tipos de modelos psicométricos y otros marcos estadísticos como son la Teoría Clásica de los Test (TCT), la TRI, el Análisis Factorial Confirmatorio (AFC), el modelado de ecuaciones estructurales y las estadísticas Beyesianas (Mislevy, 1995; Rupp, Templin & Henson, 2010).

En cuanto a la clasificación de los modelos componenciales, se pueden diferenciar entre los modelos derivados de la TRI y los MDC. Los modelos componenciales derivados de la TRI buscan descomponer los parámetros de los ítems en atributos subyacentes. Un ejemplo de estos modelos es el Modelo Logístico Lineal de Rasgo Latente (LLTM por sus siglas en inglés) de Fischer (1973, 1995). Por su parte, los MDC clasifican a los examinados en *estados de conocimiento* (Tatsuoka, 1995 en Romero, 2010) con respecto al dominio que presentan los sujetos en cada uno de los atributos. Algunos ejemplos representativos de los MDC son el Rule Space Method (RSM; Tatsuoka, 1983), los modelos estadísticos de redes Bayesianas (Mislevy, 1995) y los modelos *Deterministic Input, Noisy And Gate* (DINA; Junker & Sijtsma, 2001), entre otros.

Con respecto a los modelos componenciales derivados de la TRI, estos pueden a su vez clasificarse en modelos *unidimensionales*, modelos *multidimensionales*, modelos de *estrategias de respuesta* y modelos *mixtos* (Romero, 2010). Los modelos *unidimensionales* descomponen los parámetros de los ítems en componentes cognitivos. Un ejemplo de dicho modelo es el LLTM (Fischer, 1973). Por su parte, los modelos *multidimensionales* descomponen la habilidad para resolver ítems dicotómicos o politómicos en atributos cognitivos como en el caso del Modelo General de Rasgo Latente (GLTM; Embretson, 1984). En el caso de los modelos de *estrategias de respuesta*; un ejemplo de estos es el propuesto por Mislevy y Verhelst (1990). En los modelos *mixtos* se tiene la particularidad de vincular la TRI con los modelos de clase latente; un ejemplo de los modelos *mixtos* es el modelo MIRA (Rost, 1990).

Asimismo, los MDC en general proporcionan información específica sobre el grado de dominio de los atributos (de la Torre, 2008a y 2008b). Algunos ejemplos relativamente novedosos de dichos modelos son el DINA (en inglés deterministic inputs, noisy ‘‘and’’ gate), el NIDA (en inglés noisy inputs, deterministic ‘‘and’’ gate), y el modelo reparametrizado unificado (de la Torre y Douglas, 2004; Junker & Sijtsma, 2001). También dentro de los MDC se encuentra el RSM (en inglés *Rule Space Model*; Tatsuoka, 1983) y el enfoque en el que se aplican las redes Bayesianas de Mislevy (1995).

Los diversos modelos componenciales presentados hasta el momento proporcionan diferentes aproximaciones analíticas y propuestas de validación cognitiva. Sin embargo, en esta tesis son dos los modelos componenciales que se utilizaron para el análisis de evidencias de validez basadas en la estructura del modelo cognitivo, el LLTM

y el Método de las Distancias Mínimo-Cuadráticas (LSDM por sus siglas en inglés) de Dimitrov (2007). Ambos modelos se derivan de la TRI y se complementan perfectamente (Romero, 2010). Según Dimitrov (2007), la mayor parte de los modelos componenciales, independientemente de sus bases matemáticas y teóricas, requieren información sobre las puntuaciones de los examinados y no presentan de forma independiente evidencias de validez basadas en la estructura del modelo cognitivo. Para responder a las dos necesidades mencionadas, Dimitrov (2007, en Romero, 2010) propone el LSDM.

El LLTM es utilizado con frecuencia para predecir la dificultad de los ítems ante los atributos cognitivos y ver si estos son significativos o no (Embretson, 1994; Embretson, 1998; Gorin & Embretson, 2013). Por su parte, el LSDM se propone como una aproximación para la validación de atributos cognitivos requeridos en la solución de ítems binarios (Dimitrov, 2007; Dimitrov, Romero, Ponsoda & Ximénez, 2006; Romero, 2010; Romero, Ordoñez, López & Navarro, 2009). Este es un método recientemente desarrollado con el cual se proporciona información sobre la dificultad relativa de los atributos (operaciones cognitivas), de las curvas de probabilidad de dichos atributos y de la validez relacionada a cada uno de los ítems con respecto a los atributos determinados. Como otro beneficio más de utilizar el LLTM y el LSDM juntos, es la posibilidad de realizar análisis de validez cruzada. Lo anterior hace del uso de los dos modelos mencionados un camino provechoso para obtener evidencias de validez de la estructura del modelo cognitivo de una prueba.

2.5.1. Modelo logístico lineal de rasgo latente de Fisher

Por su parte, los modelos componenciales de la TRI han facilitado a la psicometría la medición de procesos estudiados por la psicología cognitiva (Embretson, 1994). Algunos de estos modelos (Mislevy, 1996), son el reflejo de un gran esfuerzo por mezclar modelos ampliados de la TRI y modelos de medición cognitivos basándose en una descomposición lineal de $\boldsymbol{\beta_j}$ o $\boldsymbol{\theta_i}$. Un ejemplo de ello es el LLTM, el cual permite estimar la dificultad de los ítems y la contribución de los diferentes componentes establecidos previamente por las teorías cognitivas a dicha dificultad y decidir si estos son significativos o no. La descripción detallada del fundamento psicométrico de este modelo, se puede encontrar en las obras de Fischer y Molenaar (1995) y en las de Van der Linden y Hambleton (1997).

Por su parte, en el análisis del LLTM (por ejemplo, Draney, Pirolli & Wilson, 1995; Fischer, 1995), $\boldsymbol{\beta_j}$ se reescribe como una combinación lineal de K parámetros básicos $\boldsymbol{\eta_k}$ con pesos $\boldsymbol{q_{jk}}$ y lógito

$$P_j(\theta_i) = \theta_i - \sum_{k=1}^{K} q_{jk}\eta_k \,, \tag{1}$$

donde $Q = [\boldsymbol{q_{jk}}]$ es una matriz generalmente obtenida a priori basada en un análisis de los ítems sobre los atributos cognitivos necesarios para resolverlos, y $\boldsymbol{\eta_k}$ es la

contribución del atributo *K* a la dificultad de los ítems de dicho atributo (Junker y Sijtsma, 2001).

Otros modelos de la TRI, como los modelos compensatorios multidimensionales (p. ej., Adams, Wilson & Wang, 1997; Reckase, 1997), siguen la tradición del análisis factorial. En dichos modelos, se descompone el parámetro θ_i de unidimensionalidad en una combinación lineal del ítem-dependiente a los rasgos subyacentes con lógito:

$$P_j(\theta_i) = \sum_{k=1}^{K} B_{jk}\,\theta_{ik} - B_j\,, \tag{2}$$

Los modelos compensatorios de la TRI, como es el caso de los modelos de análisis factorial, pueden ser sensibles a componentes relativamente grandes que presentan variación en θ. Sin embargo, estos modelos generalmente no están diseñados para distinguir los componentes más finos de variación entre los examinados que suelen ser de interés en la EDC. Por otra parte, el modelo LLTM puede ser sensible a estos componentes más finos de variación entre los ítems, pero tiene la particularidad de no estar diseñado para ser sensible a los componentes de la variación entre los examinados (Junker & Sijtsma, 2001).

El LLTM se obtiene del modelo de RASCH dicotómico

$$P(x = 1|\theta,\delta_1) = \frac{e^{(\theta-\delta_1)}}{1 + e^{(\theta-\delta_1)}}\,, \tag{3}$$

donde el parámetro de dificultad del ítem δ_1 se descompone de forma lineal en los parámetros de dificultad de cada uno de los componentes (Fischer & Molenaar, 1995):

$$\delta_i = \sum_{k=1}^{p} q_{ik}\ \alpha_k + c, \qquad (4)$$

En la ecuación (**4**) δ_1 es el parámetro de dificultad del *i*-ésimo ítem en el modelo de Rasch; α_k, *k*=1,…, *p*, son los parámetros básicos del LLTM y corresponden a las dificultades de cada componente *k*; q_{ik} son los pesos dados de los parámetros básicos α_k, representando la complejidad correspondiente al ítem *i* en el componente *k*-ésimo y; *c* es una constante de normalización. Cabe señalar que la aplicación del modelo LLTM mediante la ecuación (**4**) tiene sentido solo si el modelo de RASCH se ajusta suficientemente bien a los datos. Por otra parte, si los componentes considerados explican de manera exhaustiva las diferencias entre los ítems, se deberían recuperar a través de dicha ecuación estimaciones δ_1 similares a las obtenidas directamente del modelo de RASCH, lo que implicaría una alta correlación entre los parámetros estimados bajo ambos modelos.

Ahora bien, para cada ítem *i*, se debe definir a priori un vector de pesos q_i = (q_{i1}, q, …, q_{lk} …. q_{im}) donde *K* es la ocurrencia del atributo o componente en la respuesta. La matriz formada por todos los vectores de los ítems integra la matriz Q comúnmente compuesta de unos ($q_{ik} = 1$) y ceros ($q_{ik} = 0$). Tal composición, se estructura en función

de la presencia o ausencia de una determinada operación cognitiva en la resolución de un ítem. A partir de dicha matriz y de las respuestas de los sujetos, se estiman los valores α_k, que sirven para calcular el parámetro de dificultad δ de la formula (**4**).

Por otra parte, la aplicación del LLTM requiere de un modelo cognitivo que aporte un marco de referencia para la explicación de los requerimientos de procesamiento cognitivo de los ítems y de la variabilidad de la dificultad entre los mismos. Con ello, la adecuación del LLTM se fundamenta en la plausibilidad y en la validez sustantiva del modelo cognitivo, representado formalmente en la matriz Q. La validación de la matriz Q es un requerimiento previo sin el cual la estimación de los parámetros no tiene valor. Es por ello que, las estimaciones de δ_1 deben contener la información valida que refleje fielmente el modo en que los individuos resuelven los ítems (Fisher, 1995). En definitiva, la pertinencia del LLTM depende estrechamente de la adecuación de la matriz Q en referencia a la teoría cognitiva subyacente a los ítems de la prueba.

2.5.2. Método de las distancias mínimo-cuadráticas de Dimitrov

Los diversos modelos presentados hasta el momento, con sus ventajas y limitaciones, proporcionan diferentes perspectivas de análisis y validación cognitiva. Por ejemplo, los MDC como el DINA o NIDA proporcionan información sobre los perfiles de los examinados en el dominio de los atributos latentes, usando modelos probabilísticos complejos. Por otra parte, los modelos derivados de la TRI, como el LLTM, se enfocan en la predicción de la dificultad del ítem a partir de los atributos cognitivos, pero no profundizan en las relaciones cognitivas entre ítems y no proporcionan información

diagnóstica sobre el nivel de los examinados en cada uno de los componentes. La mayor parte de los modelos anteriormente presentados, independientemente de sus bases matemáticas y teóricas, requieren información sobre las puntuaciones de los examinados en los ítems y no presentan evidencia independiente sobre la validez de los atributos cognitivos y de la estructura del modelo cognitivo de la prueba. Para responder a estas dos necesidades, Dimitrov (2007) propone una aproximación para la validación de atributos cognitivos requeridos en la solución de ítems binarios, el Método de las Distancias Mínimo-Cuadráticas (LSDM por sus siglas en inglés).

El LDSM (Dimitrov, 2007; Dimitrov, Romero, Ponsoda & Ximénez, 2006) es un modelo componencial derivado de la TRI de tipo conjuntivo con características particulares. A diferencia del modelo DINA (y otros modelos no paramétricos), el LSDM se basa en los datos de calibración de los ítems obtenidos bajo los modelos unidimensionales de la TRI de un parámetro (o RASCH), de dos parámetros o de tres parámetros (1PL, 2PL o 3PL). Entre muchas otras cosas, el LSDM proporciona información acerca de la relativa dificultad de los atributos cognitivos, sus curvas de probabilidad y la validez de la relación entre cada uno de los ítems de una prueba y los atributos específicos. Además, un rasgo característico que diferencia al LSDM de cualquier otro modelo psicométrico componencial es que para hacer sus estimaciones no requiere los puntajes obtenidos por los examinados tras la aplicación de los ítems para hacer sus estimaciones.

Para aplicar el LSDM se requieren los parámetros de dificultad de cualquiera de los modelos unidimensionales de la TRI y la matriz Q con las especificaciones de la teoría cognitiva que subyace a los ítems. Tanto los parámetros como las especificaciones

de la matriz Q son utilizados por el LSDM para estimar la probabilidad del funcionamiento correcto del atributo a través de los niveles de habilidad fija (en la escala logarítmica). La información resultante de dicha aplicación puede ser utilizada para obtener la curva de probabilidad para cada atributo y para diagnosticar la validez de las especificaciones en la matriz Q de cada uno de los ítems a través de los niveles de habilidad fija. De tal forma que desde el LDSM, la probabilidad de responder correctamente el ítem se presenta como un producto de las probabilidades de dominio de los atributos, es decir,

$$\mathrm{P}\left(\mathrm{X_{ij}}=1|\theta_i\right)=\prod_{k=1}^{K}P[A_k=1|\theta_i]^{q_{jk}} \qquad (\mathbf{5})$$

donde: X_{ij} es la respuesta binaria (1/0) de un individuo *i* en el ítem *j*; θ_i es la puntuación del rasgo (en escala logarítmica) del individuo *i*; $P[A_k=1|\theta_i]$ es la probabilidad de poseer el atributo A_k para una persona con habilidad θ_i; y q_{jk} es un elemento 0/1 de la matriz *Q* que relaciona al ítem *j* con el atributo A_k.

Si tomamos el logaritmo natural de ambos lados de la ecuación (**5**), nos conduce a un sistema de ecuaciones que toma la siguiente forma matricial para cualquier valor fijo de habilidad θ: **L=QX**, donde **L** es un vector (conocido) con elementos en P_{ij}, **Q** es la matriz de *Q* (conocida), y **X** es el vector (desconocido) con elementos $X_k=\ln P[A_k=1|\theta_i]$. Dimitrov (2007) incorporó para las soluciones del vector X la minimización de la norma euclidiana del vector $\|\mathbf{QX}-\mathbf{L}\|$ usando el método de las distancias mínimo-cuadráticas. Asimismo, si tomamos en cuenta las soluciones para X_k,

la probabilidad de poseer el atributo A_k para una persona con habilidad θ_i se encuentra directamente, $\ln P[A_k = 1|\theta_i] = \exp(X_k)$, permitiendo obtener la curva de probabilidad para cada atributo a través de los niveles de habilidad fijos en la escala logarítmica de -4.0 a 4.0 con un intervalo de 0.5.

Según Dimitrov (2007), para aplicar el LSDM se deben realizar los siguientes pasos: (a) obtener a través de la calibración de TRI o de otras fuentes (por ejemplo, informes y resultados de investigación) los parámetros de dificultad de los ítems a analizar; (b) seleccionar un conjunto de valores θ para cubrir un intervalo específico en la escala logarítmica (digamos de -4.0 a 4.0. con una diferencia de 0.5); (c) calcular la probabilidad $P = (X = 1|\theta)$ para responder correctamente el ítem para cada valor de θ utilizando el parámetro de dificultad estimado desde cualquier modelo unidimensional de la TRI (1PL, 2P o 3PL); (d) tomando los logaritmos naturales de ambos lados de la ecuación (**5**), generar una norma de sistemas de ecuaciones lineales $\|\mathbf{QX} - \mathbf{L}\|$ para cada valor θ, donde **Q** es la matriz *Q*, **X** es un vector de elementos desconocidos $\ln P(A_k = 1|\theta_i)$; **L** es un vector de elementos conocidos $\ln P(X_{ij} = 1|\theta_i)$; (e) minimizar la norma $\|\mathbf{QX} - \mathbf{L}\|$ por medio de las distancias mínimo-cuadráticas y usar la solución X_k para obtener la probabilidad de poseer el atributo A_k: $P[A_k = 1|\theta] = \exp(X_k)$; (f) tabular y graficar las probabilidades de los atributos a través de los valores θ para obtener las Curvas de Probabilidades de los Atributos (CPA); y (g) recuperar las Curvas Características de los Ítems (CCI) mediante el producto de las probabilidades de los

atributos que subyacen a los ítems para evaluar posibles especificaciones erróneas en la matriz de *Q*.

Diferentes autores (por ejemplo, Ma, Çetin & Green, 2009; Romero, 2010; Romero, Ordóñez, López & Navarro, 2009) han encontrado múltiples usos y beneficios de la aplicación del LSDM. Uno de los usos es la validación de los atributos con base en la calibración de los ítems mediante la TRI. Dicha calibración se puede realizar antes de la administración de la prueba. Otro de los beneficios de la aplicación del LSDM es la información proporcionada sobre las CPA en la escala logarítmica. Con ello se facilita el análisis de los ítems y de los atributos dentro de una escala ampliamente conocida por psicómetras y por desarrolladores de pruebas (Dimitrov, 2007; Romero, 2010).

III. MÉTODO

En este capítulo se presenta el método que se siguió para el análisis del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. En el primer sub-apartado se describe la adscripción teórica del presente estudio de validez, así como el modelo teórico–metodológico que fue adaptado para llevarlo a cabo. Para el segundo sub-apartado se describen cada una de las diferentes fases, etapas y actividades de dicho modelo que fueron desarrolladas a lo largo del procedimiento de la investigación.

3.1. Modelo teórico-metodológico para analizar la validez del modelo cognitivo del área de HC del EXHCOBA

De forma concreta, para el diseño y para la adaptación del modelo teórico–metodológico del presente estudio de validez (ver Tabla 3.1) se tomaron en cuenta los aspectos principales del Enfoque Sistémico de Diseño Cognitivo (ESDC) propuesto por Embretson (*The cognitive design system approach* en inglés, 1994) y de la perspectiva *top-down* para pruebas con GAÍ (ver Bejar, 2002, 2010; Gorin y Embretson, 2013; Messick, 1989b). De manera general, la adaptación del modelo radicó en focalizar los trabajos analíticos para obtener evidencias del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja.

Cabe señalar que, aunque cada una de las versiones del EXHCOBA, incluyendo la versión con generación automática de ítems de respuesta compleja se desarrollaron con apoyo de diferentes teorías cognitivas (Backhoff & Larrazolo, 2012; Backhoff & Tirado, 1992; 1994; Tirado, 1986), la estructura interna de cada una de ellas se encuentra construida bajo un modelo de medición de *redes nomológicas*. Retomando lo mencionado por Messick (1989b) en el capítulo anterior, no siempre el proceso de medición empieza desde una estructura del modelo cognitivo. En ocasiones, como en el caso de la presente tesis, se requiere trabajar desde una pesquisa inductiva (Campbell, 1976) para estructurar el modelo teórico o cognitivo subyacente a la prueba tomando en cuenta los procesos naturales de los examinados. Una vez obtenida la teoría sustantiva de la prueba, se puede tener una mejor comprensión e interpretación de sus resultados a la luz de las recientes propuestas teóricas del concepto de validez (Bejar, 2013; Borsboom & Mellenbergh, 2007; Kane, 2007; 2008; Gorin & Embretson, 2013; Messick, 1989; 1995; Yang & Embretson, 2007).

En la Tabla 3.1 se puede observar el modelo teórico-metodológico adaptado para el análisis del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA. Como se mencionó en el marco teórico, en el modelo propuesto por Embretson (1994) y en el enfoque *top-dow* para la GAÍ propuesto por Bejar (2002) se establece que al inicio del desarrollo de una prueba es ideal partir de un modelo cognitivo bien definido en términos del proceso de respuesta con el cual se guie el diseño de los ítems y la GAÍ (Yang & Embretson, 2007). Es por ello que las primeras fases del modelo teórico-metodológico adaptado se encuentran encaminadas a la aplicación de estudios cognitivos con el fin de definir el modelo cognitivo de las dos versiones estudiadas.

Tabla 3.1. Modelo teórico-metodológico para el análisis del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA

Fases	Etapas	Actividades
Fase I Diseño y pilotaje del estudio cognitivo	1.1 Selección del tipo de estudio cognitivo	-Determinar el conjunto de ítems para su análisis, verificar las áreas de membrecía o dominio y analizar las características particulares de los ítems de la prueba. -Seleccionar el tipo de estudio cognitivo y definir los métodos y las técnicas específicas para el análisis cognitivo con base en las características de los ítems y de los participantes.
	1.2 Diseño de los estudios cognitivos	-Adaptar el conjunto de ítems y tareas evaluativas para el piloteo del estudio cognitivo. -Definir los procedimientos de operación del estudio cognitivo. -Determinar el tipo de herramientas, materiales e instrumentos tecnológicos requeridos para la captura de los datos del estudio cognitivo. -Adaptar un laboratorio cognitivo acorde a las necesidades específicas del estudio.
	1.3 Piloteo de los estudios cognitivos	-Establecer los criterios de selección de los participantes del estudio piloto. -Establecer los criterios de selección del grupo de participantes del estudio cognitivo. -Seleccionar y capacitar a los participantes del estudio piloto. -Pilotear y probar las técnicas, estrategias y materiales del estudio cognitivo. -Modificar y adaptar las técnicas, estrategias y materiales del estudio cognitivo con base en los resultados del pilotaje.
Fase II Aplicación del estudio cognitivo	2.1 Selección del grupo de participantes	-Seleccionar al grupo de participantes del estudio cognitivo -Confirmar el consentimiento informado y recabar datos de identificación. -Establecer cronograma de actividades y citas con los participantes del estudio cognitivo.
	2.2 Aplicación en forma del estudio cognitivo	-Entrenar a los participantes del estudio cognitivo. -Aplicar a los participantes el estudio cognitivo. -Recopilar la información obtenida durante el estudio cognitivo.
Fase III Desarrollo y definición del modelo cognitivo	3.1. Análisis de los datos obtenidos durante el estudio cognitivo	-Capacitar a expertos en el análisis del proceso de respuesta de los examinados y en la elaboración de la estructura del modelo cognitivo de la prueba. -**Identificar los procesos y atributos cognitivos subyacentes a los ítems***. -**Analizar y evaluar el diseño de los ítems*** para identificar posible varianza irrelevante introducida por la interfaz de la prueba.
	3.2. Desarrollo y definición del modelo cognitivo de la prueba	-**Modelar el proceso cognitivo requerido para responder a los ítems.*** -Determinar la cantidad, tipo y relaciones entre los ítems de la prueba y los atributos u operaciones cognitivas definidas por los expertos. **-Elaborar la matriz Q con base en el modelo cognitivo de la prueba.***
Fase IV Aplicación del análisis componencial	4.1 Revisión de la estructura interna bajo el modelo de *redes nomológicas*	**-Calibrar con la aplicación del modelo de la TCT los ítems de la prueba.**** **-Analizar la estructura interna bajo el *modelo de redes nomológicas*****. **-Aplicar las pruebas de ajuste entre los distintos modelos psicométricos aplicados.****
	4.2 Revisión de la estructura del *modelo cognitivo* de la prueba	-Elegir y aplicar los modelos psicométricos componenciales acordes a las características del modelo cognitivo estructurado. **-Analizar la estructura del *modelo cognitivo de la prueba*****. **-Realizar una análisis de validez cruzada entre los distintos modelos componenciales aplicados a la prueba.**** -Considerar posibles mejoras en el diseño de la prueba con base en los resultados de los estudios de validez.
	4.3 Interpretación de los resultados de los examinados	-Asignar puntuaciones a los examinados sobre los componentes que se están evaluando a través de la prueba. -Observar que las puntuaciones asignadas a los examinados sobre los componentes cognitivos evaluados a través de la prueba sean informativas. -Elaborar reportes acorde a las necesidades de los diferentes usuarios de la prueba.

(*) Actividades relacionadas con evidencias de validez basadas en el proceso de respuesta requerido para resolver los ítems de la prueba.
(**) Actividades relacionadas con evidencias de validez basadas en la estructura del modelo cognitivo de la prueba.

Un aspecto importante a resaltar de la Tabla 3.1 son las actividades relacionadas específicamente con las distintas evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo. Nótese que en la Fase III del modelo teórico-metodológico adaptado se marcan con un asterisco (*****) las diferentes actividades encaminadas a obtener las evidencias basadas en el proceso de respuesta. Por su parte, en la Fase IV las actividades definidas para obtener las diferentes evidencias de la estructura del modelo cognitivo se encuentran marcadas con dos asteriscos (**).

También, cabe mencionar que la tercera etapa (Interpretación de los resultados de los examinados) de la Fase IV no se desarrolló en la presente tesis. Dicha etapa es importante en el contexto del modelo global para el desarrollo de evaluaciones cognitivas propuesto por Embretson (1994). Sin embargo, el logro de dicha actividad queda fuera de los objetivos comprometidos en el presente estudio que concluye con las actividades de la segunda etapa de la Fase IV relacionada con la revisión de la estructura del modelo cognitivo de la prueba.

3.2. Procedimiento del estudio de validez

Para la aplicación del presente estudio de validez se procuró que todos los procedimientos se apegaran a las prescripciones establecidas en el modelo teórico-metodológico adaptado. Sin embargo, en el transcurso de las actividades del estudio se tuvieron que readaptar algunas actividades y procedimientos, pero siempre procurando focalizar los esfuerzos en el aspecto sustantivo de la validez de constructo de la prueba. Así, en el presente apartado de procedimientos del estudio de validez se describen las

actividades realizadas en cada una de las fases y etapas del modelos teórico-metodológico adaptado.

Fase I. Diseño y pilotaje de los estudios cognitivos

Para la Fase I del modelo teórico-metodológico adaptado de la presente investigación, se diseñaron y se pilotearon los estudios cognitivos destinados para el análisis de validez basada en el proceso de respuesta de los examinados ante los ítems de las dos versiones del área de HC del EXHCOBA. Para ello, se llevaron a cabo tres grandes etapas: la primera se refiere a la selección del tipo de estudio cognitivo; la segunda es sobre el diseño y la adaptación de las técnicas, los instrumentos y los materiales del estudio cognitivo; la tercera etapa es su piloteo.

Etapa 1.1. Selección de los tipos de estudios cognitivos

Como primera actividad a desarrollar en la Etapa 1.1, el modelo teórico-metodológico para análisis del aspecto sustantivo de la validez de constructo de la prueba analizada, se determinó y seleccionó el conjunto de ítems a analizar; asimismo, se identificó su área de dominio en el contexto de la prueba. Una vez realizada dicha actividad, se seleccionó el tipo de estudio cognitivo y se definieron los métodos y las técnicas específicas para el análisis con base en las características de los ítems seleccionados.

Así, para este estudio se decidió delimitar el trabajo analítico en los ítems del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. Por su parte, la Versión con Ítems de Opción Múltiple

(V-ÍOM) contiene 30 ítems y la Versión con Ítems de Respuesta Compleja (V-ÍRC) 20 ítems. A modo de justificación, es necesario recordar que las áreas y los nodos del EXHCOBA presentan una gran diversidad y amplitud de contenidos y temáticas que ameritan un abordaje con extensos recursos de tiempo y operación, lo cual quedó fuera del alcance del presente estudio doctoral. Dado lo anterior, se consideró más apropiado y sin menoscabo de la trascendencia del estudio, el determinar un conjunto de ítems asequible para su manejo.

También, se procuró que el área elegida tuviera más posibilidad de ser contrastada con otros estudios y otras experiencias parecidas en otros países líderes en este tipo de estudios (por ejemplo, Brown & Burton, 1978; Chen & Macdonald, 2011; Gierl et al., 2009; Ma, Çetin & Green, 2009; Revuelta & Ponsoda, 1998; Romero, Ponsoda & Ximénez, 2008). Con ello, los resultados de la presente tesis aportarán a la discusión y al debate en el ámbito disciplinar de pertinencia. Aunado a ello, se procuró establecer las bases y las condiciones para que, en futuros estudios, el modelo teórico-metodológico aquí ilustrado se aplique no sólo a todas las áreas y las versiones del EXHCOBA, sino también en otros contextos de evaluación a nivel nacional e internacional.

Por otra parte, es importante señalar que la V-ÍRC del área de HC del EXHCOBA es una prueba basada en la Generación Automática de Ítems de Respuesta Compleja (GAÍRC). Como ya se mencionó en capítulos anteriores, dicha prueba se basa en modelos de tareas con los cuales se producen *ítems base* (también conocidos como *ítems padre)* que a su vez generan ítems isomorfos (también conocidos como *ítems hijos*), los cuales deben presentar propiedades psicométricas similares. En la Tabla 3.2

se observa el área de HC en el contexto del nivel primaria de la prueba. A su vez, en la Tabla 3.3 se pueden observar los contenidos curriculares evaluados en el área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta construida.

Tabla 3.2. Áreas del EXHCOBA seleccionadas para el análisis del aspecto sustantivo de la validez de constructo

Versión	Nivel	Sección	Área	ítems
V-ÍOM	Primaria	Habilidades básicas (60 contenidos curriculares representados)	Habilidades Verbales	30
			Habilidades Cuantitativas*	30
V-ÍRC	Primaria	Habilidades básicas (40 contenidos curriculares representados)	Habilidades Verbales	20
			Habilidades Cuantitativas*	20

(*) Áreas seleccionadas

Tabla 3.3. Contenidos curriculares evaluados en el área de HC del EXHCOBA

Contenidos curriculares evaluados en el área de HC			
No. Ítem	**V-ÍOM**	**No. Ítem**	**V-ÍRC**
01	Sumas algebraicas	**01**	Sucesiones aritméticas
02	Secuencias lógicas	**02**	Localización en la recta numérica
03	Solución de problemas algebraicos	**03**	Comparación de números decimales
04	Unidad, decena y centena	**04**	División
05	Décima, centésima y milésima	**05**	Representación gráfica de fracciones
06	Multiplicación	**06**	Operaciones básicas con números fraccionarios
07	División	**07**	Elementos de la circunferencia
08	Exponentes	**08**	Cálculo de perímetros de circunferencias
09	Equivalencias	**09**	Cálculo de área
10	Solución de problemas con fracciones	**10**	Cálculo de volumen
11	Concepto de fracción	**11**	Conversión de unidades de volumen
12	Gráficas y fracciones	**12**	Conversión de unidades lineales
13	Gráficas y decimales	**13**	Escalas
14	Suma de fracciones	**14**	Interpretación de graficas circulares
15	Resta de decimales	**15**	Cálculo de porcentajes
16	Multiplicación de decimales	**16**	Regla de tres simple
17	División de decimales	**17**	Cálculo de probabilidad
18	Regla de tres simple	**18**	Interpretación de tablas
19	Cálculo de longitud	**19**	Interpretación de gráficas
20	Cálculo de áreas	**20**	Relación frecuencia-probabilidad
21	Cálculo de volumen		
22	Cálculo de unidades de peso y masa		
23	Cálculo de unidades de tiempo		
24	Equivalencias		
25	Solución de problemas y regla de tres simple		
26	Regla de tres inversa		
27	Suma de los ángulos		
28	Probabilidad con enteros		
29	Probabilidad con decimales		
30	Media estadística		

Por su parte, para seleccionar el tipo de estudio cognitivo y las técnicas específicas a utilizar en la obtención de evidencias de validez basadas en el proceso de respuesta, se analizaron las características particulares de cada uno de los ítems seleccionados. Cabe recordar que los ítems del área HC evalúan contenidos de la asignatura de matemáticas del nivel primaria de educación. También, que tanto la V-ÍOM como la V-ÍRC se presentan en formato computarizado y, en especial, la segunda versión mencionada contiene ítems de respuesta compleja con tres tipos de tareas operativas: (a) selección de elementos (ver Figura 2), (b) arrastre de elementos (ver Figura 3) y (c) escritura numérica y algebraica (ver Figura 4).

Tomando en cuenta lo anterior, para la definición del modelo cognitivo de las dos versiones analizadas se seleccionó el método de *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983), el cual se apoyó con el *análisis de expertos* (Rupp, Templin, & Henson, 2010) en el área de dominio de la prueba. Por su parte, para evaluar el diseño del interfaz de los ítems y para verificar el modelo del proceso de respuesta elaborado por los expertos, se utilizó la *técnica de pensamiento de voz alta* con el *análisis de protocolos* concurrentes y retrospectivos (Ericsson & Simon, 1984, 1993; Leighton, 2009; Leighton & Gierl, 2007b). Aunado a ello, se siguieron las recomendaciones hechas por algunos autores (por ejemplo, Snow & Lohman, 1989; Sternberg, 1977) en el campo de la psicología cognitiva referentes a acompañar el *análisis de protocolos* con el *análisis del sendero de la vista* (Newell & Simon, 1972) y el *análisis cronométrico* o de tiempo de latencia de respuesta (Fredericksen, 1980; Posner, 1978; Posner & Rogers, 1978). Con dicho acompañamiento de técnicas cognitivas, se procuró tener información a la mano para

los casos de los *análisis verbales* en donde se presentaron procesos que se suscitaron en tan sólo algunos segundos y que, por lo tanto, no era posible su introspección (Sternberg, 1977). La obtención de información sobre las diferencias en los tiempos de respuesta e inferir etapas, pasos o ciclos del proceso cognitivo en los casos de informes verbales con información muy pobre o escasa (Messick, 1989b) ayudó a una mejor verificación entre el modelo cognitivo elaborado por los expertos y los procesos cognitivos naturales utilizados por los examinados para responder los ítems del área de HC del EXHCOBA.

Etapa 1.2. Diseño de los estudios cognitivos

Seleccionados el conjunto de ítems de interés, el área de la prueba a analizar y los tipos de estudios cognitivos a aplicar, se procedió a realizar las actividades prescritas en la segunda etapa (Etapa 1.2) de la Fase I. Para ello, se determinaron los procesos operativos de los estudios cognitivos seleccionados, se determinaron el tipo de herramientas e instrumentos tecnológicos requeridos para la captura de los datos y se adaptó un laboratorio cognitivo pensando en las necesidades específicas para el análisis del proceso de respuesta de los examinados.

En cuanto a la definición de los procesos de operación de los estudios cognitivos aplicados, se establecieron diferentes momentos. Primero, se aplicó el método de *modelado matemático de sub-tareas de respuesta* propuesto, el cual se apoyó con el *análisis de expertos*. Después se aplicaron las *técnicas de pensamiento de voz alta* con el *análisis de protocolos* concurrentes y retrospectivos y su acompañamiento con el *análisis del sendero de la vista* y el *análisis cronométrico*. En

la Tabla 3.4 se pueden observar las fases y las etapas de los estudios cognitivos aplicados para analizar las evidencias de validez, basadas en el proceso de respuesta de los examinados ante los ítems de las dos versiones del área de HC estudiadas.

Tabla 3.4. Procesos de operación de los estudios cognitivos

Métodos cognitivos	Procesos de operación
I. Aplicación del método de *modelado matemático de sub-tareas de respuesta*	1.1. Entrenamiento de los expertos previo a los *análisis verbales*
	1.2. Aplicación del *análisis verbal* concurrente y retrospectivo
	1.3. Análisis del proceso de respuesta subyacente a los ítems
	1.4. Desarrollo de un modelo del proceso de respuesta subyacente a cada uno de los ítems
	1.5. Elaboración de una declaración verbal general sobre el contenido evaluado por el ítem basada en el modelo del proceso de respuesta desarrollado
	1.6. Definición de los atributos cognitivos en el proceso de respuesta que ayuden a explicar la dificultad presentada por los ítems de la prueba
	1.7. Elaboración del modelo cognitivo global de la prueba y de la matriz Q de cada una de las versiones estudiadas
II. Aplicación del *análisis de protocolos* concurrentes y retrospectivos	2.1. Entrenamiento de los examinados previo a los *análisis de protocolos* concurrentes y retrospectivos
	2.2. Aplicación del *análisis de protocolos* concurrentes y retrospectivos
	2.3. Elaboración de los reportes verbales
	2.4. Evaluación del diseño del interfaz de los ítems de la prueba
	2.5. Verificación sobre si los examinados utilizan los procesos cognitivos determinados por expertos en el modelo del proceso de respuesta

Cabe señalar que, para la evaluación del diseño del interfaz de los ítems de la prueba, se trabajó con algunos de los elementos de análisis del modelo para la Evaluación del Diseño Universal (EDU) propuestos por Thompson, Johnstone y Thurlow (2002). En investigaciones basadas en la EDU se ha encontrado que los

diseñadores de pruebas pueden desarrollar evaluaciones más accesibles para los examinados mediante su aplicación (Johnstone, 2003). También, se ha encontrado que con la aplicación de dicha evaluación y con el apego en estrategias de diseño eficaz, se puede minimizar la varianza irrelevante del constructo dada por problemas en el diseño y por problemas en el formato de los ítems (Haladyna, Downing & Rodríguez, 2002).

Asimismo, con la aplicación de la EDU se puede evaluar el diseño de los ítems y aumentar la validez de la información que se deduce de los resultados de las pruebas desarrolladas. Por consiguiente, se aplicó la EDU con el fin de evaluar el diseño del interfaz de los ítems del EXHCOBA. Para ello, se tomaron en cuenta los elementos de la EDU propuestos por Thompson y colaboradores (2002): (a) inclusión poblacional, (b) definición precisa del constructo (c) accesibilidad e imparcialidad (d) acomodación flexible de los contenidos, (e) procedimientos e instrucciones simples, claras e intuitivas, (f) comprensibilidad y máxima legibilidad, y (g) máxima legibilidad.

Por otra parte, en las dos últimas actividades de la etapa dos del modelo teórico-metodológico adaptado relacionada con el diseño, la selección y la adaptación de los materiales e instrumentos del estudio cognitivo, se determinó el tipo de herramientas e instrumentos tecnológicos requeridos para la captura de los datos y se montó un laboratorio cognitivo acorde a las necesidades específicas del estudio. Tomando en cuenta las características del interfaz del EXHCOBA, se utilizó el software CAMTASIA STUDIO versión 5 (TechSmith, s.f.). Principalmente, se seleccionó dicho software debido a que permite grabar las verbalizaciones de los examinados, la imagen del interfaz de la prueba junto con todas las acciones ocurridas

en ella durante los reportes verbales, el sendero del indicador del mouse y el tiempo de latencia de cada una de las actividades realizadas por el examinado. Además, al final de la aplicación de las *técnicas de pensamiento en voz alta*, se puede obtener un video con todos los datos mencionados.

Para el análisis de la versión con ítems de respuesta compleja fue de gran ayuda el uso del software CAMTASIA STUDIO versión 5 (TechSmith, s.f.). Lo anterior debido a que se pudieron capturar finamente las acciones realizadas por los examinados en ítems donde se requiere de tareas operativas de selección (), escritura numérica y algebraica () y arrastre de elementos (). Gracias a las ventajas de dicho software, sólo fue necesario conseguir un espacio libre de interrupciones y de ruido para el montaje del laboratorio cognitivo.

Tomando en cuenta el tipo de evidencias para el argumento de validez basado en el proceso de respuesta, los ocho pasos generales para la aplicación de técnicas de pensamiento en voz alta recomendados por Leighton (2009), las características del interfaz de los ítems de las dos versiones estudiadas y los elementos de la EDU (Thompson, Johnstone y Thurlow, 2002), se definieron los pasos y procesos operativos de la *técnica de pensamiento en voz alta* con el *análisis de protocolos* (ver Tabla 3.5 y Apéndice 2), así también, se elaboraron y estructuraron los formatos de aplicación con las indicaciones y preguntas a aplicarse en dicha técnica (ver Apéndice 3).

Tabla 3.5. Pasos para el desarrollo de los procesos operativos del *análisis de protocolos*

Tipo de análisis de protocolo		Pasos
Momento de recolección de datos	**Análisis de protocolos concurrentes**	1er paso: Presentación
		2do paso: Firma del consentimiento informado y captura de los datos de identificación
		3er paso: Revisión del laboratorio cognitivo y de los materiales
		4to paso: Presentación a modo de guía de los tipos de reactivos
		5to paso: Entrenamiento para la técnica de pensamiento en voz alta
		6to paso: Entrenamiento para el seguimiento del indicador del mouse
	Análisis de protocolos retrospectivos	7mo paso: Aplicación de los análisis de protocolos concurrentes
		8vo paso: Aplicación de los análisis de protocolos retrospectivos y las entrevistas de salida.
		9no paso: Agradecimientos y cierre de la sesión

Etapa 1.3. Piloteo de los estudios cognitivos

Una vez que se diseñaron y adaptaron las técnicas, las estrategias y los materiales del estudio cognitivo, se procedió a su piloteo. Para ello, se establecieron los criterios de selección de los participantes y, posteriormente, se realizó el pilotaje a cuatro participantes voluntarios. Después, con base en los resultados del pilotaje, se modificaron y se adaptaron las técnicas, las estrategias y los materiales diseñados.

Ahora bien, para el establecimiento de los criterios de selección de los participantes de los estudios cognitivos se tomaron en cuenta las recomendaciones de Ericsson y Simon (1984, 1993). Dichos autores proponen incorporar al análisis del proceso de respuesta tanto a novatos como a expertos en el dominio de interés. Por lo tanto, se establecieron como criterios para la selección de los participantes las variables de rendimiento escolar, grado educativo y la recomendación del profesor. Además, se estableció que del total de participantes 50% fueran hombres y 50% fueran mujeres.

Por su parte, para la selección de estudiantes novatos y expertos se estableció la selección de estudiantes de tercer grado de secundaria que presentaran un promedio mayor a 8.5 y que, además, fueran referidos por los profesores como estudiantes sobresalientes en el dominio de las matemáticas. Dichos estudiantes conformaron el grupo de estudiantes expertos. De igual forma, se estableció la selección de estudiantes de tercer grado de secundaria que presentaran un promedio mayor a 6.0 pero menor a 8.0 y que fueran referidos por los profesores como estudiantes con bajo desempeño o no sobresalientes en el dominio de las matemáticas. Dichos estudiantes conformaron el grupo de estudiantes novatos. De la misma forma, se determinó que del total de participantes 50% fueran estudiantes expertos y 50% fueran estudiantes novatos.

Por su parte, en cuanto a los criterios de selección de los participantes del estudio cognitivo referente al análisis por expertos, se determinó trabajar con especialistas en el área de matemáticas. Consecuentemente, se definieron diferentes perfiles que deberían presentar los miembros del grupo de participantes para el

análisis por expertos. De forma puntual, para el grupo de expertos en el área se determinó seleccionar a profesores de primaria, secundaria y licenciatura de la asignatura de matemáticas, a profesionales con licenciatura o maestría en el área de matemáticas, a desarrolladores de ítems del área de HC del EXHCOBA y a un especialista en los análisis cognitivos.

Como ya se comentó, una vez definidos los criterios de selección para los participantes, se procedió al piloteo de las técnicas, las estrategias y los materiales del estudio cognitivo. Durante dicho piloteo se identificaron, principalmente, problemas con el tiempo requerido para la aplicación de los *análisis de protocolos* a estudiantes de secundaria. Las primeras aplicaciones del piloteo de los reportes verbales se dieron en un rango de duración entre 90 y 120 minutos en promedio. Ello dificultó recopilar a profundidad los procesos de respuesta de los participantes por factores relacionados con el cansancio. Para solucionar dicho problema, se acortaron y se especificaron tanto las indicaciones de las técnicas como los rubros y elementos de los ítems a analizar, reduciendo así el tiempo de ejecución para quedar entre 60 y 90 minutos en promedio. Con los resultados del piloteo, se procedió a modificar y adaptar las técnicas, las estrategias y los materiales con base en los resultados del pilotaje. En la Tabla 3.6 se pueden observar los procedimientos y el rango de duración para cada uno de ellos.

Tabla 3.6. Rangos de tiempo promedio de ejecución de los cuatro participantes voluntarios en cada procedimiento del estudio piloto del análisis de protocolo

Procedimientos	Rangos de tiempo promedio de ejecución
1.1. Entrenamiento de los examinados previo a los *análisis de protocolos*	20 a 30 minutos
1.2. Aplicación del *análisis de protocolos* concurrentes	30 a 45 minutos
1.3. Aplicación del *análisis de protocolos* retrospectivos	10 a 15 minutos
Total	60 a 90 minutos

Fase II. Aplicación de los estudios cognitivos

Para la Fase II del modelo teórico-metodológico adaptado de la presente tesis, se aplicaron los estudios cognitivos destinados para el análisis del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA. De forma puntual, para esta fase se llevaron a cabo dos grandes etapas relacionadas con la selección del grupo de participantes y la aplicación de los estudios cognitivos seleccionados en **la Etapa 1.1.**

Etapa 2.1. Selección del grupo de participantes para los estudios cognitivos

Para la primera etapa de la Fase II se seleccionó al grupo de participantes de los dos tipos de estudios cognitivos aplicados. Para el método de *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983) se trabajó en total con cuatro especialistas en el área de matemáticas y un especialista en análisis cognitivo. En resumen, se trabajó con una profesora de la asignatura de matemáticas en el nivel secundaria y nivel medio superior, con un licenciado en matemáticas con doctorado en

métodos de investigación y especialista en modelos psicométricos componenciales, y con dos licenciadas en matemáticas con maestría en ciencias educativas y que además son miembros del comité técnico del EXHCOBA. Para la selección de los expertos se siguieron las recomendaciones de Rupp, Templin y Henson (2010) con respecto a la elección de profesionales con un profundo conocimiento de los procesos de respuesta que utilizan los individuos en el dominio de interés, con el conocimiento de diferentes caminos para el desarrollo de los componentes o atributos y con el conocimiento de los contextos en los que los examinados adquieren y utilizan dichos atributos.

Por su parte, para la *técnica de pensamiento en voz alta* con el *análisis de protocolos* se seleccionaron a 24 estudiantes voluntarios de tercero de secundaria, 12 participantes (6 mujeres y 6 hombres) para el *análisis de protocolos* con los 30 ítems de la versión de opción múltiple y otros 12 participantes (6 mujeres y 6 hombres) para el *análisis de protocolos* con los 20 ítems de la versión de respuesta compleja de la misma área. En cuanto a la estimación de la cantidad de participantes requeridos para el análisis de protocolos, Nielson (1994) menciona que puede ser variada según sea el propósito del estudio. Para el análisis de protocolos llevado a cabo en esta investigación sólo se requirió de un grupo pequeño de participantes que aportara suficiente información sobre sus procesos de respuesta ante los ítems estudiados con el fin de verificar si estos están representados en el modelo cognitivo, elaborado previamente por los expertos.

Una vez seleccionados los participantes para los dos estudios cognitivos aplicados, se procedió a la confirmación del consentimiento informado y a la

recolección de los datos de identificación. Así, se establecieron el cronograma de actividades y las citas para la aplicación de los estudios cognitivos.

Etapa 2.2. Aplicación en forma de los estudios cognitivos

Para la aplicación de los dos tipos de estudios cognitivos seleccionados se desarrollaron los procedimientos tal cual se determinaron después de su adaptación y corrección al final del piloteo. Con ello, en la primera actividad relacionada con el *modelado matemático de sub-tareas de respuesta* se aplicó a expertos la *técnica de pensamiento en voz alta* con *análisis verbales* concurrentes y retrospectivos. La razón de la aplicación de dicha técnica fue el ayudar a los expertos a hacer más conscientes los procesos mentales involucrados para responder a cada uno de los ítems de las dos versiones analizadas. Inmediatamente, después de los *análisis verbales,* se realizó el *modelado matemático de sub-tareas de respuesta* con ayuda de los expertos. Cabe recordar que el objetivo de dicha técnica es definir el modelo cognitivo de las dos versiones mencionadas de la prueba. Por su parte, los procedimientos seguidos para la aplicación de dicho método se describen puntualmente en el siguiente apartado.

Por otra parte, el rango promedio de duración de los *análisis verbales* a los expertos fue de 50 a 60 minutos, lo cual representó menos tiempo que el utilizado por los estudiantes de secundaria para realizar el total de procedimientos de los *análisis de protocolos*. Durante la ejecución de los *análisis verbales* a los expertos, se

recolectaron los datos con la ayuda del software CAMTASIA STUDIO versión 5 (TechSmith, s.f.), tal como fue prescrito.

Asimismo, previo a la aplicación de los análisis de protocolo se entrenó a los 24 estudiantes de secundaria. Dicho entrenamiento duró en promedio entre 20 y 30 minutos como se tenía contemplado. Después, se aplicaron en forma los análisis de protocolos concurrentes y retrospectivos, los cuales tuvieron una duración entre 30 a 45 minutos y 10 a 15 minutos, respectivamente. De igual forma que en los análisis verbales a expertos, durante la ejecución de los análisis de protocolos a estudiantes de secundaria se recolectaron los datos con la ayuda del software CAMTASIA STUDIO versión 5 (TechSmith, s.f.), tal como se contempló durante las etapas del diseño y el piloteo de los estudios cognitivos.

Es importante recordar que durante la operación de los *análisis de protocolos* concurrentes se evaluó el diseño del interfaz de los ítems; también se verificó que el proceso de respuesta utilizado por los participantes ante los ítems de la prueba estuviera representado en el modelo cognitivo elaborado por los expertos. Para ello, en los *análisis de protocolos* retrospectivos se realizaron preguntas (ver Apéndice 2 y 3) a los participantes inmediatamente después de contestar el ítem con el fin de complementar la información obtenida en los *análisis de protocolos* concurrentes.

Fase III. Desarrollo y definición del modelo cognitivo

Para la Fase III, relacionada con el desarrollo, la definición y la verificación del modelo cognitivo de la prueba, se llevaron a cabo dos etapas. Una primera etapa relacionada

con el análisis de los datos obtenidos en los estudios cognitivos aplicados y una segunda etapa relacionada con el desarrollo y definición del modelo cognitivo de cada una de las versiones del área de HC del EXHCOBA.

Etapa 3.1. Análisis de los datos obtenidos durante el estudio cognitivo

De forma general, durante la primera etapa de la Fase III se capacitó a los expertos del área de matemáticas seleccionados para identificar y para categorizar los procesos y atributos cognitivos subyacentes a los ítems y, con ello, desarrollar los diferentes modelos del proceso de respuesta de cada uno de los ítems y de los modelos cognitivos de las dos versiones analizadas.

Ahora bien, para el análisis cognitivo por parte de los expertos se realizaron diversas actividades. Primero, los expertos identificaron y categorizaron los procesos y atributos cognitivos subyacentes a los ítems. Después, realizaron un contraste entre los resultados obtenidos durante los reportes verbales de los estudiantes de tercero de secundaria y el resultado de su propio análisis cognitivo de los ítems del área de HC. Con ello, se especificó el dominio de los contenidos evaluados en cada ítem de las dos versiones de la prueba analizada, se elaboraron los modelos de los procesos de respuesta utilizados por los examinados ante cada uno de los ítems de la prueba y se definieron las operaciones cognitivas sustantivas en cada uno de los ítems. En la Figura 9 se puede observar un ejemplo del modelo cognitivo del ítem dos de la V-ÍOM del área HC del EXHCOBA, el cual evalúa según los expertos la *obtención del valor faltante en secuencias numéricas* (ver Apéndice 4).

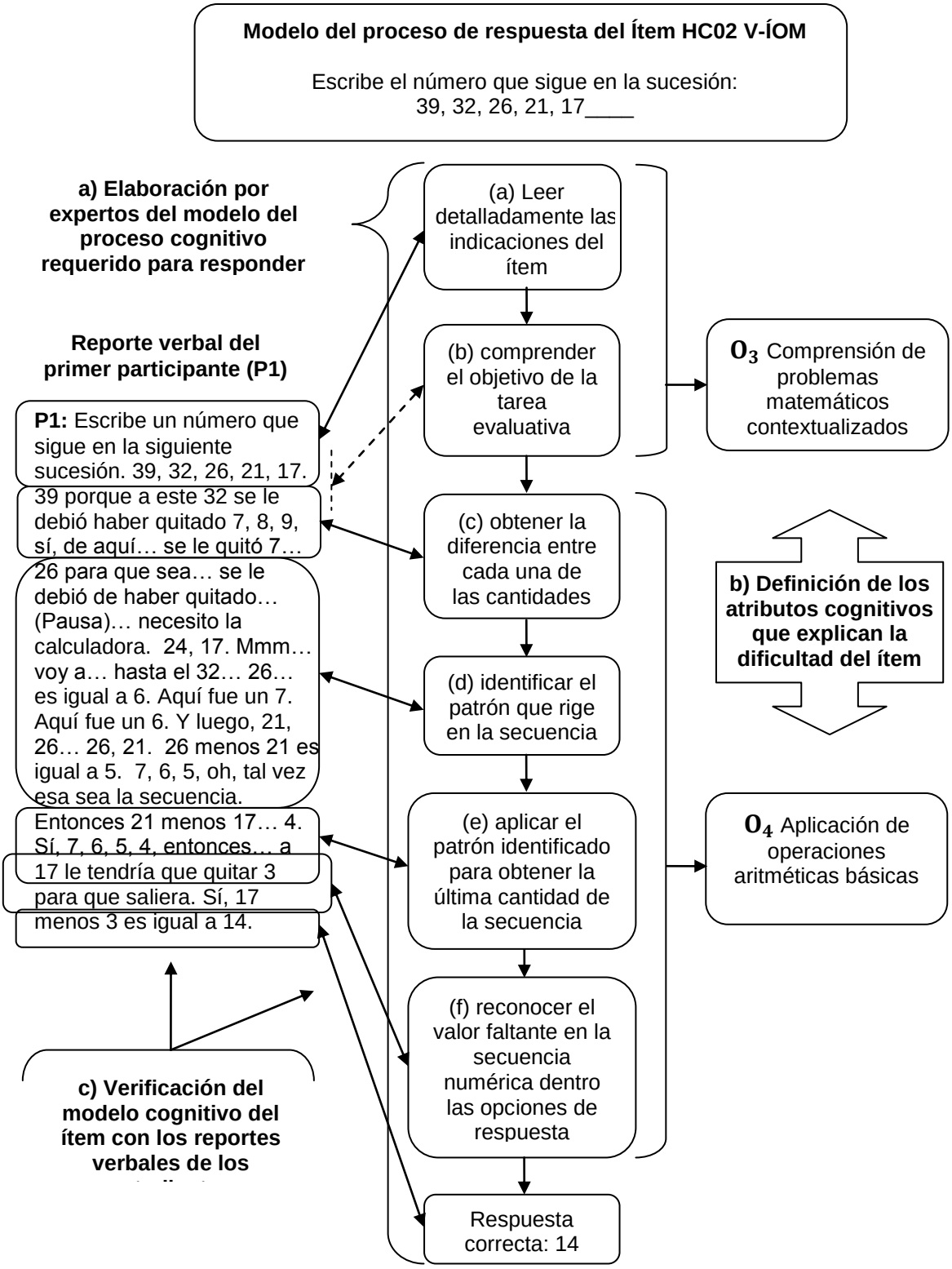


Figura 9. Ejemplo del modelo cognitivo del ítem dos de la V-ÍOM del área HC del EXHCOBA

En el ejemplo presente en la Figura 9 que muestra el modelo cognitivo del ítem dos de la V-ÍOM se pueden identificar: (a) El modelo del proceso cognitivo requerido para responder el ítem elaborado por expertos, (b) los atributos cognitivos que explican la dificultad del ítem y, (c) la verificación del modelo cognitivo del ítem con los reportes verbales de los estudiantes. En especial, se puede ver en dicho ejemplo los diferentes niveles de granularidad de las representaciones de los procesos cognitivos requeridos por el estudiante de secundaria para responder al ítem HC02. También, se puede verificar los pasos del modelo de respuesta elaborado por los expertos y el proceso de respuesta natural presente en el reporte verbal del estudiante. Nótese que fueron dos los atributos que los expertos definieron como posibles fuentes de dificultad del ítem.

De esta manera, después de la elaboración del modelo cognitivo de los ítems de la prueba se realizó la verificación del modelo de respuesta elaborado por los expertos con el proceso de respuesta natural presente en el reporte verbal de los estudiantes participantes. Cabe señalar que durante los primeros reportes verbales resultantes del *análisis de protocolos* aplicados a los estudiantes de tercero de secundaria se pudo identificar que varios de estos presentaban poca información útil para ser analizada. Especialmente, algunos de los datos obtenidos en los *análisis de protocolos* aplicados a los estudiantes *novatos* del área de matemáticas presentaron información superficial y escasa sobre sus procesos de respuesta ante los ítems de la prueba. Tomando en cuenta lo anterior, se decidió descartar 8 de los reportes verbales aplicados. En definitiva, se analizaron solo 8 de los reportes verbales resultantes de

los *análisis de protocolos* para la V-ÍOM y 8 reportes verbales aplicados con los ítems de la V-ÍRC.

Lo anterior revela que no fue afortunada la selección de participantes para el *análisis de protocolos*, lo que resultó en una pérdida de tiempo valioso tanto para los investigadores como para los examinados. Sin embargo, con los datos recopilados de los 16 reportes verbales restantes se pudo obtener información suficiente para proseguir con la evaluación el diseño del interfaz de los ítems con el fin de identificar problemas de varianza irrelevante en el proceso de respuesta. La evaluación del diseño del interfaz dio como resultado la identificación de diferentes problemas relacionados con: (a) respuesta por adivinación o azar por parte de los examinados, (b) dificultades de comprensión y/o legibilidad de las indicaciones del ítem, (c) dificultades de comprensión y/o legibilidad de la base del ítem, (d) usabilidad de las operaciones de respuesta del ítem, (e) la estructura y/o formato del ítem y (f) dificultades de comprensión de las opciones de respuesta del ítem. Los resultados puntuales de la evaluación del diseño de los ítems del área HC del EXHCOBA se presentan en el primer apartado del capítulo de resultados.

Etapa 3.2. Desarrollo y definición del modelo cognitivo de la prueba

Durante la Etapa 3.2 del modelo teórico-metodológico se definieron y estructuraron los modelos cognitivos de cada una de las versiones de la prueba aquí analizadas. Concretamente se definieron 14 operaciones para la V-ÍOM y 9 operaciones para la V-ÍRC. También, se estructuró el modeló cognitivo y la matriz Q tanto de la V-ÍOM como

de la V-ÍRC. Cada uno de los productos resultantes de los estudios cognitivos se muestran con amplitud en el capítulo de resultados.

Para estructurar el modelo cognitivo y la matriz Q se trabajó con el apoyo de expertos quienes determinaron la cantidad y tipo de relaciones entre los ítems de la prueba y los atributos u operaciones cognitivas. Una vez desarrollada la primera versión de la matriz Q, se llevó a cabo una segunda vuelta para su revisión o, si fuese el caso, su corrección. Al final de dicho proceso de revisión y corrección se obtuvo una matriz Q 30_iX14_k para la V-ÍOM y una matriz Q 20_iX9_k para la V-ÍRC.

Fase IV. Aplicación del modelo componencial

Para la cuarta y última fase del modelo teórico-metodológico se desarrollaron tres etapas generales. Dos de dichas etapas están relacionadas con la selección y aplicación de los modelos componenciales, así como con la revisión de la estructura del modelo cognitivo de la prueba. Cabe recordar que la tercera etapa de la Fase IV, que se encuentra relacionada con la interpretación de los resultados de los examinados, no se llevó a cabo debido a que se encuentra fuera de los objetivos de investigación de la presente tesis.

Etapa 4.1. Selección y aplicación de los modelos componenciales

Una de las primeras actividades de la primera etapa de la Fase III consistió en la selección de los modelos componenciales necesarios para el análisis de la validez de la estructura del modelo cognitivo de las dos versiones estudiadas del EXHCOBA. Los

modelos componenciales seleccionados fueron el Modelo Logístico Lineal de Rasgo Latente (LLTM, por sus siglas en inglés) de Fischer (1973; 1995) y el Método de las Distancias Mínimo-Cuadráticas (LSDM, por sus siglas en inglés) de Dimitrov (2007). En especial, el LLTM es propuesto por Embretson (1994) para realizar el análisis psicométrico dentro de su enfoque sistémico de diseño cognitivo. Con la aplicación de dicho modelo, se puede estimar la dificultad de los ítems, medir la contribución de los diferentes componentes a dicha dificultad y decidir si estos son significativos.

Por su parte el LSDM (Dimitrov, 2007), fue seleccionado para complementar y validar los hallazgos resultantes del LLTM aportando evidencias en el proceso integral de validación con el análisis de la estructura del modelo cognitivo (Romero, 2010). Cabe señalar que tanto el LLTM como el LSDM son modelos anidados en la TRI, resultando su uso en conjunto un buen complemento para obtener evidencias de validez basadas en la estructura del modelo cognitivo de la prueba. Sin embargo, para asegurar que los modelos mencionados representan una buena elección, se comparó su ajuste con el modelo DINA (Junker & Sijtsma, 2001).

Previo a la aplicación de los modelos componenciales, se realizó un análisis psicométrico básico de las dos versiones del área de HC del EXHCOBA. Dicho análisis consistió en la calibración y aplicación del modelo de la TCT, en la aplicación del modelo de Análisis Factorial Confirmatorio (AFC) de Fraser (1988) y en la aplicación del modelo de RASCH unidimesional. Además, se realizaron distintas pruebas de ajuste entre los diferentes modelos psicométricos aplicados.

Por su parte, para la calibración de la prueba se utilizó el software Ítem and Test Analysis Program (ITEMAN) (tm) Versión 3.50 (1993), el cual es un producto de la

Assessment Systems Corporation. Dicho software se desarrolló desde la perspectiva de la Teoría Clásica de los Test (Crocker y Algina, 1986). También se utilizó para la calibración el programa libre R 2.15.1. (Ihaka, R. & Gentleman, R., 1996).

La finalidad del análisis psicométrico de los ítems de las dos versiones del área HC con la TCT es calibrarlos y estimarlos a la luz de los *estándares* clásicos de calidad técnica. Los indicadores psicométricos que puntualmente se analizaron son el *índice de dificultad*, el *índice de discriminación* (altos-bajos), el *coeficiente de correlación puntual-biserial* (r_{pbis}) y el *coeficiente de consistencia interna* (**α** de Cronbach). El procedimiento para la obtención de los indicadores psicométricos mencionados consistió de cuatro ecuaciones principales. La primera que se requirió para la obtención del *índice de dificultad* del reactivo fue la ecuación (**6**):

$$p_i = \frac{A_i}{N_i} \tag{6}$$

En esta ecuación (**6**) p_i es el *índice de dificultad* del reactivo, A_i es la cantidad de aciertos en el reactivo y N_i es la cantidad de aciertos más la cantidad de errores en el reactivo. La ecuación (**7**) que se utilizó para obtener el *índice de discriminación* (altos-bajos) fue:

$$D_i = \frac{GA_i - GB_i}{N_{grupo\ mayor}} \tag{7}$$

En esta ecuación (**7**) D_i es el *índice de discriminación* del reactivo i, GA_i es la cantidad de aciertos del reactivo i del 27% de examinados que obtuvieron las puntuaciones más altas en el examen, GB_i la cantidad de aciertos del reactivo i del 27% de examinados que obtuvieron las puntuaciones más bajas en el examen, y N es la cantidad de personas en el grupo más cuantioso (GA_i o GB_i). La ecuación (**8**) que se utilizó para obtener el *coeficiente de correlación puntual-biserial* fue:

$$r_{pbis} = \frac{\bar{x}_1 - \bar{x}_0}{s_x} * \sqrt{\frac{n_1 n_0}{n(n-1)}} \qquad (\mathbf{8})$$

En esta ecuación (**8**) $\bar{x}_1$ es la media de las puntuaciones totales de aquellos que respondieron correctamente el ítem, $\bar{x}_0$ es la media de las puntuaciones totales de aquellos que respondieron incorrectamente el ítem, S_x es la desviación estándar de las puntuaciones totales, n_1 es la cantidad de casos que respondieron correctamente al ítem, n_0 es la cantidad de casos que respondieron incorrectamente al ítem y n es igual a $n_1 + n_0$. Por último, la ecuación (**9**) que se utilizó para obtener el *coeficiente de consistencia interna* (**α** de Cronbach) del instrumento fue:

$$\alpha = \left(\frac{n}{n-1}\right)\left(\frac{\sigma_t^2 - \Sigma\sigma_i^2}{\sigma_t^2}\right) \qquad (\mathbf{9})$$

En esta ecuación (**9**) α es el coeficiente de consistencia interna, n es la cantidad de ítems de la prueba, σ_i^2 es la varianza de las puntuaciones de la prueba y $\Sigma\sigma_i^2$ es la sumatoria de las varianzas de los reactivos.

Por su parte, para la aplicación del modelo de AFC de Fraser (1988) se utilizó el programa NOHARM (Fraser, McDonald & *Vandermeulen*, 2012). Dicho programa estima la matriz de varianza-covarianza residuales después de ajustar el modelo con un número determinado de dimensiones y da la raíz de la media de los residuos al cuadrado (RMSR). Los residuos altos (por ejemplo, superiores a 0.025) indican que se incumple el supuesto de independencia local (recordar que en la matriz de varianzas-covarianzas el máximo valor es 0.25). Si el RMSR es superior al error típico de los residuos (que es $4/\sqrt{N}$ siendo N el tamaño de la muestra), es un indicio de que no se ajusta bien. Podemos observar si el RMSR baja cuando añadimos más factores. Considérese primero la ojiva normal unidimensional que, utilizando la notación de los Lores, el modelo quedaría expresado como:

$$P\{y_j = 1\,|\theta\} = c_j + (1 - c_j)\,N[a_j(\theta - b_j)] \quad (\mathbf{10})$$

Donde θ es la variable latente; a_j es el parámetro de discriminación; b_j es el parámetro de dificultad; c_j es el parámetro de adivinación; $N[.]$ es la función de la distribución normal. De tal manera que si dejamos que $f_{j0} = -a_j b_j$ and $f_{j1} = a_j$, entonces la formula anterior (10) se puede reescribir como:

$$P\{y_j = 1\,|\theta\} = c_j + (1 - c_j)\,N[f_{j0} + f_{j1}\,\theta] \quad (\mathbf{11})$$

Como ya se pudo apreciar, para los diferentes análisis psicométricos se utilizaron distintos paquetes estadísticos especializados. En lo que respecta a la aplicación del modelo de RASCH unidimesional y la aplicación de los modelos componenciales LLTM, LSDM y DINA, se utilizó el programa libre R 2.15.1. (Ihaka, R. & Gentleman, R., 1996). De forma especial, los resultados del análisis del LSDM obtenidos con el programa libre R 2.15.1 se complementaron con los resultados del programa LPCM Win 1.0 de Fischer y Ponocny-Seliger (1998).

Por otra parte, para los distintos análisis psicométricos aplicados en la presente tesis se utilizaron dos bases de datos, una con datos dicotómicos con la respuestas de 2801 graduados de preparatoria ante la V-ÍOM y otra con la respuesta de 702 estudiantes del sexto semestre de preparatoria correspondientes a la V-ÍRC de la misma área. Por su parte, los puntajes brutos obtenidos de la V-ÍRC presentan datos tanto dicotómicos como de crédito parcial (Ferreyra, 2013). Sin embargo, para el análisis psicométrico se decidió trabajar con datos dicotómicos, lo que conllevó a condensar los datos de crédito parcial. La regla general para transformar y para condensar los datos de crédito parcial a dicotómicos se basó en la propuesta de López (2005). Esta regla consiste en condensar las categorías o créditos tomando en cuenta la naturaleza cuantitativa de los atributos psicológicos. Para ello, es necesario definir el punto de corte en la estructura ordinal y aditiva de las magnitudes de esos atributos (Michell, 1999). Dado lo anterior, se decidió asignar 1 cuando el puntaje resultante del ítem fuese igual o mayor del 80 % del crédito total; es decir, en el caso de ítems que presentan cinco créditos, si el examinado obtuvo cuatro de ellos se asignó 1 y, si el examinado obtuvo sólo 3 créditos de cinco posibles, se asignó 0.

Etapa 4.2. Revisión de la estructura del modelo cognitivo de la prueba

Para la segunda etapa de la última fase del modelo teórico-metodológico se desarrollaron otras tantas actividades encaminadas a la revisión de los resultados del análisis componencial con base en el modelo de la estructura del modelo cognitivo definida por los expertos. También, se analizaron las evidencias de validez basadas en la estructura del modelo cognitivo de la prueba y se consideraron posibles mejoras de su diseño.

De forma puntual, para la revisión de los resultados del análisis componencial con base en la estructura del modelo cognitivo definida por los expertos se evaluó con ayuda del LLTM, se evaluó la contribución ($\boldsymbol{\alpha_k}$) de los diferentes componentes u operaciones cognitivas a la dificultad de los ítems ($\boldsymbol{b}$) y se decidió si estos fueron significativos. Cuidadosamente, se revisó la contribución y la significancia de las operaciones cognitivas a la dificultad de los ítems de las dos versiones analizadas.

También, se revisó si se cumplía con lo esperado en relación a la complejidad de las operaciones cognitivas definidas por los expertos. Para ello, se contrastó la complejidad cognitiva definida por los expertos y el peso ($\boldsymbol{\alpha_k}$) obtenido mediante la aplicación del LLTM atribuido a cada una de las operaciones cognitivas de la prueba. Asimismo, se revisó si el LSDM complementó y validó los hallazgos resultantes del LLTM, aportando evidencias en el proceso integral de validación de la estructura del modelo cognitivo de la prueba (Romero, 2010).

Además, en la última actividad del estudio se realizó un trabajo de reconfiguración de la matriz Q con base en la mejora de las estimaciones tanto del

LLTM como del LSDM. El criterio básico para reconfigurar la matriz Q de cada una de las dos versiones analizadas fue mantener la consistencia de la matriz Q reconfigurada con el modelo cognitivo de la prueba, el cual fue definido al inicio por los expertos. Una vez que se consiguió la matriz Q de las dos versiones del área de HC de la prueba con las mejores estimaciones posibles, se procedió al análisis de la validez cruzada y al análisis comparativo del ordenamiento de las operaciones cognitivas subyacentes a las dos versiones del área de HC, según la dificultad relativa entre los modelos LLTM y LSDM.

IV. RESULTADOS

En este capítulo se presentan las actividades de las Fases III y IV con respecto al análisis de los resultados de la aplicación del modelo teórico-metodológico que se adaptó para obtener evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo del área de HC del EXHCOBA. Para ello, el capítulo se estructura en tres secciones que dan cuenta de dichos resultados. En la primera sección se describen los resultados del análisis de las evidencias de validez basadas en el proceso de respuesta, los cuales, se presentan en dos sub-apartados; para el primer sub-apartado se muestran los resultados del análisis y la definición por expertos del modelo cognitivo de los ítems del área de HC del EXHCOBA; en el segundo sub-apartado se describen los resultados de la evaluación del diseño de la interfaz de los ítems y de la verificación del modelo del proceso de respuesta elaborado por los expertos con el proceso de respuesta utilizado por los estudiantes de tercero de secundaria para responder a los ítems de la prueba.

Por su parte, en la segunda sección de este capítulo se presentan los resultados del análisis psicométrico básico bajo la Teoría Clásica de los Test (TCT), del análisis de dimensionalidad y del ajuste de los modelos RASCH y LLTM de Fischer (1973; 1995). Mientras que en la tercera sección se presentan los resultados del análisis de las evidencias de validez basadas en la estructura del modelo cognitivo mediante la aplicación de los modelos componenciales LLTM y el LSDM de Dimitrov (2007); también, en dicha sección se presentan los resultados del proceso de

reconfiguración de la matriz Q de las dos versiones estudiadas, y los resultados de la validación cruzada entre los modelos componenciales aplicados.

4.1. Resultados del análisis de las evidencias de validez basadas en el proceso de respuesta

4.1.1. Análisis y definición por expertos del modelo cognitivo de los ítems del área de HC del EXHCOBA

Como resultado de la aplicación del método de *modelado matemático de sub-tareas de respuesta* (Embretson, 1983) apoyado con el *análisis de expertos* (Rupp, Templin, & Henson, 2010) se obtuvieron diferentes productos que aportan a la definición integral del modelo cognitivo de los ítems del área de HC del EXHCOBA. Un primer producto de la aplicación de dichos métodos fue el modelo del proceso de respuesta de cada uno de los ítems de la prueba (ver Apéndices 4 y 5). Dichos modelos se encuentran desarrollados bajo un esquema lineal en términos de los pasos a seguir para responder correctamente los ítems de la prueba. Por ejemplo, para el ítem dos de la V-ÍOM los expertos definieron que la secuencia de pasos a seguir para contestarlo correctamente es: (a) leer detalladamente las indicaciones del ítem, (b) comprender el objetivo de la tarea evaluativa, (c) obtener la diferencia entre cada una de las cantidades, (d) identificar el patrón que rige en la secuencia, (e) aplicar el patrón identificado para obtener la última cantidad de la secuencia y (f) reconocer el valor faltante en la secuencia numérica dentro las opciones de respuesta. Para una mejor visualización de dicha secuencia de pasos se puede modelar, en caso de ser necesario, un diagrama de cajas y flechas como el que se muestra en la Figura 10.

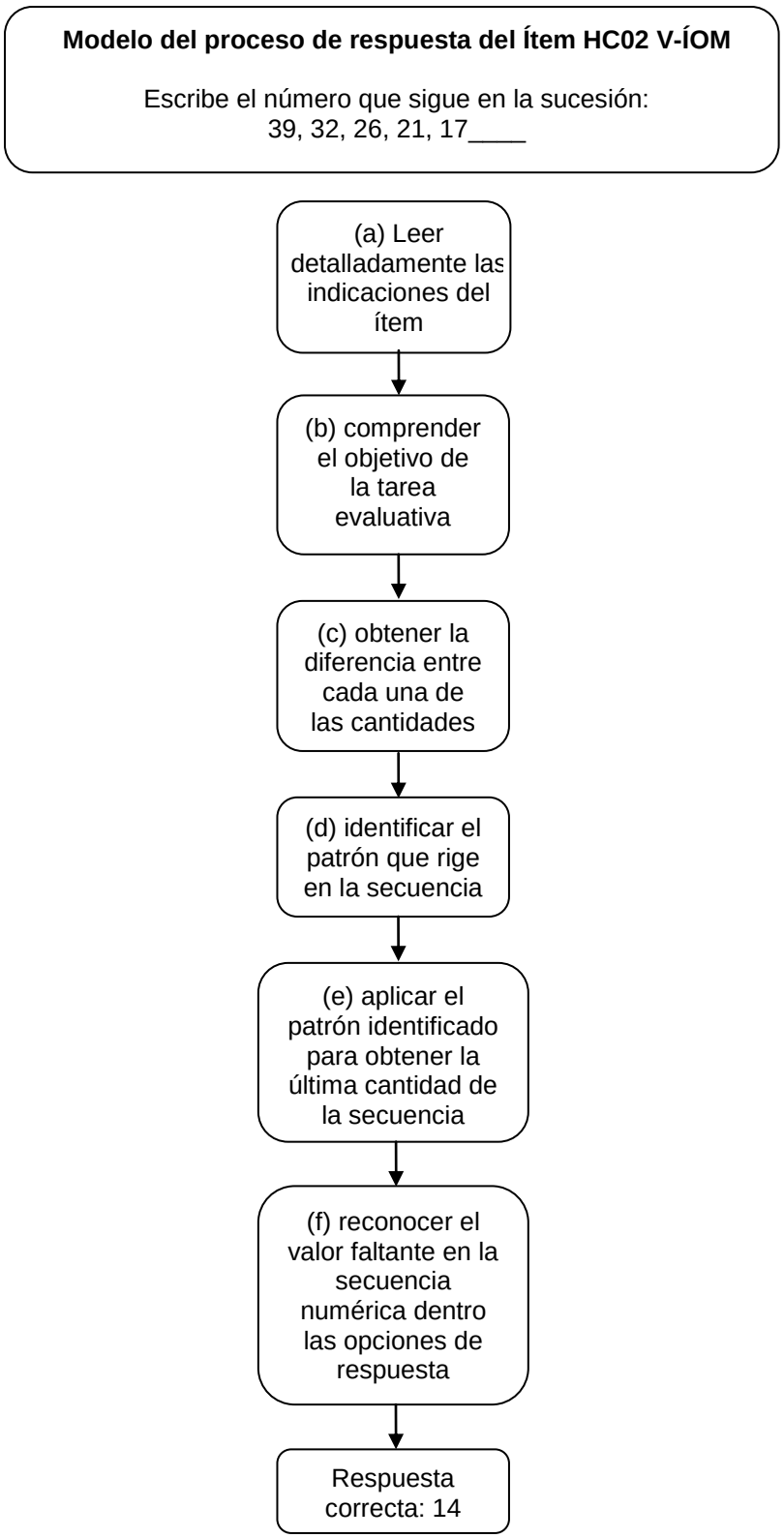


Figura 10. Ejemplo del modelo del proceso cognitivo del ítem dos de la V-ÍOM del área HC del EXHCOBA

Asimismo, un producto obtenido inmediatamente después de los modelos de respuesta es una lista definida por los expertos con las operaciones cognitivas que presentan mayor probabilidad de explicar la dificultad de los ítems de las dos versiones de la prueba. En la Tabla 4.1 y la Tabla 4.2 se pueden observar que los expertos identificaron para los 30 ítems de la V-ÍOM catorce operaciones sustantivas y para los 20 ítems de la V-ÍRC identificaron nueve de ellas. Cabe señalar que el orden descendente presente en la columna del código responde al grado de complejidad de las operaciones cognitivas. Por ejemplo, en el caso de la V-ÍOM, la operación relacionada con *Conocer el nombre y ubicación de los dígitos* ($\mathbf{O_1}$) es la operación cognitiva con menor grado de complejidad en opinión de los expertos y la operación relacionada con la *Aplicación de expresiones algebraicas* ($\mathbf{O_{14}}$) la de mayor grado de complejidad. Para la V-ÍRC, la operación relacionada con la *Representación de lugares geométricos* ($\mathbf{O_1}$) es la de menor complejidad, mientras que la operación relacionada con *Representación de modelos matemático-geométricos* ($\mathbf{O_9}$) presenta una mayor complejidad.

Tabla 4.1. Operaciones cognitivas del área de HC del EXHCOBA de la V-ÍOM

Código	Operaciones cognitivas	Descripción
$\mathbf{O_1}$	Conocer el nombre y ubicación de los dígitos	Comprender los nombres de los dígitos dentro del sistema decimal y su ubicación
$\mathbf{O_2}$	Fracción y visualización de figuras geométricas	Fraccionar y visualizar figuras geométricas poligonales como el triángulo, el cuadrilátero y el pentágono
$\mathbf{O_3}$	Comprensión de problemas matemáticos contextualizados	Comprender los problemas matemáticos planteados en lenguaje común y contextualizado
$\mathbf{O_4}$	Aplicación de operaciones aritméticas básicas	Aplicar operaciones aritméticas básicas de =,+, -,*, /, signos y números
$\mathbf{O_5}$	Adición de ángulos de un triángulo	Aplicar la suma de los ángulos interiores y exteriores de un triángulo
$\mathbf{O_6}$	Aplicación de fracciones	Comprender y resolver fracciones, así como su equivalencia con números decimales
$\mathbf{O_7}$	Identificación de patrones de secuencias numéricas	Identificar patrones de secuencia numérica de suma, resta y mixta
$\mathbf{O_8}$	Representación de modelos exponenciales	Representar las reglas básicas de las operaciones exponenciales
$\mathbf{O_9}$	Representación del modelo matemático del perímetro, el área y el volumen	Representar y aplicar el modelo matemático del perímetro, el área y el volumen de una figura geométrica
$\mathbf{O_{10}}$	Uniformidad de unidades de medida diferentes	Uniformizar y comparar las unidades de medida de litros a kilogramos
$\mathbf{O_{11}}$	Aplicación de la regla de tres simple	Representar y aplicar el modelo matemático-aritmético de la regla de tres simple
$\mathbf{O_{12}}$	Suma de fracciones	Aplicar la suma de fracciones con mínimo común múltiplo
$\mathbf{O_{13}}$	Aplicación del modelo matemático de probabilidades	Representar modelos probabilísticos y aplicar la adición de probabilidades y para los sucesos mutuamente excluyentes
$\mathbf{O_{14}}$	Aplicación de expresiones algebraicas	Representar y aplicar el modelo matemático de las expresiones algebraicas

Tabla 4.2. Operaciones cognitivas del área de HC del EXHCOBA de la V-ÍRC

Código	Operaciones cognitivas	Descripción
$\mathbf{O_1}$	Representación de lugares geométricos	Representar y visualizar figuras geométricas planas y volumétricas, así como sus elementos (puntos, líneas, vértices, ángulos, etcétera)
$\mathbf{O_2}$	Posicionamiento y ubicación de valores	Posicionar y ubicar valores en una recta numérica o en una secuencia de menor a mayor que
$\mathbf{O_3}$	Aplicación de operaciones aritméticas básicas	Aplicar operaciones aritméticas básicas de =,+, -,*, /, signos y números
$\mathbf{O_4}$	Aplicación de escalas gráficas en mapas	Aplicar el modelo matemático Escala=D (medidas en el dibujo o mapa)/R (medidas reales) y calcular una aproximación de la distancia entre distintos lugares en un mapa
$\mathbf{O_5}$	Identificación de patrones de secuencias numéricas	Identificar patrones de secuencia numérica de suma, resta y mixta
$\mathbf{O_6}$	Interpretación de la información del problema	Comprender la descripción contextual del problema y e inferir valores con base en la información de gráficas, mapas y tablas
$\mathbf{O_7}$	Representación de modelos matemático-aritméticos	Representar modelos de división con galera con números decimales, probabilísticos, números racionales, MCM, regla de tres y de porcentajes
$\mathbf{O_8}$	Cálculo de equivalencias de unidades de medida	Calcular equivalencias de volumen y de longitud con base en el sistema internacional de unidades de medida y el sistema inglés
$\mathbf{O_9}$	Representación de modelos matemático-geométricos	Representar modelos matemáticos–geométricos para perímetros, áreas y volúmenes

Durante la definición del nivel de complejidad de las operaciones cognitivas, los expertos mostraron algunas divergencias. Para la versión del área de HC con ítems de opción múltiple, algunos expertos mencionaron que las operaciones *Comprensión de problemas matemáticos contextualizados* ($\mathbf{O_3}$) e *Identificación de patrones de secuencias numéricas* ($\mathbf{O_7}$) deberían colocarse en un nivel anterior de complejidad, es decir, algunos los expertos discutían la posibilidad de que dichas operaciones

cognitivas no representaran el nivel de complejidad asignado, sino uno menor. Por otra parte, algunos expertos mencionaban que era necesario condensar distintas operaciones en un atributo. Por ejemplo, se discutió que las operaciones cognitivas *Representación de modelos exponenciales* ($\mathbf{O_8}$) y *Aplicación de la regla de tres simple* ($\mathbf{O_{11}}$) de la V-ÍOM deberían condensarse junto con la operación $\mathbf{O_4}$ en un atributo relacionado con la *Aplicación de modelos mate*m*ático-aritméticos*. Otros expertos propusieron la desagregación de algunas operaciones. Ese fue el caso de operaciones como la *Suma de fracciones* ($\mathbf{O_{12}}$). Concretamente, se propuso que los ítems asociados con dicho atributo se relacionaran de forma desagregada con las operaciones *Aplicación de operaciones aritméticas básicas* ($\mathbf{O_4}$) y *Aplicación de fracciones* ($\mathbf{O_6}$).

Por su parte, para la definición del nivel de complejidad de las operaciones cognitivas de la versión del área de HC con ítems de respuesta compleja, también se presentaron algunas divergencias entre los expertos. Uno de los tres expertos que analizaron los ítems de respuesta compleja sugirió que la operación relacionada con la *Aplicación de escalas gráficas en mapas* ($\mathbf{O_4}$) debería estar entre las operaciones de menor dominio requerido por parte de los examinados a la par de la operación $\mathbf{O_2}$ relacionada con el *Posicionamiento y ubicación de valores* que se encuentra en segundo lugar. También, algunos expertos presentaron problemas para definir la operación cognitiva *Representación de modelos matemáticos-aritméticos* ($\mathbf{O_7}$) como

diferente o separada de la operación *Aplicación de operaciones aritméticas básicas* ($\mathbf{O_3}$).

La problemática mencionada en el párrafo anterior es común en la definición de modelos cognitivos, con los cuales se intentan explicar los procesos de respuesta necesarios ante una tarea evaluativa compleja. Estudios cognitivos como los realizados por Ericsson y Simon (1984, 1993), en donde se analizan los procesos cognitivos de novatos y expertos, presentan un buen ejemplo de ello. No obstante, ante dicha dificultad para condensar o estratificar diferentes operaciones cognitivas complejas o simples, se ha encontrado que el análisis y el juicio por expertos es un excelente método para solucionar tal problemática.

Otro de los productos resultado del análisis por expertos fue la matriz Q para cada una de las del área HC analizadas. Dicha matriz presenta la relación entre los ítems y las operaciones cognitivas que subyacen a estos (ver Tabla 4.3 y 4.4). Nótese que los expertos estructuraron con las 14 operaciones cognitivas definidas para la V-ÍOM una matriz Q 30iX14k y con las 9 operaciones cognitivas definidas para los ítems de la V-ÍRC una matriz Q 20iX9k. En la última columna de las tablas de cada matriz Q mencionadas se incluye la cantidad total de operaciones requeridas por cada ítem *compuesto*. En caso de tratarse de un ítem *individual* se presenta sólo la operación cognitiva que pretende evaluar. Cabe recordar que el número 1 se asigna a la condición del ítem cuando se requiere de la operación cognitiva señalada y 0 en el caso contrario.

Tabla 4.3. Matriz Q 30i X14k de la V-ÍOM del área HC del EXHCOBA

No. ítem	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	O_{10}	O_{11}	O_{12}	O_{13}	O_{14}	Total
01	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
02	0	0	0	1	0	0	1	0	0	0	0	0	0	0	2
03	0	0	1	1	0	0	0	0	0	0	0	0	0	1	3
04	1	0	1	0	0	0	0	0	0	0	0	0	0	0	2
05*	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
06	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
07	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
08	0	0	1	0	0	0	0	1	0	0	0	0	0	0	3
09	0	0	0	1	0	1	0	0	0	0	0	0	0	0	2
10	0	0	1	1	0	1	0	0	0	0	1	0	0	0	3
11	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
12	0	1	0	0	0	1	0	0	0	0	0	0	0	0	2
13	0	1	0	1	0	1	0	0	0	0	0	0	0	0	3
14	0	1	0	1	0	1	0	0	0	0	0	1	0	0	3
15	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
16	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
17	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
18	0	0	1	1	0	0	0	0	0	0	1	0	0	0	3
19	0	0	0	1	0	0	0	0	1	0	0	0	0	0	3
20	0	0	1	1	0	0	0	0	1	0	0	0	0	0	3
21	0	1	0	1	0	0	0	0	1	0	0	0	0	0	3
22	0	0	1	1	0	0	0	0	0	1	0	0	0	0	3
23	0	0	1	1	0	0	0	0	0	0	0	0	0	1	3
24	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3
25	0	0	1	1	0	1	0	0	0	0	1	0	0	0	4
26	0	0	1	1	0	0	0	0	0	0	1	0	0	0	3
27*	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
28	0	0	1	1	0	0	0	0	0	0	0	0	1	0	3
29	0	0	1	1	0	0	0	0	0	0	0	0	1	0	4
30	0	0	1	1	0	0	0	0	0	0	0	0	0	0	3

(*) ítems que miden una sola operación

Tabla 4.4. Matriz Q 20iX9k de la V-ÍRC del área HC del EXHCOBA

No. Ítem	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_8	O_9	Total
01	0	0	1	0	1	0	0	0	0	2
02*	0	1	0	0	0	0	0	0	0	1
03*	0	1	0	0	0	0	0	0	0	1
04	0	0	1	0	0	1	1	0	0	3
05*	0	0	0	0	0	0	1	0	0	1
06	0	0	1	0	0	1	1	0	0	3
07*	1	0	0	0	0	0	0	0	0	1
08	1	0	1	0	0	1	0	0	1	4
09	1	0	1	0	0	1	0	0	1	4
10	1	0	1	0	0	1	0	0	1	4
11	0	0	1	0	0	1	1	1	0	4
12	0	0	1	0	0	1	1	1	0	4
13	0	0	0	1	0	1	0	0	0	2
14	0	0	1	0	0	1	1	0	0	3
15	0	0	1	0	0	1	1	0	0	3
16	0	0	1	0	0	1	1	0	0	3
17	0	0	1	0	0	1	1	0	0	3
18	0	0	1	0	0	1	1	0	0	3
19	0	1	0	0	0	1	0	0	0	2
20	0	0	1	0	0	1	1	0	0	3

(*) ítems que miden una sola operación

Tomando en cuenta los modelos del proceso cognitivo para responder a los ítems de la prueba, se elaboraron descripciones resumidas y generales del contenido para cada uno de los ítems analizados (ver Tabla 4.5). Con ello y con cada uno de los productos resultado del *análisis por expertos* se aportan en cierta medida, además de las evidencias basadas en el proceso de respuesta, evidencias de validez que se basan en el contenido (Yang & Embretson, 2007). Con los estudios cognitivos aquí aplicados se aporta información valiosa que puede evaluar la adecuación de la definición y de la descripción del contenido de la prueba. Además, se pueden identificar errores en el desarrollo de los ítems, mejorar la definición del constructo y a su vez enriquecer el propio desarrollo de las especificaciones de los ítems del área HC del EXHCOBA.

Tabla 4.5. Descripción de los contenidos de la V-ÍOM y V-ÍRC del área de HC del EXHCOBA resultantes del análisis del proceso de respuesta por expertos

No. Ítem	V-ÍOM	No. Ítem	V-ÍRC
01	Adición y sustracción con apoyo en la recta numérica	**01**	Obtención del valor faltante en secuencias numéricas
02	Obtención del valor faltante en secuencias numéricas	**02**	Ubicación de fracciones en la recta numérica
03	Construcción de una expresiones algebraica	**03**	Ordenación de números decimales de menor y mayor en espacios vacíos
04	Identificación del valor posicional en números enteros	**04**	División de números decimales
05	Identificación posicional de números decimales	**05**	Representación de fracciones en figuras geométricas
06	Juicio situacional de la representación de un número	**06**	Solución de suma y resta de fracciones en un contexto
07	Aplicación de módulos en un contexto	**07**	Identificación de elementos de la circunferencia
08	Representación de exponentes en un contexto	**08**	Cálculo de perímetros de círculos en un contexto
09	Equivalencias de decimales a fracciones	**09**	Cálculo de áreas de triángulos rectángulos
10	Representación de la fracción en un contexto	**10**	Cálculo de volúmenes de prismas rectangulares
11	Identificación de la descripción de las partes de una fracción	**11**	Cálculo de equivalencias de unidades de volumen
12	Identificación de una fracción en una figura	**12**	Cálculo de equivalencias de unidades de longitud
13	Identificación de una fracción en una figura para su conversión a decimal	**13**	Cálculo de distancias en mapas con uso de escalas
14	Identificación de fracciones en una figura para su suma	**14**	Representación numérica de porcentajes menores que 100 en un contexto
15	Suma y resta de números decimales	**15**	Representación numérica de porcentajes mayores que 100 en un contexto
16	Multiplicación de números decimales	**16**	Cálculo de la regla de tres simple en un contexto
17	División de números decimales	**17**	Cálculo de probabilidades expresada en fracciones
18	Aplicación de la regla de tres simple	**18**	Inferencia y cálculo de un valor dada una tabla de proporcionalidad directa
19	Cálculo de perímetros de círculos	**19**	Inferencia de valores dada una gráfica poligonal en un contexto
20	Cálculo de áreas de las caras de un prisma	**20**	Estimación de frecuencias de ocurrencias de eventos dada la probabilidad
21	Relación de volúmenes		
22	Comparación de unidades entre unidades de medida distintas		
23	Aplicación de la regla de tres simple		
24	Relaciones de proporcionalidad		
25	Regla de tres simple con fracciones		
26	Aplicación de la regla de tres inversa en un contexto		
27	Conocimiento de la suma de los ángulos internos de un triángulo		
28	Probabilidad de un evento en un contexto		
29	Adición de probabilidades para eventos		
30	Cálculo de la media aritmética		

4.1.2. Análisis del proceso de respuesta de los examinados ante los ítems del área de HC del EXHCOBA

Con la aplicación de las *técnicas de pensamiento en voz alta* con *análisis de protocolos* a estudiantes de tercero de secundaria se obtuvieron tres tipos de resultados: (a) los procesos de respuesta que utilizan los estudiantes de tercero de secundaria ante los ítems de la prueba, (b) la verificación de la similitud entre los modelos del proceso de respuesta definidos por expertos y procesos de respuesta utilizados por los estudiantes de tercero de secundaria ante los ítems de la prueba y, (c) la evaluación del diseño del interfaz de los ítems.

En la Figura 11 se muestra un contraste entre el modelo del proceso de respuesta del ítem dos de la V-ÍOM definido por los expertos y por el proceso de respuesta utilizado por el participante uno con el fin de verificar sus similitudes. A la izquierda de la figura se puede ver una sección del reporte verbal del participante uno ante el ítem mencionado y, al lado derecho, el modelo del proceso de respuesta elaborado por los expertos. Nótese que para la mayoría de los pasos previamente establecidos por los expertos se encuentra evidencia en el reporte verbal sobre su uso por parte del examinado. Sin embargo, el paso (d), relacionado con el proceso de *comprender el objetivo de la tarea evaluativa*, requirió de un análisis más profundo, donde se utilizó, a la par, información del *análisis de protocolo retrospectivo*.

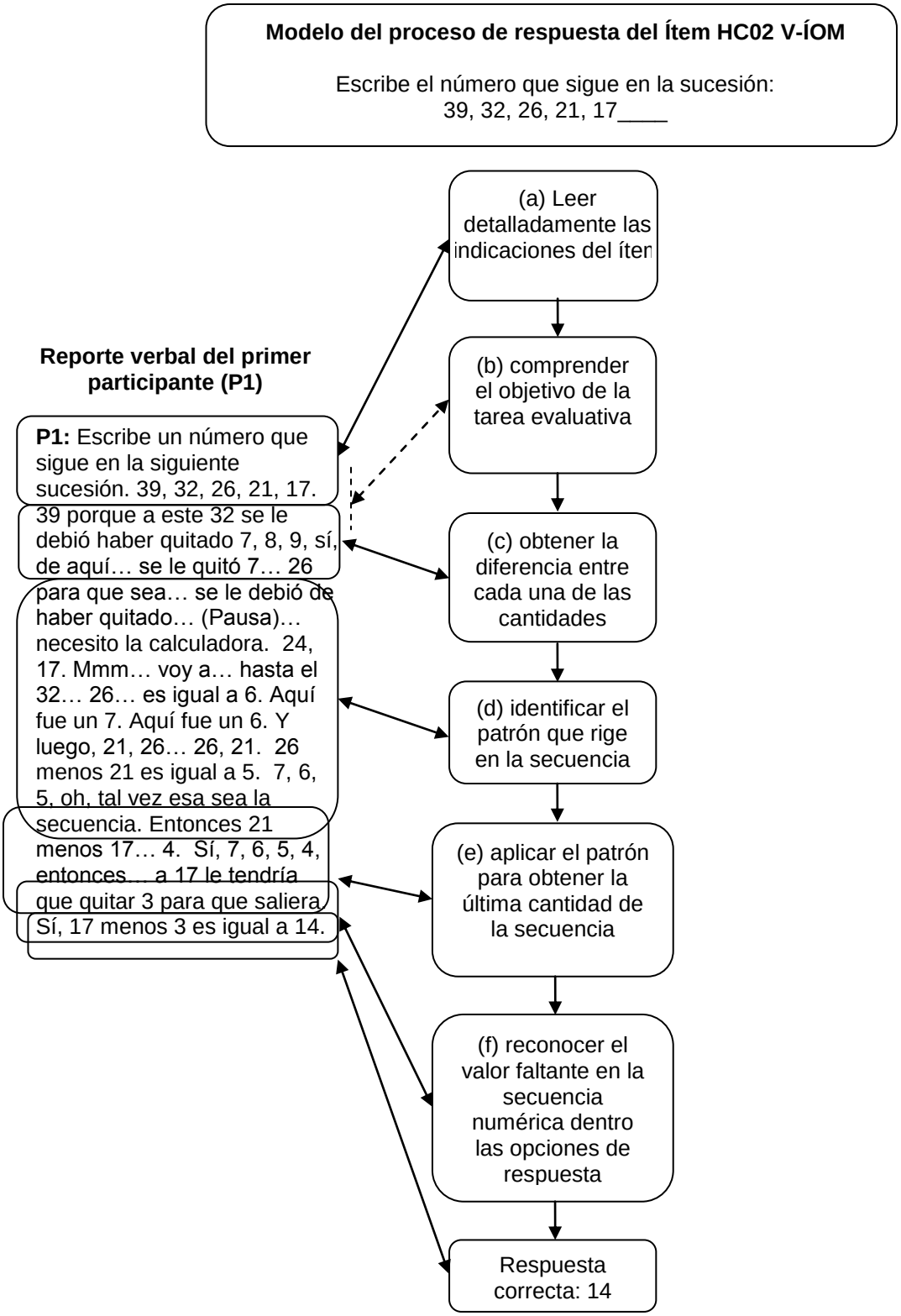


Figura 11. Verificación de las similitudes entre el modelo del proceso cognitivo definido por los expertos y por el proceso de respuesta de los examinados ante el ítem dos de la V-ÍOM del área HC del EXHCOBA

Dados los resultados de la verificación de las similitudes entre el modelo del proceso cognitivo definido por los expertos y por el proceso de respuesta de los examinados ante los ítems de la prueba, se analizaron una serie de problemas asociados a diferentes factores, tanto del interfaz como del proprio estudiante (ver Tabla 4.6). Es necesario recordar que para la evaluación del diseño del interfaz de los ítems de la prueba se analizaron diferentes características particulares (uso de elementos gráficos, botones y controles de ayuda, operaciones de selección y arrastre de elementos, y escritura numérica y algebraica). A modo de recordatorio, las áreas evaluadas del diseño del interfaz de los ítems analizados fueron: (a) problemas de comprensión y/o legibilidad de las indicaciones del ítem, (b) problemas de comprensión y/o legibilidad de la base del ítem, (c) problemas de usabilidad de las operaciones de respuesta del ítem, (d) problemas con la estructura y/o formato del ítem y (e) problemas con la comprensión de las opciones de respuesta del ítem.

En la Tabla 4.6 se muestran las frecuencias de los examinados que durante las *técnicas de pensamiento en voz alta* con *análisis de protocolos* realizaron su proceso de respuesta, tal como lo modelaron los expertos. También se presenta la cantidad de examinados que presentaron procesos diferentes a los prescritos en los modelos de respuesta elaborados por los expertos, y el tipo de problema asociado. Nótese que dentro de la V-ÍOM se identificaron 11 ítems de 30 con más de un problema en el diseño de su interfaz o con un problema de diseño, que presentaron tanto los examinados que obtuvieron su respuesta correcta como los que contestaron incorrectamente. Por su parte, dentro de la V-ÍRC se identificaron 8 ítems con más de un problema en el diseño de su interfaz.

Tabla 4.6. Problemas presentados en los procesos de respuesta de los examinados

No. Ítem	N=8 V-ÍMO				No. Ítem	N=8 V-ÍRC			
	PR correctos	PP	Respuesta incorrecta	PP		PR correctos	PP	Respuesta incorrecta	PP
1*	6	4B y 3C	2	2B y 2C	**1****	5	-	3	1A
2**	6	-	2	1A	**2***	3	3D	5	1A y 5D
3*	3	3C	5	3A y 5C	**3***	4	2D	4	1A, 2D
4	6	-	2	1B	**4**	2	-	6	1A y 2B
5	4	-	4	1B	**5***	5	2D	3	3D
6*	7	3C	1	1A y 1C	**6**	3	-	6	2A y 1B
7*	5	4C	3	1A, 3C y 1B	**7***	5	4D y 1C	3	2D
8*	3	2C	5	2A y 5C	**8****	3	-	5	-
9**	3	-	5	2A	**9***	3	1C	5	4C
10**	6	1C	2	-	**10****	4	-	4	-
11*	5	1F	3	1A y 1F	**11****	2	-	6	-
12	5	-	3	1A y 1C	**12***	3	1C	5	1A y 3C
13	3	-	5	1A y 1C	**13***	4	1C	4	2A y 3C
14	6	-	2	1A y 1C	**14****	6	-	2	-
15**	5	-	3	-	**15****	4	-	4	1A
16**	7	-	1	-	**16****	6	-	2	-
17**	4	-	4	1A	**17**	3	-	5	1A y 1C
18**	3	-	5	1A	**18**	5	-	3	1A y 1B
19*	2	2C	6	2A y 5C	**19***	3	2C	5	1A, 1B, 3C y 2E
20*	2	2C	6	2A, 1B y 5C	**20**	5	-	3	1C
21*	4	1C	4	1A y 2C					
22*	4	1F	4	2C y 1F					
23**	5	-	3	-					
24	5	-	3	1C					
25	5	1C	3	-					
26*	2	1C	6	2A, 3B y 4C					
27	5	-	3	1C					
28**	3	-	5	2A					
29**	7	-	1	1A					
30**	5	-	3	1A					

(*) Ítems con más de un problema en el diseño de su interfaz e ítems con un problema de diseño, el cual presentaron tanto los examinados que obtuvieron su respuesta correcta como los que contestaron incorrectamente.

(**) Ítems sin problemas significativos en su diseño de interfaz.

Problemas Presentados (PP) durante el Proceso de Respuesta (PR):	**Frecuencia**
A. Problemas de respuesta por adivinación o azar	40
B. Problemas de comprensión y/o legibilidad de las indicaciones del ítem	18
C. Problemas de comprensión y/o legibilidad de la base del ítem	86
D. Problemas de usabilidad de las operaciones de respuesta del ítem	23
E. Problemas con la estructura y/o formato del ítem	2
F. Problemas con la comprensión de las opciones de respuesta del ítem	4
Total	173

En cuanto a los tipos de problemas asociados a los procesos de respuesta de los estudiantes de tercero de secundaria ante los ítems de la prueba, se puede decir que con la ayuda de las *técnicas de pensamiento en voz alta* con *análisis de protocolos* se encontró una variedad de ellos. En resumen, los tipos de problema con mayor frecuencia encontrados en el interfaz de los ítems de la V-ÍOM fueron los relacionados con la *Comprensión y/o legibilidad de la base del ítem* y los relacionados con *Respuestas al azar o por adivinación* con una frecuencia de 67 (60%) y 27 (24%), respectivamente. Asimismo, los tipos de problemas en el procesos de respuesta de los examinados con menor frecuencia fueron los relacionados con la *Comprensión y/o legibilidad de las indicaciones del ítem* y con la *Comprensión de las opciones de respuesta*, siendo la frecuencia de 13 (12%) y 4 (4%), respectivamente.

Por su parte, los tipos de problema con mayor frecuencia presentes en los procesos de respuesta de los examinados ante los ítems de la V-ÍRC fueron los relacionados con la *Usabilidad de las operaciones de respuesta del ítem* y con la *Comprensión y/o legibilidad de la base del ítem*, siendo la frecuencia de 23 (36%) y 21 (33%), respectivamente. En contra parte, los tipos de problemas en el procesos de respuesta de los examinados con menor frecuencia fueron los relacionados con *Respuestas al azar o por adivinación* por parte de los examinados, donde la frecuencia fue de 13 (20%); los relacionados con la *Comprensión y/o legibilidad de las indicaciones del ítem*, donde la frecuencia fue de 5 (8%); y los relacionados con la *Estructura y/o formato del ítem*, donde ocurrieron 2 (3%) frecuencias.

4.2. Resultados del análisis psicométrico básico y de unidimensionalidad de la prueba

4.2.1. Análisis psicométrico básico aplicado al área de HC del EXHCOBA

Para los resultados del análisis psicométrico básico se presentan en este apartado los resultados de la aplicación del modelo de la TCT. De forma puntual, para los resultados de la calibración de los ítems de las dos versiones analizadas se presenta el total de aciertos por ítem, el índice de dificultad, el índice de discriminación bajos-altos, y el coeficiente de correlación puntual-biserial (*Rpbis*). Más adelante, se presenta la descripción del coeficiente de consistencia interna de las dos versiones de la prueba, así como una síntesis de los indicadores de calidad técnica aquí mencionados.

Al final de la primera columna en el último renglón de la Tabla 4.7 se muestra el promedio del índice de dificultad de los ítems que integran las dos versiones de la prueba. Dicho promedio fue de 0.54 para la V-ÍOM y 0.36 para la V-ÍRC. La proporción de aciertos de los ítems de V-ÍOM oscila entre 0.27 (alta dificultad) y 0.87 (baja dificultad). Por su parte, la proporción de aciertos de los ítems de V-ÍRC oscila entre 0.07 (elevada dificultad) y 0.64 (media dificultad). Concretamente, dos ítems (11 y 12) de la V-ÍRC resultaron ser elevadamente difíciles de responder por lo que no cumplen con el estándar P > 0.05 y <0.95 (ver Tabla 4.8).

Ahora bien, el promedio del *índice de discriminación* (altos-bajos) fue de 0.54 para la V-ÍOM y 0.47 para la V-ÍRC, respectivamente. Sin embargo, tres de los ítems (7, 11 y 12) de la V-ÍRC presentaron un *índice de discriminación* (altos-bajos) por

debajo del estándar establecido (***D***=>0.2). Cabe señalar, que debido a que la aplicación de los modelos componenciales LLTM y LSDM requiere de un buen ajuste tanto al modelo de la TCT como al modelo de RASCH, los ítems que no cumplen los *estándares* básicos se descartaron para los subsecuentes procedimientos relacionados con el análisis de la estructura del modelo cognitivo. En lo que respecta al promedio del coeficiente de correlación puntual-biserial (*Rpbis*), el valor promedio fue de 0.47 para la V-ÍOM y de 0.43 para la V-ÍRC. Los coeficientes obtenidos para las dos versiones es adecuado según el estándar técnico de calidad establecido (*Rpbis* => 0.20).

Tabla 4.7. Resultado del análisis con TCT de la V-ÍOM y la V-ÍRC del área de HC del EXHCOBA

No. ítem	EXHCOBA V-ÍOM K=30			No. ítem	EXHCOBA V-ÍRC K=20		
	p	D	*Rpbis*		p	D	*Rpbis*
1	0.62	0.45	0.38	1	0.63	0.41	0.35
2	0.76	0.46	0.44	2	0.21	0.44	0.46
3	0.44	0.43	0.36	3	0.53	0.56	0.46
4	0.73	0.46	0.43	4	0.33	0.47	0.42
5	0.43	0.58	0.49	5	0.55	0.62	0.50
6	0.87	0.22	0.27	6	0.23	0.50	0.52
7	0.50	0.47	0.39	7	0.41	0.19*	0.21
8	0.45	0.49	0.42	8	0.25	0.48	0.47
9	0.46	0.70	0.59	9	0.24	0.34	0.34
10	0.75	0.46	0.43	10	0.27	0.58	0.54
11	0.50	0.62	0.52	11	0.07*	0.12*	0.25
12	0.67	0.73	0.62	12	0.07*	0.13*	0.26
13	0.49	0.76	0.61	13	0.45	0.40	0.36
14	0.55	0.77	0.62	14	0.49	0.69	0.54
15	0.66	0.70	0.58	15	0.32	0.66	0.61
16	0.71	0.47	0.42	16	0.46	0.73	0.58
17	0.34	0.60	0.52	17	0.38	0.48	0.41
18	0.39	0.47	0.41	18	0.64	0.50	0.43
19	0.27	0.50	0.48	19	0.33	0.42	0.38
20	0.29	0.50	0.47	20	0.43	0.69	0.57
21	0.54	0.61	0.50		-	-	-
22	0.47	0.49	0.40		-	-	-
23	0.60	0.58	0.48		-	-	-
24	0.59	0.56	0.46		-	-	-
25	0.57	0.63	0.53		-	-	-
26	0.34	0.34	0.31		-	-	-
27	0.60	0.77	0.62		-	-	-
28	0.38	0.55	0.48		-	-	-
29	0.84	0.36	0.40		-	-	-
30	0.51	0.61	0.50		-	-	-
Promedios:	**0.54**	**0.54**	**0.47**		**0.36**	**0.47**	**0.43**

TA: Total de aciertos por ítem
D: Índice de discriminación
***** No satisfacen los *estándares* de calidad
p: Índice de dificultad
***Rpbis*:** Coeficiente de correlación puntual-biserial

En la Tabla 4.8, se puede observar una síntesis de los indicadores de calidad técnica del examen hasta ahora descritos, además del coeficiente de confiabilidad de las dos versiones de la prueba analizadas. Los estadígrafos utilizados para obtener dicho coeficiente fueron el índice de consistencia interna **α** de Cronbach y el KR-21 de Kuder Richardson. El **α** de confiabilidad de la V-ÍOM fue de 0.880 y de la V-ÍRC fue de 0.776. En el caso del **α** de confiabilidad de la V-ÍRC se puede observar que se encuentra por debajo del estándar establecido (**α** => 0.85). Es necesario aclarar que los problemas de consistencia interna de la prueba están asociados a los problemas que presentan los índices de discriminación de los ítems. Así que, una vez corregidos dichos problemas en un proceso de mejora continua es probable que incremente la confiabilidad de las dos versiones sin necesidad de agregar más ítems a la prueba.

Tabla 4.8. Estándares de calidad técnica e indicadores psicométricos de la V-ÍOM y V-ÍRC del área de HC del EXHCOBA

Estándares de calidad	Indicadores psicométricos	Observaciones
p > 0.05 y <0.95	$\bar{x}$ de p= 0.54 para V-ÍOM y 0.36 para V-ÍRC.	-El promedio del índice de dificultad de la V-ÍRC presenta mayor nivel de dificultad; sin embargo, es adecuado para un examen de selección. Sólo dos ítems (11 y 12) resultaron altamente difíciles.
D => 0.2	$\bar{x}$ de D= 0.54 para V-ÍOM y 0.47 para V-ÍRC.	-El promedio del *índice de discriminación* (altos-bajos) es adecuado; sin embargo, en los ítems 7, 11 y 12 de la V-ÍRC fue positiva, pero baja.
Rpbis => 0.20	$\bar{x}$ de *Rpbis* = 0.47 para V-ÍOM y 0.43 para V-ÍRC.	-El promedio de la *Rpbis* es adecuado y de igual forma en cada uno de los ítems de las dos versiones de la prueba.
α => 0.85	El índice α = 0.880 para V-ÍOM y 0.776 para V-ÍRC.	-El coeficiente de confiabilidad α de Cronbach de la V-ÍOM es adecuado; sin embargo, en la V-ÍRC se encuentra ligeramente por debajo del criterio.

Como ya se mencionó en párrafos anteriores, para los subsecuentes análisis se descartaron los ítems 7, 11 y 12 de la V-ÍRC. La principal razón de dicha decisión es mejorar el ajuste de la prueba, el cual es un requisito previo para aplicar los modelos componenciales LLTM y LSDM. Cabe señalar que, una vez descartados los ítems que presentaron baja calidad técnica, el **α** de confiabilidad de la V-ÍRC mejoró ligeramente (α = 0.786), acercándose al estándar establecido.

4.2.2. Análisis de unidimensionalidad y del ajuste entre el modelo RASCH unidimensional y el LLTM

En cuanto a la dimensionalidad de la V-ÍOM y V-ÍRC, se puede observar en la Tabla 4.9 los resultados de la aplicación del modelo de AFC de Fraser (1988). Obsérvese que las cargas factoriales son todas positivas, y grandes para los ítems de las dos versiones. Los valores más bajos corresponden al ítem 6 de la V-ÍOM y a los ítems 1 y 13 de la V-ÍRC del EXHCOBA. Tomando en cuenta todos los valores obtenidos, se puede decir que los ítems analizados de las dos versiones contribuyen adecuadamente a la medida del rasgo latente de la prueba. Además, se obtuvo un Índice de Tanaka de bondad de ajuste (Tanaka & Huba, 1985) de 0.993 con un valor RMSR=0.01 para la versión con 30 ítems de opción múltiple y para la versión con ítems de respuesta compleja reducida a 17 ítems se obtuvo un índice de 0.986 con un RMSR=0.01, lo cual confirma que los ítems de las dos versiones miden una sola dimensión, respectivamente. En relación a los valores obtenidos en el índice de Tanaka, algunos investigadores como Gierl, Tan y Wang (2005) mencionan que, aunque no hay normas de interpretación generalmente aceptadas para dicho índice,

se ha propuesto que valores arriba de 0,9 indican un buen ajuste (Rouse, Finger & Butcher, 1999), cumpliéndose así el supuesto de unidimensionalidad.

Tabla 4.9. Cargas factoriales de los ítems de la V-ÍOM y la V-ÍRC del área de HC del EXHCOBA

No. ítem	Cargas factoriales	
	V-ÍOM	V-ÍRC
1	0.455	0.394
2	0.592	0.844
3	0.455	0.572
4	0.566	0.563
5	0.664	0.635
6	0.398	0.915
7	0.483	*d*
8	0.542	0.745
9	0.800	0.542
10	0.580	0.875
11	0.676	*d*
12	0.861	*d*
13	0.837	0.387
14	0.831	0.754
15	0.774	0.947
16	0.537	0.787
17	0.784	0.494
18	0.545	0.523
19	0.846	0.452
20	0.758	0.785
21	0.645	
22	0.506	
23	0.616	
24	0.581	
25	0.683	Índice de Tanaka: 0.9933805 V-ÍOM 0.9862326 V-ÍRC
26	0.421	
27	0.824	
28	0.684	
29	0.598	
30	0.651	

(*d*) Ítems descartados para análisis subsecuentes debido a su baja calidad técnica y bajo ajuste.

En cuanto a la aplicación del modelo unidimensional de RASCH, los rangos de los parámetros de dificultad de cada uno de los ítems de la V-ÍOM van desde -2.021 (fácil) hasta 1.550 (difícil), mientras que para la V-ÍRC van desde -1.461 (fácil) hasta 2.246 (difícil) (ver Tabla 4.10). Tomando en cuenta el criterio de calidad establecido de -1.8 a 1.8 para los valores **δi** de dificultad de los ítems de la prueba, se encontró que el ítem 6 de la V-ÍOM y los ítems 11 y 12 de la V-ÍRC no cumplen con dicho criterio. En el ajuste gráfico de los ítems de la prueba al modelo unidimensional de RASCH (véase Figura 12 y 13) se pueden observar las Curvas Características de los Ítems (CCI) que presentan bajo ajuste al modelo.

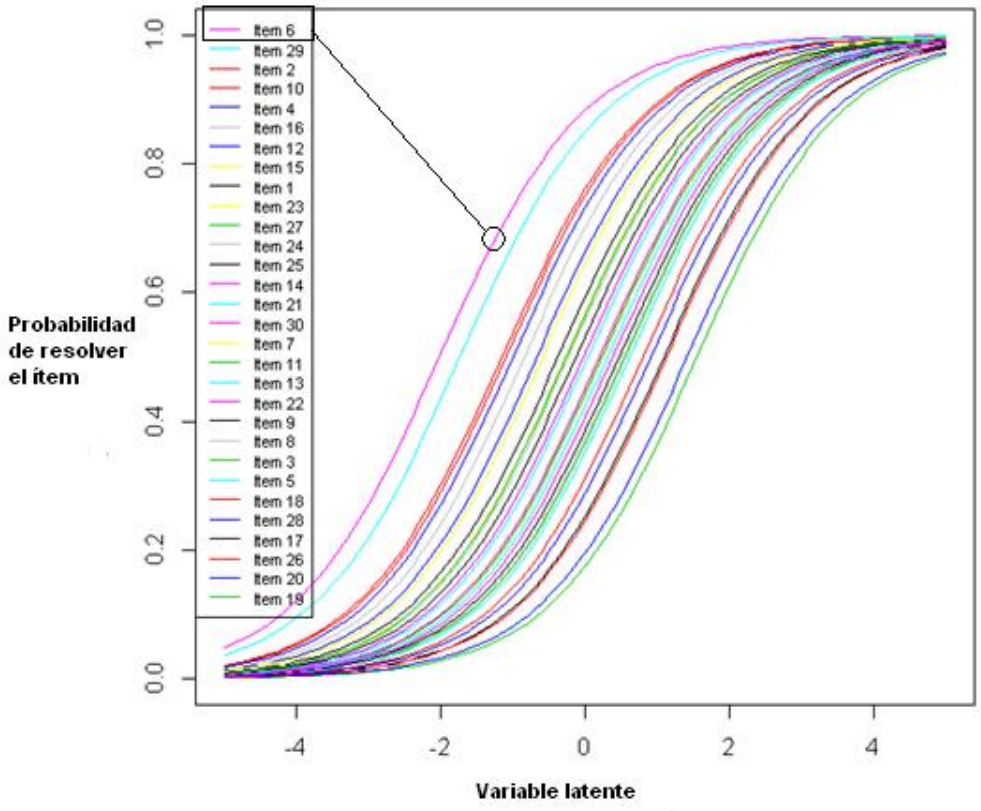


Figura 12. Curvas características de los ítems de la V-ÍOM del área de HC del EXHCOBA

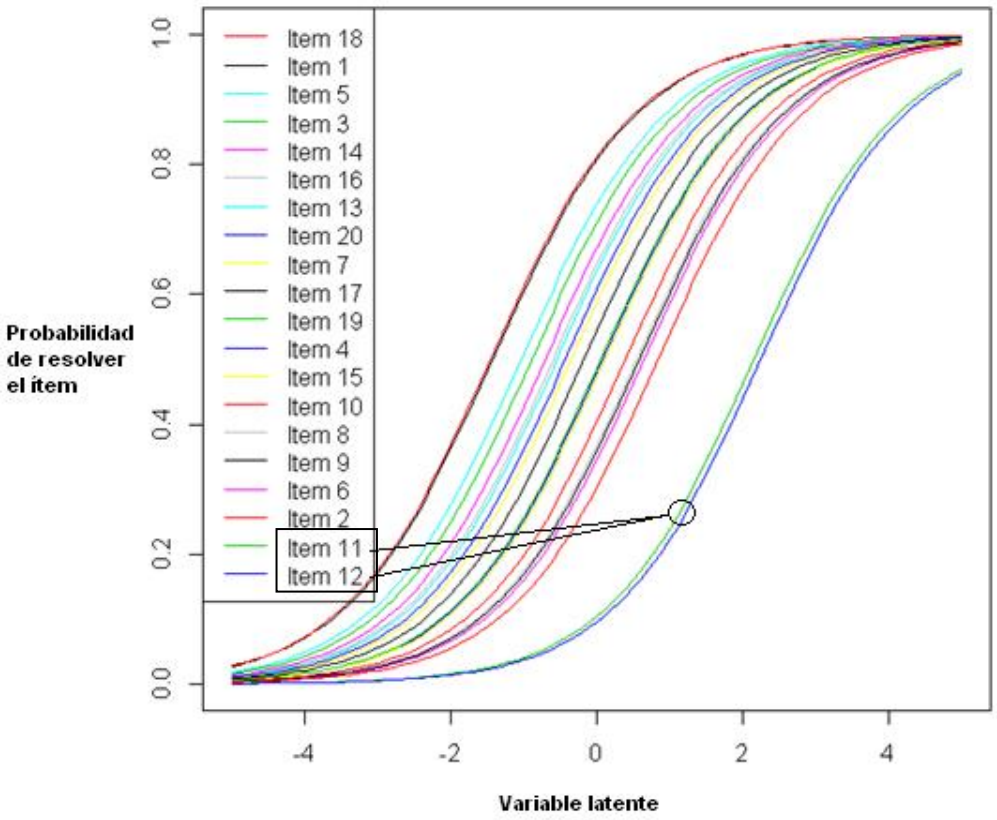


Figura 13. Curvas características de los ítems de la V-ÍRC del área de HC del EXHCOBA

Por su parte, el estadístico de la razón de verosimilitudes condicional CLR (Fischer y Ponocny-Seliger, 1998) resultó significativo (gl=29; X^2=827.367) para la V-ÍOM y, de igual forma, para la V-ÍRC (gl=19; X^2=130.319); por lo tanto, no se confirma el ajuste (ver Figura 14 y 15). Esto no es sorprendente, dado el rigor divulgado de la prueba CLR en la literatura (Fischer, 1995, p. 147). En la Tabla 4.9 se detalla, para cada uno de los ítems de las dos versiones analizadas, la estimación de los parámetros del modelo de RASCH, el error estándar de esta estimación y el valor Anderson-Darling X^2.

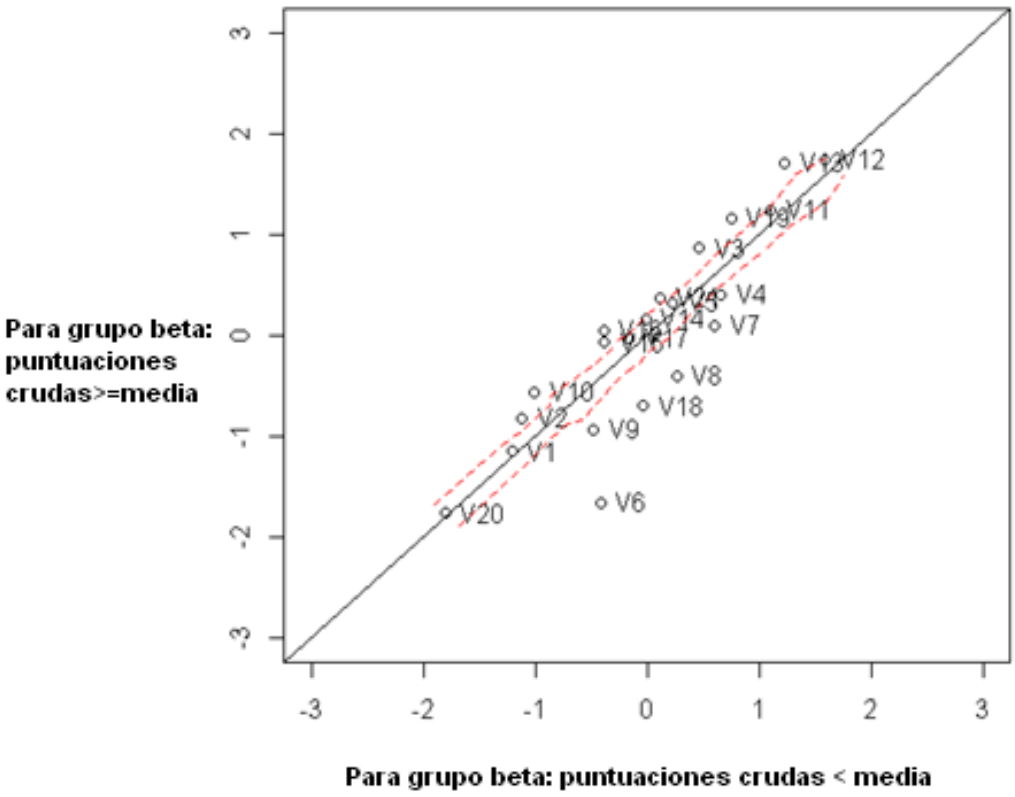


Figura 14. Prueba de ajuste gráfico del modelo RASCH de la V-ÍOM del área de HC del EXHCOBA

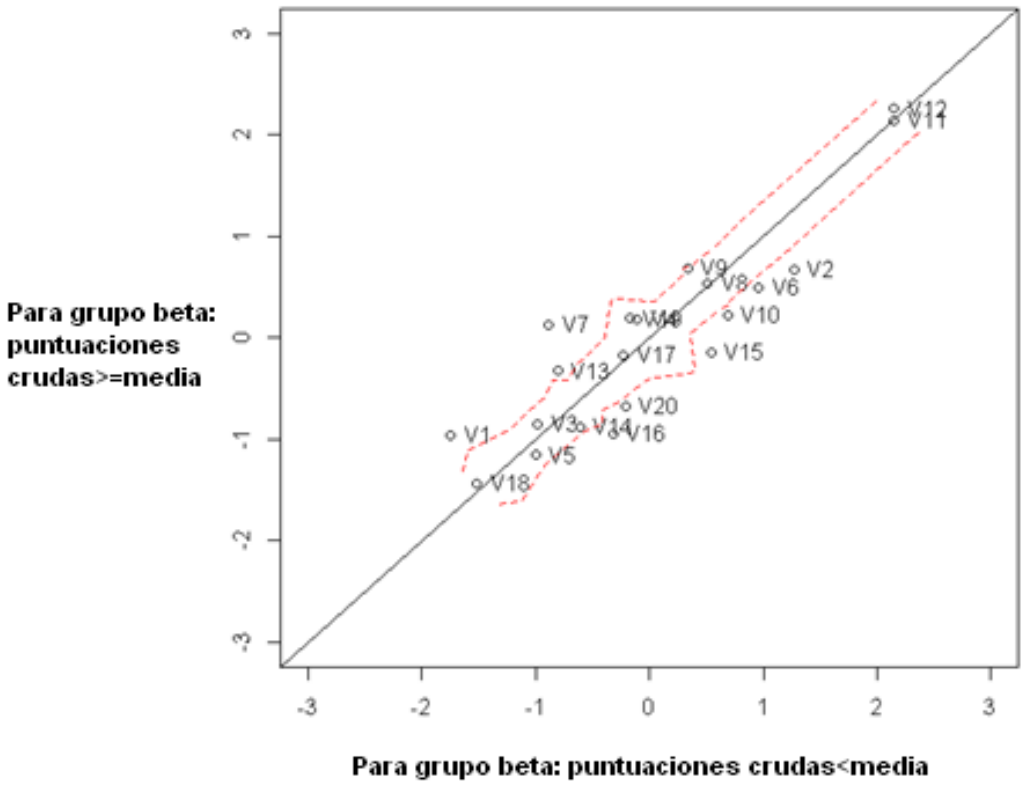


Figura 15. Prueba de ajuste gráfico del modelo RASCH de la V-ÍRC del área de HC del EXHCOBA

Tabla 4.10. Estimación de los parámetros del modelo de RASCH y el valor Anderson-Darling $\mathcal{X}^2$ de la V-ÍOM y la V-ÍRC del área de HC del EXHCOBA

No. ítem	Estimaciones de la V-ÍOM				No. ítem	Estimaciones de la V-ÍRC			
	δi	SE(δi)	$\mathcal{X}^2$	$p(\mathcal{X}^2)$		δi	SE(δi)	$\mathcal{X}^2$	$p(\mathcal{X}^2)$
1	-0.375	0.043	3208.126	0.000*	1	-1.432	0.085	788.675	0.007*
2	-1.156	0.047	2679.194	0.927	2	0.836	0.098	552.991	1.000
3	0.550	0.043	3494.702	0.000*	3	-0.898	0.082	689.667	0.529
4	-1.006	0.046	2741.596	0.727	4	0.078	0.086	740.793	0.101
5	0.613	0.043	2874.761	0.121	5	-1.038	0.083	643.612	0.910
6	-2.021	0.058	3142.886	0.000*	6	0.645	0.094	528.404	1.000
7	0.225	0.042	3350.812	0.000*	7	-0.348	0.083	1021.501	0.000*
8	0.524	0.043	3108.573	0.000*	8	0.545	0.092	573.561	1.000
9	0.470	0.042	2323.536	1.000	9	0.581	0.093	759.771	0.039
10	-1.101	0.047	3135.814	0.000*	10	0.390	0.09	542.695	1.000
11	0.234	0.042	2635.049	0.981	11	2.153	0.145	963.522	0.000*
12	-0.674	0.044	1834.346	1.000	12	2.246	0.150	750.321	0.065
13	0.311	0.042	2220.251	1.000	13	-0.551	0.082	851.181	0.000*
14	-0.023	0.042	2255.841	1.000	14	-0.711	0.082	615.544	0.984
15	-0.593	0.044	2091.906	1.000	15	0.117	0.086	496.275	1.000
16	-0.850	0.045	2791.835	0.471	16	-0.600	0.082	573.989	1.000
17	1.085	0.045	2610.022	0.992	17	-0.175	0.084	788.061	0.007*
18	0.798	0.043	3190.346	0.000*	18	-1.461	0.085	701.596	0.402
19	1.550	0.048	2650.447	0.968	19	0.055	0.086	743.114	0.091
20	1.414	0.047	2797.163	0.442	20	-0.432	0.083	548.674	1.000
21	0.038	0.042	2795.346	0.452	-	-	-	-	-
22	0.383	0.042	3363.689	0.000*	-	-	-	-	-
23	-0.271	0.043	2714.154	0.835	-	-	-	-	-
24	-0.228	0.043	2843.678	0.223	-	-	-	-	-
25	-0.126	0.042	2541.894	1.000	-	-	-	-	-
26	1.124	0.045	3753.892	0.000*	-	-	-	-	-
27	-0.262	0.043	2030.827	1.000	-	-	-	-	-
28	0.903	0.044	2851.752	0.192	-	-	-	-	-
29	-1.736	0.053	2370.339	1.000	-	-	-	-	-
30	0.199	0.042	2638.296	0.978	-	-	-	-	-

(*) Ítems que presentan bajo ajuste (p<.01)

Sin descartar del análisis ninguno de los ítems que no aportan al buen ajuste del modelo según los valores obtenidos en la prueba Anderson-Darling X^2, la correlación entre los parámetros del modelo de RASCH y LLTM de las dos versiones analizadas es de 0.561 para el caso de la V-ÍOM y 0.870 para la V-ÍRC. Asimismo, tomando en cuenta las operaciones cognitivas propuestas por el panel de expertos, se explica un 32% de las dificultades de los ítems de la V-ÍOM y un 76% de las dificultades de los ítems de la V-ÍRC. Lo anterior, indica un ajuste bajo entre los modelos RASCH y LLTM de la V-ÍOM y alto para la V-ÍRC.

Sin embargo, como ya se mencionó en el capítulo de método, el buen ajuste de los ítems de la prueba al modelo de RASCH es un requisito previo para la aplicación de los modelos componenciales anidados en la TRI, como son el LLTM y el LSDM. Para cumplir con dicho requisito, se descartaron los ítems de las dos versiones analizadas que obtuvieron valores de $p(X^2)$ significativos (p<.01) en la prueba de ajuste Anderson-Darling X^2. Así que, para la V-ÍOM se descartaron los ítems 1, 3, 6, 7, 8, 10, 18, 22 y 26 y para la V-ÍRC se descartaron los ítems 1, 7, 11, 12 (al no cumplir con el criterio -1.8 a 1.8 del parámetro de dificultad en el modelo de RASCH), así como los ítems 13 y 17. Con ello se logró un mejor ajuste de las dos versiones de la prueba al modelo de RASCH, obteniendo un estadístico de la razón de verosimilitudes condicional CLR de 399.284 con 20 grados de libertad para la V-ÍOM, reducida a 21 ítems, y de 69.141 con 13 grados de libertad para la V-ÍRC, reducida a 14 ítems.

Dada la relativa mejora del ajuste al modelo de RASCH de las dos versiones de la prueba, se realizó un análisis de correlación entre los modelos RASCH con el LLTM.

Para dicha prueba, se descartaron las operaciones cognitivas de cada matriz Q de las dos versiones de la prueba que ya no tenían ítems que representar. También se decidió eliminar la $\mathbf{O_{12}}$ de la V-ÍOM relacionada con la *Suma de fracciones,* la cual sólo está designada para el ítem 14. Dicha decisión se tomó así porque para responder al ítem 14 es suficiente dominar la *fracción y visualización de figuras geométricas*, la *aplicación de operaciones aritméticas básicas* y la *aplicación de fracciones*. La matriz Q de la V-ÍOM se configuró en 21i X11k, mientras que la matriz Q de la V-ÍRC en 14i X 6k (ver Tablas 4.11 y 4.12).

Tabla 4.11. Matriz Q 21i X11k de la V-ÍOM y proporción de aciertos media por subconjunto de ítems de k

No. ítem	$\mathbf{O_1}$	$\mathbf{O_2}$	$\mathbf{O_3}$	$\mathbf{O_4}$	$\mathbf{O_5}$	$\mathbf{O_6}$	$\mathbf{O_7}$	$\mathbf{O_9}$	$\mathbf{O_{11}}$	$\mathbf{O_{13}}$	$\mathbf{O_{14}}$	Total
02	0	0	0	1	0	0	1	0	0	0	0	2
04	1	0	1	0	0	0	0	0	0	0	0	2
05	1	0	0	0	0	0	0	0	0	0	0	1
09	0	0	0	1	0	1	0	0	0	0	0	2
11	0	0	0	0	0	1	0	0	0	0	0	1
12	0	1	0	0	0	1	0	0	0	0	0	2
13	0	1	0	1	0	1	0	0	0	0	0	3
14	0	1	0	1	0	1	0	0	0	0	0	3
15	0	0	0	1	0	0	0	0	0	0	0	1
16	0	0	0	1	0	0	0	0	0	0	0	1
17	0	0	0	1	0	0	0	0	0	0	0	1
19	0	0	0	1	0	0	0	1	0	0	0	2
20	0	0	1	1	0	0	0	1	0	0	0	3
21	0	1	0	1	0	0	0	1	0	0	0	3
23	0	0	1	1	0	0	0	0	0	0	1	3
24	0	0	1	1	0	0	0	0	1	0	0	3
25	0	0	1	1	0	1	0	0	1	0	0	4
27	0	0	0	0	1	0	0	0	0	0	0	1
28	0	0	1	1	0	0	0	0	0	1	0	3
29	0	0	1	1	0	0	0	0	0	1	0	3
30	0	0	1	1	0	0	0	0	0	0	0	2
$\bar{x}$ **de** p_i	0.58	0.56	0.56	0.56	0.60	0.54	0.76	0.37	0.58	0.61	0.60	

Tabla 4.12. Matriz Q 14*i* X 6*k* de la V-ÍRC y proporción de aciertos media por subconjunto de ítems de *k*

No. Ítem	O_1	O_2	O_3	O_6	O_7	O_9	Total
02	0	1	0	0	0	0	1
03	0	1	0	0	0	0	1
04	0	0	1	1	1	0	3
05	0	0	0	0	1	0	1
06	0	0	1	1	1	0	3
08	1	0	1	1	0	1	4
09	1	0	1	1	0	1	4
10	1	0	1	1	0	1	4
14	0	0	1	1	1	0	3
15	0	0	1	1	1	0	3
16	0	0	1	1	1	0	3
18	0	0	1	1	1	0	3
19	0	1	0	1	0	0	2
20	0	0	1	1	1	0	3
$\bar{x}$ **de** p_i	0.25	0.36	0.39	0.40	0.43	0.25	

Por su parte, en la Tabla 4.11 se puede observar que el promedio de la proporción de respuestas correctas ($\bar{x}$ **de** p_i) más alto (dificultad baja) del subconjunto de ítems de la V-ÍOM son los relacionados con la *Identificación de patrones de secuencias numéricas* (**O7**). La proporción media de aciertos de dicho subconjunto de ítems es de 0.76. Por su parte, los subconjuntos de ítems con dificultad media ordenados de más fácil a más difícil son los asociados a las operaciones **O7**, **O13**, **O5**, **O14**, **O1**, **O11**, **O2**, **O3**, **O4** y **O6**. Por otro lado, el subconjunto de ítems asociados a la operación cognitiva de *Representación del modelo matemático del perímetro, el área y el volumen* (**O9**) tiene una proporción de aciertos de 0.37, resultando el más difícil. Este ordenamiento no se esperaba de acuerdo al modelo cognitivo propuesto por el panel de expertos (ver Tabla 4.1). Se esperaba que **O11**, **O13** y **O14** presentaran mayor

dificultad al contener procesos cognitivos más complejos en comparación a **O1**, **O2**, **O3**, **O4** y **O6**.

De igual forma, se puede observar en la Tabla 4.12 que el promedio de la proporción de respuestas correctas ($\bar{x}$ **de** p_i) más alto (dificultad media) del subconjunto de ítems de la V-ÍRC son los relacionados con el *Representación de modelos matemático-aritméticos* (**O7**). La proporción media de aciertos de dicho subconjunto de ítems es de 0.43, seguido del subconjunto de ítems asociados a la *Interpretación de la información del problema* (**O6**) con una proporción media de 0.40. De igual forma, los subconjuntos de ítems relacionados respectivamente a **O1, O2** y **O3** presentan una proporción media de aciertos entre 0.25 y 0.39. Por otro lado, el subconjunto de ítems asociados a la operación cognitiva de *representación de modelos matemático-geométricos* (**O9**) tiene una proporción de aciertos de 0.25, que junto al subconjunto de ítems asociados a la **O1**, resultan los más difíciles de dominar. Este ordenamiento tampoco se esperaba (ver Tabla 4.2), pues la **O7** y **O6** deberían presentar mayor dificultad debido a que, según los expertos, incluyen procesos cognitivos más complejos en comparación a **O1**, **O2** y **O3**. Lo anterior indica que para esta versión se subestimó la complejidad cognitiva de las operaciones subyacentes a algunos ítems de la prueba.

Después de configurar la matriz Q para cada una de las versiones se procedió al análisis del ajuste entre el modelo de RASCH y el LLTM. De forma específica, se aplicó una prueba de ajuste de correlación entre los parámetros de los modelos mencionados. De dicha prueba se obtuvo un valor de correlación de 0.829 para la V-ÍOM y 0.777 para la V-ÍRC (ver Figuras 16 y 17). Como se puede observar, aun con el

descarte de los ítems de las dos versiones que presentaron un bajo ajuste, es necesario mejorar el ajuste de dichos ítems al modelo de RASCH. Sin embargo, para los fines de la presente tesis se decidió que era suficiente con el ajuste obtenido para proseguir con los subsecuentes análisis planteados.

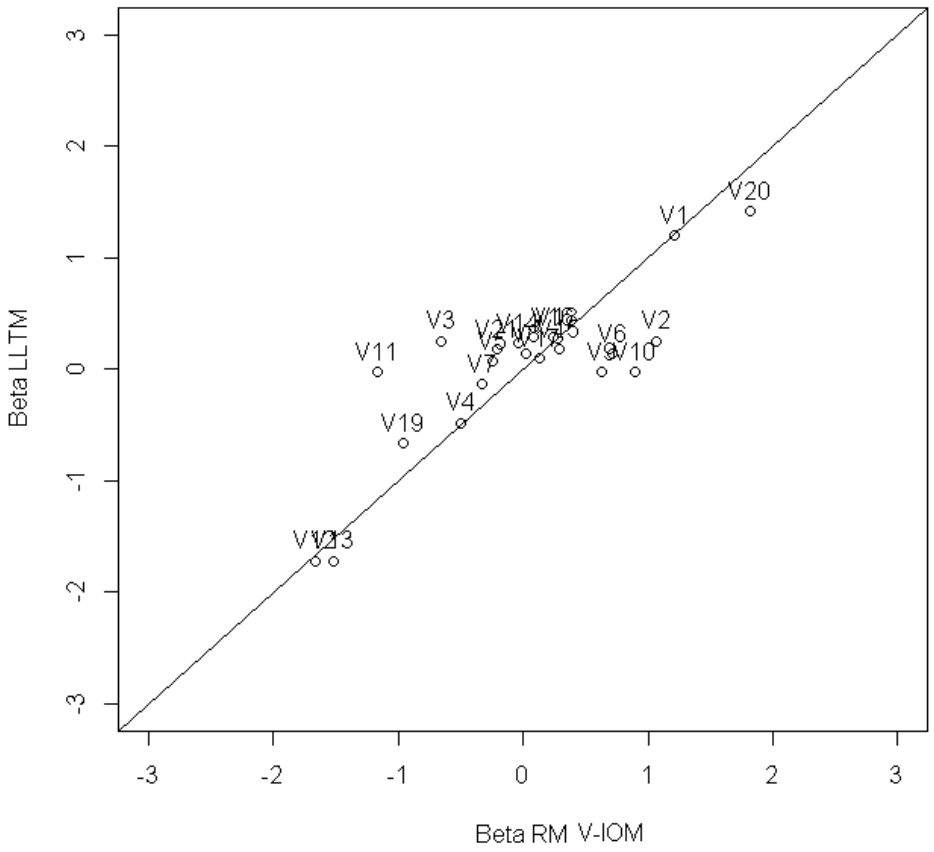


Figura 16. Prueba de ajuste gráfico del LLTM al modelo de RASCH de la V-ÍOM del área de HC del EXHCOBA

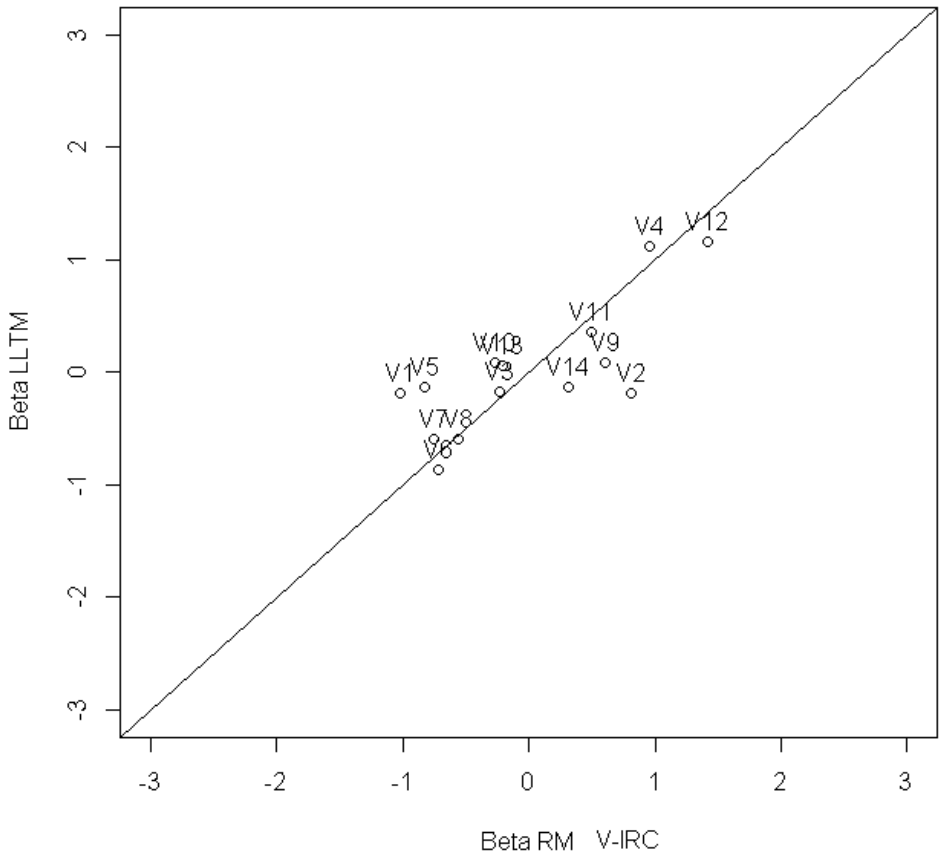


Figura 17. Prueba de ajuste gráfico del LLTM al modelo de RASCH de la V- ÍRC del área de HC del EXHCOBA

4.3. Resultados del análisis de las evidencias de validez basadas en la estructura del modelo cognitivo

En este apartado se presentan los resultados de la aplicación de los modelos LLTM y LSDM para obtener evidencias de validez, basadas en la estructura del modelo cognitivo del área de HC del EXHCOBA, en dos de sus versiones: una con ítems de opción múltiple y otra con ítems de respuesta compleja. Para ello, antes de presentar los resultados de la aplicación del LLTM, se muestra un análisis comparativo del ajuste

entre los modelos componenciales LLTM y DINA. Consecutivamente, se presentan los parámetros ***b*** recuperados por el LLTM correspondientes a cada ítem y su contraste con los parámetros del modelo RASCH; enseguida se muestran las estimaciones de los parámetros básicos del LLTM para cada operación cognitiva $(\boldsymbol{\alpha_k})$.

Después de la aplicación del modelo LLTM, se presentan los resultados obtenidos del análisis con el LSDM. Concretamente, el resultado de un proceso de reconfiguración de la matriz Q con base en la estimación de las Curvas de Probabilidades de los Atributos (CPA) y los valores MAD (en inglés *Mean Absolute Difference*) de cada una de las versiones analizadas. Asimismo, ya con la versión de la matriz Q resultante del proceso de reconfiguración se presentan los resultados del análisis de los residuales LSD (en inglés *Least Squares Distance*) y el comportamiento de las CPA, la adecuación de los vínculos entre ítems y operaciones cognitivas mediante el análisis de la recuperación de las CCI por medio del análisis de los valores MAD, MAD ponderado (WMAD, por sus siglas en inglés), la Raíz de la Diferencia Media al Cuadrado (RMSD, por sus siglas en inglés), y el trazo de límites superiores e inferiores a la curva. También, se presenta la validación cruzada de los resultados obtenidos por los modelos LLTM (Fischer, 1973) y LSDM (Dimitrov, 2007).

4.3.1. Aplicación del modelo componencial LLTM a los ítems del área de HC del EXHCOBA

Como ya se mencionó, antes de la aplicación del LLTM se presentan los resultados del análisis comparativo del ajuste entre los modelos componenciales LLTM y DINA. Es necesario recordar que desde el establecimiento del modelo de RASCH como

apropiado para los datos (ver apartado 4.2.2.), la preocupación principal era el hecho de haber elegido el modelo componencial con mejor ajuste. Para tal fin, se eligió como contraste el modelo no anidado *Deterministic Inputs, Noisy And Gate* (en inglés, DINA), propuesto inicialmente por Macready y Dayton (1977) y adaptado para propósitos de diagnóstico por Junker y Sijtsma (2001). Pero, el modelo de RASCH fue el de mejor ajuste para las dos versiones del área de HC analizadas. Dicho modelo presentó el valor de Criterio de Información de Akaike (CIA) más bajo (V-ÍOM: $-2\ell nL$=48384.32, CIA= 48424.32; V-ÍRC: $2\ell nL$=-3810.083, CIA=7646.166). El LLTM, con 11 estimaciones de las variables del modelo de la V-ÍOM y 6 de la V-ÍRC, fue el segundo con mejor ajuste (V-ÍOM: $-2\ell nL$=50802.24, CIA= 50824.24; V-ÍRC: $2\ell nL$=-3992.465, CIA=7996.930). Finalmente, el modelo DINA produjo el peor ajuste (V-ÍOM: $-2\ell nL$=65271.70, CIA= 69449.70; V-ÍRC: $2\ell nL$=-5580.146, CIA=11342.29).

Con lo que respecta a los resultados de la aplicación del modelo LLTM y a su contraste con los parámetros del modelo RASCH, en las Tablas 4.13 y 4.14 se pueden observar los 21 ítems de la V-ÍOM y 14 ítems de la V-ÍRC re-calibrados y, los parámetros ***b*** recuperados por el LLTM correspondientes a cada ítem. Con respecto a la dificultad (***b***) del LLTM de los ítems de la V-ÍOM, se puede observar que los ítems más fáciles (parámetros ***b*** negativos) han sido 13 en total. Entre algunos de los más fáciles se encuentran los ítems 29 (expresiones algebraicas y probabilidad), 2 (secuencias numéricas), 24 (regla de tres simple y análisis de proporciones) y 27 (suma de ángulos de un triángulo). Por su parte, el total de ítems más difíciles (parámetros ***b*** positivos) de dicha versión es 8, entre los que se encuentran los ítems

19 (cálculo del perímetro), 20 (cálculo del área), 28 (reglas básicas de probabilidad) y 9 (fracciones y decimales) (ver Tabla 4.13).

Por su parte, los valores ***b*** más bajos (fáciles) de la V-ÍRC le pertenecen a los ítems 5 (números racionales y fracciones de figuras) y 18 (análisis de datos en tablas), siendo estos los únicos de dicha versión que presentan parámetros ***b*** negativos. Los ítems de la V-ÍRC que presentan parámetros ***b*** positivos (difíciles) son 12 en total. Algunos de los ítems más difíciles según el LLTM de la V-ÍRC son el 8 (cálculo de perímetros), 9 (cálculo de áreas), 10 (cálculo de volúmenes), 2 (fracciones y recta numérica), 3 (Menor y mayor que), 4 (división de decimales), 20 (análisis de tablas de proporciones) y 19 (análisis de datos de graficas poligonales) (ver Tabla 4.14).

Tabla 4.13. Parámetros ***B*** y error típico calibrado con el modelo de RASCH y, parámetros ***b*** recuperados por el LLTM de la V-ÍOM

No. Ítem	Contenido	RASCH *B*	e(*B*)	LLTM *b*	e(*b*)
01	Adición y sustracción con apoyo en la recta numérica	*			
02	Obtención del valor faltante en secuencias numéricas	-1.216	0.048	-1.216	0.087
03	Construcción de una expresiones algebraica	*			
04	Identificación del valor posicional en números enteros	-1.060	0.047	-0.272	0.062
05	Identificación posicional de números decimales	0.655	0.044	-0.272	0.062
06	Juicio situacional de la representación de un número	*			
07	Aplicación de módulos en un contexto	*			
08	Representación de exponentes en un contexto	*			
09	Equivalencias de decimales a fracciones	0.501	0.044	0.463	0.056
10	Representación de la fracción en un contexto	*			
11	Identificación de la descripción de las partes de una fracción	0.248	0.043	-0.085	0.089
12	Identificación de una fracción en una figura	-0.713	0.045	-0.159	0.071
13	Identificación de una fracción en una figura para su conversión a decimal	0.331	0.043	0.116	0.090
14	Identificación de fracciones en una figura para su suma	-0.025	0.043	-0.159	0.071
15	Suma y resta de números decimales	-0.628	0.045	0.004	0.080
16	Multiplicación de números decimales	-0.897	0.046	0.004	0.080
17	División de números decimales	1.164	0.046	0.004	0.080
18	Aplicación de la regla de tres simple	*			
19	Cálculo de perímetros de círculos	1.667	0.049	1.702	0.081
20	Cálculo de áreas de las caras de un prisma	1.520	0.048	1.702	0.081
21	Relación de volúmenes	0.039	0.043	-0.249	0.080
22	Comparación de unidades entre unidades de medida distintas	*			
23	Aplicación de la regla de tres simple	-0.289	0.044	-0.194	0.076
24	Relaciones de proporcionalidad	-0.243	0.043	-0.305	0.068
25	Regla de tres simple con fracciones	-0.135	0.043	-0.118	0.091
26	Aplicación de la regla de tres inversa en un contexto	*			
27	Conocimiento de la suma de los ángulos internos de un triángulo	-0.279	0.044	-0.299	0.084
28	Probabilidad de un evento en un contexto	0.967	0.045	0.643	0.080
29	Adición de probabilidades para eventos	-1.816	0.054	-1.432	0.087
30	Cálculo de la media aritmética	0.211	0.043	-0.194	0.076

* ($p<0.01$) ítems descartados para la segunda calibración

Tabla 4.14. Parámetros ***B*** y error típico calibrado con el modelo de RASCH y, parámetros ***b*** recuperados por el LLTM de la V-ÍRC

No. Ítem	Contenido	RASCH *B*	e(*B*)	LLTM *b*	e(*b*)
01	Obtención del valor faltante en secuencias numéricas	*			
02	Ubicación de fracciones en la recta numérica	1.024	0.819	1.103	0.092
03	Ordenación de números decimales de menor y mayor en espacios vacíos	-0.809	0.819	1.103	0.092
04	División de números decimales	0.230	0.688	1.081	0.093
05	Representación de fracciones en figuras geométricas	-0.958	1.445	-0.210	0.075
06	Solución de suma y resta de fracciones en un contexto	0.824	0.172	1.041	0.085
07	Identificación de elementos de la circunferencia	*			
08	Cálculo de perímetros de círculos en un contexto	0.720	0.034	1.777	0.125
09	Cálculo de áreas de triángulos rectángulos	0.757	0.098	1.505	0.103
10	Cálculo de volúmenes de prismas rectangulares	0.558	0.098	1.505	0.103
11	Cálculo de equivalencias de unidades de volumen	*			
12	Cálculo de equivalencias de unidades de longitud	*			
13	Cálculo de distancias en mapas con uso de escalas	*			
14	Representación numérica de porcentajes menores que 100 en un contexto	-0.609	0.930	0.830	0.121
15	Representación numérica de porcentajes mayores que 100 en un contexto	0.270	0.930	0.830	0.121
16	Cálculo de la regla de tres simple en un contexto	-0.490	1.445	0.558	0.120
17	Cálculo de probabilidades expresada en fracciones	*			
18	Inferencia y cálculo de un valor dada una tabla de proporcionalidad directa	-1.412	2.133	-0.251	0.070
19	Inferencia de valores dada una gráfica poligonal en un contexto	0.206	0.304	0.852	0.132
20	Estimación de frecuencias de ocurrencias de eventos dada la probabilidad	-0.311	0.688	1.041	0.085

* ($p<0.01$) ítems descartados para la segunda calibración

En las Tablas 4.15 y 4.16 se pueden observar las estimaciones de los parámetros básicos de cada operación cognitiva ($\boldsymbol{\alpha_k}$), sus respectivos errores estándar, así como valores mínimos y máximos obtenidos de los ítems de las dos versiones estudiadas del área de HC de la prueba. En general, véase que cada uno de los parámetros básicos, de las dos versiones analizadas, son significativos, indicando que las operaciones correspondientes contribuyen a la dificultad de los ítems. En

resumen, son tres las operaciones que contribuyen a la dificultad de los ítems (parámetros básicos negativos) tanto para la V-ÍOM como para la V-ÍRC. En cuanto a las operaciones que contribuyen a la "facilidad" (parámetro básico positivo) se pueden contabilizar 8 operaciones para la V-ÍOM y 3 para la V-ÍRC.

Analizando a mayor profundidad, en los resultados que se muestran en la Tabla 4.15 se puede observar que las operaciones cognitivas que introducen mayor dificultad a los ítems de la V-ÍOM son: la *comprensión de problemas matemáticos contextualizados* (**O3**), la *aplicación de operaciones aritméticas básicas* (**O4**) y la *aplicación de fracciones* (**O6**). En lo que respecta a la V-ÍRC, las operaciones cognitivas que introducen mayor dificultad son: la *aplicación de operaciones aritméticas* (**O3**), la *representación de valores en sistemas posicionales* (**O2**) y la *representación de modelos matemáticos-geométricos* (**O9**) (ver Tabla 4.16).

Tabla 4.15. Parámetros básicos del LLTM de la V-ÍOM

Código	Operaciones	α_k	$SE\ (\alpha_k)$	Min.	Max.	*t*	*p*
O1	Comprensión del sistema decimal (dígitos)	0.272	0.062	4.387	0.151	0.393	(p<.01)
O2	Fracción y visualización de figuras geométricas	0.346	0.054	6.407	0.24	0.453	(p<.01)
O3	Comprensión de problemas matemáticos contextualizados	-1.605	0.069	-23.261	-1.74	-1.469	(p<.01)
O4	Aplicación de operaciones aritméticas básicas	-0.275	0.031	-8.871	-0.336	-0.215	(p<.01)
O5	Adición de ángulos de un triángulo	0.299	0.084	3.560	0.134	0.464	(p<.01)
O6	Aplicación de fracciones	-0.187	0.038	-4.921	-0.261	-0.113	(p<.01)
O7	Desarrollo de sucesiones aritméticas	1.492	0.076	19.632	1.343	1.64	(p<.01)
O8	Representación de modelos exponenciales	*					
O9	Representación del modelo del área y volumen	0.179	0.05	3.580	0.08	0.278	(p<.01)
O10	Aplicación de conversiones básicas	*					
O11	Aplicación de la regla de tres simple	2.185	0.059	37.034	2.07	2.3	(p<.01)
O12	Suma de fracciones	**					
O13	Aplicación de las reglas básicas de la probabilidad	1.237	0.049	25.245	1.142	1.333	(p<.01)
O14	Aplicación de expresiones algebraicas	2.074	0.052	39.885	1.972	2.176	(p<.01)
(*) Operaciones descartadas (**) Operaciones condensadas						α=0.01, *t*=>2.33 y α=0.05, *t*=>1.96	

Tabla 4.16. Parámetros básicos del LLTM de la V-ÍRC

Código	Operaciones	α_k	$SE\ (\alpha_k)$	Min.	Max.	*t*	*p*
O1	Representación y visualización de figuras geométricas	0.272	0.077	0.121	0.423	3.532	(p<.01)
O2	Posicionamiento y ubicación de valores	-1.103	0.092	-1.283	-0.923	-11.989	(p<.01)
O3	Aplicación de operaciones aritméticas básicas	-1.291	0.079	-1.446	-1.137	-16.342	(p<.01)
O4	Aplicación de escalas gráficas en mapas	*					
O5	Identificación de patrones de secuencias numéricas	*					
O6	Interpretación de la información del problema	0.251	0.070	0.113	0.389	3.586	(p<.01)
O7	Representación de modelos matemático-aritméticos	0.210	0.075	0.064	0.357	2.800	(p<.01)
O8	Cálculo de equivalencias de unidades de medida	*					
O9	Representación de modelos matemático-geométricos	-0.486	0.130	-0.740	-0.232	-3.738	(p<.01)

(*)Operaciones descartadas α=0.01, *t*=>2.33 y α=0.05, *t*=>1.96

4.3.2. Aplicación del modelo componencial LSDM a los ítems del área de HC del EXHCOBA

En lo concerniente a los resultados de la aplicación del LSDM, en las Figuras 18 y 19 se presentan las CPA de la matriz Q obtenidas con el LSDM para la V-ÍOM y V-ÍRC, respectivamente. En la Figura 18 se pueden observar las CPA de la matriz Q (21i X11*k*) de la V-ÍOM en donde las probabilidades correspondientes a las operaciones **O11, O13** y **O14** son una constante de 1 para todos los niveles de rasgo. Según en estudios de simulación realizados por Romero (2010), dicho resultado indica que el dominio de la operación es irrelevante para la solución de ítems analizados, o que la habilidad representada en la operación cognitiva está contenida en otro componente y, por lo tanto, hay una dependencia secuencial con este.

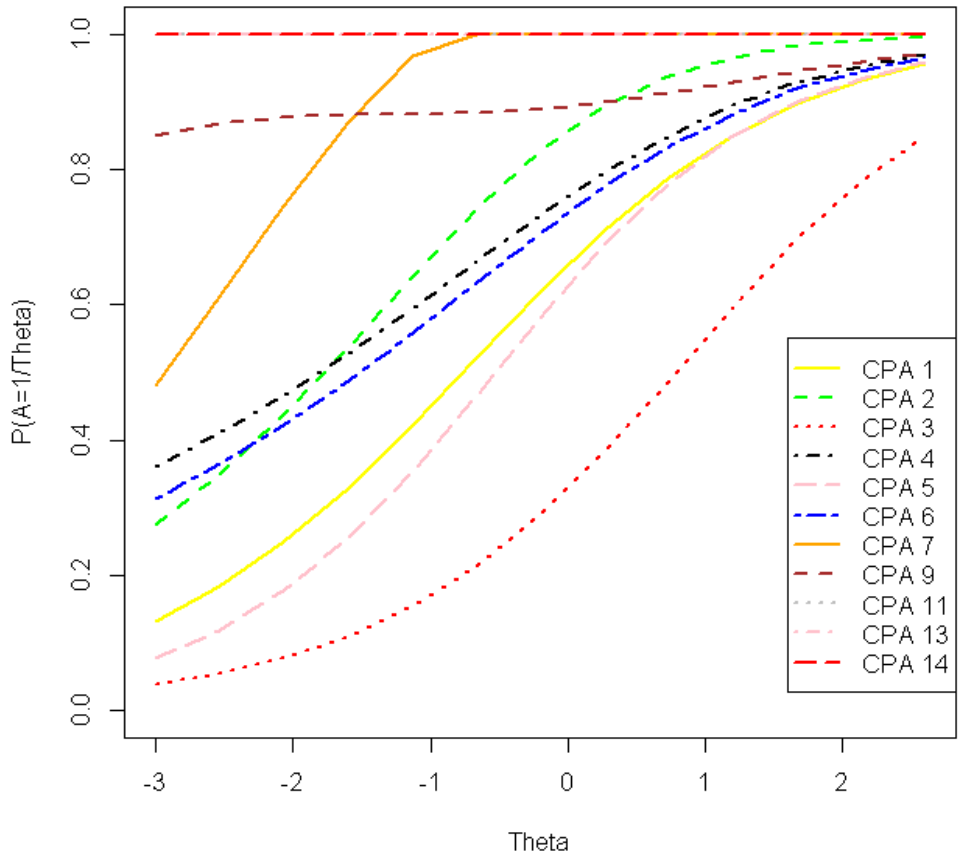


Figura 18. Curvas de probabilidad de las operaciones de la matriz Q (21i X11k) reconfigurada de la V-ÍOM

Por su parte, en la Figura 19 se pueden observar las CPA de la matriz Q (14i X6k) de la V-ÍRC en donde las probabilidades correspondientes a la operación **O**$_{1}$, es al igual que en el caso de las operaciones **O**$_{11}$**, O**$_{13}$ y **O**$_{14}$ de la V-ÍOM una constante de 1 para todos los niveles de rasgo. Además, los valores LSD estimados para la V-ÍOM se encuentran entre 0.123 (θ = −3) y 0.006 (θ = 3), mientras que para la V-ÍRC se encuentra entre o.234 (θ = −3) y 0.051 (θ = 3).

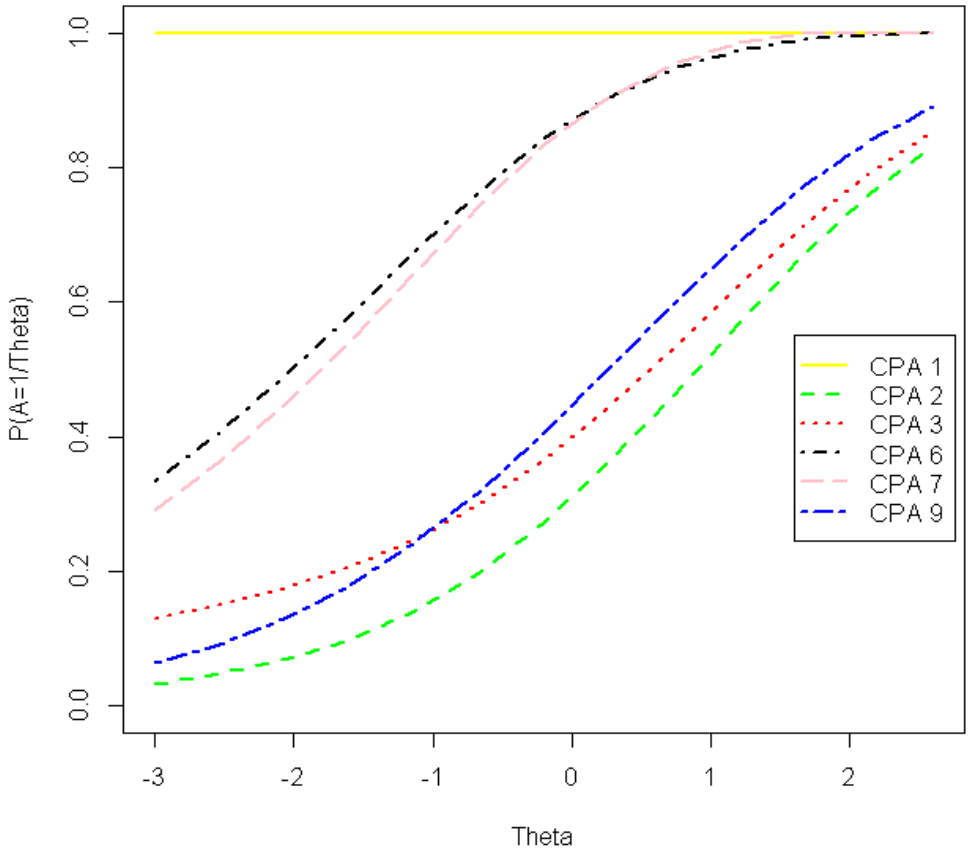


Figura 19. Curvas de probabilidad de las operaciones de la matriz Q reconfigurada (14i X6k) de la V-ÍRC

4.3.3. Proceso de reconfiguración de la matriz Q de dos versiones del área de HC del EXHCOBA

Como ya se comentó en el método, debido a que no se pudo obtener una estimación de varias operaciones cognitivas y a que se presentaron valores MAD superiores a 0.1 en las dos versiones aquí analizadas, se decidió reconfigurar reiteradamente la matriz Q hasta alcanzar una mejor estimación (ver Romero, 2010; Loye, 2008). Cabe recordar que, para las distintas reconfiguraciones de la matriz Q, el principal criterio fue que el modelo cognitivo reconfigurado de cada uno de los ítems mantuviera su

consistencia básica con el modelo definido por el panel de expertos. Al final del proceso de reconfiguración de la matriz Q se establecieron 10 operaciones cognitivas para la V-ÍOM y 5 operaciones para la V-ÍRC. Cabe señalar que el orden de los atributos presente en las Tablas 4.17 y 4.18 está basado en los resultados de la aplicación de los modelos LLTM y LSDM, después de la reconfiguración de la matriz Q de cada una de las versiones analizadas. También, que la **O_{14}, O_{13}**, **O_1** y **O_5** de la V-ÍOM se redefinieron (*), cambiando su descripción sobre los procesos cognitivos que las sustentan y redimensionaron, colocándolas en un diferente nivel de complejidad al igual que la operación **O_3** de la V-ÍRC.

Tabla 4.17. Operaciones cognitivas resultantes del proceso de reconfiguración de la matriz Q de la V-ÍOM del área de HC del EXHCOBA

Código	Operaciones cognitivas	Descripción
O_2	Fracción y visualización de figuras geométricas	Fraccionar y visualizar figuras geométricas poligonales como el triángulo, el cuadrilátero y el pentágono
O_{14}*	Sustitución de valores en expresiones algebraicas	Aplicar la substitución de valores en expresiones algebraicas
O_9	Representación del modelo matemático del área y del volumen	Representar y aplicar el modelo matemático del área para obtener el área de una figura geométrica
O_6	Aplicación de fracciones	Aplicar fracciones, su equivalencia con números decimales y su aritmética
O_{13}*	Comprensión de las reglas básicas de la probabilidad	Comprensión de la regla general para la adición de probabilidades y para los sucesos mutuamente excluyentes
O_7	Desarrollo de sucesiones aritméticas	Desarrollar sucesiones de números reales monótonas y acotadas
O_1*	Comprensión del sistema decimal (dígitos) en problemas matemáticos contextualizados	Comprender en problemas matemáticos contextualizados las nociones básicas, su ubicación y los nombres de los dígitos dentro del sistema decimal
O_3	Comprensión de problemas matemáticos contextualizados	Comprender los problemas matemáticos planteados en lenguaje común y contextualizado
O_4	Aplicación de operaciones y modelos matemático-aritméticos	Aplicación de operaciones de +, -, *, /,=, >, <, signos y números y, de modelos como 2·n, 2(término), (término)², n/2, la regla de tres simple y la raíz cuadrada
O_5*	Comprensión y aplicación de la adición de ángulos de un triángulo	Comprensión y aplicación de la suma de los ángulos interiores y exteriores de un triángulo
(*) Atributos redefinidos y redimensionados		

Tabla 4.18. Operaciones cognitivas resultado del proceso de reconfiguración de la matriz Q de la V-ÍRC del área de HC del EXHCOBA

Código	Operaciones cognitivas	Descripción
O_6	Interpretación de la información del problema	Comprender la descripción contextual del problema y e inferir valores con base en la información de gráficas, mapas y tablas.
O_7	Representación de modelos matemático-aritméticos	Representar modelos de división con galera con números decimales, probabilísticos, números racionales, MCM, regla de tres y de porcentajes.
O_9	Representación de modelos matemático-geométricos	Representar modelos matemáticos–geométricos para perímetros, áreas y volúmenes.
O_2	Posicionamiento y ubicación de valores	Posicionar y ubicar valores en una recta numérica o en un una secuencia de menor a mayor que.
O_3*	Aplicación de operaciones y modelos matemático-aritméticos	Aplicación de operaciones de +, -, *, /,=, >, <, signos y números y, de modelos como 2·n, 2(término), (término)², n/2, la regla de tres simple y la raíz cuadrada.
(*) Atributos redefinidos y redimensionados		

A su vez, se muestra en las Tablas 4.19 y 4.20 la reconfiguración de la matriz Q de cada una de las dos versiones analizadas. Nótese que del proceso de reconfiguración de la matriz Q para la V-ÍOM resultó un total de 10 operaciones cognitivas para 21 ítems (Matriz Q, 21i X10k) y 5 operaciones cognitivas para 14 ítems (Matriz Q, 14i X5k) de la V-ÍRC. En la última columna de dichas tablas se incluye el número total de operaciones cognitivas requeridas por cada ítem, así como la proporción media de aciertos por ítem y por subconjunto de ítems asociado a cada operación cognitiva.

Tabla 4.19. Matriz Q (21i X10k) de la V-ÍOM y proporción media de aciertos por ítems de cada operación cognitiva

No. Ítem	O_1	O_2	O_3	O_4	O_5	O_6	O_7	O_9	O_{13}	O_{14}	Total	p
02	0	0	0	0	0	0	1	0	0	0	1	0.76
04	1	0	0	0	0	0	0	0	0	0	1	0.73
05	1	0	1	0	0	0	0	0	0	0	2	0.43
09	0	0	0	1	0	1	0	0	0	0	2	0.46
11	1	0	0	0	0	1	0	0	0	0	2	0.50
12	0	1	1	0	0	0	0	0	0	0	2	0.67
13	0	1	0	1	0	1	0	0	0	0	3	0.49
14	0	1	0	1	0	1	0	0	0	0	3	0.55
15	0	0	0	1	0	0	0	0	0	0	1	0.66
16	0	0	0	1	0	0	0	0	0	0	1	0.71
17	1	0	0	1	0	0	0	0	0	0	2	0.34
19	0	1	1	1	0	0	0	0	0	0	3	0.27
20	1	0	1	0	0	0	0	1	0	0	3	0.29
21	0	1	0	1	0	0	0	1	0	0	3	0.54
23	0	0	1	0	0	0	0	0	0	1	2	0.60
24	0	0	1	0	0	1	0	0	0	0	2	0.59
25	0	0	1	0	0	1	0	0	0	0	2	0.57
27	0	0	0	0	1	0	0	0	0	0	1	0.60
28	0	0	1	0	0	0	0	0	1	0	2	0.38
29	0	0	0	0	0	0	0	0	1	1	2	0.84
30	0	0	0	1	0	0	0	0	0	1	2	0.51
$\bar{x}$ **de** p_i	0.46	0.50	0.48	0.50	0.60	0.53	0.76	0.42	0.61	0.65	$\bar{x}$ **de** p	0.55

Tabla 4.20. Matriz Q (14i X5k) de la V-ÍRC y proporción media de aciertos por ítems y por subconjunto de ítems asociado a cada operación cognitiva

No. Ítem	O_2	O_3	O_6	O_7	O_9	Total	p
2	1	1	0	0	0	2	0.21
3	1	0	0	0	0	1	0.53
4	0	1	0	1	0	2	0.33
5	0	0	0	1	0	1	0.55
6	0	1	1	0	0	2	0.23
8	0	1	0	0	1	2	0.25
9	0	1	0	0	1	2	0.24
10	0	1	0	0	1	2	0.27
14	0	0	1	1	0	2	0.49
15	0	0	1	1	0	2	0.32
16	0	0	1	1	0	2	0.46
18	0	0	1	0	0	1	0.64
19	1	0	1	0	0	2	0.33
20	0	1	1	0	0	2	0.43
$\bar{x}$ **de** p_i	0.37	0.28	0.41	0.43	0.25	$\bar{x}$ **de** p	0.38

Es necesario mencionar, que aun después de varios ejercicios de reconfiguración de la Matriz Q, el ítem 16 de la V-ÍOM presentó un valor MAD mayor a 0.1 (recuperación algo pobre). Por su parte, los ítems de la V-ÍRC presentaron valores MAD menores a 0.1 (ver Tabla 4.21). En lo que respecta a los valores WMAD y RSMD de las dos versiones, estos presentan buena recuperación bastante, similar a los valores MAD. Sin embargo, Romero (2010) menciona que para poder realizar una valoración precisa de la recuperación de las CCI con estos dos estadísticos, es necesario estudiar su distribución y encontrar puntos de corte adecuados.

Tabla 4.21. Estadísticos de recuperación de las CCI con el LSDM de la V-ÍOM y de la V-ÍRC

No. ítem	V-ÍOM			V-ÍRC		
	WMAD	**RMSD**	**MAD**	**WMAD**	**RMSD**	**MAD**
01	-	-	-	-	-	-
02	0.000	0.000	0.000	0.025	0.034	0.026
03	-	-	-	0.048	0.062	0.052
04	0.084	0.105	0.090	0.026	0.033	0.028
05	0.042	0.050	0.045	0.044	0.063	0.052
06	-	-	-	0.057	0.070	0.062
07	-	-	-	-	-	-
08	-	-	-	0.005	0.006	0.005
09	0.035	0.042	0.037	0.009	0.012	0.010
10	-	-	-	0.015	0.019	0.016
11	0.047	0.054	0.051	-	-	-
12	0.071	0.088	0.076	-	-	-
13	0.021	0.027	0.023	-	-	-
14	0.025	0.029	0.027	0.043	0.053	0.049
15	0.083	0.103	0.089	0.069	0.087	0.075
16	0.118	0.141	**0.127 ***	0.028	0.034	0.032
17	0.017	0.022	0.018	-	-	-
18	-	-	-	0.036	0.056	0.043
19	0.063	0.081	0.067	0.004	0.005	0.005
20	0.030	0.040	0.032	0.085	0.107	0.093
21	0.049	0.062	0.052	-	-	-
22	-	-	-	-	-	-
23	0.010	0.012	0.011	-	-	-
24	0.051	0.061	0.055	-	-	-
25	0.037	0.044	0.040	-	-	-
26	-	-	-	-	-	-
27	0.000	0.000	0.000	-	-	-
28	0.028	0.033	0.030	-	-	-
29	0.086	0.110	0.094	-	-	-
30	0.047	0.056	0.051	-	-	-
Promedio	0.045	0.055	0.048	0.035	0.046	0.039

(*) MAD > 0.1

Continuando con la presentación de resultados de la aplicación del LSDM, en las *Figuras* 20 y 21 se presentan las CPA resultantes de la última reconfiguración de la matriz Q para la V-ÍOM y la V-ÍRC, respectivamente. Como se esperaba, se ha podido obtener una estimación de todas las operaciones cognitivas en las dos versiones. También se puede observar en dichas figuras que el comportamiento de algunas de las CPA de las operaciones de la V-ÍOM no presentan buena monotonicidad. Sin embargo, en el caso del comportamiento de las CPA de las operaciones de la V-ÍRC, estas presentan buena monotonicidad, lo que da cuenta de un buen nivel de discriminación

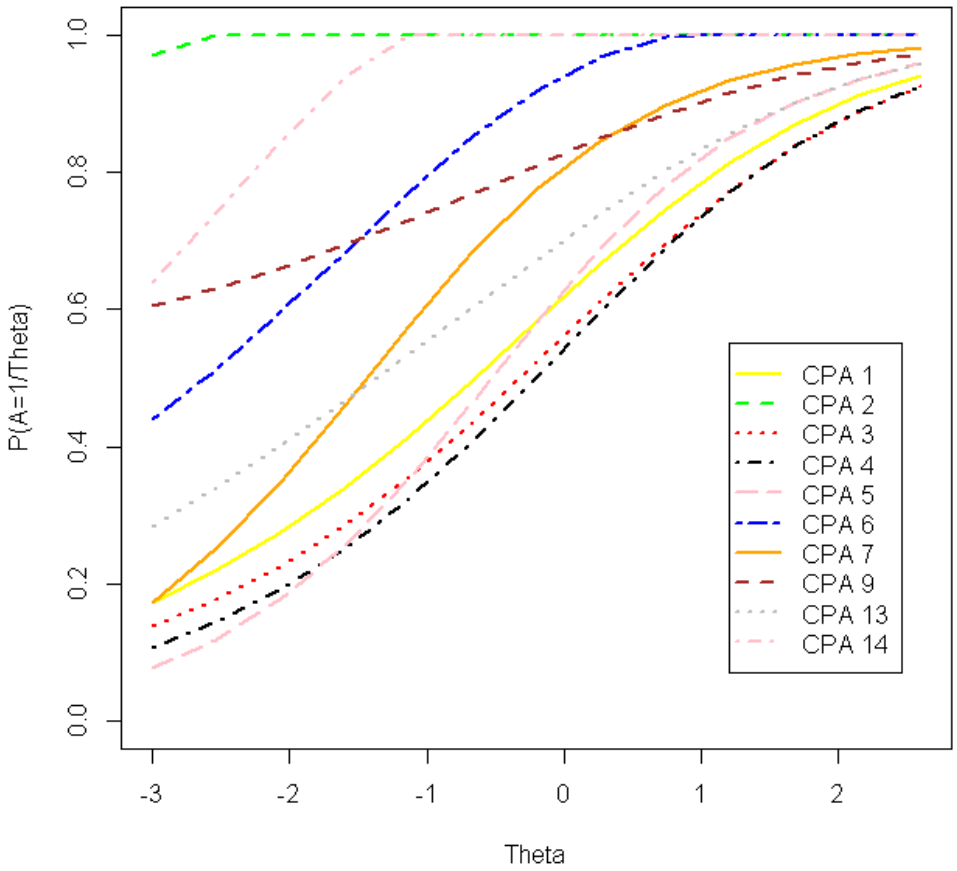


Figura 20. Curvas de probabilidad de los atributos de la Matriz Q reconfigurada de la V-ÍOM

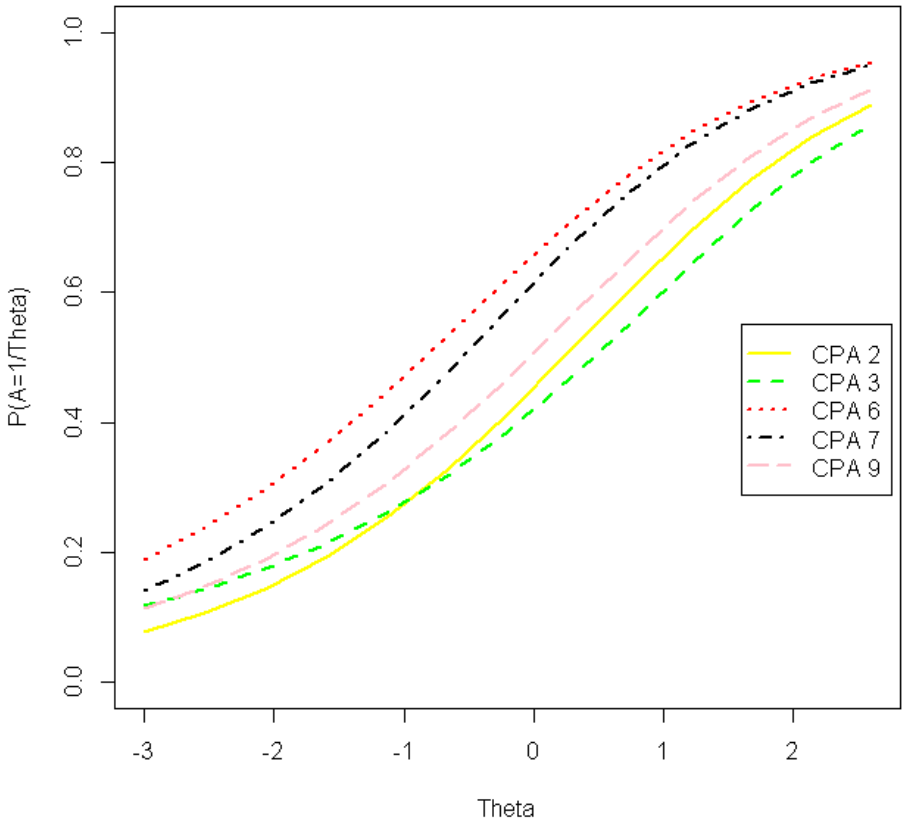


Figura 21. Curvas de probabilidad de los atributos de la Matriz Q reconfigurada de la V-ÍRC

En las Figuras 22, 23, 24 y 25 se representan las CCI y límites de algunos ítems de la versión V-ÍOM que presentaron la recuperación más baja. Puntualmente, en la Figura *22* se presentan las CCI y límites del ítems 12 (MAD = 0.76) de la V-ÍOM; en la Figura *23* se presenta la CCI y límites del ítem 16 (MAD = 0.127) de la misma versión. Dichos ítems presentan una recuperación algo buena y algo pobre, respectivamente. Obsérvese en las Figuras *24* y *25* que en particular la CCI estimada por el LSDM para los ítems 15 y 18 se encuentra dentro de los límites en todos los niveles de habilidad, indicando que se trata de un ítem con recuperación muy buena.

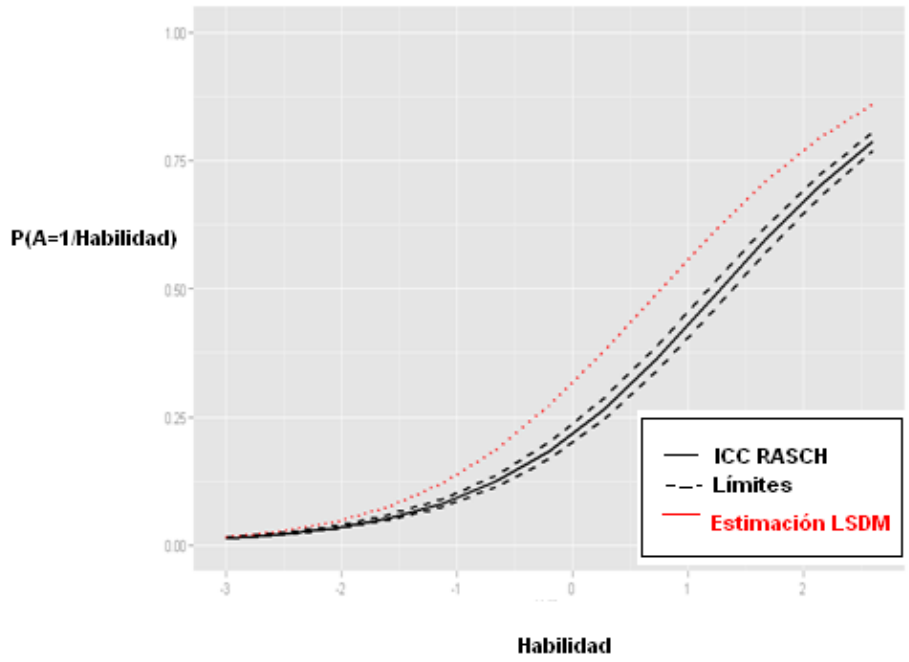


Figura 22. CCI original y recuperada con límites para el ítem 12 de la V-ÍOM

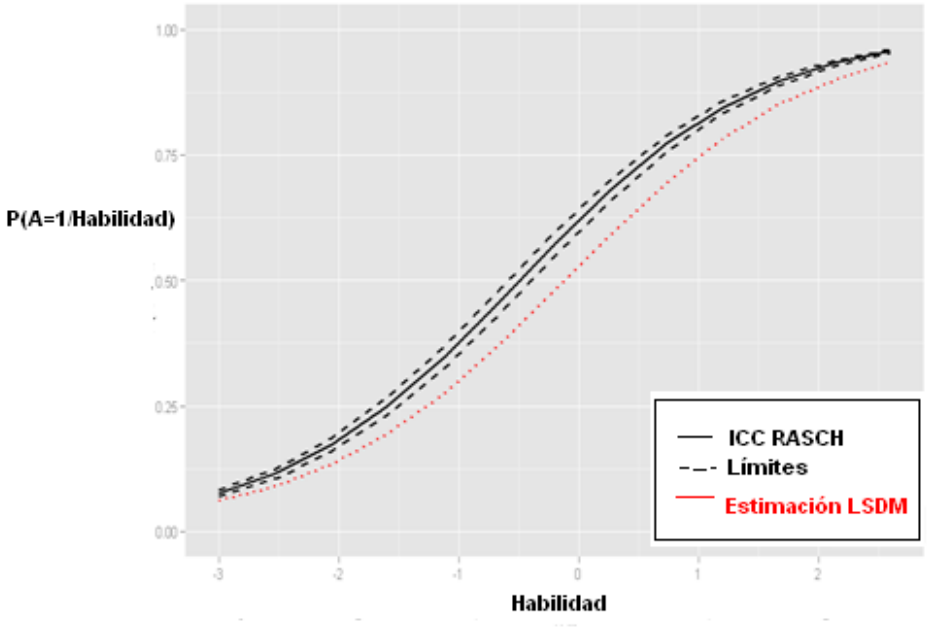


Figura 23. CCI original y recuperada con límites para el ítem 16 de la V-ÍOM

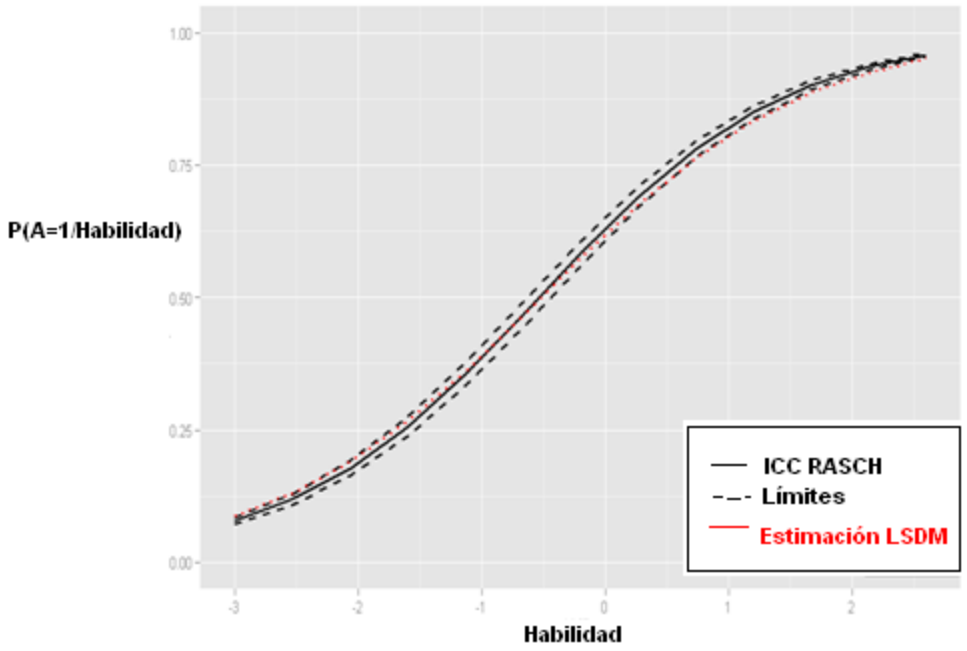


Figura 24. CCI original y recuperada con límites para el ítem 15 de la V-ÍOM

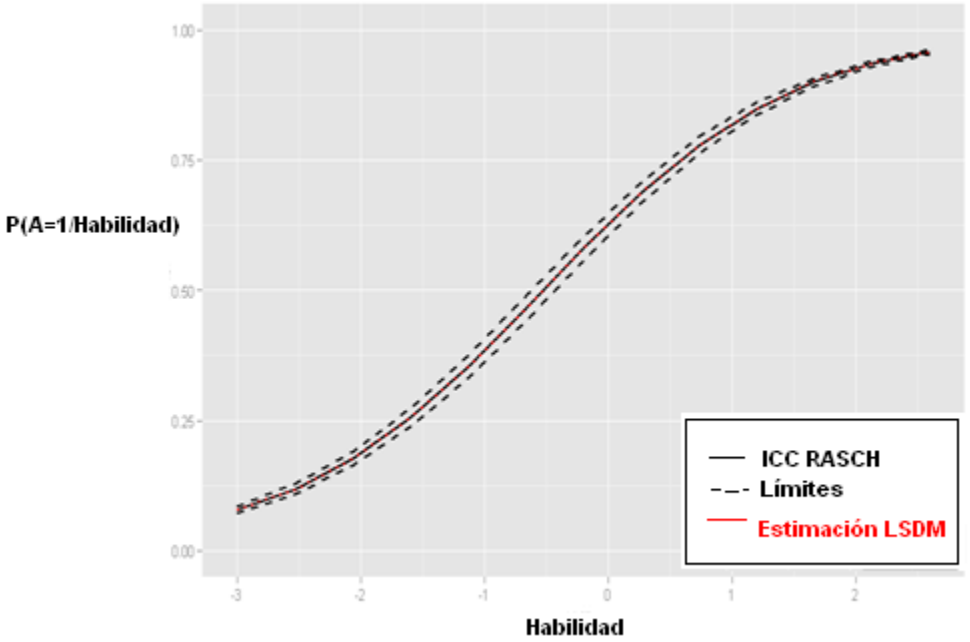


Figura 25. CCI original y recuperada con límites para el ítem 18 de la V-ÍOM

De igual forma, en las Figuras 26 y 27 se pueden observar los ítems de la V-ÍRC con menos recuperación de las CCI. Concretamente, las CCI y límites de los ítems 15 (MAD = 0.08) y del ítem 20 (MAD = 0.09) presentan una recuperación algo buena. Nótese, que la CCI estimada por el LSDM para los dos ítems recientemente mencionados sobrepasa ligeramente los límites en todos los niveles de habilidad, indicando que se trata de ítems con recuperación algo buena. En cambio, en las Figuras 28 y 29 se presenta la situación contraria, pues se trata de ítems con muy buena recuperación (0 ≤ MAD ≤ 0.02), cuya CCI se encuentra dentro de los límites establecidos.

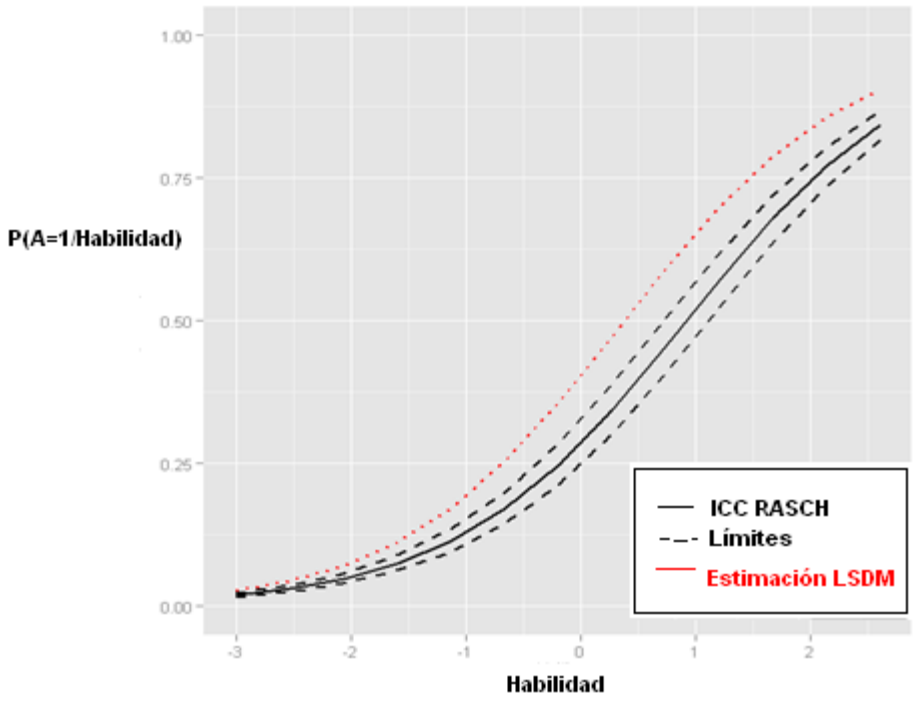


Figura 26. CCI original y recuperada con límites para el ítem 15 de la V-ÍRC

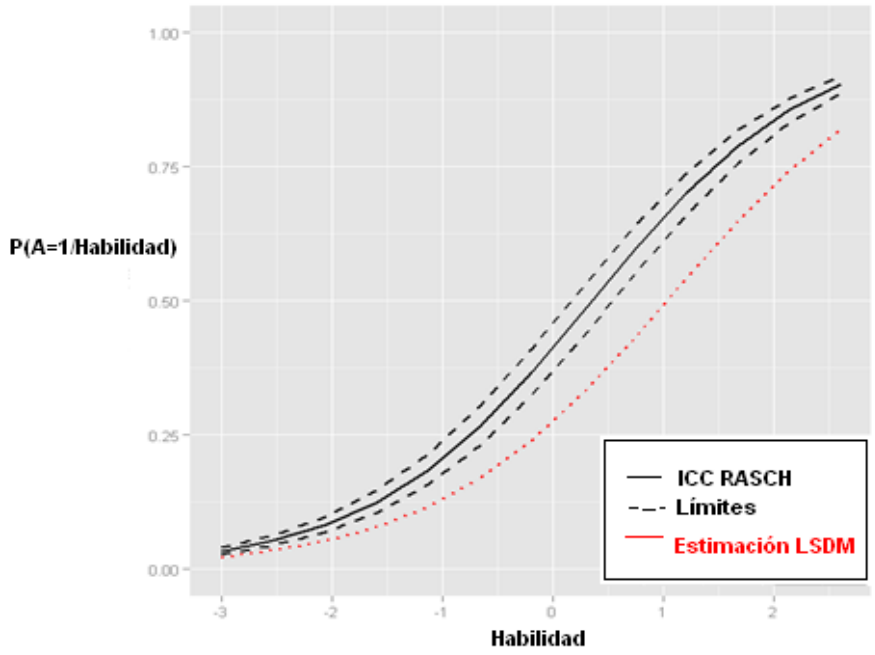


Figura 27. CCI original y recuperada con límites para el ítem 20 de la V-ÍRC

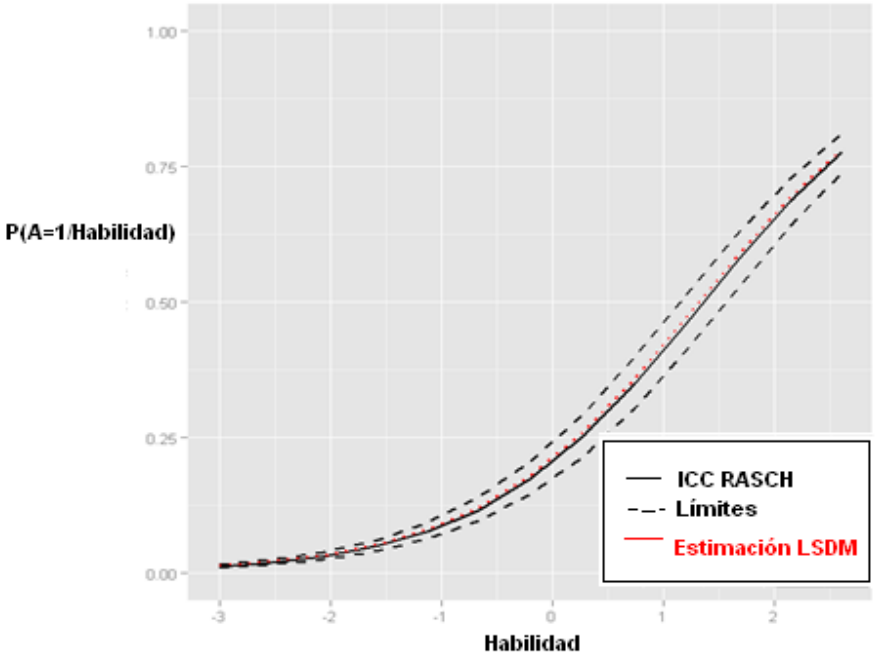


Figura 28. CCI original y recuperada con límites para el ítem 8 de la V-ÍRC

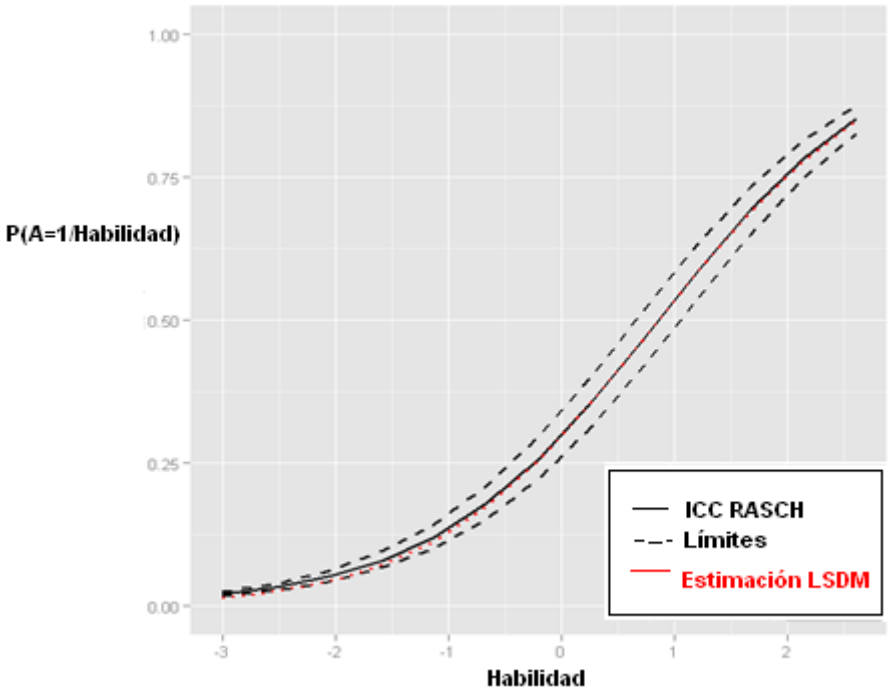


Figura 29. CCI original y recuperada con límites para el ítem 19 de la V-ÍRC

4.3.4. Análisis de la validación cruzada con los modelos LLTM y LSDM

Para avanzar con mayor seguridad en el análisis de la validación cruzada entre los modelos LLTM y LSDM, se presentan antes los resultados de la prueba de ajuste del LLTM con la matriz Q reconfigurada y el modelo de RASCH. Para la V-ÍOM mejoró mucho la correlación entre los parámetros del LLTM y los de RASCH de 0.829 a 0.95. De igual forma, para la V-ÍRC mejoró el valor de ajuste con una correlación de 0.91 en comparación al valor anterior de 0.78. También, disminuyó el CIA, dando cuenta de un mejor ajuste del modelo LLTM reconfigurado al modelo de RASCH de las dos versiones estudiadas (ver Tabla 4.22).

Tabla 4.22. Comparación de la mejora en el ajuste entre los modelos RASCH, LLTM y LLTM reconfigurado de las dos versiones del área de HC

Versión	Modelos	Criterio de información de Akaike		Correlación *b* RASCH - *b* LLTM
		2ln*L*	CIA	
V-ÍOM	RASCH	48384.32	48424.32	
	LLTM	50802.24	50824.24	0.83
	LLTM*	49079.46	49099.46	0.95
V-ÍRC	RASCH	-3810.083	7646.166	
	LLTM	-3992.465	7996.930	0.78
	LLTM*	-3886.066	7784.132	0.91

LLTM* Modelo con matriz Q reconfigurada

Complementando los análisis de ajuste entre los modelos RASCH y LSDM, en las Figuras 30 y 31 se presentan las pruebas gráficas de ajuste con el contraste de las curvas características de cada modelo. Nótese que en la prueba gráfica de ajuste con el contraste entre el modelo RASCH y el LSDM de la V-ÍOM se alcanza un excelente ajuste entre los dos modelos. Por otra parte, en la prueba gráfica de ajuste con el contraste entre el modelo RASCH y el LSDM de la V-ÍRC se presentan, en los niveles altos de dificultad y de dominio de los atributos, algunos problemas. Estos problemas de ajuste no son fuertes, pero sí es recomendable trabajar en una mejor estimación de los MAD de dicha versión de la prueba.

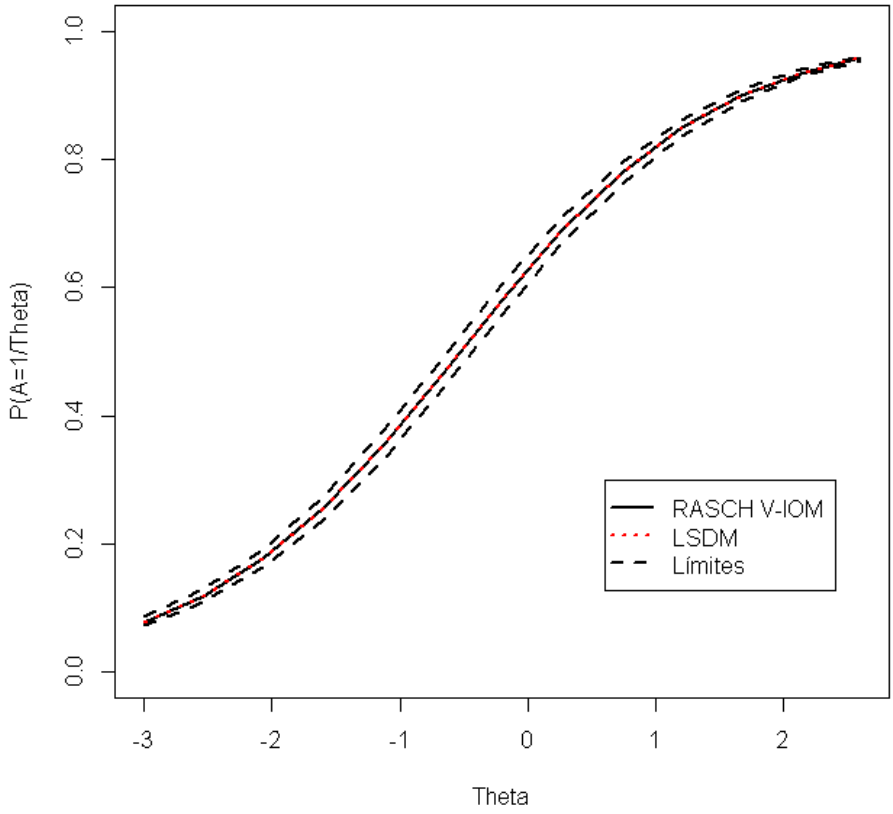


Figura 30. Prueba gráfica de ajuste con el contraste entre el modelo RASCH y el LSDM de la V-ÍO

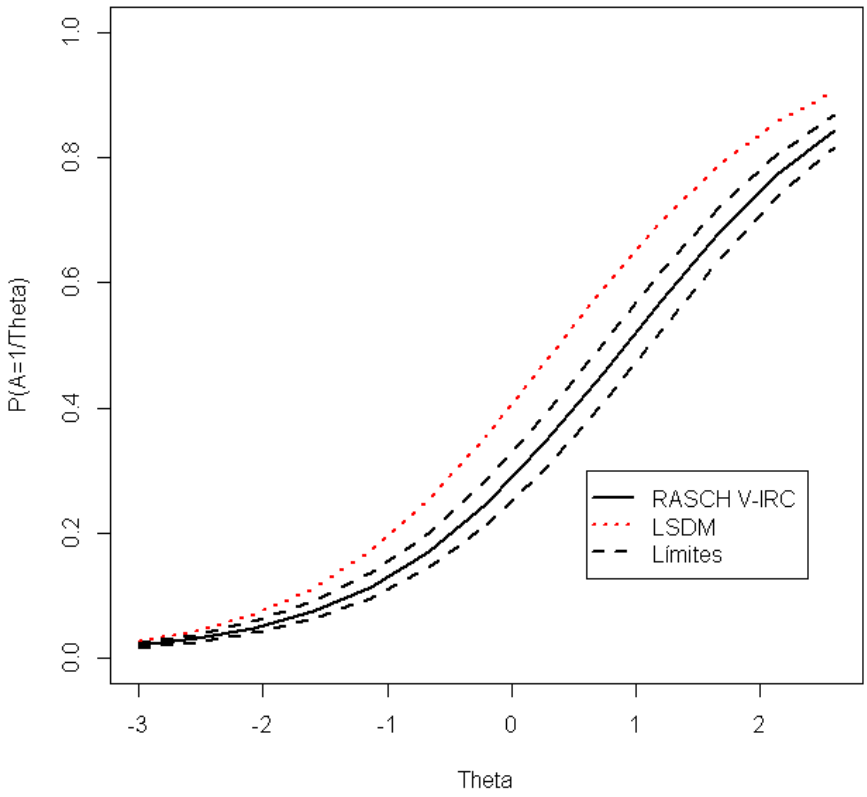


Figura 31. Prueba gráfica de ajuste con el contraste entre el modelo RASCH y el LSDM de la V-ÍRC

Para la validación cruzada de los resultados de los dos modelos utilizados, se puede decir que los valores de los parámetros básicos, estimados con el LLTM, ordenados de fácil a difícil coinciden perfectamente con el orden de las CPA estimadas con el LSDM (ver Tabla 4.23). Además, todos los parámetros básicos del LLTM en las dos versiones estudiadas, son significativamente diferentes de 0 ($p < 0.01$), indicando así que todas las operaciones cognitivas contribuyen a la explicación de la dificultad de los ítems.

En la Tabla 4.23 se presenta un resumen de la comparación del ordenamiento de las operaciones cognitivas de las dos versiones del área de HC, según su dificultad relativa entre los modelos LLTM y LSDM. También se muestra en dicha tabla la proporción de dominio observada empíricamente mediante el promedio de la proporción de aciertos del subconjunto de ítems relacionados a cada operación cognitiva. El ordenamiento producido por los modelos LLTM y LSDM para la V-ÍOM coincide bien entre ambos, pero con cierta incertidumbre, ya que algunas CPA recuperadas por el LSDM no presentan monotonía adecuada. Incluso, el ordenamiento obtenido con los dos modelos no concuerda con lo esperado teóricamente por los expertos ni con la proporción de dominio empírico. Por otro lado, el ordenamiento producido por los modelos LLTM y LSDM para la V-ÍRC coincide entre ellos perfectamente. Con ello, se obtiene una importante evidencia de validez y consistencia de los resultados encontrados.

Sin embargo, al igual que en la V-ÍOM el ordenamiento obtenido por los modelos LLTM y LSDM no concuerdan con lo esperado en el modelo cognitivo planteado por los expertos. Dado lo anterior, es importante señalar la necesidad de revisar el modelo cognitivo que se planteó de inicio. Es posible, con el buen ajuste que hay entre los parámetros de los modelos componenciales, que se trate de una sobre simplificación o subestimación de algunas operaciones cognitivas. Es posible que la coincidencia que hay en los dos modelos de que las operaciones cognitivas planteadas desde un inicio como las más fáciles (V-ÍOM: O_1, O_3, O_4 y O_5; V-ÍRC: O_2

y O_3) y que resultan siendo las más complejas tanto en los resultados del LLTM como en los del LSDM sea una evidencia de la subestimación mencionada (ver Tabla 4.23).

Tabla 4.23. Comparación del orden de la dificultad relativa de las operaciones cognitivas reconfiguradas de la V-ÍOM y de la V-ÍRC del área de HC

Versión	Modelos	Operaciones cognitivas reconfiguradas									
		Fácil →									Difícil
V-ÍOM	Observada	O_7	O_{14}	O_{13}	O_5	O_6	O_2	O_4	O_3	O_1	O_9
	LLTM	O_2	O_{14}	O_9	O_6	O_{13}	O_7	O_1	O_3	O_4	O_5
	LSDM	O_2	O_{14}	O_9*	O_6*	O_{13}*	O_7*	O_1*	O_3	O_4	O_5*
		Fácil →				Difícil					
V-ÍRC	Observada	O_7	O_6	O_2	O_3	O_9					
	LLTM	O_6	O_7	O_9	O_2	O_3					
	LSDM	O_6	O_7	O_9	O_2	O_3					

(*) Operaciones cognitivas que presentan cierta ambigüedad con su nivel de dominio.

V. CONCLUSIONES

En el presente capitulo se muestran las conclusiones de la investigación. Además, se discuten los logros y limitaciones de los diferentes procesos analíticos, sus productos y los resultados obtenidos. También se reconsidera el modelo teórico-metodológico adaptado que guio el análisis del aspecto sustantivo de la validez de constructo del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. Las conclusiones del estudio se estructuran en tres apartados relacionados con la discusión de los logros y de las aportaciones de la tesis, las limitaciones de la misma y las recomendaciones para futuras investigaciones.

5.1. Discusión de los logros y de las aportaciones de la tesis

Realizar estudios de validez y mejorar la calidad de las pruebas, ayuda al uso justo y ético de sus resultados en los distintos contextos y procesos de aplicación (AERA, APA & NCME, 1999). Además, la interpretación de los resultados de las pruebas fundamentada en evidencias científicas, y en el uso adecuado, puede ayudar a sus usuarios en la toma de decisiones. Por su parte, los estudios sobre el aspecto sustantivo de la validez de constructo de pruebas psicológicas y educativas, además de beneficiar al desarrollo y validación de estas, beneficia directamente a otros contextos como la realimentación y la operación del currículum, así como a la mejor

comprensión del proceso de enseñanza-aprendizaje (Cortada de Kohan, 2000; Messick, 1989b; Snow & Lohman, 1989).

En especial, las evidencias basadas en el aspecto sustantivo de la validez de constructo toman relevancia para las pruebas de selección —como en el caso del EXHCOBA—, debido a que estas idealmente deben desarrollarse con base en los constructos que mejor predigan el rendimiento y el éxito escolar de los examinados. Visto desde la teoría de validez propuesta por Messick (1989b) o la propuesta por Borsboom, Mellenbergh y van Heerden (2004), ésta no debe encontrarse inherente al tipo de evaluación, sino que el significado debe analizarse con respecto al efecto causal de los atributos en el puntaje de la prueba, es decir, dichos tipos de evaluaciones, aun definiéndose previamente como normativas, continúan requiriendo evidencias relacionadas con la fidelidad de los procesos de respuesta ante la prueba y el constructo supuestamente medido.

Con ello, un examen que mide las variables con mayor poder de predicción del éxito escolar, y que tiene fuertes evidencias del aspecto sustantivo de la validez, presenta mayor certidumbre para su uso en la selección de aspirantes a ingresar al nivel educativo de interés (Tirado, et al. 1997). Del mismo modo, toma mayor relevancia el aspecto sustantivo de la validez cuando el constructo de interés tiene que ver con demostrar mediante métodos o mediciones cognitivas, que la justificación para el uso de la prueba (sea para selección o de diagnóstico) o la interpretación de sus puntuaciones (criterial o normativa) depende de las premisas teóricas de los procesos

psicológicos o de las operaciones cognitivas utilizadas por los examinados (AERA, APA & NCME, 1999).

Es por ello que el hecho de haber adaptado un modelo teórico-metodológico con enfoque *top-down* para obtener evidencias de validez, basadas en el proceso de respuesta y en la estructura del modelo cognitivo del EXHCOBA, fue una de las aportaciones que se consideran de mayor impacto y alcance de esta investigación. Primero, porque el modelo teórico-metodológico adaptado favorece a los desarrolladores de pruebas y de investigadores en los campos de la medición y de la evaluación con el desarrollo de una plataforma operativa para la obtención de diferentes evidencias de validez basadas en una teoría *fuerte* (Griel & Lai, 2013). Segundo, porque dicho modelo puede ser aplicado en una gran diversidad de contextos evaluativos para guiar el fortalecimiento de la validez de instrumentos que iniciaron su desarrollo desde un modelo teórico *de redes nomológicas,* como en el caso de la mayoría de las pruebas nacionales e internacionales de aprendizaje. Tercero y último, porque con la aplicación de dicho modelo en el examen, se logró un análisis a profundidad de las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo del área de HC del EXHCOBA en dos de sus versiones, una con ítems de opción múltiple y otra con ítems de respuesta compleja. Con ello, se tiene mayor seguridad de que la aplicación del modelo teórico–metodológico adaptado en las demás áreas de la prueba podrá alcanzar las expectativas en relación a la obtención de evidencias del aspecto sustantivo de la validez de constructo.

Además, el contar con un modelo teórico-metodológico adaptado que ayude en la obtención de evidencias del aspecto sustantivo de validez del EXHCOBA, no sólo aporta a su calidad técnica, sino que también abre la posibilidad de diagnosticar a profundidad las fortalezas y las debilidades de sus examinados. Lo anterior da una gran ventaja a los usuarios de los resultados, dando información más allá del simple puntaje total. Dicho lo anterior, para profundizar en las diferentes aportaciones de la presente tesis, se muestran a consideración algunos de los logros puntuales en relación a los objetivos específicos de investigación.

En cuanto al objetivo específico de *documentar los fundamentos teóricos del análisis de evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas psicológicas y educativas*, se puede decir que se logró una revisión a profundidad en distintas bases de datos y en revistas científicas de alto impacto sobre los fundamentos teóricos y técnicos de las teorías de la validez. Con ello, se obtuvo información valiosa sobre los antecedentes y fundamentos teóricos relacionados al tema de investigación, para lo cual, el capítulo del marco teórico es una evidencia de ello. Sin embargo, en especial lo relacionado con información sobre estudios y experiencias asociadas a la obtención de evidencias del aspecto sustantivo de validez de constructo se encontraron serias limitaciones, las cuales se mencionan en el siguiente apartado.

En cuanto al objetivo de *adaptar y aplicar un modelo teórico-metodológico con enfoque top-down para obtener y analizar las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo de pruebas psicológicas*

y educativas computarizadas, se considera que fue uno de los objetivos en donde el logro superó las expectativas iniciales de este estudio. Los beneficios y las aportaciones de tal logro presentan un alto alcance teórico-práctico. Principalmente, es un gran aporte para el campo del desarrollo y de la validación de pruebas educativas y psicológicas el ofrecer un marco operativo para obtener evidencias de validez relacionadas con las más innovadoras propuestas teóricas de validez y de la evaluación cognitiva, dada la escasez de experiencias similares. Además, con ello se coloca a la vanguardia el EXHCOBA al incorporar estudios de primer nivel que dan evidencia de su calidad técnica, ayudándole a seguir posicionada como una de las pruebas educativas con mayores índices de calidad en México y compitiendo con pruebas como PISA y EXCALE.

Además, la magnitud de la aportación que presentó la adaptación del modelo teórico-metodológico utilizado en la presente tesis fue crucial para el logro de los objetivos específicos de la investigación. Principalmente con aquellos objetivos relacionados con identificar y definir los modelos de los procesos cognitivos utilizados por los examinados para resolver los ítems del área de HC del EXHCOBA, y con el análisis de su validez mediante los modelos LLTM y LSDM.

Gracias a la aplicación del método de *modelado matemático de sub-tareas de respuesta* propuesto por Embretson (1983), y a su acompañamiento con el *análisis de expertos* y con las *técnicas de pensamiento en voz alta*, se logró definir un modelo cognitivo que muestra a gran profundidad las operaciones cognitivas y atributos sustantivos que utilizan los evaluados para contestar los ítems de la prueba. La

aplicación de dichas técnicas resultó una estrategia con muchas ventajas. Una de las principales fue la obtención de información valiosa para construir y definir la estructura del modelo cognitivo de la prueba que, a su vez, sirvió como insumo para la aplicación de los modelos componenciales. Asimismo, con la obtención del modelo cognitivo subyacente a los ítems del área de HC del EXHCOBA se puede realimentar de forma puntual a los usuarios de la prueba sobre los procesos de respuesta utilizados por los examinados y, con ello, aportar a la mejora del proceso enseñanza y aprendizaje.

Otra de las ventajas relacionada con la aplicación de los *análisis de protocolos* fue la posibilidad de contar con información para identificar varianza irrelevante presente en los elementos del diseño del interfaz computarizada del EXHCOBA. Asimismo, se pudo identificar aquellos problemas en el diseño de la interfaz de los ítems que pudieran afectar la validez de las inferencias y de los resultados de la prueba. Con todo ello, se logró un programa integral de validez fundamentado desde el aspecto sustantivo de la prueba.

Ahora bien, para el objetivo relacionado al *análisis de las evidencias de validez basadas en la estructura del modelo cognitivo del área de HC del EXHCOBA mediante los modelos LLTM y LSDM*, se puede decir que se alcanzaron varios logros. Primero, con la aplicación del LLTM, se pudo conocer la contribución de las operaciones cognitivas estructuradas en la matriz Q a la dificultad de los ítems de las dos versiones analizadas (Romero, Ponsoda & Ximénez, 2006). También, se encontró que era necesario un refinamiento de los catorce atributos propuestos por los expertos. Tal resultado fue confirmado con el modelo LSDM, por lo que se decidió incorporar a esta

investigación un proceso reiterativo de reconfiguración de la matriz Q de las dos versiones estudiadas. Segundo, con la aplicación del LSDM se pudo complementar el análisis de las evidencias de validez basadas en la estructura del modelo cognitivo y, por lo tanto, se obtuvieron evidencias sustantivas con un buen nivel de integración para el argumento de validez. Al respecto, el análisis de validación cruzada entre los resultados de los dos modelos componenciales utilizados da cuenta de ello.

Como un beneficio alterno, con la definición de la estructura del modelo cognitivo subyacente a los ítems, se fortaleció el modelo de la GAÍ del EXHCOBA. Lo anterior presenta un gran impacto en el incremento de su validez, fundamentándola en una teoría cognitiva *fuerte* (Griel & Lai, 2013; Gorin & Embretson, 2013). Con lo anterior, se puede tener un mejor uso y confianza en los resultados obtenidos por la GAÍ del EXHCOBA (Bejar, 1993; 1998; 2010; Embretson, 1998).

5.2. Limitaciones del estudio

En contraste con los logros y aportaciones de la tesis, se puede decir que hay varias limitantes en el estudio por mencionar. De forma general, se presentaron limitaciones relacionadas con la relativa novedad que presentan los estudios sobre el aspecto sustantivo de validez de pruebas psicológicas y educativas basadas en la GAÍ y, por lo tanto, cierta escasez de experiencias, aplicaciones metodológicas y publicaciones que pudieran aportar un ejemplo claro y cercano que sirviera como guía para este estudio. También, se presentaron limitaciones relacionadas a la escasa información del proceso cognitivo ante los ítems de la prueba arrojada en los *análisis de protocolos* de

examinados novatos en el constructo de interés. Además, los modelos componenciales presentaron algunos problemas referentes a sus requerimientos específicos de aplicación.

En la actualidad, aunque cada vez hay una mayor cantidad de publicaciones relacionadas al tema del aspecto sustantivo de la validez de constructo, es aun insipiente su producción, comparada con otros tipos de análisis psicométricos clásicos como los relacionados a las evidencias basadas en los aspectos de validez de contenido, de estructura interna y de criterio. Por su parte, se encontraron pocas investigaciones que aportaran una visión más clara sobre los procedimientos a seguir en la obtención de evidencias basadas en el aspecto sustantivo de la validez de pruebas con GAÍRC, como en el caso del EXHCOBA. Tal hecho afectó en cierta medida a la posibilidad de predecir errores en el procedimiento metodológico desarrollado en la presente tesis.

De igual forma, las *técnicas de pensamiento en voz alta con análisis de protocolos* presentaron dificultadas asociadas con la información recabada para el análisis del proceso cognitivo de los examinados ante los ítems de la prueba y su verificación con el modelo cognitivo elaborado por los expertos. Para verificar el modelo cognitivo elaborado por los expertos, es importante disponer de reportes verbales de participantes novatos y expertos en las áreas de interés que presenten información suficiente y representativa sobre los procesos cognitivos necesarios para resolver los ítems de la prueba. Sin embargo, sólo se pudieron analizar 16 de 24 reportes verbales obtenidos mediante las técnicas mencionadas, debido a que estos

no contenían información suficiente para ser analizados. La causa de dicho problema está relacionada con la incorporación de estudiantes de secundaria novatos en el área de interés (con un muy bajo rendimiento escolar en matemáticas) que presentaron dificultad para verbalizar en voz alta sus pensamientos, aún después del entrenamiento. Con ello, fue casi imposible utilizar 8 reportes verbales con escasa o casi nula información sobre los procesos cognitivos requeridos por los examinados para responder a los ítems analizados.

Sin embargo, el haber analizado relativamente pocos reportes verbales ayudó a aligerar la complejidad del trabajo de verificación. Con ello, el modelo del proceso cognitivo para resolver los ítems elaborado por los expertos se pudo verificar de forma más rápida con los procesos de respuesta utilizados por los examinados para resolver los ítems. Cabe recordar que los *análisis de protocolos* al igual que el *análisis verbal* son técnicas complejas de corte cualitativo en donde se recopilan una infinidad de datos kinestésicos, visuales y verbales, y que si no hay una delimitación clara de los datos a analizar, el investigador puede presentar fuertes dificultades para alcanzar los objetivos del análisis. Lo anterior no sustituye la necesidad de contar con una muestra suficiente de participantes que aporten información suficiente y representativa para verificar si sus procesos cognitivos requeridos para contestar a los ítems de una prueba se vinculan con algún modelo cognitivo, dado por la teoría sustantiva o por otros métodos cognitivos. Es por ello que lo descrito en este párrafo se considera una fuerte limitación del estudio.

En cuanto a la relativa novedad de los estudios relacionados con el análisis de las evidencias de validez basadas en la estructura del modelo cognitivo, se puede decir que la principal limitación fue el escaso abanico de modelos componenciales que ayudaran a agilizar el análisis de las propiedades técnicas de validez de la estructura del modelo cognitivo elaborado por los expertos. El hecho de que el estudio de la calidad de los modelos componenciales es aún reciente presenta como limitación el que todavía algunos modelos se encuentran en etapa de consolidación y revisión, como en el caso del LSDM. Otra limitación asociada con los modelos componenciales, es que en la actualidad hay pocos programas computarizados de uso sencillo que ayuden en la estimación de los parámetros y estadísticos necesarios. Asimismo, para la aplicación de algunos modelos componenciales sólo hay algunas aplicaciones computarizadas que presentan cierto grado de complejidad, lo que hace inaccesible su uso para aquellos desarrolladores de pruebas e investigadores que adolezcan de su manejo.

También, otra limitación asociada con los modelos componenciales fue su aplicación. Por su parte, el modelo LLTM presentó algunas limitaciones relacionadas con: (a) la dificultad para cumplir con el requisito previo de unidimensionalidad y de ajuste, (b) la falta de una asíntota inferior y de un parámetro de discriminación a partir del contenido del ítem (Revuelta & Ponsoda, 1998), y (c) el carácter compensatorio lineal de la descomposición de los parámetros de dificultad (Romero, Ponsoda & Ximénez, 2006). Es por las dos últimas limitaciones mencionadas que se complementó el análisis componencial con el modelo LSDM, extendiéndose más el

proceso analítico de la validez de evidencias basadas en la estructura del modelo cognitivo. Sin embargo, dicha complementación también fue un beneficio, ya que se pudieron obtener evidencias de validez cruzada entre los dos modelos componenciales.

5.3. Recomendaciones para futuras investigaciones

Finalmente, se proponen recomendaciones para futuras investigaciones básicamente dirigidas a mejorar la aplicación del modelo teórico-metodológico aquí adaptado. En resumen, se plantea la necesidad de aplicar dicho modelo a todas las áreas del EXHCOBA y, en especial, a las áreas de la versión con base en la GAÍ. Además, se propone probar la aplicación del modelo teórico-metodológico en otras pruebas y en otros procesos de medición en donde se deseen atender los *estándares* de calidad relacionados con las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo. Dicho lo anterior, se presentan algunas recomendaciones puntuales para futuras investigaciones:

- Aplicar el criterio de saturación teórica para determinar la cantidad de participantes en el *análisis de protocolos* y, a su vez, tomar en cuenta tres niveles de pericia (novato, medio y experto). Lo anterior, con el fin de lograr la mejor representación de los procesos de respuesta de los examinados con la cantidad mínima requerida de participantes.
- Ampliar la cantidad de expertos de cada área de contenido a cuatro individuos, además de un psicólogo cognitivo que ayude en el ejercicio de análisis cognitivo de los ítems de las otras áreas del EXHCOBA.

- Aplicar el modelo teórico–metodológico adaptado y mejorado en todas las áreas del EXHCOBA para analizar las evidencias de validez basadas en el proceso de respuesta y en la estructura del modelo cognitivo.
- Implementar nuevos análisis componenciales con mayor adecuación a la GAÍRC del EXHCOBA que actualmente se aplica.
- Elaborar modelos cognitivos complejos de tipo multidireccional para cada uno de los ítems de las distintas áreas del EXHCOBA.
- Definir los atributos subyacentes a los ítems de cada una de las áreas del EXHCOBA para el análisis psicométrico componencial y la validez de la GAÍRC.
- Fundamentar el desarrollo de las versiones con ítems de opción múltiple e ítems de respuesta compleja del EXHCOBA en una teoría *fuerte*.
- Realizar análisis psicométricos con datos de crédito parcial en todas las áreas de la V-ÍRC del EXHCOBA.
- Realizar un estudio de evaluación diagnóstica con el fin de conocer las fortalezas y debilidades de los examinados ante las diferentes áreas del EXHCOBA.
- Implementar el modelo teórico–metodológico adaptado en otras pruebas y en otros procesos de medición con el fin de evaluar su aplicabilidad, su pertinencia y sus posibles aportes.

REFERENCIAS

American Educationai Research Association, American Psychological Association & National Council on Measurement in Education. (1985). *Standards for educational and psychological testing*. Washington: American Psychological Association.

American Educational Research Association [AERA], American Psychological Association [APA] & National Council on Measurement in Education [NCME]. (1999). *Standards for Educational and Psychological Testing*. Washington: AERA.

Anastasi, A. (1967). Psychology, psichologits, and psychological testing. *American Psychologist, 22*, 297-306.

Anastasi, A. (1986). Evolving concepts of test validation. *Annual Review of Psycholigy, 37*, 1-15.

Ayala, M., Ayala, C. & Shavelson, R. (2001). *On the Cognitive Interpretation of Performance Assessment Scores*. *Center for the Study of Evaluation. Technical Report 546*, Vol. 1522. Los Ángeles.

Anderso, J. (1976). *Languaje, memory, and thought.* Hillsdale: Lawrence Erlbaum.

Angoff, W. (1998). Validity: An evolving concept. En H. Wainer & H. Braum (Eds.), *Test validity*, pp. 9-13. Hillsdale: Lawrence Erlbaum.

Backhoff, E., Ibarra, M.A. y Rosas, M. (1995). Sistema Computarizado de Exámenes (SICODEX). *Revista Mexicana de Psicología, 12* (1), 55-62.

Backhoff, E. Ibarra, M.A. y Rosas, M. (1994, julio). *Versión Computarizada del Examen de Habilidades y Conocimientos Básicos*. Trabajo presentado en el 23vo Congreso Internacional de Psicología Aplicada. Madrid.

Backhoff, E. y Larrazolo, N. (2012, 7 de noviembre). *Evaluación de competencias escolares*. Trabajo presentado en el 1er Foro Iberoamericano de Evaluación Educativa de la Red Iberoamericana de Medición y Evaluación de Sistemas Educativos, Baja California. Consultado en http://www.ustream.tv/recorded/26785327

Backhoff, E., Aguilar, J. y Larrazolo, N. (2006). Metodologia para la Validacion de Contenidos de Examenes Normativos. *Revista mexicana de Psicología*, *23* (1), 79-86.

Backhoff, E. y Tirado, F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista de la Educación Superior*, XXI, *3* (83).

Backhoff, E. y Tirado, F. (1994). Estructura y lógica del Examen de Habilidades y Conocimientos Básicos. *Revista Sonorense de Psicología*, *8* (1), 21-33.

Backhoff, E., Tirado, F. y Larrazolo, N. (2001). Ponderación diferencial de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa, 3* (1). Consultado en http://redie.ens.uabc.mx/vol3no1/contenido-tirado.html

Backhoff, E., Larrazolo, N. y Rosas, M. (2000). Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Electrónica de Investigación Educativa*, *2* (1). Consultado en http://redie.uabc.mx/vol2no1/contenido-backhoff.html

Baddeley, A. (2006). Working memory: an overview. En *Pickering S. Working Memory and Education*, pp. 1-31. New York: Academic Press.

Bejar, I. (1990). A generative approach to the modeling of a three-dimensional spatial task. *Applied Psychological Measurement*, *14* (3), 137-245.

Bejar, I. (1993). A Generative Approach to Psychological and Educational Measurement. En N. Frederiksen, J. R. Mislevy, R.J. y. Bejar I. (eds.). *Test Theory for a New Generation of Tests*. Hillsdale: Lawrence Erlbaum Associates.

Bejar, I. (2002). Item generation: From conception to implementation. In S. H. Irvine and P. C. Kyllonen (Eds.), *Item heneration for test development*, pp. 199-218. Mahwah: Lawrence Erbaum Associates.

Bejar, I. (2010). Item Generation. Implications for a Validity Argument. In Gierl, Mark J.; Haladyna, Thomas M. (eds.) *Automatic Item Generation: Theory and Practice*, pp. 40-56. New York: Routledge.

Bejar, I. & Yocom, P. (1986). A generative approach to the development of hidden-figure items. *Research Report* Nº RR-150-531. Princeton: Educational Testing Service.

Borsboom, D. (2006). The attack of the psychometricians. *Psychometrika, 71,* 425–440.

Borsboom, D. & Mellenberg, G. (2007). Test validity in cognitive assessment. In Leighton, J. & Griel, M. (Edit.), *Cognitive diagnostic assessment for education: Theory and applications*, pp. 85-118. Cambridge: Cambridge University Press.

Borsboom, D., Mellenbergh, G. & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111,* 1061–1071.

Brown, J. & Burton, R. (1978). Diagnostic models for procedural bugs in basic mathematical skills. *Cognitive Science, 2*, 155-192.

Campbell, J. (1960). Recommendation for APA test standards regarding construct, trait, or discriminant validity. *American Psychologist, 15,* 546-553.

Campbell, J. (1976). Psychometric theory. In M.D. Dunnette (Ed.), *Handbook of industrial and organizational psychology*, pp. 185-222. Chicago: Rand McNally.

Chi, M.T.H. (1997). Quantifying qualitative analyses of verbal data: A practical guide. *The Journal of the Learning Sciences, 6*, 271-315.

Chi, M., Glaser, R, & Farr, M. (1988). *The nature of expertise.* Hillsdale: Earlbaum.

Chen, Y. & Macdonald, G. (2011). Validating Cognitive Sources of Mathematics Item Difficulty: Application of the LLTM to Fraction Conceptual Items. *Psychological Assessment*, *7*, 74–93.

Colins, A. & Lofttus, E. (1975). A spreading-action theory of sementic processing. *Psychological Review, 82,* 407-428.

Collis, J., Tapsfield, P., Irvine, S., Dann, P. & Wright D. (1995). The British Army Recruit Battery Goes Operational: from Theory to Practice in Computer-Based Testing Using Item-Generation Techniques. *International Journal of Selectio and Assessment*, *3* (2), 96-104.

Contreras, L.A. (2000). *Desarrollo y Pilotaje de un Examen de Español para la Educación Primaria en Baja California.* Tesis doctoral. Ensenada: Instituto de Investigación y Desarrollo Educativo/UABC.

CRESST (1994). *Assesstment Profile-State Summary. Evaluation Comment*. National Center for Research on Evaluation, Standars and Student Testing.

Conbrach, L. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational Measurement*, 2da ed., pp. 443-507. Washington: American Council on Education.

Crocker, L. y Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston.

Cronbach, L. (1990). *Essentials of psychological testing*. Nueva York: Harper Collins Publishers.

Cronbach, L. y Meehl, P. (1955). Construct validity in psychological test. *Psychological Bulletin*, *52*, 281-302.

Cureton, E. (1951). Validity. In E. F. Lindquist (Ed.), *Educational measurement*, pp. 621-694. Washington: American Council on Education.

De la Torre, J. & Douglas, J. (2004). Higher-order latent trait models for cognitive diagnosis. *Psychometrika*, *69*, 333–353.

De la Torre, J. (2008a). An Empirically Based Method of Q-Matrix Validation for the DINA Model: Development and Applications, *Journal of Educational Measurement, 45(4), 343- 362.*

De la Torre, J. (2008b). DINA model and parameter estimation: a didactic. *Journal of Educational and Behavioral Statistics*, *34*, 115–130.

Dehn, N. & Schank, R. (1982). Artificial and human intelligence. In R. J. Sternberg (Ed.), *Hanbook of human intelligence*, pp. 352-391. New York: Cambridge University Press.

DiBello, L. V., Stout, W. F. & Roussos, L. A. (1995). Unified cognitive/psychometric diagnostic assessment likelihood-based classification techniques. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment,* pp. 361-389. Hillsdale: Lawrence Erlbaum Associates.

Dimitrov, D. M. (2007). Least squares distance method of cognitive validation and analysis for binary items using their item response theory parameters. *Applied Psychological Measurement, 31*(5), 367-387.

Dimitrov, D., Romero, S., Ponsoda, V. & Ximénez, C. (2006, October). *Psychometric analysis of cognitive operations underlying student performance on basic arithmetic operations: An application of the Least Squares Distance Method.* Paper presented at the 2006 Annual Meeting of the Mid-Western Educational Research Association (MWERA), Columbus.

Draney, K., Pirolli, P. & Wilson, M. (1995). A measurement model for a complex cognitive skill. In P. D. Nichols, S. F. Chipman, R. L. Brennan (Eds.), *Cognitively diagnostic assessment*, pp. 103–125. Hillsdale: Erlbaum.

Drasgow, F., Luecht, R. & Bennett, R. (2006). Technology and testing. In R. L. Brenan (Edit.), *Educational measurement* (4ta ed.), pp. 471-516. Washington: American Council on Education.

Dunker, K. (1945). Problem solving. *Psichological Monograps*. *53* (5, 270).

Ebel, R. (1961). Must all tests be valid? *American Psychologist*, *16*, 640-647.

Embretson, S. (1983). Construct validity: Construct representation versus nomothetic span. *Psychological Bulletin*, *93*, 179–197.

Embretson, S. (1984). A general latent trait model for response processes. *Psychometrika, 49*(2), 175-186.

Embretson, S. (Ed.). (1985). *Test design: Developments in psychology and psychometrics.* Orlando: Academic Press.

Embretson, S. (1994). Applications of Cognitive Design Systems to Test Development. In C. R. Reynolds (Ed.), *Cognitive Assessment A Multidisciplinary Perspective* (Plenum Pre.). New York.

Embretson, S. & Reise, S. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Embretson, S. & Wetzel, C. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement, 11(2), 175-193.*

Embretson, S. (1995). Developments toward a cognitive design system for psychological and educational tests. In D. Lubinsky and R. Dawis (Eds.), *Assessing individual differences in human behavior: New concepts, methods andfindings,* pp. 17-48. Palo Alto: Consulting Psychologist Press.

Embretson, S. (1998). A cognitive design system approach to generating valid tests: Application to abstract reasoning. *Psychological Methods, 3(3), 380-396*

Embretson, S. (1999). Cognitive psychology applied to testing. In F. T. Durso (Ed.), *Handbook of applied cognition,* pp. 629-660. New York: John Wiley & Sons Ltd.

Ericsson, K.A. (2006). Protocol analysis and expert thought: concurrent verbalizations of thinking during experts' performance on representative tasks. In K.A. Ericsson, N. Charness, P.J. Feltovich, R.R. Hoffman (Eds.), *The Cambridge handbook of expertise and expert performance,* pp. 223-241. Cambridge: Cambridge University Press.

Embretson, S. & Gorin, J. (2001). Improving Construct Validity With Cognitive Psychology Principles. *Journal of Educational Measurement. 4,* pp. 343-368.

Ericsson, K. & Charness, N. (1994). Expert performance, its structure and acquisition. *American Psychologist, 49,* 725-747.

Ericsson, K. & Simon, H. (1984). *Protocol analisys: verbal reports as data. Cambridge: MIT Press.*

Ericsson, K. & Simon, H. (1993). *Protocol Analysis*. Cambridge: The MIT Press.

Ericsson, K. & Simon, H. (1998). How to study thinking in everyday life: Contrasting think-aloud protocols with descriptions and explanations of thinking. *Mind, Culture & Activity*, *5*(3), 178-186.

Ferreyra, F. (2013). *Modelo para la validación empírica del EXHCOBA-R, producido por un generador automático de reactivo*. Tesis de doctorado. Baja California: Instituto de Investigación y Desarrollo Educativo/UABC.

Fischer, G. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica, 37*(6), 359-374.

Fischer, G. (1995). The linear logistic test model. En G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications,* pp. 131–155. New York: Springer-Verlag.

Fischer, G. & Molenaar, I. (1995). *Rasch models. Foundations, recent developments and applications*. New York: Springer-Verlag.

Fischer, G. & Ponocny-Seliger, E. (1998). *Structural Rasch modeling. handbook of the usage of PLCM-WIN 1.0*. Groningen, the Netherlands: proGAMMA.

Fraser, C. (1988). *NOHARM: Computer software and manual.* Australia: Author.

Fraser, C., McDonald, R. & *Vandermeulen*, M. (2012). *NOHARM. Free trial.* Australia: Author. Descargado en http://noharm.niagararesearch.ca/nhweb.html

Fredericksen, N., Mislevy, R. & Bejar, I. (Eds.) (1993). *Test theory for a new generation of tests*. Hillsdale: LEA.

Fredericksen, J. (1980). Component skills in Reading: measurements of individual diferences thought chronometric analisys. In R. E. Snow, P-A. Federico & W. E. Montage (Eds.), *Aptitude, learning, and instructions: Cognitive process analyses of aptitude, Vol. 1,* (pp. 105-138). Hillsdale: Lawrence Erlbaum.

Gierl, M. J., Wang, C., & Zhou, J. (2008). *Using the Attribute Hierarchy Method to make diagnostic inferences about examinees' cognitive skills in algebra on the SAT©*. Journal of Technology, Learning, and Assessment, 6 (6). Retrieved [date] from http://www.jtla.org.

Gierl, M. y Lai, H. (2012): The Role of Item Models in Automatic Item Generation, *International Journal of Testing*, *12*(3), 273-298.

Griel, M. & Lai, H. (2013). Using weak and strong theory to create ítem models for atomatic ítem generation. Some practical guidelines with examples. In Gierl, M. y Haladyna, T. (Edit.). *Automatic item generation: Theory and practice* (pp. 26-39). New York: Routledge.

Gierl, M., Lai, H. & Turner, S. (2012). Using automatic item generation to create multiple-choice items for assessments in medical education. *Medical Education, 46,* 757-765.

Gierl, M., Leighton, J., Changjiang, W., Jiawen, Z., Rebecca, G. & Tan, A. (2009). *Validating Cognitive Models of Task Performance in Algebra on the SAT. Research Report 2009-3*. *College Board,* Research Report, 2009(3). New York.

Gierl, M., Tan, X. & Wang, Ch. (2005*). Identifying content and cognitive dimensions on the SAT*. Research Report, 2005(11). College Board.

Gierl, M., Wang, C. & Zhou, J. (2008). Using the Attribute Hierarchy Method to Make Diagnostic Inferences about Examinees' Cognitive Skills in Algebra on the SAT. *The Journal of Technology, Learning, and Assessment*, *6*(6).

Gierl, M. Zhou, J. & Alves, C. (2008). Developing a Taxonomy of Item Model Types to Promote Assessment Engineering. *The Journal of Technology, Learning, and Assessment*, *7*(2), 1-51. Recuperado en http://www.jtla.org.

González, M. (2004). *La Definición y Medición de Estándares Académicos para la Educación Superior: un estudio Formativo en la Universidad de Sonora*. Disertación Doctoral. Tucson: Departamento de Psicología Educativa, Universidad de Arizona.

Gorin, J. (2007). Test Construction and Diagnostic Testing. In Leighton, J. y Griel, M. (Edit.). *Cognitive diagnostic assessment for education: Theory and applications*, pp. 85-118. Cambridge: Cambridge University Press.

Gorin, J. & Embretson, S. (2013). Using Cognitive Psychology to generate Items and Predict Item Characteristics. In Gierl, M. y Haladyna, T. (edit.). *Automatic Item Generation: Theory and Practice*, pp. 40-56. New York: Taylor and Francis Group.

Guion, R. (1977). Content validity: The source of my discontent. *Applied Psychological Measurement, 1*, 1-10.

GUN (s.f.). *R. Free software*. Descargado en http://cran.r-project.org/bin/windows/base/

Haertel, E. (1989). Using restricted latent class models to map the skill structure of achievement items. *Journal of Educational Measurement, 26*(4), 301-321.

Hakel, M. (Edit.) (1998). *Beyond multiple choice: evaluating alternatives to traditional testing for selection*. NewJersey: LEA.

Haladyna, T. (1999). *Developing and validating multiple-choice test items*. Hillsdale: LEA.

Haladyna, T. Downing, S. M. & Rodríguez, M. C. (2002). A review of multiple-choice item writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309–334.

Hogaboam, T. & Pellegrino, J. (1978). Hunting for individual differences incognitive processes: Verbal ability and semantic processing of pictures and words. *Memory & Cognition, 6*(2), 189-193.

Hoppmann, T. (2007). Examining the “point of frustration”. The think-aloud method applied to online search tasks. *Quality Quantity*, *43*(2), 211–224.

Hornke, L. & Habon, M. (1986). Rule Based Item Bank Construction and Evaluation with the Linear Logistic Framework. *Applied Psychological Measurement*, *10*(4), 369-380.

Huff, K., Alves, C., Pellegrino, J. & Kaliski, P. (2013). Using Evidence-Centered Design Task Models in Automatic Item Generation. In Gierl, Mark J.; Haladyna, Thomas M. (eds.) *Automatic Item Generation: Theory and Practice*, pp. 40-45. New York: Routledge.

Ihaka, R. & Gentleman, R. (1996). R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics 5*, 299–314.

Johnstone, C. (2003). *Improving validity of large-scale tests: Universal design and student performance* (Technical Report 37). Minneapolis: National Center on Educational Outcomes.

Johnstone, C., Bottsford-Miller, N. & Thompson, S. (2006). *Using the think aloud method (cognitive labs) to evaulate test design for students with disabilities and English language learners* (Technical Report 44). Minneapolis: University of Minnesota, National Center on Educational Outcomes.

Junker, B. & Sijtsma, K. (2001). Cognitive assessment models with few assumptions, and connections with nonparametric item response theory. *Applied Psychological Measurement, 25(3), 258-272.*

Kane, M. (1992). An argument-based approach to validation. *Psychological Bulletin, 112,* 527–535.

Kane, M. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319-342.

Kane, M. (2006). Validation. In R.L. Brennan (Ed.), *Educational measurement* (4ta. ed.), pp.17-64. Wesport: National Council on Measurement in Education and American Council on Education.

Kane, M. (2008). Terminology, Emphasis, and Utility in Validation. *American Educationai Research Association*, *37*(2), 76–82.

Leighton, J. (2004). Avoiding misconception, misuse, and missed opportunities: The collection of verbal reports in educational achievement testing. *Educational Measurement: Issues and Practice, 23*, 6–15.

Leighton, J. (2009). Two Types of Think Aloud Interviews for Educational Measurement: Protocol and Verbal Analysis Paper presented for symposium *How to Build a Cognitive Model for Educational Assessments* at the 2009 annual meeting of the National Council on Measurement in Education (NCME), April, 14-16.

Leighton, J., Cui, Y. & Cor, M. (2009). Testing Expert-Based and Student-Based Cognitive Models: An Application of the Attribute Hierarchy Method and Hierarchy Consistency Index. *Applied Measurement in Education*, *22*(3), 229-254

Leighton, J. & Gierl, M. (en prensa). *Defining and evaluating models of cognition used in educational measurement to make inferences about examinees'thinking processes*. Educational Measurement: Issues and Practice.

Leighton, J. & Gierl, M. (2004). The attribute hierarchy method for cognitive assessment: A variation on Tatsuoka's rule-space aproach. *Journal of Educational Measurement, 41*, 205-236.

Leighton, J. & Gierl, M. (edit.). (2007a). *Cognitive diagnostic assessment for education: Theory and applications,* pp. 146-172. Cambridge: Cambridge University Press.

Leighton, J. & Gierl, M. (2007b). verbal reports as data for cognitive diagnostic assessment. In Leighton, J. & Gierl, M. (edit.). *Cognitive diagnostic assessment for education: Theory and applications,* pp. 146-172. Cambridge: Cambridge University Press.

Li, M. (n.d.). *Linking Assessment to Science Achievement: A Knowledge-Based Approach Project Report Submitted to NSF By*. Washington: University of Washington.

Linn, R. L. (Ed.). (1989). *Educational measurement* (3a. ed.). Washington: American Council on Education.

Lopez, J. (2005). Ítems politómicos *vs.* dicotómicos: Un estudio metodológico. *Anales de Psicologia, 21*, 339-344.

Loevinger, J. (1957). objetive testas instruments of psychological theory. *Psychological Reports, 3,* 635-694 (Monograph Supp. 9).

Loye, N. (2008). *Elaborer la matrice Q de modeles cognitifs dans diverses conditions et definir leur impact sur sa valídate et sa fidelite* (Desarrollo de la matriz Q para evidenciar el modelo cognitivo implícito en diferentes condiciones y evaluando su impacto en la validez y fidelidad de las evaluaciones). Tesis doctoral no publicada, Otawa: University of Otawa.

Luecht, R. (2008, October). *Assessment engineering in test design, development, assembly, and scoring.* Invited keynote address at the Annual Meeting of the East Coast Organization of Language Testers (ECOLT), Washington, D.C.

Ma, L. Çetin, E. y Green, K. (2009, April). *Cognitive assessment in Mathematics with the Least Squares Distance Method*. Articulo presentado en el Congreso anual de la AERA 2009. San Diego.

Macready, G. & Dayton, C. (1977). The use of probabilistic models in the assessment of mastery. *Journal of Educational Statistics, 2*, 99-120.

Maris, E. (1999) Estimating multiple classification latent class models. *Psychometrika, 64* 187-212.

Maris, E. (1995). Psichometric latent response models. *Psychometrika, 60*, 523-547.

Martínez, F. (2001). *Evaluación Educativa y Pruebas Estandarizadas. Elementos para Enriquecer el Debate. Revista de Educación Superior*, XXX, *4*(120), 71-85.

Martínez, F., Backhoff, E., Castañeda S., De la Orden, A., Schmelkes, S., Solano-Flores, G.,Tristán, A. & Vidal, R. (2000) *Estándares de calidad para instrumentos de evaluación educativa*. México: CENEVAL.

Messick, S. (1980). Test validity and the ethics of assessment. *American Psychologist, 35,* 1012-1027.

Messick, S. (1988). The once and the future issues of validity. Assesing the meaning and consequences of measurement. In H. Wainer & H. Braun (Eds.), *Test validity*, pp. 33-45. Hillsdale: Laurance Erlbaum.

Messick, S. (1989a). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher, 18*(2)*,* 5-11.

Messick, S. (1989b). Validity. En R. L. Linn (Ed.), *Educational measurement* (3a. ed.), pp. 13-103. New York: Macmillan Publishing Co.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Michell, J. (1999). *Measurement in Psychology: A critical history of a methodological concept*. Cambridge**:** Cambridge University Press.

Mislevy, R. J. & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55*(2), 195-215.

Mislevy, R. (1994). Evidence and inference in educational assessment. *Psychometrika, 59*, 439-483.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment,* pp. 43-71. Hillsdale: Lawrence Erlbaum Associates.

Mislevy, R., Steinberg, .S. & Almond, R. (2002). Design and analisys in task-based language assessment. *Languaje Testing. Special issue: Interpretatios, intended uses, and desings in task-based language, 19*(4), 477-496.

Mislevy, R. (2007). Cognitive psychology and educational assessment. En R. L. Brennan (Ed.), *Educational measurement (*4a. ed.), pp. 257-305. Portsmouth: Grenwood.

Mislevy, R. (2009). *Validity from the model-based reasoning*. Report 752. U. of Maryland: Ed. CRESST.

Moss, P. (1998). Shisting conceptions of validity in educational measurement: Implications for performance assessment. *Review of Educational Researcher, 23*, 5-12.

National Research Council. (2001). *Knowing what students know: The science and design of educational assessment*. (R. Committee on the Foundations of Assessment. Pelligrino, J., Chudowsky, N., and Glaser, Ed.), *Social Sciences*. Washington: National Academy Press.

Newell, A., Shaw, J. y Simon, H. (1957). Problem solving in humans and computers. *Carnegie technical*, *21*(4), pp. 34-38.

Newell, A. & Simon, H. (1972). *Human problem solving*. Englewood Cliffs: Prentice Hall.

Nichols, P. (1994). A framework for developing cognitively diagnostic assessments. *Review of Educational Research, 64*(4), 575-603.

Nichols, P. (1994b). A guide for developing cognitively diagnostic assessments Review of Educational Research, 64, 575-603.

Nichols, P., Chipman, S. y Brenan, R. (1995). *Cognitively diagnostic assessment*. Hillsdale: Lawrence Erlbaum Associates.

Nielson, J. (1994). Estimating the number of subjects needed for a thinking aloud Test. *International Journal of Human-Computer Studies, 41*(3), 385–397.

Pellegrino, J., Baxter, G. & Glaser, R. (1999). Addressing the "two disciplines" problem: Linking theories of cognition and learning with assessment and instructional practice. In A. Iran-Nejad and P.D. Pearson (Eds.), *Review of research in education* (*Volume 24*, 307-353. Washington: American Educational Research Association.

Pellegrino, J. & Glaser, R. (1979). Cognitive correlates and components in the analysis of individuals diference. *Intelligence, 3,* pp. 187-214.

Pellegrino, J., Chudowsky, N. & Glaser, R. (Eds.). (2001). *Knowing what students know: The science and design of educational assessment*. Washington: National Academy Press.

Pellegrino, J. (2010). *The Design of an Assessment System for the Race to the Top: A Learning Sciences Perspective on Issues of Growth and Measurement*. Illinois: Educational Testing Service (ETS).

Pérez, J., Larrazolo, R., Backhoff, E. y Rojas, D. (2013). Análisis de la estructura cognitiva del área de habilidades cuantitativas del EXHCOBA mediante el modelo LLTM de Fisher. *The international Journal of Learning* (en proceso de publicación).

Powell, M. (1990). *Performance assessment: panacea or pandora's box?* Rockville: Montgomery Cunty Public Schools.

Posner, M. I. (1978). *Chronometric exploration of mind.* New York: Jhon Wiley.

Reid, D.; W. Hresko & H. Swanson (1996). *Enfoques Cognitivos del Aprendizaje*. Austin: Pro-ed.

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema*, *10*, 753-760.

Romero, S. (2010). *Propiedades y aplicaciones de las distancias mínimo-cuadráticas (LSDM) para la validación y análisis de atributos cognitivos*. Tesis doctoral. Madrid: UAM.

Romero, S., Ordoñez, X., López, E., y Navarro, E. (2009). Análisis de la estructura cognitiva de la competencia científica en PISA 2006 mediante el LSDM: el caso español. *Psicothema*, 21, 509-514.

Romero, S., Ponsoda, V., y Ximenez, C. (2008). Análisis de un test de aritmética mediante el modelo logístico lineal de rasgo latente 1. *Revista Latinoamericana de Psicologia*, *40*, 85–95.

Rosas, M., Ramírez, J. y Larrazolo, R. (2009). Examen de selección: sistema computarizado de exámenes SICODEX versión 3, *X Congreso Nacional de Investigación Educativa.*

Rost, J. (1990). Rasch models in latent classes: An integration of two approaches to item analysis. *Applied Psychological Measurement, 14*(3), 271-282.

Roussos, L., DiBello, L., Stout, W., Hartz, S., Henson, R. y Templin J. (2007). The Fusion Model Skills Diagnosis System. In Leighton, J. & Gierl, M. (2007a). *Cognitive diagnostic assessment for education: Theory and applications.* Cambridge: Cambridge University Press.

Rumelhart, D. (1980). Schemata: Thebuilding blocks of cognition. In R.J.Spiro, B.C. Bruce, & W. F. Brewer (Eds.), *Theoretical issues in reading comprehension*, pp. 33-57. Hilsdale: Erlbaum.

Rupp, A. (2007). The answer is in the question: A guide for describing and investigating the conceptual foundations and statistical properties of cognitive psychometric models. *International Journal of Testing, 7*(2), 95-125.

Rupp, A. & Mislevy, R. (2007). Cognitive foundations of structured item response models. En J. Leighton & M. Gierl (Eds.), *Cognitive diagnostic assessment for education: Theory and applications,* pp. 205-241). New York: Cambridge University Press.

Rupp, A. & Mislevy, R. (2007). Cognitive foundations of structured ítem responses theory models. In J. Leighton y M. J. Gierl (Eds.), *Cognitibely diagnostic assessment for education: Theory and applications,* pp. 205-241). Cambridge: Cambridge University Press.

Rupp, A., Templin, J. & Henson, R. (2010). *Diagnostic measurement: Theory, methods, and applications.* New York: The Guilford Press.

Rupp, A., Vock, M., Harsch, C. & Koller, O. (2008). *Developing standars-based assessment items for English as a first foreing language: Context, processes, and outcomes in Germany*. Munster: Waxmann.

Shepard, L. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.) *Review of research in education*, Vol. 19, 405-450. Washington: American Educational Research Association.

Shye, S., Elizur, D., y Hoffman, M. (1994). *Introduction to facet theory.* Thousand Oaks, CA: Sage Publishers.

Sireci, S. (2008, Octubre). Packing and unpacking sources of validity evidence: History repeats itself again. Paper presented at the conference *The Concept of Validity: Revisions, New Directions and Applications*, University of Maryland, College Park, MD.

Snow, R. & Lohman, D. (1989). Implications of cognitive psychology for educational measurement. In R. L. Linn (Ed.), *Educational measurement* (3a. ed.), pp. 263-331. New York: Macmillan Publishing Co.

Sternberg, R. (1977). *Intelligence, information processing, and analogical reasoning: The componential analysis of human abilities*. Oxford: Lawrence Erlbaum.

Sternberg, R. (2007). *Cognitive Psychology*. New York: The Guilford Press.

Tanaka, J. y Huba, G. (1985). A fit Index for covariance structure models under arbitrary GLS estimation. *British Journalof Mathematical and Statistical Psychology, 38*, 197-201.

Tatsuoka, K. (1983). Rule space: An approach for dealing with misconceptions based on item response theory. *Journal of Educational Measurement, 20(4), 345-354.*

Tatsuoka, K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. En P. D. Nichols, S. F. Chipman & R. L. Brennan (Eds.), *Cognitively diagnostic assessment*. (pp. 327-359). Hillsdale: Lawrence Erlbaum Associates.

Tatsuoka, K. (2009). *Cognitive Assessment an introduction to the Rule Space Method*. (Taylor & Francis Group, Ed.) *Structural Equation Modeling* (1ra ed.). New York.

Tatsuoka, K. (2009). *Cognitive assessment: An introduction to the rule-space method.* Florence: Routledge.

TechSmith. (s.f.). *CAMTASIA STUDIO. Free trial.* Michigan: Author. Descargado en http://www.techsmith.com/download/camtasia/

Templin, J. & Henson, R. (2006). Measurement of psychological disorders using cognitive diagnosis models. *Psychological Methods, 11*, 287-305.

Tirado, F. (1986). La Crítica Situación de la Educación Básica en México. *Ciencia y Desarrollo*. México: Consejo Nacional de Ciencia y Tecnología, *71*, 81-94.

Tomkins, S. & Messick, S. (Eds.). (1963). *Computer simulation of personality: Frontier of psychological theory*. New York: Jhon Wiley.

Thompson, S., Johnstone, C. & Thurlow, M.. (2002). *Universal design applied to large scale assessments* (Synthesis Report 44). Minneapolis, MN: National Center on Educational Outcomes.

Tirado, F. (2010, 7 de abril). *Reactivos estructurales constructivos*. Ponencia en la segunda reunión del consejo consultivo del EXHCOBA realizada en el Instituto de Investigación y Desarrollo Educativo/UABC.

Tirado, F., Backhoff, E., Larrazolo, N. y Rosas, M, (1997). Validez predictiva del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista Mexicana de Investigación Educativa*. *11*(3), 67-84.

van Lehn, K. (1989). Problem solving and cognitive skill acquisition. En M. I. Posner (Ed.), *Foundations of cognitive science*, pp. 527-579. Cambridge: The MIT Press.

van der Linden, W. y Glass, C. (Eds.) (2000). *Computer-adaptive testing: theory and practice.* Boston: Kluwer Academic Publishers.

van der Linden, W. y Hambleton, R. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

Wainer, H. (Ed.) (1990). *Computer adaptive testing: a primer*. Hillsdale: LEA.

Williamson, D. Mislevy, R. & Bejar, I. (Eds.). (2006). *Automated Scoring of complex performances in computer based testing.* Mahwah: Lawrence Erlbaum Associates.

Williamson, D. M., Johnson, M. S., Sinharay, S. & Bejar, I. (2002). *Hierarchical IRT examination of isomorphic equivalence of complex constructed response tasks.* Paper presented at the American Educational Research Association, New Orleans, LA.

Yang, X. & Embretson, S. (2007). Construct Validity and Cognitivy Diagnostic Assesment. In Leighton, J. y Griel, M. (Edit.). *Cognitive diagnostic assessment for education: Theory and applications*, pp. 85-118. Cambridge: Cambrige University Press.

Xu, X.,& von Davier, M (2008a).*Fitting the structured general diagnostic model to NAEP data (RR-08-27)*. Princeton: Educational Testing Service.

Xu, X.,& von Davier, M (2008b).*Liking for the general diagnostic model (RR-08-08)*. Princeton: Educational Testing Service.

APÉNDICES

Apéndice 1. Formato de especificación o del modelo para la GAÍRC del EXHCOBA

I. Datos de la elaboración

Revisores:	En este apartado se identifica a los especialistas que revisaran el avance y desarrollo del modelo para la GAÍRC del EXHCOBA
Desarrollador de la especificación:	**Fecha de redacción inicial**
En este apartado se identifica al o los desarrolladores de los modelos de la GAÍRC que en específico para el caso del EXHCOBA son profesores en el área del currículum en desarrollo capacitados en la elaboración dichos modelos.	-o-
Revisiones y correcciones	**Fechas de envío**
-o-	-o-

II. Datos de identificación del contenido a evaluar (ejemplo ficticio)

Asignatura o Área		**Nivel educativo**	
Ciencias naturales		Segundo de primaria	
Clave	**Eje temático/Ámbito/Asignatura**	**Tema**	**Subtema**
CsN-5	Los seres vivos	Animales ovíparos y vivíparos	-o-
Contenido	**Nombre**	Ejemplos de animales vivíparos y vivíparos	
	Definición	Clasificación de animales según su viviparidad e identificación de las características animales vivíparos así como de sus diferencias.	

III. Características del contenido a evaluar

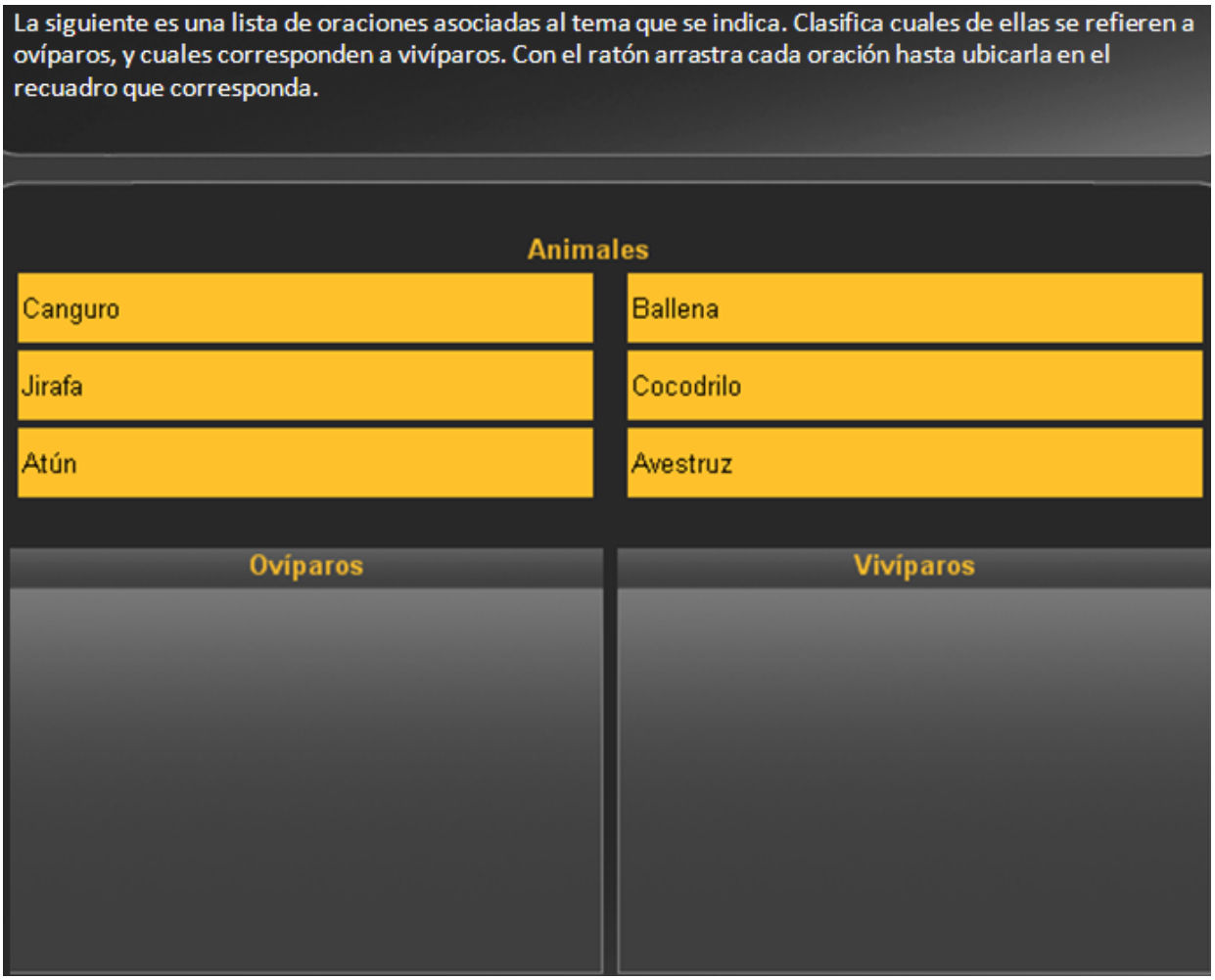
Importancia (justificación) del contenido a evaluar
En este apartado se justifica la selección del contenido elegido del currículum según su importancia en el contexto del plan de estudios.
Delimitación del contenido
En este apartado se delimita de mejor forma el contenido seleccionado en caso de estar descrito de forma ambigua en el plan de estudios. También se describen las características del contenido a evaluar pensando en la elaboración del modelo del ítem a generar.
Conocimientos y habilidades involucrados en la solución correcta del reactivo
Operaciones cognitivas que predicen la dificultad del ítem: (en este rubro, con ayuda de expertos se definen los atributos subyacentes al contenido curricular y que deberán verse reflejados en la tarea evaluativa).

IV. Especificaciones del reactivo

Estrategia de evaluación

En este apartado se describen las características del modelo de ítem en donde se deberán tomar en cuenta la representatividad del constructo y de las características de los atributos y dominios a medir.

Ejemplo del ítem generado



Datos para el programador

Procedimiento de captura: (en este rubro se especifican las actividades de edición del reactivo a realizar en el editor del EXHCOBA)

1. Elegir el tipo de plantilla del editor para categorías y elementos;
2. elegir la plantilla adecuada (posiblemente imagen área) para el caso de presentar solo tres cajas en una línea de tiempo;
3. alimentar el contenido del reactivo en los espacios determinados para la instrucción, los títulos de las categorías y las respuestas correctas;
4. determinar como regla de operación, que en el reactivo se presenten a la izquierda de la pantalla (ver ejemplo) solo cinco oraciones elegidas de manera aleatoria del banco de respuestas correctas previamente alimentado y;

5. establecer la posibilidad de que se presenten oraciones para solo una de las tres categorías del reactivo.
Elementos de cada categoría:

Animales vivíparos	**Animales ovíparos**
R1.1 Humano R1.2 Ballena R1.3 Oso Rn...	R2.1 Murciélago R2.2 Cocodrilo R2.3 Avestruz Rn...

Referencia bibliográfica
En este rubro se colocan las referencias de donde se obtuvo la información del contenido del ítem a desarrollar
Observaciones:
(R…) Elemento para proceso aleatorio

Apéndice 2. Guía de procedimientos y materiales para el análisis de protocolos del Examen de Habilidades y Conocimientos Básicos (EXHCOBA).

1er paso: Presentación

-Breve presentación personal por parte del investigador al participante.

-Breve presentación del EXHCOBA y del estudio a realizar (propósito y características generales).

-Descripción al participante de las actividades que desempeñará a lo largo del estudio. Se debe aclarar al participante que será grabada su voz y sus acciones en la pantalla.

2do paso: Firma del consentimiento informado y captura de los datos de identificación

-Entrega, lectura y en su caso firma del consentimiento informado.
-Llenado del formato de identificación.

3er paso: Revisión del laboratorio cognitivo y de los materiales

Lugar y espacio de aplicación	**Instrumental técnico**
___ Espacio adecuado ___ Mesa ___ Sillas	___ Computadora ___ Cargador eléctrico ___ Mouse ___ Micrófono ___ Router o conexión inalámbrica a internet ___ Cable conector al Router
Programas de computo	**Papelería**
___ Conexión a internet activada ___ Editor de reactivos activado ___ Camtasia Studio activado ___ Word office activado ___ Calculadora activada	___ Lista de participantes ___ Protocolos de evaluación ___ Consentimiento informado ___ Tarjeta de presentación ___ Hojas blancas ___ Lápiz ___ Sacapuntas

4to paso: Presentación a modo de guía de los tipos de reactivos

-Presentación a modo de guía de los tipos de reactivos según su formato de respuesta:

1.Imagen área
2. Español acentos
3. Categorías y elementos
4. Orden párrafos
5. Formula valores tabla
6. Grupo ecuaciones
7. Triángulos
8. Recta
9. Tabla valores
10. Dobles
11. Sucesiones
12. Orden valores
13. Volumen prisma
14. Seleccionar frases
15. Orden eventos
16. Orden etiquetas
17. Fracciones figuras
18. Graficas encuestas
19. Funciones plano

5to paso: Entrenamiento para la técnica de pensamiento en voz alta

-Practica general de entrenamiento que permita al participante familiarizarse con la técnica de "pensamiento en voz alta". Pueden utilizarse como ejemplos y guías las siguiente instrucciones:

"En este estudio estamos interesados en como piensas la solución de problemas en ítems de un examen computarizado, en este caso del EXHCOBA. Para ello, de forma ordenada voy a presentarte un conjunto de ítems y te pediré que pienses en voz alta la forma en cómo lo resuelves. Cuando te pido que pienses en voz alta, significa que quiero que me digas todo lo que pasa por tu mente desde el primer momento en que veas la pregunta del examen y hasta el final cuando llegues a su respuesta. Yo procuraré que hables en voz alta constantemente mientras resuelves el problema en el que te encuentres trabajando. Yo no quiero que me digas o trates de explicar cómo respondes o solucionas el ítem. Solo actúa como si estuvieras sola en este cuarto y hablando para ti mismo(a). No olvides que es muy importante que te mantengas hablando. Si en algún momento guardas silencio por algún periodo de tiempo de más de tres segundos, te pediré que vuelvas a pensar en voz alta.
¿Hasta el momento todo lo que te comenté es claro?
Muy bien, ahora vamos a iniciar con algunos problemas de práctica. Primero, voy a pedirte que multipliques dos números en tu cabeza. Así que piensa en voz alta como multiplicarías 24 por 34.
Muy bien, ahora quiero que multipliques dos números en silencio y me digas después cual fue tu pensamiento con el llegaste a la respuesta.
¿Cuál es el resultado de multiplicar 24 por 36?

-Práctica de ejemplo que permita al participante familiarizarse con la técnica de "pensamiento en voz alta" aplicada a un ítem en específico. Primero el investigador debe resolver un problema frente al participante para que pueda observar el comportamiento y desempeño que se espera de él. Después se pedirá al participante resolver uno (La cámara o dispositivo de grabación no debe estar prendido para la práctica de entrenamiento). Pueden utilizarse como ejemplos y guías las siguiente instrucciones:

Muy bien, ahora voy a pensar en voz alta mientras resuelvo este problema. Eso significa que voy a decir todo lo que pasa por mi mente. (Complete el problema mientras piensa en voz alta.)
"Ahora te voy a pedirte que resuelvas un problema del mismo modo. Solo di todo lo que pasa por tu mente mientras resuelves el problema."

6to paso: Entrenamiento para el seguimiento del indicador del mouse

-Práctica de ejemplo que permita al participante familiarizarse con la técnica de "pensamiento en voz alta" en asociación con el seguimiento del indicador del mouse. Primero el investigador debe resolver un problema frente al participante para que pueda observar el comportamiento y desempeño que se espera de él. Después se pedirá al participante resolver uno (La cámara o dispositivo de grabación no debe estar prendido para la práctica de entrenamiento).
-Pueden utilizarse como ejemplos y guías las siguiente instrucciones:

"Muy bien, ahora habrá un ligero cambio a la práctica anterior. Observa detenidamente, voy a pensar en voz alta y a seguir con la vista el indicador del mouse mientras resuelvo este problema. Eso significa que voy a decir todo lo que pasa por mi mente y que a donde se dirija mi vista dirigiré el indicador del mouse con la mayor exactitud posible." (Complete el problema mientras piensa en voz alta y dirige el indicador del mouse a donde se encuentra su vista.)
"Ahora te voy a pedir que resuelvas un problema del mismo modo. Solo di todo lo que pasa por tu mente mientras resuelves el problema y no olvides seguir con el indicador del mouse

tu vista.” **“Como se comentó a un inicio del entrenamiento, no estoy tan interesado en la respuesta al problema como lo estoy en la forma cómo piensas la solución del problema. ¿Tienes alguna pregunta?”** Si la respuesta es no, entonces se enciende primero la videocámara y después se le indica al participante que inicie la solución del problema.
7mo paso: Aplicación de los análisis de protocolos concurrentes
-Registrar en el formato de evaluación y codificación las observaciones pertinentes. -Registrar con copia electrónica la respuesta del participante -Elaborar un archivo y folder para cada uno de los participantes con sus respectivos formatos de respuesta y registros de audio y video.
8vo paso: Aplicación de los análisis de protocolos retrospectivos y las entrevistas de salida.
-Aplicar los reportes verbales retrospectivos y las entrevistas de salida con ayuda de los protocolos de evaluación del diseño.
9no paso: Agradecimientos y cierre de la sesión.

Apéndice 3. Formatos de aplicación de las técnicas de pensamiento en voz alta del EXHCOBA

Formato de identificación

a) Datos del participante

Fecha de aplicación:_____

ID. del participante: _____ Nombre:_____ Edad: _____

Teléfono:_____ Dirección:_____ Correo electrónico: _____

Escuela: _____ Grado escolar:_____ Promedio escolar:_____

Promedio general en: Español: _____ Matemáticas: _____

Puntajes ENLACE: Español: _____ Matemáticas: _____

Registro y No. de identificación:

a) Entrenamiento _____ b) Estudio concurrente _____ c) Estudio retrospectivo _____

b) Áreas de análisis:

Nivel básico I (primaria)	**Nivel básico II (secundaria)**		
	Lengua y matemáticas	**Ciencias Naturales**	**Ciencias Sociales**
1.1 Habilidades verbales (HV) ____ 1.2 Habilidades cuantitativas (HC) ___	2.1 Español (ESP) ___ 2.2 Matemáticas (MAT)___	2.3 Física (FIS) ___ 2.4 Química (QUI) ___ 2.5 Biología (BIO) ___ 2.6 Geografía (GEO) ___	2.7 Formación cívica y ética (FCyE) ___

Ítems analizados:

Notas y observaciones:

Formato de aplicación del análisis de protocolos concurrente de ítems del Examen de Habilidades y Conocimientos Básicos (EXHCOBA)

I. Estudio concurrente	ID. del análisis: _____
1.1. Lectura de las instrucciones y base del ítem	
Tipo de lectura de las instrucciones y de la base del ítem: ____ Lectura en voz alta (estudiante) ____ Lectura en silencio (estudiante) ____ Lectura salteada (estudiante) ____ Lectura omitida (estudiante) ____ Señalada por el aplicador ____ Lectura por el aplicador ____ Otro, ¿Cuál? _____	
Descripción de la omisión de la lectura de las instrucciones y/o de la base del ítem: _____	
1.2. Fluidez en la lectura de las instrucciones y base del ítem	
Tipo de lectura de la base del ítem:	___ El participante leyó todas las palabras correctamente ___ El participante falló en la lectura de algunas palabras ___ El participante tuvo dificultades con varias palabras
Lista de palabras leídas incorrectamente:	
Descripción de las dificultades para comprender las instrucciones y la base del ítem:	

1.3. Lectura global del ítem
Tipo de lectura global del ítem: ____ Lectura en voz alta (estudiante) ____ Lectura en silencio (estudiante) ____ Lectura salteada (estudiante) ____ Lectura omitida (estudiante) ____ Señalada por el aplicador ____ Lectura por el aplicador ____ Otro, ¿Cuál? _____
Descripción de la omisión de la lectura global del ítem:
1.4. Fluidez en la lectura global del ítem
Descripción del recorrido o la trayectoria de la lectura global del ítem:
Tipo de lectura global del ítem: ___ El participante leyó todas las palabras correctamente ___ El participante falló en la lectura de algunas palabras ___ El participante tuvo dificultades con varias palabras
Lista de palabras leídas incorrectamente:
Elementos del ítem que presentan dificultades para su comprensión y/o manejo:
Descripción de las dificultades en la lectura global del ítem:

1.5. Proceso de respuesta del ítem	
___ No se solucionó aparentemente ___ No se atendió a la solución del problema	____ Proceso de solución incorrecto ____ Aparentemente adivino
Descripción general del proceso de respuesta del problema:	
El participante se distrajo o confundió por algún elemento del ítem: ___ SI ___NO ___ No aparentemente Descripción de la distracción:	
Observaciones:	

II Estudio retrospectivo ID. del análisis: _____
2.1 Análisis de las instrucciones y base del ítem
¿El contenido o temática del ítem fue visto durante tus estudios de primaria o secundaria? SI ___ NO ___ ¿A qué profundidad?
¿La instrucción y/o base del ítem en general es clara y comprensible? SI ___ NO ___ En caso de que no sea clara describir la razón:
En la instrucción y/o base del ítem ¿hubo alguna palabra, frase o elemento que fuera confuso o difícil de comprender? SI___ NO ___ ¿Cuál(es)?
2.2 Análisis global del ítem
¿Algunos de los elementos del ítem fuera de la instrucción y/o base del reactivo (opciones de respuestas, imágenes, graficas, recuadros, botones de control, etc.) en general son claros y comprensibles? SI ___ NO ___ ¿Cuál(es) y por qué?

¿La estructura y/o formato del ítem es clara y comprensible? SI___ NO ____ ¿Por qué?
¿El manejo o uso de los elementos y controles del ítem son comprensibles y fáciles de usar? SI ___ NO ___ ¿Cuál(es) y por qué?

2.3 Análisis del proceso de respuesta del ítem
Ahora describe paso a paso y a profundidad como hiciste para responder el ítem: Diagrama del proceso de respuesta:
Observaciones finales:

Apéndice 4. Procesos de respuesta subyacentes a los ítems de opción múltiple del área HC del EXHCOBA definidos por los expertos

No. Ítem	Contenido	Procesos de respuesta	No. Ítem	Contenido	Procesos de respuesta	No. Ítem	Contenido	Procesos de respuesta
01	Adición y sustracción con apoyo en la recta numérica.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) comprender el uso de la recta numérica en la adición y sustracción de números enteros, d) identificar el valor resultante de la adición y sustracción de números enteros en la recta y, e) reconocer el valor solicitado dentro de las opciones de respuesta.	**11**	Identificación de la descripción de las partes de una fracción.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las partes (numerador y denominador) de una fracción, d) representar la parte de abajo de la fracción (denominador) y la parte de arriba (numerador) y, e) reconocer la descripción correcta de fracción dada dentro de las opciones de respuesta.	**21**	Relación de volúmenes.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar y seleccionar los elementos (área de la base y altura) requeridos para obtener los volúmenes de dos cubos, d) calcular la relación de un volumen respecto del otro y, e) reconocer el resultado en las opciones de respuesta.
02	Obtención del valor faltante en secuencias numéricas.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) obtener la diferencia entre cada una de las cantidades, d) identificar el patrón que rige en la secuencia, e) aplicar el patrón identificado para obtener la última cantidad de la secuencia y, f) reconocer el valor faltante en la secuencia numérica dentro las opciones de respuesta.	**12**	Identificación de una fracción en una figura.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) encontrar el denominador de la fracción que representa la parte de la figura dada y, d) reconocer dicha representación dentro de las opciones de respuesta.	**22**	Comparación de unidades entre unidades de medida distintas.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las unidades de medida involucradas, e) uniformizar las unidades de medida, f) comparar las unidades de medida y, g) reconocer dentro de las opciones de respuesta el dato solicitado.
03	Construcción de una expresiones algebraica	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las variables presentes en el contexto del problema, e) identificar la regla aritmética para representar y construir el modelo algebraico, f) calcular el valor deseado utilizando las operaciones aritméticas básicas y, g) reconocer el valor solicitado dentro de las opciones de respuesta.	**13**	Identificación de una fracción en una figura para su conversión a decimal.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) encontrar el denominador de la fracción que representa la parte de la figura dada, d) convertir la fracción a número decimal y, e) reconocer la representación correspondiente del número decimal dentro de las opciones de respuesta.	**23**	Aplicación de la regla de tres simple.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) construir el modelo de la regla de tres simple, e) calcular el valor deseado utilizando las operaciones aritméticas básicas y, f) reconocer dicho valor dentro de las opciones de respuesta.
04	Identificación del valor posicional en números enteros.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa y, d) reconocer el número en la cantidad en enteros presentada y, e) reconocer de las opciones de respuesta la que corresponda con el valor posicional.	**14**	Identificación de fracciones en una figura para su suma.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) encontrar los denominadores de las fracciones que representan las partes de la figura dada, d) representar y sumar las fracciones y, e) reconocer el resultado dentro de las opciones de respuesta.	**24**	Relaciones de proporcionalidad.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar y seleccionar los elementos para representar el modelo de proporcionalidad, e) calcular el valor deseado utilizando las operaciones aritméticas básicas y, f) reconocer dicho valor en las opciones de respuesta.
05	Identificación posicional de números decimales.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa y, c) reconocer dentro de las opciones de respuesta la que corresponda con la posición del número decimal.	**15**	Suma y resta de números decimales.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) construir el modelo de la suma y resta de números decimales, d) calcular el valor de la suma y resta de números decimales utilizando las operaciones aritméticas básicas y, e) reconocer el resultado dentro de las opciones de respuesta.	**25**	Regla de tres simple con fracciones.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) sumar las fracciones dadas, e) representar el valor de referencia en fracción, f) construir el modelo de la regla de tres simple, g) calcular el valor deseado con las operaciones aritméticas básicas y, h) reconocer dicho valor dentro de las opciones de respuesta.
06	Juicio situacional de la representación de un número.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) evaluar la situación y, e) elegir dentro de las opciones de respuesta la representación más sencilla.	**16**	Multiplicación de números decimales.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) construir el modelo de la multiplicación de números decimales, d) calcular el valor de la multiplicación de números decimales utilizando las operaciones aritméticas básicas y, e) reconocer el resultado dentro de las opciones de respuesta.	**26**	Aplicación de la regla de tres inversa en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) construir el modelo de la regla de tres simple, e) calcular el valor deseado utilizando las operaciones aritméticas básicas y, f) reconocer dicho valor solicitado dentro de las opciones de respuesta.
07	Aplicación de módulos en un contexto	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las variables presentes en el contexto del problema, e) calcular el residuo del módulo utilizando las operaciones aritméticas básicas y, f) reconocer el residuo dentro de las opciones de respuesta.	**17**	División de números decimales.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) construir el modelo de la división de números decimales, d) calcular el valor de la división de números decimales utilizando las operaciones aritméticas básicas y, e) reconocer el resultado dentro de las opciones de respuesta.	**27**	Conocimiento de la suma de los ángulos internos de un triángulo.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar en las opciones de respuesta el valor resultante de la suma de los ángulos interiores de un triángulo, independientemente del valor de sus lados.
08	Representación de exponentes en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las variables (base y exponente) presentes en el contexto del problema y, d) reconocer la representación del modelo exponencial dentro de las opciones de respuesta.	**18**	Aplicación de la regla de tres simple	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) construir el modelo de la regla de tres simple, e) calcular el valor deseado con las operaciones aritméticas básicas y, f) reconocer el resultado dentro de las opciones de respuesta.	**28**	Probabilidad de un evento en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las variables presentes en el contexto del problema, e) construir el modelo probabilístico de la ocurrencia de un evento, e) calcular el valor deseado utilizando las operaciones aritméticas básicas y, f) reconocer el valor solicitado dentro de las opciones de respuesta.
09	Equivalencias de decimales a fracciones.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) expresar el número decimal en una fracción y, d) obtener una fracción equivalente y, f) reconocer la fracción equivalente obtenida dentro de las opciones de respuesta.	**19**	Cálculo de perímetros de círculos.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) conocer la fórmula adecuada a las variables presentes en el problema, d) sustituir los valores en la formula, e) calcular el perímetro de la circunferencia utilizando las operaciones aritméticas básicas y, f) reconocer el resultado dentro de las opciones de respuesta.	**29**	Adición de probabilidades para eventos.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las variables presentes en el contexto del problema, e) construir el modelo Adición de probabilidades para eventos, e) calcular el valor deseado utilizando las operaciones aritméticas básicas y, f) reconocer dicho valor dentro de las opciones de respuesta.
10	Representación de la fracción en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar la unidad de la variable y su fracción en un contexto, d) representar el modelo matemático (regla de tres), e) calcular el valor de la fracción requerida utilizando las operaciones aritméticas básicas y, f) reconocer la fracción obtenida dentro de las opciones de respuesta.	**20**	Cálculo de áreas de las caras de un prisma.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar y seleccionar los elementos (largo y ancho) requeridos para obtener el área de cada una de las caras del prisma, e) identificar el número de caras involucradas para el cálculo del área, f) sumar el área de las caras requeridas y, g) reconocer el resultado dentro de las opciones de respuesta.	**30**	Cálculo de la media aritmética	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar las variables presentes y sus frecuencias en el contexto del problema, e) calcular la media aritmética con las operaciones aritméticas básicas y, f) reconocer dicho valor dentro de las opciones de respuesta.

Apéndice 5. Procesos de respuesta subyacentes a los ítems de respuesta compleja del área HC del EXHCOBA definidos por los expertos

No. Ítem	Contenido	Procesos de respuesta
01	Obtención del valor faltante en secuencias numéricas.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) obtener la diferencia entre cada una de las cantidades, d) identificar el patrón que rige en la secuencia, e) aplicar el patrón identificado para obtener la última cantidad de la secuencia y, f) escribir el valor faltante de la secuencia numérica dada.
02	Ubicación de fracciones en la recta numérica.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar en cuantas partes se dividiría la unidad de medida en la recta de acuerdo al denominador de la fracción a ubicar y, d) posicionar las fracciones en la recta numérica. Otros procedimientos previos a la división de la recta que se pueden utilizar para responder a la tarea evaluativa son: el uso del Mínimo Común Múltiplo (MCM) y representar las fracciones en su valor decimal.
03	Ordenación de números decimales de menor y mayor en espacios vacíos.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar que la parte entera de los números presentados es la misma, d) contrastar los décimos de las cantidades presentadas para e) identificar el menor, si existen decimos iguales se continua con f) el contraste de centésimos para de la misma forma g) identificar el menor y, si existen centésimos iguales, h) se realiza el mismo proceso anterior para milésimos. Debe tomarse en cuenta que el procedimiento mencionado se realiza dependiendo la presencia de decimos, centésimos o milésimos en los números.
04	División de números decimales.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las partes de la división (dividendo, divisor, cociente y residuo), d) aplicar cualquier modelo de la división con números decimales y, e) escribir el resultado.
05	Representación de fracciones en figuras geométricas.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) calcular una fracción equivalente con un denominador igual a la cantidad de partes en que se encuentra dividida la figura e, d) iluminar la cantidad de partes de la figura que representa el numerador de la fracción.
06	Solución de suma y resta de fracciones en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) representar las fracciones con un mismo denominador, e) sumar o restar el numerador según sea el caso y, f) escribir el valor faltante de la respuesta.
07	Identificación de elementos de la circunferencia.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) distinguir los elementos de la circunferencia y d) colocar el nombre en la figura dada de acuerdo a las características que definen a dichos elementos.
08	Cálculo de perímetros de círculos en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las variables presentes en el contexto del problema para el cálculo de perímetros, d) identificar la fórmula adecuada a las variables presentes en el problema, e) sustituir los valores en la formula, f) realizar el cálculo con las operaciones aritméticas básicas para obtener el perímetro de la circunferencia y, g) escribir el valor del perímetro solicitado.
09	Cálculo de áreas de triángulos rectángulos.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar y seleccionar los elementos (base y altura) requeridos para obtener el área del triángulo rectángulo, d) identificar la fórmula adecuada a las variables, e) sustituir los valores en la formula, f) realizar el cálculo con las operaciones aritméticas básicas para obtener el área del triángulo rectángulo y, g) escribir el valor del área solicitada.
10	Cálculo de volúmenes de prismas rectangulares.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar y seleccionar los elementos (área de la base y altura) requeridos para obtener el volumen del prisma rectangular, d) identificar la fórmula adecuada a las variables, e) sustituir los valores en la formula, f) realizar el cálculo con las operaciones aritméticas básicas para obtener el volumen del prisma rectangular y, g) escribir el valor del volumen solicitado.
11	Cálculo de equivalencias de unidades de volumen	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las unidades de volumen, d) identificar la equivalencia necesaria para su cálculo, e) representar un modelo matemático-aritmético, f) aplicar la regla tres simple y, g) escribir el valor solicitado.
12	Cálculo de equivalencias de unidades de longitud	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las unidades de longitud, d) identificar la equivalencia necesaria para su cálculo, e) representar un modelo matemático-aritmético, f) aplicar la regla tres simple y, g) escribir el valor solicitado.
13	Cálculo de distancias en mapas con uso de escalas.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) observar el mapa para ubicar los lugares que se solicitan, d) identificar la unidad representada en la escala, e) calcular la distancia entre ambos lugares utilizando la escala y, e) escribir el valor solicitado.
14	Representación numérica de porcentajes menores que 100 en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) identificar la variable en la gráfica de pastel, e) calcular el valor numérico del porcentaje representado y, f) escribir el valor solicitado.
15	Representación numérica de porcentajes mayores que 100 en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, b) identificar las variables en el contexto del problema, d) calcular el valor del porcentaje agregado a una cantidad neta, e) sumar el valor del porcentaje agregado a la cantidad neta y, f) escribir el valor solicitado. Otro procedimiento es obtener el personaje de manera directa. Es decir, aplicar el modelo matemático-aritmético de la regla de tres simple.
16	Cálculo de la regla de tres simple en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) identificar las variables en el contexto del problema, d) representar el modelo matemático, e) aplicar la regla de tres simple y, f) escribir el valor solicitado.
17	Cálculo de probabilidades expresada en fracciones	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) conocer el número total de los posibles eventos para obtener el denominador, d) determinar las posibilidades de ocurrencia del evento específico (numerador) y, e) escribir el valor solicitado.
18	Inferencia y cálculo de un valor dada una tabla de proporcionalidad directa.	a) Leer detalladamente las indicaciones del ítem, b) comprender el objetivo de la tarea evaluativa, c) interpretar la información representada en tabla de proporcionalidad directa, d) identificar el dato solicitado y el proporcionado en el problema, e) representar el modelo matemático con el dato proporcionado y los datos de la tabla necesarios para f) aplicar la regla de tres simple y, g) escribir el valor solicitado.
19	Inferencia de valores dada una gráfica poligonal en un contexto.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema y el objetivo de la tarea evaluativa, d) interpretar la información representada en la gráfica poligonal, e) identificar los datos solicitados en el eje horizontal y, f) proporcionar y escribir el valor correspondiente en el eje vertical.
20	Estimación de frecuencias de ocurrencias de eventos dada la probabilidad.	a) Leer detalladamente las indicaciones del ítem, b) comprender el contexto del problema, c) comprender el objetivo de la tarea evaluativa, d) interpretar la información representada en la tabla, e) representar el modelo matemático con los datos proporcionados y el dato respectivo de la tabla necesarios para f) aplicar la regla de tres simple y, g) escribir el valor solicitado.