Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Evidencias de validez de contenido de un generador automático de ítems

TESIS

Que para obtener el grado de

DOCTORA EN CIENCIAS EDUCATIVAS

Presenta

Citlalli Sánchez Alvarez

Ensenada, Baja California, México, diciembre de 2015

INSTITUTO DE INVESTIGACIÓN Y DESARROLLO EDUCATIVO

Universidad Autónoma de Baja California

Instituto de Investigación y Desarrollo Educativo

Doctorado en Ciencias Educativas

Evidencias de validez de contenido de un generador automático de ítems

TESIS

Que para obtener el grado de

DOCTORA EN CIENCIAS EDUCATIVAS

Presenta

Citlalli Sánchez Alvarez

APROBADA POR:

Dr. Eduardo Backhoff Escudero
Director de tesis

Dra. Norma Larrazolo Reyna
Sinodal

Dr. Guillermo Solano Flores
Sinodal

Dr. Felipe Tirado Segura
Sinodal

Dra. Virginia Velasco Ariza
Sinodal

DCE
DOCTORADO EN CIENCIAS EDUCATIVAS

Ensenada, B.C. a 27 de noviembre de 2015

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

Dra. Alicia Alelí Chaparro Caso López
Coordinadora del Doctorado en Ciencias Educativas
Presente.

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

"Evidencias de validez de contenido de un Generador Automático de Ítems"

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Eduardo Backhoff Escudero

Ensenada, B.C. a 27 de noviembre de 2015

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

Dra. Alicia Alelí Chaparro Caso López
Coordinadora del Doctorado en Ciencias Educativas
Presente.

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

"Evidencias de validez de contenido de un Generador Automático de Ítems"

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dra. Norma Larrazolo Reyna

Ensenada, B.C. a 27 de noviembre de 2015

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

Dra. Alicia Alelí Chaparro Caso López
Coordinadora del Doctorado en Ciencias Educativas
Presente.

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

"Evidencias de validez de contenido de un Generador Automático de Ítems"

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Guillermo Solano Flores

Ensenada, B.C. a 27 de noviembre de 2015

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

Dra. Alicia Alelí Chaparro Caso López
Coordinadora del Doctorado en Ciencias Educativas
Presente.

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

"Evidencias de validez de contenido de un Generador Automático de Ítems"

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dr. Felipe Tirado Segura

Ensenada, B.C. a 27 de noviembre de 2015

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

Dra. Alicia Alelí Chaparro Caso López
Coordinadora del Doctorado en Ciencias Educativas
Presente.

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

"Evidencias de validez de contenido de un Generador Automático de Ítems"

Esperando reciba el presente de conformidad, quedo de Usted.

Atentamente

Dra. Virginia Velasco Ariza

Dedicatoria

A ti, mamá.

Por enseñarme mediante el ejemplo, consistencia y acompañamiento que la superación es fiel e inseparable compañera del trabajo y esfuerzo.

Por mostrarme desde muy temprana edad lo que significa mantener la mirada siempre al frente y las alas listas para emprender el vuelo.

Por tu gran sabiduría, conocimientos y experiencia profesional, con los cuales contribuiste de manera muy importante al desarrollo de este trabajo de tesis.

Agradecimientos

A mis hermanas Ariadna y Eva Luz, quienes me brindaron apoyo familiar incondicional y sabiduría personal y profesional para culminar esta labor.

A Eduardo Backhoff Escudero, Director de tesis, quien mostró gran visión al guiar mi trabajo. Su apoyo, paciencia y calidez fueron determinantes en la realización de la tesis. Mi más profunda admiración, respeto y cariño.

A los miembros de mi comité de tesis, quienes orientaron el contenido de este trabajo y permitieron su mejoramiento. Norma Larrazolo Reyna, por su cálido e incondicional acompañamiento; Felipe Tirado Segura, por compartir sus conocimientos sobre teoría cognitiva; Virginia Velasco Ariza, por su experiencia en el campo de desarrollo de instrumentos de evaluación; y Guillermo Solano Flores, por compartir conmigo su amplio conocimiento sobre la generación automática de ítems y haber realizado sabias observaciones que brindaron firmeza teórica a esta tesis.

A Ayita Ruiz-Primo, además de contribuir con aspectos conceptuales importantes, por haberme abierto las puertas de su hogar. En especial por haberme prestado su casco y bicicleta durante mi estancia de investigación, sin ellos no hubiera descubierto caminos que me llevaron a mi destino.

A los investigadores del IIDE, mis profesores, quienes son ejemplos a seguir y parte imprescindible de mi formación. Particularmente a Luis Ángel contreras por haberme presentado el mundo de la investigación, transmitido sin restricciones su conocimiento y depositado su confianza desde mi llegada al IIDE.

A mis amigos y compañeros de doctorado: Fabiana Ferreyra Martínez, Juan Carlos Pérez Morán, Laura Fierro López, Mónica Monsivais Almada, Mónica López Ortega, Antelmo Castro López, Erika Reyes Piñuelas y Karla Castillo Villapudua. El buen humor, gratas conversaciones, apoyo y solidaridad permitieron crear un equipo de colegas y grandes amigos que siempre tendrán una pieza interesante del rompecabezas.

A todo el personal que conforma el equipo administrativo del IIDE, en especial a Iván Contreras Espinoza por su amistad y gran calidad humana.

A mis amigos de la vida, quienes vivieron mi presencia intermitente y largas ausencias durante estos años, pero que siempre estuvieron ahí. Gracias por las sonrisas, palabras de apoyo, y sobre todo, por las labores de rescate realizadas.

A la danza, a través de la cual siempre he respirado.

Índices

Índice general

Índice general i
Índice de tablas iii
Índice de figuras iv
Resumen vii
1. Introducción 1
1.1. Generador de exámenes Excoba 5
1.2. Planteamiento del problema 7
1.3. Preguntas de investigación 11
1.4. Objetivos 11
1.5. Justificación 12
1.6. Contenido de la tesis 15
2. Antecedentes teóricos: evaluación auténtica, generadores automáticos de ítems y validación de contenidos 18
2.1. Evaluación auténtica. Conceptualización y usos 18
2.2. Ingeniería de tests 25
2.3. Generación automática de ítems 30
2.3.1. Antecedentes 30
2.3.2. Modelos de ítems. Conceptualización 34
2.3.3. GAI mediante teoría fuerte y teoría débil 41
2.4. Validez de instrumentos de evaluación a gran escala 47
2.4.1. Evidencias de validez relacionadas con el criterio 52
2.4.2. Evidencias de validez relacionadas con el constructo 52
2.4.3. Evidencias de validez relacionadas con el contenido 54
2.4.4. Aproximaciones a la medición de la validez de contenido 60
2.4.5. Análisis de las opiniones de los expertos 65

3. Evidencias de validez de contenido del Excoba: estructura del examen 67

Fase I: Documentación del proceso de diseño y elaboración del Excoba 70

3.1. Etapa 1 del MVCE: Descripción del Modelo de Elaboración del Excoba (MEE) 75

3.2. Resultados 82

3.3. Modelos de ítems del Excoba 89

3.4. Familias de ítems, ítems padre e ítems hijos 98

4. Evidencias de validez de contenido del Excoba: modelos de ítems 101

4.1. Fase II del MVCE. Obtención de evidencias de validez de contenido de los modelos de ítems del Excoba 101

4.2. Resultados de la Fase II del MVCE 109

4.2.1. Modelos de ítems del área de Matemáticas 112

4.2.2. Modelos de ítems del área de Historia 140

4.2.3. Modelos de ítems del área de Química 166

4.2.4. Modelos de ítems del área de Español 199

5. Discusión y conclusiones 244

5.1. Síntesis de los resultados de la fase 2 del MVCE, etapa 1: análisis de los modelos según las distintas categorías de clasificación 250

5.2. Síntesis de los resultados del trabajo realizado con los paneles de expertos 253

5.3. Alcances y limitaciones de la metodología utilizada 256

5.4. Sugerencias para nuevas líneas de investigación 259

Referencias 263

Apéndices 282

Apéndice 1. Modelo de Elaboración del Examen de competencias Básicas (Excoba) 282

Apéndice 2. Formato de currículum vitae 285

Apéndice 3. Protocolo para la evaluación de los modelos de ítems 287

Apéndice 4. Formato de compromiso de confidencialidad 292

Índice de tablas

Tabla 2.1	Valores mínimos de CVR y CVRt, de acuerdo a lo propuesto por Lawshe….	62
Tabla 3.1	Modelo de Elaboración del Excoba (MEE) ……	75
Tabla 3.2	Clasificación de los modelos de ítems en el sistema GenerEx del Excoba …	88
Tabla 4.1	Distribución de modelos de ítems de Matemáticas por eje temático curricular ……	113
Tabla 4.2	Clasificación y distribución de modelos de ítems de Matemáticas en el GenerEx ……	114
Tabla 4.3	Distribución de modelos de ítems de Matemáticas, según el tipo de ejecución que demandan del estudiante …….	116
Tabla 4.4	Tipo de conocimiento que evalúan los modelos de ítems de Matemáticas	117
Tabla 4.5	Resultados globales del proceso de evaluación del panel de expertos del área de Matemáticas del Excoba ……	120
Tabla 4.6	Modelos de ítems de Matemáticas y regularidades que presentaron …….	137
Tabla 4.7	Modelos de ítems, ejes temáticos e indicadores no validados con mayor frecuencia en Matemáticas ……	139
Tabla 4.8	Distribución de modelos de ítems de Historia por bloque temático …….	144
Tabla 4.9	Clasificación y distribución de modelos de ítems de Historia en el sistema editor de reactivos ……	145
Tabla 4.10	Resultados globales del proceso de evaluación del panel de expertos del área de Historia del Excoba ……	154
Tabla 4.11	Distribución de modelos de ítems de Química por bloque temático curricular ……	168
Tabla 4.12	Clasificación y distribución de modelos de ítems de Química en el sistema GenerEx ……	169
Tabla 4.13	Distribución de modelos de ítems de Química según el tipo de ejecución que demandan del estudiante ……	171
Tabla 4.14	Tipo de conocimiento que evalúan los modelos de ítems de Química …….	171
Tabla 4.15	Resultados globales del proceso de evaluación del panel de expertos del área de Química del Excoba ……	175
Tabla 4.16	Modelos de ítems de Química y regularidades que presentaron ……	194
Tabla 4.17	Modelos de ítems, bloques temáticos e indicadores no validados con mayor frecuencia en Química ……	197
Tabla 4.18	Distribución de modelos de ítems de Español por ámbito de estudio …….	203
Tabla 4.19	Clasificación y distribución de modelos de ítems de Español en el GenerEx	204
Tabla 4.20	Distribución de modelos de ítems de Español según el tipo de ejecución que demandan del estudiante ……	206
Tabla 4.21	Tipo de conocimiento que evalúan los modelos de ítems de Español ……	207
Tabla 4.22	Resultados globales del proceso de evaluación del panel de expertos del área de Español del Excoba ……	210
Tabla 4.23	Modelos de ítems de Español y regularidades que presentaron ……	228
Tabla 4.24	Modelos de ítems, ámbitos curriculares e indicadores no validados con mayor frecuencia en Español ……	231

Índice de figuras

Figura 2.1	Modelo para obtención de evidencias de validez de contenido de Lynn (1986) …….	57
Figura 3.1	Modelo para la obtención de evidencias de validez de contenido de los modelos de ítems del Excoba ……	68
Figura 3.2	Ejemplo de un reactivo de arrastre ……	72
Figura 3.3	Estructura conceptual del Excoba ……	84
Figura 3.4	Ejemplo de una plantilla del Excoba ……	85
Figura 3.5	Módulos de captura del GenerEx ……	87
Figura 3.6	Secciones que conforman al Modelo de ítem del Excoba ……	90
Figura 3.7	Ejemplo de modelo de ítems. Sección datos de identificación del contenido a evaluar ……	91
Figura 3.8	Ejemplo de modelo de ítems. Sección de características del contenido a evaluar ……	92
Figura 3.9	Ejemplo de modelo de ítems. Generador con algoritmos y reglas para generar ítems ……	94
Figura 3.10	Ejemplo de modelo de ítems. Banco de información curricular que se utilizará para generar ítems que evalúen el manejo del concepto de fracciones. ……	96
Figura 3.11	Ítems hijos de matemáticas, provenientes de un modelo de ítems con dos familias: sombreado y escritura ……	98
Figura 3.12	Familias de ítems, ítems padre e ítems hijos en un modelo de ítems para evaluar el manejo del concepto de fracciones ……	99
Figura 4.1	Proceso de obtención de evidencias de validez de contenido de los modelos de ítems del Excoba. Trabajo con paneles de expertos ……	105
Figura 4.2	Porcentaje de modelos de ítems de Matemáticas no validados por los expertos en los indicadores ……	122
Figura 4.3	Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Matemáticas ……	127
Figura 4.4	Ejemplo de ítem hijo del modelo MAT03: sucesiones aritméticas ……	128
Figura 4.5	Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT03 ……	129
Figura 4.6	Ejemplo de ítem hijo del modelo MAT16: conteo ……	131
Figura 4.7	Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT03 ……	132
Figura 4.8	Ejemplo de ítem hijo del modelo MAT20: gráfica de una parábola ……	133
Figura 4.9	Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT20 ……	134
Figura 4.10	Ejemplo de ítem hijo del modelo MAT06: Sistemas de ecuaciones ……	135
Figura 4.11	Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT06 ……	135
Figura 4.12	Estructura del programa de estudios de la asignatura de Historia ……	143

Figura 4.13	Ejemplo del procedimiento de calificación del sistema informático en reactivos de tipo *Elemento categoría* …….	146
Figura 4.14	Ejemplo del procedimiento de calificación del sistema informático en reactivos de tipo *Elemento imagen* …….	148
Figura 4.15	Ejemplos de dos distintos tipos de ítems hijos en los que se utiliza ejecución de *Arrastre* para emitir las respuestas ……	150
Figura 4.16	Porcentaje de modelos de ítems de Historia no validados por los expertos en los indicadores ……	156
Figura 4.17	Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Historia ……	160
Figura 4.18	Ejemplo de ítem hijo del modelo HIS06: Viajes de exploración, hegemonía europea y colonización ……	161
Figura 4.19	Opinión de los expertos en los 27 indicadores validados del modelo de ítems HIS06 ……...	162
Figura 4.20	Porcentaje de modelos de ítems de Química no validados por los expertos en los indicadores …….	177
Figura 4.21	Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Química ……	181
Figura 4.22	Ejemplo de ítem hijo del modelo QUI13: Propiedades intensivas y extensivas de la materia …….	182
Figura 4.23	Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI13 …….	183
Figura 4.24	Ejemplo de ítem hijo del modelo QUI20: Características del método científico ……	186
Figura 4.25	Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI20 …….	187
Figura 4.26	Ejemplo de ítem hijo del modelo QUI14: Mezclas homogéneas y heterogéneas …….	188
Figura 4.27	Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI14 …….	189
Figura 4.28	Ejemplo de ítem hijo del modelo QUI17: El enlace químico y la valencia ..	191
Figura 4.29	Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI17 …….	192
Figura 4.30	Distribución de los contenidos de la asignatura Español ……	202
Figura 4.31	Porcentaje de modelos de ítems de Español no validados por los expertos en los indicadores …….	212
Figura 4.32	Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Español …….	216
Figura 4.33	Ejemplo de ítem hijo del modelo ESP10: Géneros literarios ……	217
Figura 4.34	Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP10 …….	218
Figura 4.35	Ejemplo de ítem hijo del modelo ESP15: Reglamento o instructivo ……	221
Figura 4.36	Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP15 …….	221

Figura 4.37	Ejemplo de ítem hijo del modelo ESP03: Gráficas, esquemas y diagramas en textos ……..……..	223
Figura 4.38	Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP03 …….....	224
Figura 4.39	Ejemplo de ítem hijo del modelo MAT06: Sistemas de ecuaciones …….	225
Figura 4.40	Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP01 …….....	226
Figura 4.41	Distribución de los 56 modelos de ítems del Excoba por tipo de programación en el GenerEx ……	234
Figura 4.42	Distribución de los 56 modelos de ítems del Excoba por tipo de ejecución	237
Figura 4.43	Distribución de los 56 modelos de ítems del Excoba por tipo de conocimiento ……	238
Figura 4.44	Porcentaje de modelos de ítems no validados por los expertos en cada indicador evaluado ……	240

Resumen

Esta tesis aborda un campo emergente en la evaluación educativa: la generación automática de ítems (GAI). La GAI representa una forma eficiente de resolver el problema de desgaste que sufren los exámenes que se utilizan en forma intensiva, como es el caso de los exámenes de admisión. Sin embargo, también presenta nuevos retos para la psicometría y la evaluación educativa. Uno de ellos es la forma de validar los modelos de ítems, que se diseñan para construir decenas de reactivos isomorfos y, con ellos, una cantidad importante de exámenes paralelos o equivalentes.

Aunque los primeros intentos por automatizar la construcción de ítems datan de los años sesenta del siglo pasado, el advenimiento de las ciencias computacionales y el desarrollo de las ciencias cognitivas en los años recientes, han dado un impulso renovador a la GAI en la última década. Los desarrollos más importantes en este campo de la medición psicológica y educativa han sido encabezados principalmente por investigadores norteamericanos, tales como Gutman, Lynn, Shavelson y Haladyna.

En México sólo se conoce un proyecto que utiliza los principios de la GAI para desarrollar exámenes de admisión a las instituciones de educación media superior y superior. Este proyecto desarrollado por Backhoff, Larrazolo, Ramírez, Rosas y Tirado (2015), consta de: el Examen de Competencias Básicas (Excoba) y el Generador Automático de Exámenes (GenerEx). El primero consiste en el desarrollo conceptual (basado en teoría débil) de un examen que evalúa los aprendizajes básicos del currículo de la educación obligatoria mexicano, a través de reactivos de respuesta construida y semiconstruida, con lo que se acerca a la evaluación considerada como auténtica. El segundo consiste en el desarrollo de un sistema informático que: 1) genera los reactivos y exámenes del Excoba en forma automática; 2) administra las versiones de los exámenes generados vía computadora y califica las respuestas de los estudiantes de manera objetiva. Actualmente, el

Excoba/GenerEx se utiliza periódicamente en varias instituciones educativas mexicanas, para seleccionar a miles de estudiantes de educación media superior.

Uno de los problemas que presenta el Excoba/GenerEx, como cualquier GAI, es tener que validar cientos o miles de reactivos que se generan a través de sus modelos de ítems. Por consiguiente, es necesario contar con una metodología capaz de validar familias de ítems, y no de reactivos en lo individual. Hasta el momento se han desarrollado dos tesis doctorales para atender este problema en el Excoba/GenerEx: una relacionada con la validez cognitiva (Pérez, 2013), y otra que atiende la validez de constructo (Ferreyra, 2014) del Excoba/GenerEx. Entre otras necesidades, hacía falta desarrollar un modelo con el cual generar evidencias de validez de contenido de las familias de ítems de este examen.

Por lo anterior, el propósito central de este trabajo fue diseñar y probar un modelo metodológico para validar los modelos o familias de reactivos de la versión para educación media superior del Excoba/GenerEx.

El método seleccionado constó de dos fases. La primera tuvo el propósito de documentar el proceso de diseño del examen, la elaboración de los 120 modelos de ítems que lo conforman y la forma en que opera el instrumento computarizado para evaluar al estudiante. La segunda fase se centró en recabar evidencias de validez de contenido de algunos modelos de ítems del Excoba/GenerEx, mediante la adaptación del procedimiento de jueceo propuesto por Lynn (1986) para evaluar ítems individuales, según las necesidades de las familias de reactivos de la GAI. En este trabajo se evaluó casi la mitad de los 120 modelos de ítems de los que consta el examen computarizado: 20 de Matemáticas, 20 de Español, 8 de Química y 8 de Historia.

Para el análisis individual de los modelos de ítems se consideraron los siguientes aspectos: 1) los modelos de ítems, según el contenido curricular que evalúan; 2) su clasificación en el sistema informático del Excoba/GenerEx; 3) el tipo de ejecución que se le

solicita al examinado (arrastre, selección, escritura o respuesta mixta), y 4) el tipo de conocimiento que exploran (declarativo, procedimental, esquemático o estratégico). Adicionalmente se analizaron los modelos de ítems de acuerdo con los problemas detectados por los expertos, tratando de encontrar regularidades: 1) según las sugerencias emanadas de los paneles de expertos para mejorarlos, y 2) por la relación encontrada entre los problemas detectados y los contenidos evaluados por los modelos de ítems.

Los resultados de esta tesis proporcionan información sobre el Excoba/GenerEx, respecto a:

1) La precisión y delimitación de su definición conceptual y los contenidos que evalúa.

2) La alineación y pertinencia curricular de los elementos conceptuales incluidos en los modelos de ítems.

3) El nivel de dificultad aparente de los elementos que se intercambian en la base de los reactivos y son utilizados para generar los ítems que conformarán las distintas versiones del instrumento.

4) El nivel de demanda cognitiva bajo el cual se conciben y elaboran los ítems; (5) los tipos de ejecución que realiza el estudiante al responder y su proximidad con la forma en que se lleva a cabo su proceso de enseñanza-aprendizaje.

6) El uso de distintos tipos y formatos de ítems como aproximaciones a la evaluación autentica en evaluaciones de gran escala.

Con este método para el análisis de los modelos de ítems del Excoba, las aportaciones realizadas se pueden enmarcar en dos grandes vertientes. Por una parte la posibilidad de identificar los errores, las fortalezas y debilidades del Excoba en términos de su contenido, con la finalidad de apoyar a los autores en su mejoramiento. Por la otra, la propuesta

metodológica en términos de lograr una nueva aproximación para obtener evidencias de validez de contenido en instrumentos de evaluación del aprendizaje a gran escala, específicamente de los modelos de ítems elaborados mediante la aplicación de los principios de la GAI de teoría débil.

La tesis se divide en cinco capítulos, un apartado de referencias bibliográficas y una sección de apéndices. El primer capítulo introduce el tema de la GAI y del Excoba/GenerEx. El segundo presenta el planteamiento teórico de la tesis, al analizar los trabajos sobre la evaluación auténtica, la ingeniería de los tests, la evolución de la GAI y los tipos de modelos de ítems; así como el concepto de validez de las evaluaciones, con especial énfasis en la validez de contenido. El tercer capítulo se centra en la metodología empleada para encontrar evidencias de validez de contenido del Excoba/GenerEx, relacionada con el procedimiento para diseñar y construir el examen. El cuarto capítulo se dedica a dar cuenta de la recolección de evidencias proporcionadas por diferentes grupos de expertos que juzgan diferentes propiedades de los modelos de ítems, y los reactivos que con ellos se generan. En el quinto capítulo se sintetizan los resultados de mayor importancia y se concluye sobre las bondades del método utilizado.

Palabras clave: Generación automática de Ítems, ingeniería de tests, validez de contenido, modelos de ítems, Excoba, GenerEx.

1. Introducción

Debido a la alta y creciente demanda que existe para ingresar a las Instituciones de educación media superior (IEMS), el uso de mecanismos evaluativos para la selección de estudiantes ha cobrado especial importancia. En México existen muchas instituciones educativas que utilizan sistemas de admisión con instrumentos para evaluar a los estudiantes, que proporcionan información que impacta de manera directa en su vida, ya que los resultados determinan el ingreso o la exclusión a este nivel educativo de muchos jóvenes. Por tanto, es de suma importancia garantizar su calidad técnica.

Todo instrumento de evaluación debe cumplir con una serie de criterios de calidad. Dichos criterios son establecidos por organismos nacionales e internacionales, los cuales garantizan, entre otras cosas, un elevado nivel técnico en los distintos tipos de evidencias de validez y confiabilidad, así como una evaluación objetiva e imparcial de los conocimientos, las habilidades y aptitudes para los que fueron diseñados (Consejo Asesor Externo del Centro Nacional de Evaluación para la Educación Superior [Ceneval], 2000).

Actualmente se cuenta con una amplia experiencia en el ámbito internacional en el desarrollo de instrumentos de evaluación del aprendizaje. En México, organismos como el Centro Nacional de Evaluación para la Educación Superior (Ceneval), el Instituto Nacional para la Evaluación de la Educación (INEE) y la Secretaría de Educación Pública (SEP), entre otros, han desarrollado diversos exámenes de gran escala siguiendo los criterios técnicos mencionados.

Algunos de esos instrumentos son de tipo criterial, y se aplican con la finalidad de conocer el nivel de dominio de los estudiantes respecto al currículum de la educación básica. Tal es el caso de los Exámenes de Calidad y Logro Educativo (Excale) y el Plan Nacional para la Evaluación de los Aprendizajes (Planea). Otros, de tipo normativo, se

utilizan para seleccionar a aquellos estudiantes que son más aptos para cursar el bachillerato o los estudios de nivel superior.

Entre los instrumentos que se consideran de mayor importancia debido a la diversidad de instituciones que los utilizan y a la gran cantidad de estudiantes que impactan, se encuentran: el examen de ingreso a la Universidad Nacional Autónoma de México (UNAM, 2014); el Examen Nacional de Ingreso a la Educación Media Superior (EXANI-I) y el Examen Nacional de Ingreso a la Educación Superior (EXANI-II), desarrollados por el Ceneval (2013); la Prueba de Aptitud Académica (PAA) y la Prueba para el Ingreso al Nivel de Educación Media Superior (PIENSE II), elaborados por el College Board (2011a; 2011b; 2011c; 2012); y el Examen de Habilidades y Conocimientos Básicos (EXHCOBA), diseñado y elaborado por Backhoff y Tirado en 1992.

El EXHCOBA es un instrumento que se utilizó para la selección y el ingreso de estudiantes al bachillerato y la universidad. En la Universidad Autónoma de Baja California (UABC) se empleó como requisito de ingreso durante más de 20 años y hasta el año 2014 se utilizaba en 15 Instituciones de educación media superior (IEMS) y de educación superior (IES) del país, con la evaluación de más de 120,000 estudiantes en ese año. Su formato original fue en papel, para su llenado con lápiz, con reactivos de opción múltiple que se calificaban utilizando un lector óptico. A partir de 1993 se administró y calificó en forma computarizada mediante una plataforma informática llamada Sistema de Exámenes Computarizados (SICODEX), la cual fue creada ex profeso (Backhoff, Ibarra y Rosas, 1996).

El EXHCOBA mide los conocimientos y las habilidades básicas e indispensables que un estudiante requiere para acceder a la educación de nivel superior. Cuenta con siete versiones paralelas, que derivan de la desarrollada originalmente en 1992. Todas las versiones han sido objeto de diversos estudios para la obtención de evidencias de su validez

y confiabilidad (Tirado, Backhoff, Larrazolo y Rosas, 1997; Backhoff, Larrazolo y Rosas, 2000; González-Montesinos, 2004).

Debido a los cambios curriculares que se han dado en los programas de estudio de la educación básica y superior en el país durante los últimos años, así como a los avances científicos y tecnológicos en materia de instrumentos de evaluación, los autores del EXHCOBA se dieron a la tarea de replantear la estructura conceptual del instrumento y redefinir su diseño. Para ello, utilizaron los principios de la ingeniería de tests (IT) y la generación automática de ítems (GAI). Esta última es una aproximación de diseño y construcción de instrumentos de evaluación que permite la elaboración de decenas y hasta cientos de exámenes similares, mediante un procedimiento iterativo.

La GAI ofrece ventajas sobre la forma tradicional de elaborar instrumentos, ya que los ítems que se elaboran surgen a partir de documentos llamados *modelos de ítems*. Estos documentos contienen información detallada respecto a las reglas y restricciones que se seguirán para generar los reactivos. Asimismo, incluyen un amplio banco de elementos conceptuales y contenidos curriculares que se utilizan para generar grandes cantidades de ítems y elaborar varias versiones del mismo examen.

Bajo esta aproximación fue necesario modificar el formato de los ítems que hasta el momento se habían utilizado en el EXHCOBA, ya que los de opción múltiple presentaban algunas limitaciones, tales como el constate desgaste asociado a su uso, su forma artificial de evaluar el aprendizaje y la presencia del factor de adivinación en las respuestas de los examinados.

Como resultado de este replanteamiento se diseñó y elaboró una nueva generación de exámenes que sustituyó al EXHCOBA, llamada Excoba (Examen de competencias

básicas).[1] Esta nueva generación se fundamenta en el currículum de educación básica y media superior mexicanos, y permite elaborar exámenes con características estructurales y psicométricas similares, mediante ítems de respuesta construida y semiconstruida. A la par se desarrolló el sistema informático Generador Automático de Exámenes (GenerEx), como una plataforma digital que permite la generación de ítems equivalentes (isomorfos) y con ello, múltiples versiones del examen (Tirado, Backhoff y Larrazolo, 2014).

El Excoba es un instrumento de nueva creación, que representa una aproximación a la evaluación auténtica del aprendizaje a gran escala mediante el uso de la GAI, cuyos resultados impactan de manera directa la vida de miles de estudiantes. Debido a esas características es imprescindible contar con estudios que aporten evidencias respecto a sus indicadores de calidad técnica, particularmente sobre la validez de los contenidos de los ítems que se generan mediante sus modelos.

Por ello, la presente investigación se realizó con el objetivo de proponer un modelo para la obtención de evidencias de validez de contenido de los modelos de ítems que conforman cuatro áreas de la estructura conceptual del Excoba: Matemáticas, Historia, Química y Español, de educación secundaria. Es importante mencionar que, debido a la etapa de desarrollo en la que se encuentra el instrumento, este estudio se centra en la versión utilizada para seleccionar estudiantes en el nivel de educación media superior.

Para esclarecer el problema que creó la necesidad de realizar esta investigación, en la siguiente sección se presenta una semblanza de lo que es el Excoba. En ella se mencionan los aspectos que orientaron la toma de decisiones y el trabajo de los desarrolladores del Examen al incursionar en un campo novedoso en la elaboración de instrumentos de

[1] A partir de este momento se mencionarán de manera separada el Excoba y el GenerEx. El primero como la estructura conceptual y los contenidos curriculares que permiten generar distintas versiones del Examen, y el segundo como el sistema informático creado para generar de manera automática los ítems que se especifican en el Excoba.

evaluación, llamado *generación automática de ítems* (GAI). Asimismo, se presentan los objetivos de la investigación, las preguntas que la guiaron y los motivos que justificaron su realización.

1.1. Generador de exámenes Excoba

El Examen de Competencias Básicas (Excoba) es un instrumento que incluye tres grandes cambios respecto al modelo original:

1) Su construcción se fundamenta en las premisas de la GAI de teoría débil, transitando de las especificaciones a los modelos de ítems, los cuales permiten generar una gran cantidad de versiones de ítems equivalentes en contenido y propiedades psicométricas.

2) Se sustenta en el currículum y en los programas de estudio vigentes de la educación básica mexicana, de donde se seleccionan los contenidos que aborda la evaluación de las competencias básicas y de los aprendizajes esperados en los estudiantes que concluyeron sus estudios de educación básica.

3) Introduce reactivos en los que la respuesta es construida y semiconstruida.

En los siguientes párrafos se presentan estos tres aspectos, los cuales se describen con mayor profundidad en los capítulos dos y tres.

Generación automática de ítems (GAI). Se trata de un campo emergente en el diseño y la elaboración automatizada de instrumentos de evaluación, en donde se utilizan modelos de ítems y tecnología computarizada para generar grandes cantidades de ítems isomorfos. En el caso del Excoba se creó el sistema informático GenerEx para alojar toda la información necesaria para construir ítems de manera aleatoria y automática. El sistema contiene los algoritmos o procedimientos que permiten utilizar las reglas, restricciones y los elementos

conceptuales para hacer combinaciones y generar los ítems que conformarán cada versión del examen. Así, el Excoba puede generar decenas o cientos de versiones de un instrumento de evaluación, de tal manera que no se le considera como un solo examen sino como un generador automático de exámenes.

Sustento del Excoba. A diferencia de su antecesor, el Excoba se basa en el currículum mexicano de educación básica. Mediante el trabajo con comités de expertos en las distintas disciplinas que forman parte de la estructura conceptual del Examen, se seleccionaron las competencias y los aprendizajes esenciales del currículum. Se detectaron las habilidades necesarias para dominar los contenidos, y los tipos de estrategias y materiales utilizados en el proceso de enseñanza-aprendizaje. Esto con la finalidad de elaborar un instrumento que se aproximara a una evaluación auténtica sobre el nivel de dominio de los estudiantes respecto a los contenidos que marca el currículum.

Ítems del Excoba. Los ítems que genera son de respuesta construida y semiconstruida, de tipo complejo. En ellos el estudiante no emite una sola respuesta correcta por ítem, sino múltiples, con lo cual se modifica la calificación dicotómica por la de crédito parcial. Aunque no se ha designado un término definitivo para nombrarlos, los ítems fueron denominados como Reactivos Estructurales Constructivos (ReEsCo) por Tirado en 2010, quien consideró que son la representación de estructuras semánticas que dan cuenta de la forma en la que los estudiantes se han apropiado de los conocimientos, y que a su vez proporcionan información sobre cómo construyen las estructuras necesarias para su comprensión. Bajo este modelo se evalúan campos del conocimiento que aglutinan habilidades, conocimientos y competencias para el manejo de contenidos curriculares específicos, que requieren del uso del razonamiento y la aplicación del conocimiento.

Lo anterior representa una transición del formato de reactivos de opción múltiple al de ítems de respuesta compleja. La finalidad de estos últimos es lograr un acercamiento a

la evaluación de competencias de manera auténtica. Los ítems del Excoba requieren que al responder, el estudiante utilice el juicio y la innovación, ya que se presenta la posibilidad de más de una respuesta correcta, con lo cual se requiere que el estudiante tome decisiones informadas (Wiggins, como se cita en Svinicki, 2004).

El Excoba fue creado para atender las limitaciones de su antecesor aprovechando los avances logrados en materia de diseño de instrumentos de evaluación. Como generador de exámenes creado bajo las premisas de la GAI, representa una aproximación innovadora en la elaboración de exámenes normativos de gran escala y alto impacto. Sus ítems son novedosos por el tipo de respuestas que pueden emitir los estudiantes, por la forma en la que se califican, y porque son generados en grandes cantidades mediante modelos de ítems.

1.2. Planteamiento del problema

En el apartado anterior se habló del desarrollo de un examen normativo, de gran escala y alto impacto, elaborado bajo los principios de la GAI de teoría débil. A diferencia de su antecesor, el Excoba es un generador de exámenes que utiliza modelos de ítems que contienen elementos conceptuales que son intercambiados aleatoriamente y en forma automatizada para crear ítems equivalentes, tanto en contenido como en sus propiedades psicométricas.

Todo instrumento de evaluación, especialmente aquel cuyos resultados impactan de manera importante la vida de estudiantes, requiere cumplir con diversos criterios de calidad. Debe estar respaldado por un riguroso proceso de desarrollo y construcción en el que se lleve a cabo un estricto apego a los estándares de calidad establecidos por organismos internacionales. A su vez, debe ser susceptible a que con dichos estándares se valide la interpretación que se haga de sus resultados.

En este sentido, el Comité Técnico del Excoba ha resuelto que para evaluar la calidad del instrumento son imprescindibles tres tipos de evidencias de validez: a) de su estructura interna, b) de los procesos cognitivos subyacentes a las respuestas de los examinados, y c) de sus modelos de ítems, ya que en ellos se encuentran los contenidos seleccionados del currículum, así como las reglas y los elementos necesarios para generar los ítems que conformarán las distintas versiones del instrumento.

El proceso para obtener evidencias de validez de contenido del Excoba representa un reto, por tratarse de un generador de exámenes y no de un examen único. Lynn (1986) indica que una parte muy importante de las evidencias de validez de contenido radica en el proceso de diseño y construcción del instrumento. En este sentido, es importante revisar el modelo que se siguió para elaborar el Excoba, y documentarlo. Este mismo autor propuso una aproximación para la revisión de los contenidos incluidos en la evaluación, mediante el uso de paneles de expertos.

Por su parte, Porter (2002) señala la importancia de conocer el grado de alineación entre los aprendizajes esperados, los estipulados en el currículum y los exámenes creados para evaluar el dominio por parte de los estudiantes. Este autor propuso una forma de obtener evidencias de ello, mediante la aplicación de encuestas a docentes. Utilizó una escala para asignar valores al tiempo que dedican a la enseñanza de los contenidos, así como al énfasis otorgado en el aula. Su propuesta sugiere la relevancia de revisar no solamente la alineación entre el examen y el currículum, sino entre lo que ocurre en el aula cuando el docente selecciona materiales y emplea su propio estilo instruccional, y el mismo examen. De esta manera se puede conocer objetiva y sistemáticamente el grado de alineación entre la instrucción, la evaluación y los contenidos que marca el currículum.

El Excoba, al ser un generador de exámenes que propone un modelo para desarrollar instrumentos de evaluación que miden el nivel de adquisición de las competencias básicas

marcadas en el currículum de educación básica mexicano, requiere de un modelo que proporcione, en primera instancia, evidencias de validez de los contenidos que fueron seleccionados y valorados como sustanciales, y que se obtuvieron a partir de un riguroso análisis curricular.

También es indispensable someter a un proceso de validación la estructura y los contenidos de los modelos de ítems utilizados para generar los reactivos que conformarán las distintas versiones de los exámenes que surjan a partir de él, ya que como afirma Haladyna (2012), un buen diseño de modelo de ítems es suficiente para garantizar a priori que los ítems creados mediante la GAI son similares en cuanto a su estructura y sus propiedades psicométricas. Esto implica que un aspecto fundamental en el diseño de un generador de exámenes es la elaboración cuidadosa de los modelos de ítems y, por ende, en su revisión mediante un proceso de validación de contenido.

Actualmente los estudios para obtener evidencias de validez de los GAI se han centrado en los modelos y métodos estadísticos que han aportado información de cómo abordar la validez de la estructura interna de este tipo de instrumentos (Embretson, 1999; Ferreyra, 2014; Geerlings, Glas y Van der Linden, 2011; Glas y Van der Linden, 2003; Holling, Bertling y Zeuch, 2009; Hombo y Dresher, 2001; Sinharay y Johnson, 2012).

También se han llevado a cabo investigaciones para obtener evidencias de validez de la estructura cognitiva de los GAI (Brown y Burton, 1978; Chen y Macdonald, 2011; Gierl et al., 2009; Ma, Çetin y Green, 2009; Pérez, 2013; Revuelta y Ponsoda, 1998; Romero, Ponsoda y Ximénez, 2008).

Algunos autores como Solano-Flores, Shavelson y Schneider (2001) han documentado las ventajas y los beneficios de utilizar moldes de ítems. Solano-Flores (1991) y Bejar (2002) han mencionado la importancia de establecer un procedimiento riguroso y sistemático para la elaboración de los ítems de un examen. Otros también han propuesto el uso de modelos

de ítems para elaborar ítems isomorfos, con la descripción de procedimientos para su generación automática (La Duca, Staples, Templeton y Holzman, 1986; Singley y Bennett, 2002). Sin embargo, en la literatura revisada no se encontraron trabajos que hayan desarrollado métodos para obtener evidencias de validez de los contenidos de los mismos.

El Excoba enfrenta una serie de retos que aún no han sido abordados, particularmente en lo concerniente a la obtención de evidencias de su validez y confiabilidad. Es necesario desarrollar estudios que posibiliten la obtención de este tipo de información, ya que el campo de la GAI se encuentra aún en desarrollo. Hasta el momento existen dos estudios que proponen aproximaciones para obtener evidencias de validez de la estructura interna de los distintos exámenes y reactivos que se producen mediante el generador de exámenes Excoba, así como de los procesos cognitivos que utilizan los examinados al responder los ítems (Ferreyra, 2014; Pérez, 2013). Los resultados de dichos estudios derivaron en la propuesta de dos aproximaciones metodológicas para obtener evidencias de validez en instrumentos creados mediante la GAI de teoría débil.

En cuanto a la validación de sus contenidos, es importante considerar que no se trata de someter a juicio únicamente los contenidos seleccionados del currículum para conformar la estructura conceptual del Excoba, sino de sus modelos de ítems, familias e ítems hijos. Por eso, el problema que plantea este estudio es el de proponer una metodología que permita obtener evidencias de validez de contenido del Excoba en función de dos aspectos: la *representatividad y pertinencia de los contenidos* que fueron seleccionados del currículum para conformar la estructura conceptual del Excoba; así como la *estructura*, *funcionalidad y pertinencia de los modelos de ítems*, con especial énfasis en las familias de ítems y los elementos conceptuales utilizados para generar los ítems hijos que conforman las distintas versiones del Excoba.

1.3. Preguntas de investigación

De acuerdo con lo expuesto, surge la siguiente pregunta de investigación:

- ¿Cómo se pueden obtener evidencias de validez de contenido de un instrumento elaborado mediante los principios de la generación automática de ítems (GAI) de teoría débil?

De esta pregunta se desprenden los siguientes cuestionamientos específicos:

- ¿Cómo se pueden examinar los modelos de ítems utilizados por el Excoba?
- ¿Qué tipo de problemas presentan los modelos de ítems en cuanto a su estructura, funcionalidad y alineación curricular?
- ¿En qué medida se aproximan a la evaluación auténtica los ítems del Excoba que se generan en forma automática?
- ¿En qué proporción los contenidos de Matemáticas, Historia, Química y Español del Excoba, representan los contenidos curriculares más importantes que se espera que dominen los estudiantes al concluir la secundaria?

1.4. Objetivos

Con el propósito de responder a las preguntas planteadas, se proponen los siguientes objetivos:

Objetivos

- Proponer una metodología para obtener evidencias de validez de contenido de un generador de exámenes, diseñado bajo los principios de la GAI de teoría débil.
- Documentar el proceso de diseño y elaboración del generador de exámenes Excoba, instrumentado para la selección de estudiantes en el nivel de educación media superior.

- Formular recomendaciones para el mejoramiento de la estructura del generador de exámenes Excoba.

1.5. Justificación

Una de las preocupaciones fundamentales asociadas al diseño y el desarrollo de instrumentos de evaluación es su validez. Esta se refiere al grado en el que las evidencias y la teoría dan sustento a las interpretaciones propuestas por las puntuaciones derivadas de una prueba (AERA, APA y NCME, 2014). Como se ha visto, el Excoba y el GenerEx son dos instrumentos que trabajan de manera paralela, y cuyo origen técnico descansa en un campo novedoso para el desarrollo y la elaboración automática de instrumentos de evaluación, llamada *generación automática de ítems* (GAI). Ambos instrumentos fueron elaborados con estrictos cuidados técnicos y metodológicos, sustentados en un modelo para la elaboración de instrumentos, que ha sido probado en diversos contextos evaluativos nacionales e internacionales.

La participación de cuerpos colegiados de expertos en la elaboración del Excoba ha proporcionado objetividad en su construcción. Por su parte, los estudios que se han realizado hasta el momento han aportado evidencias de validez de su estructura interna y de los procesos cognitivos subyacentes a la ejecución de los estudiantes. Sin embargo, al tratarse de un nuevo desarrollo es imprescindible garantizar su cumplimiento con estándares de calidad adicionales, establecidos para todo instrumento de evaluación.

Debido a que se trata del primer generador de exámenes normativos de gran escala elaborado en México, cuya aplicación se realiza mediante computadora, utilizando ítems de respuesta construida y semiconstruida, es necesario realizar un estudio para obtener evidencias de validez de los contenidos que fueron seleccionados del currículum. Asimismo, se requieren evidencias de validez de la estructura y la operación de los modelos de ítems que lo conforman, para determinar si el instrumento mide lo que dice medir y si se acerca

a la forma en que se lleva a cabo el proceso de enseñanza-aprendizaje. Es decir, es necesaria una evaluación auténtica.

Por lo anterior, las razones que justifican este estudio se enmarcan en dos tipos. Por un lado, se encuentran las de tipo técnico. El Excoba debe contar con evidencias de que los contenidos que fueron seleccionados del currículum mexicano para su conformación son los que reflejan las habilidades y competencias básicas que necesitan los estudiantes para ascender de nivel académico. El sistema de selección de estudiantes mexicanos que se utiliza en la actualidad se caracteriza principalmente por el empleo de instrumentos en formato de lápiz y papel, así como el uso de ítems de opción múltiple. Dicho sistema presenta limitaciones que impiden evaluaciones auténticas que hagan uso de las virtudes de la tecnología y que a su vez representen un bajo costo en cuanto a su desarrollo y actualización.

El avance tecnológico, así como el avance en el desarrollo de instrumentos de evaluación permiten emplear nuevas herramientas para abordar la evaluación de procesos de pensamiento más complejos, o estructuras de conocimiento en las que los estudiantes asimilan, acomodan y construyen sus propias respuestas (Tirado et al., 2014). El Excoba es un instrumento de vanguardia en dos sentidos: no se trata de un examen, sino de un generador de exámenes; además, utiliza ítems que evalúan competencias y aprendizajes esperados en los estudiantes, mediante estrategias que se asemejan a la forma en que aprenden en el aula.

La obtención de evidencias de validez de los contenidos del Excoba es útil para sus desarrolladores porque no solamente fortalece los procesos de obtención y cumplimiento de los indicadores de calidad técnica que todo instrumento debe tener, sino que lo hace dentro del campo de la GAI. También aporta elementos que fortalecen la toma de decisiones de los usuarios del Excoba, ya que se basan en los resultados de un instrumento

cuyas evidencias indican la medida en la que el Excoba evalúa los contenidos sustanciales del currículum de la educación básica mexicana.

Asimismo, considerando los altos costos (dinero, tiempo y recursos humanos) asociados a un proceso de validación en el cual se capacita, contrata y trabaja con personal especializado de diversas áreas del conocimiento, la propuesta de esta investigación para obtener evidencias de validez de contenido de un generador automático de exámenes aporta un método eficiente para la reducción de dichos costos al mínimo, ya que en tanto no se modifique la estructura conceptual del instrumento, no será necesario realizar más que un solo proceso.

Por otro lado, en la literatura revisada para este estudio no se encontraron trabajos estudios realizados en México en los que se propongan procedimientos para abordar la obtención de evidencias de validez de contenido de instrumentos de carácter normativo, diseñados bajo los principios de la GAI de teoría débil, y que se aproximen a la evaluación auténtica de las competencias adquiridas por los estudiantes durante su proceso de aprendizaje. Esto agrega un valor teórico sustancial a esta investigación.

En primera instancia, el presente trabajo permite establecer bases documentales del diseño, la construcción y operación de un instrumento de vanguardia. También permite establecer un primer modelo metodológico para la obtención de evidencias de validez de contenido dentro del campo de la GAI, y de manera implícita, una forma de analizar las distintas estrategias evaluativas contenidas en los modelos de ítems y su proximidad con la forma en que se enseña en el aula.

La segunda razón que justifica esta investigación es de carácter social. El propósito central de este estudio es aportar evidencias de validez de los contenidos que fueron seleccionados por paneles de expertos de distintas áreas del conocimiento, cuyo origen y sustento es el currículum oficial de la educación básica mexicana. Los resultados que

deriven del Excoba tienen un gran impacto social debido a que hasta el momento se aplica a más de 11,000 estudiantes al año, en cuatro instituciones de educación media superior y superior del país.

Dicha cobertura implica una gran responsabilidad por parte de los desarrolladores del instrumento, ya que una gran cantidad de estudiantes quedan eliminados del proceso de ingreso a la institución de su elección, quienes, en el mejor de los casos, recurren a otras instituciones educativas que originalmente no tenían contempladas como la opción para continuar sus estudios. Por ello, es imprescindible garantizar que la interpretación que se haga de los resultados esté fundamentada en un instrumento justo, equitativo y que mida lo que se supone que debe medir.

1.6. Contenido de la tesis

El documento está organizado en cinco capítulos y una serie de apéndices. Este primer capítulo, presenta un panorama general del tipo de evaluación que se realiza en México, así como algunos de los instrumentos de gran escala utilizados para la selección de estudiantes que concluyeron sus estudios de secundaria y desean continuar el nivel de bachillerato. Se plantea el problema de investigación y las preguntas que guiaron el camino para su resolución. También se presentan los objetivos que guiaron las acciones realizadas en el estudio y se justifican las razones por las cuales fue importante llevarlo a cabo.

El capítulo dos trata el tema de la importancia que reviste un buen diseño evaluativo basado en la obtención de evidencias que brinden información sobre cómo es que los estudiantes organizan la información, la transfieren y aplican a diversas situaciones. Se exploran los conceptos de *evaluación auténtica*, *evaluación de desempeño* y *evaluación alterna*, estableciendo similitudes y diferencias entre ellas, así como sus limitaciones. También se revisa la literatura existente en el tema de la ingeniería de tests y la generación

automática de ítems, bajo las premisas de teoría fuerte y teoría débil. Se describe qué es un modelo de ítems, cómo están conformadas cada una de sus secciones y la lógica detrás de la estructura y generación automática de ítems. Por último, el capítulo presenta la fundamentación teórica del tema de validez de instrumentos de evaluación, y las distintas formas de obtener dichas evidencias, particularmente la de contenido.

En el capítulo tres se presenta la primera de dos fases del modelo utilizado para la obtención de resultados de esta investigación, la cual consistió en documentar el procedimiento de diseño y elaboración del Excoba. Se muestra la estructura conceptual, la fundamentación teórica y metodológica del Excoba, y se describen cada una de las fases que conforman el modelo de planeación y diseño utilizado para su elaboración. También se hace una descripción detallada del Excoba como un generador automático de ítems, junto con las características de sus modelos de ítems.

El capítulo cuatro describe la segunda fase del modelo propuesto en esta investigación. Se explica con detalle la manera en que se abordó el proceso para la obtención de evidencias de validez de contenido del Excoba y sus modelos de ítems. Se presentan las etapas desarrolladas en esta fase para realizar el trabajo con paneles de expertos, y los resultados se organizan en cuatro distintos apartados, dedicados a cada una de las asignaturas evaluadas por el Excoba que fueron consideradas en esta investigación. Se incluye una descripción de la forma en que se llevó a cabo el trabajo con los distintos grupos de expertos y se analizan los resultados obtenidos en Matemáticas, Historia, Química y Español. Se presta especial atención a aquellos modelos de ítems que a juicio de los expertos presentaron mayor cantidad de problemas técnicos, estructurales y de diseño, así como las sugerencias que hicieron para mejorar el área evaluada.

El capítulo cinco se dedica a la discusión de los hallazgos y la conclusión del trabajo de investigación. Representa un análisis integral de la información obtenida, bajo la luz de la

literatura revisada e incluye recomendaciones, sugerencias y reflexiones que permiten la creación de futuras líneas de investigación en materia de evaluación.

2. Antecedentes teóricos: evaluación auténtica, generadores automáticos de ítems y validación de contenidos

El propósito de este capítulo es brindar un panorama respecto a los orígenes y las aplicaciones de la generación automática de ítems (GAI), campo emergente de la evaluación educativa. En la primera sección se describe qué es la evaluación auténtica y las distintas acepciones del término. En el segundo apartado se aborda el tema de ingeniería de tests, como un antecedente de la GAI. La tercera sección presenta los antecedentes históricos que llevaron al surgimiento de la GAI, así como las aproximaciones y experiencias existentes en torno a la elaboración de modelos de ítems. En la cuarta sección se presentan las dos aproximaciones existentes en la elaboración de instrumentos, mediante los principios de la GAI: teoría fuerte y teoría débil. Finalmente se aborda el concepto de validez de las evaluaciones, con especial énfasis en la validez de contenido.

2.1. Evaluación auténtica. Conceptualización y usos

Aún no existe consenso respecto a la definición de *evaluación auténtica*, ya que se utilizan diferentes términos para referirse a ella; entre los más comunes se encuentran: *evaluación auténtica*, *evaluación del desempeño* y *evaluación alterna*; sin embargo, a pesar de que se llegan a utilizar de manera indistinta, sí existen algunas diferencias importantes. Por ejemplo, la *evaluación auténtica* generalmente se refiere a aquellas actividades que debe realizar el estudiante y que se acercan mucho a las que realizaría en la vida real; la *evaluación del desempeño* se refiere más a las acciones que una persona realiza como prueba de que cuenta con las competencias para resolver problemas de la vida cotidiana; mientras que el término *evaluación alterna* se asocia a exámenes que se alejan de las preguntas de opción múltiple y utilizan alguna otra estrategia para captar lo que el estudiante sabe o no sabe hacer (Miller, Linn y Gronlund, 2008).

En la década de los años noventa del siglo pasado, en Estados Unidos se llevó a cabo una reforma al sistema educativo en la cual se abordó, entre otras cosas, el método que se utilizaba para evaluar el desempeño de los estudiantes. Esta reforma trajo consigo un gran reto: establecer sistemas de evaluación rigurosos, desarrollados y validados de manera sistemática, alineados al sistema de enseñanza basado en la promoción de habilidades y competencias que le ayuden a los estudiantes a enfrentar retos similares a los que se les presentan en la vida diaria. Es decir, surgió de la necesidad de contar con mecanismos que permitieran evaluar el nivel de dominio de las habilidades necesarias para la vida (Bullens, 2002).

Hasta entonces, los principales métodos de evaluación eran las pruebas estandarizadas y de opción múltiple, pero este tipo de instrumentos carecían de las características que exigía la reforma educativa norteamericana, ya que evaluaban principalmente mediante reactivos de opción múltiple, proveyendo solamente un método para demostrar el conocimiento adquirido por los estudiantes. La evaluación auténtica permite utilizar distintos métodos para obtener un panorama adecuado de lo que los estudiantes saben. Las herramientas para este tipo de evaluación son variadas y permiten al estudiante demostrar que sabe utilizar las habilidades que ha aprendido.

Algunos estudios (Costa y Kallick, 1992; Goodrich, 2001; Kallic, 1992; Wiggins, 1990) han demostrado que los métodos alternos de evaluación como las preguntas abiertas de respuesta múltiple, son más valiosos que los métodos de evaluación tradicional, porque involucran situaciones en las que el uso de habilidades como la creatividad y la solución de problemas reales son requeridas. Sin embargo, esos mismos trabajos señalan que esta forma de evaluación se realiza durante el proceso de enseñanza-aprendizaje y no al final del mismo (Costa y Kallick, 1992).

A pesar de que en las últimas décadas ha surgido un fuerte interés en utilizar sistemas de evaluación alternos a los exámenes de opción múltiple, se sabe poco respecto a su diseño, distribución, calidad e impacto (Baker, 2010). La década de los noventa del siglo pasado fue un *parteaguas* para la evaluación educativa, ya que a partir de ese momento surgió el interés por saber en qué medida las estrategias evaluativas que se utilizaban hasta entonces se aproximaban a escenarios auténticos. Surgió la necesidad de definir dentro del contexto evaluativo el término *autenticidad* y su aplicación dentro del ámbito educativo (Keyser y Howell, 2008).

Como se mencionó, aún no existe consenso en la definición del término *autenticidad*, pero diversos autores coinciden en los elementos esenciales que dan forma a este concepto. Por ejemplo, para Archbald y Newmann (1988) *autenticidad* se refiere al conocimiento básico y sustancial dentro de un campo del conocimiento. Estos autores mencionaron que se trata de evaluar tareas que reflejen situaciones de la vida real, más que un cúmulo de conocimientos adquiridos a través de actividades basadas en el currículum.

Wiggins (como se cita en Svinicki, 2004) delimitó las características que debe tener una evaluación para considerarse auténtica: a) ser realista y reflejar la forma en la que la información y las habilidades serán utilizadas en el mundo real; b) implicar el uso de juicio e innovación porque se resuelven problemas no estructurados en los cuales existe más de una respuesta correcta, al provocar que el estudiante tome decisiones informadas; c) solicitar al estudiante que realice los procedimientos necesarios para la ejecución de las tareas; d) llevar a cabo la evaluación en situaciones lo más parecidas al contexto real; e) solicitar al estudiante que demuestre el dominio de una variedad de habilidades relacionadas a problemas complejos, incluyendo el juicio, y f) permitir la retroalimentación, práctica y las oportunidades adicionales para solucionar los problemas.

En este sentido, a diferencia de los exámenes tradicionales, para Wiggins (como se cita en Svinicki, 2004) la evaluación auténtica se centra en la calidad de las respuestas y su justificación.

En 2010 Baker desarrolló una explicación de las distintas acepciones del término *evaluación auténtica*. En primer lugar, precisó que la *evaluación alterna* es cualquier tipo de evaluación distinta a la de opción múltiple o de reactivos con respuesta dicotómica (cierto-falso), que implique actividades contextualizadas que simulen o utilicen ejecuciones derivadas de tareas de la vida diaria. En este sentido, la evaluación proporciona muestras más genuinas y representativas del trabajo de los estudiantes, porque tiene significado implícito para ellos.

Baker (2010) también propuso que la *evaluación del desempeño* puede llegar a involucrar actividades que requieran observación del comportamiento del estudiante ante la solución de un problema. Esto no sólo aporta una definición precisa de lo que es la evaluación auténtica, sino que destaca la importancia que debe delimitar la definición del tipo de habilidades intelectuales que se están evaluando. Esta definición propuesta por Baker se condensa en lo que el Committee on the Foundations of Assessment (Comité para los Fundamentos de la Evaluación) del Consejo Nacional de Investigación en Estados Unidos afirmó en el año 2001, respecto a que la evaluación del desempeño requiere que los estudiantes ejecuten tareas más auténticas, lo cual describió como actividades en las que se involucra el uso conjunto de conocimientos y habilidades dentro del contexto de proyectos reales.

Para Miller et al. (2008), hay tres tipos de evaluación estrechamente relacionados: a) la *evaluación auténtica*, que se refiere a todos aquellos elementos que se alinean y tienen una fuerte similitud a las tareas de la vida diaria; b) la *evaluación del desempeño*, que evalúa la forma en que el estudiante muestra su nivel de competencia, mediante la ejecución de

actividades de la vida real, y c) la *evaluación alterna*, que al igual que la definición de Baker, implica cualquier tipo de evaluación distinta a la de opción múltiple.

Según los hallazgos de Gulikers, Bastiaens y Kirshner (2004), hay una distinción muy clara entre la evaluación auténtica y la del desempeño. Ellos señalan que a diferencia de la segunda, cuando se trata de llevar a cabo una tarea la evaluación auténtica exige un alto nivel de fidelidad o apego a la realización de la misma, así como a las condiciones del contexto en el que normalmente se llevaría a cabo la ejecución de dicha tarea. Estos autores mencionan que la evaluación auténtica exige de los estudiantes el uso de las mismas competencias o combinaciones de conocimientos, habilidades y actitudes que se necesitan en situaciones de la vida profesional. Para ellos, el nivel de autenticidad de una evaluación se define por el grado de similitud que tienen las tareas presentadas en ésta, con la realidad profesional.

Gulikers et al. (2004) elaboraron un marco de referencia que incluye dimensiones de evaluación, que se refiere a cinco niveles de autenticidad: a) la tarea evaluativa, b) el contexto físico, c) el contexto social, d) la forma o resultado de la evaluación, y e) los criterios o estándares evaluativos. La utilidad de este modelo depende de las necesidades evaluativas y de aprendizaje que se tengan en el contexto escolar.

Savery y Duffy (2001), por su parte, afirman que la autenticidad de una evaluación implica una similitud entre las demandas cognitivas de la evaluación y las de una situación que simula o refleja un escenario de la vida real. Cumming y Maxwell (1999) señalan cuatro componentes dinámicos de la evaluación auténtica: ejecución, contexto, complejidad y competencia. Estos autores afirman que la *ejecución* del estudiante debe ser evaluada mientras este se encuentre llevándola a cabo en escenarios reales. Para ellos, las tareas evaluativas que se alejan del contexto del mundo real disminuyen la integridad de los resultados que se derivan de la evaluación.

En lo que respecta al componente *contexto*, los autores indican que tanto el proceso de enseñanza, como la evaluación, deben llevarse a cabo dentro de contextos reales, ya que generalmente los estudiantes tienen muy poca habilidad para transferir los conocimientos del aula al mundo real. Afirman también que de no realizarse de esa manera no existiría garantía de que la ejecución que tenga el estudiante dentro de un contexto se realizará de la misma forma en otro.

En cuanto al componente *complejidad,* sugieren que los estudiantes tendrán mayor capacidad para desarrollar habilidades para la solución de problemas complejos, en la medida en que se les presenten escenarios de aprendizaje y evaluación igualmente complejos, y que imiten contextos auténticos.

El cuarto y último componente propuesto por Cumming y Maxwell es la *competencia,* definida como cualquier actividad realizada en el aula, que al ser transferida al mundo real requiere que el estudiante realice cambios o ajustes para poder ejecutarla debidamente.

Por su parte, Rule (2006), después de una exhaustiva revisión encontró que en educación superior las actividades que resultan de un aprendizaje auténtico y son objeto de evaluación implican cuatro características: a) involucran problemas de la vida real que imitan el trabajo profesional dentro de distintas disciplinas y, a su vez, implican la presentación de la información o los resultados a destinatarios externos al salón de clases; b) incluyen preguntas abiertas, habilidades de pensamiento y metacognición; c) introducen a los estudiantes al discurso y el aprendizaje social dentro de comunidades de aprendizaje, y d) habilitan a los estudiantes a elegir el rumbo de su aprendizaje en contextos donde realicen proyectos relevantes.

Otros autores como Newman, Secada y Wehlage (como se cita en Svinicki, 2004), han señalado que la evaluación auténtica debe dar evidencias de que el estudiante construye su conocimiento. Es decir, que es capaz de organizar información y considerar alternativas.

Afirmaron que la evaluación auténtica debe hacer una indagación disciplinada, en la que se exploren contenidos que impliquen conocimientos y procesos centrales al campo de estudio, así como proporcionar mecanismos de comunicación escrita para crear condiciones que propicien la comprensión. También declararon que este tipo de evaluación debe poseer valor en contextos distintos al escolar, permitiendo el establecimiento de una conexión entre los problemas, el mundo y audiencias más allá del salón de clases.

La evaluación auténtica se ha utilizado en diversos campos del conocimiento, mediante distintos métodos y estrategias, como: el uso de portafolios, la simulación de escenarios reales, el desarrollo de proyectos reales y ficticios, y la elaboración de reportes de investigación y rúbricas, entre otros (Snavely y Wright, 2003; Wellington, Thomas, Powell y Clarke, 2002; Randal, 2004). Pero estos instrumentos y estrategias evaluativas conllevan una serie de retos logísticos y pedagógicos, que diversos autores han documentado y referido como difíciles de validar y complicados para operar cuando se desea llevar a cabo evaluaciones individuales o a gran escala (Barton, 1999; Boyd-Batstone, 2004; Lowyck y Poysa, 2001; Quellmalz, Schank, Hinojosa y Padilla, 1999).

El proceso educativo debe ayudar a los estudiantes a desarrollar y ejercer habilidades de orden superior como son: la evaluación, síntesis, análisis y aplicación del conocimiento, con la finalidad de solucionar problemas en contextos inesperados. Por ello, el sistema de evaluación debe medir qué tanto se comprenden los conceptos que se enseñan en la escuela, y no solamente conceptos memorizados que difícilmente servirán para transferir el conocimiento a distintos tipos de problemas (Vendlinski, Underdahl y Simpson, 2002).

Los constantes avances en materia tecnológica, así como el desarrollo en el área de las ciencias cognitivas, han permitido centrar la atención en la evaluación de los constructos implicados en el dominio de los conocimientos requeridos para enfrentar y solucionar problemas, de tal forma que lo que ahora se busca evaluar es si el estudiante se puede

desempeñar de manera competente en determinados campos o dominios (Committee on the foundations of Assessment, 2001). De manera implícita, es necesario contar con un sistema eficiente de diseño y elaboración de instrumentos de evaluación de dichas competencias.

Las concepciones, descripciones y los componentes de la evaluación auténtica aquí presentados, permiten obtener un panorama respecto a la importancia que reviste para el campo de la evaluación educativa la detección de habilidades, competencias y destrezas con las que cuenta el estudiante, y que ha transferido de un escenario escolar a uno de la vida real. Aunque el uso de instrumentos automatizados de gran escala que utilicen estrategias evaluativas auténticas es aún muy limitado, sí existen esfuerzos significativos que han producido avances en materia de enseñanza y evaluación auténtica en pequeños grupos. Uno de ellos es la instrumentación de proyectos para la creación de simuladores y software informático, en los que los estudiantes realizan tareas, resuelven problemas e interactúan con personas y objetos, tal como lo harían en el mundo real (Barab, Thomas, Dodge, Carteaux y Tuzun, 2005). Sin embargo, para dar el paso a una evaluación auténtica a gran escala, aún queda camino por recorrer.

En el siguiente apartado se describe una aproximación metodológica llamada ingeniería de tests (IT), que ha revolucionado la forma en que se concibe el diseño de instrumentos de evaluación educativa, así como la generación eficiente y eficaz de reactivos.

2.2. Ingeniería de tests

Los avances tecnológicos a partir de finales del siglo pasado han proporcionado a diversos campos del conocimiento oportunidades de mejoría sustancial. La tecnología digital está presente en prácticamente toda actividad humana y ha permitido la realización de actividades interdisciplinarias, cuyos productos han generado nuevos caminos,

propuestas y descubrimientos significativos en materia de salud, ciencias exactas y ciencias sociales. El campo de la evaluación no es la excepción, ya que como consecuencia de los desarrollos tecnológicos y el trabajo de diversos especialistas dentro de las ciencias cognitivas, estadística, psicología educativa y ciencias computacionales, se han generado muchos cambios y mejoras en el campo de la medición, particularmente la educativa (Gierl, Zhou y Alves, 2008).

La fuerte necesidad de contar con alternativas eficientes y eficaces para la elaboración de instrumentos de medición ha tenido como consecuencia la creación de nuevas líneas de investigación. Este es el caso del campo de la ingeniería de los tests (IT) o *assessment engineering*, (Luecht, 2006a, 2006b, 2007a, 2008a, 2008b, 2009, 2010, 2011), la cual es una aproximación innovadora y sofisticada dentro del campo de la medición, que utiliza los principios del diseño de pruebas para dirigir el diseño y desarrollo de instrumentos, la calificación y el análisis de resultados de una evaluación, así como de la elaboración de los informes correspondientes (Zhou, 2009).

No se trata de una tecnología específica ni un modelo psicométrico, sino de un vasto y complejo marco conceptual que sirve para concebir, bajo una nueva perspectiva, la forma en que se diseñan y elaboran reactivos utilizando principios, reglas o procedimientos específicos. Esto permite generar una gran cantidad de ítems, tareas evaluativas e instrumentos que proporcionan información consistente y útil para los propósitos específicos de una evaluación.

Dicho marco conceptual se apoya en el uso de software y sistemas informáticos. En él se aplican los principios del diseño ingenieril junto con herramientas psicométricas, con la finalidad de diseñar instrumentos que contengan mecanismos claros y sistemáticos para identificar si el producto final (los ítems) se apega al diseño del instrumento (Luecht, 2008a; Luecht, 2012; Luecht, Burke y Devore, 2009).

Los objetivos que persigue la IT son cuatro: 1) diseñar y medir consistentemente y a través del tiempo el mismo constructo, ya sea con propósitos formativos o sumativos; 2) automatizar total o parcialmente los procedimientos que usualmente son complejos y muy costosos en tiempo y dinero, y que generalmente son realizados por seres humanos; 3) proporcionar métodos eficientes, adaptables y de bajo costo para la producción de grandes cantidades de ítems que no requieran ser piloteados, calibrados y/o descartados después de haber sido utilizados, y 4) reducir la dependencia que hasta el momento se ha tenido en los modelos de análisis y calibración psicométrica, en cuanto a la medición de los constructos evaluados por los instrumentos (Luecht, 2012).

La IT ha establecido un cambio de fondo en la naturaleza de los datos derivados de una evaluación. Uno de esos cambios es el diseño intencional y dirigido, mediante el cual se sigue controles de calidad estrictos en el proceso de diseño, desarrollo, elaboración, calificación, análisis de datos y elaboración de informes de los resultados de una evaluación. Lo anterior, según la propuesta de Luecht (2012), implica cuatro procesos fundamentales que se describen a continuación.

El primero se refiere al diseño ordenado y sistemático de un mapa que describa, de manera precisa, los conocimientos y habilidades (constructos) que se pretende evaluar, y que a su vez explicite el significado de cada uno de los niveles de dominio establecidos en la escala de puntuación diseñada para tales efectos. Esto significa que desde el momento inicial del diseño del instrumento se debe visualizar y articular de manera detallada cuál será el orden jerárquico de los conocimientos y las habilidades que se medirán. También implica que se debe tener una cabal comprensión de la complejidad cognitiva de las tareas que el examinado deberá ejecutar, de tal manera que se pueda dar soporte a los argumentos y las interpretaciones que se hagan en torno a los niveles de ejecución, de manera que la interpretación de los resultados sea congruente. A este proceso, Wilson (2005) le llamó *mapeo del constructo*, y lo consideró como un proceso iterativo en el cual

se pueden realizar tantas modificaciones como sean necesarias, con la finalidad de establecer claramente el significado de cada uno de los niveles de ejecución definidos en la escala de calificación, así como su relación con los constructos medidos.

El segundo proceso es la construcción de *modelos de tareas* (MT) y sus correspondientes *mapas de modelos* (MM). Esto implica una elaboración cuidadosa y sistemática de especificaciones para clases o familias de ítems. Su diseño representa una forma alterna al modelo convencional de una especificación, ya que la intención no es describir la forma en que se elabora un ítem, sino ubicar dentro de un *mapa del constructo* (MC) el nivel de domino en el que se encuentra un examinado, de acuerdo con su ejecución en una tarea (operacionalizada mediante un ítem).

La diferencia principal entre la especificación tradicional y la del MT estriba en que su orientación es de tipo cognitiva, es decir, integra de manera sistémica: a) información acerca de los componentes del tipo de conocimiento evaluado y los distintos niveles cognitivos involucrados en su dominio; b) las relaciones que se dan entre dichos niveles y las habilidades cognitivas que los subyacen, y c) los contenidos relevantes, contextos y elementos auxiliares que afectan el nivel de complejidad cognitiva de la tarea que deberá realizar el examinado.

En este sentido, la información contenida en el MT ayudará a identificar qué lugar ocupa la respuesta de un examinado dentro de los distintos niveles de dominio determinados en el mapa del constructo. Además, debido a que todos los ítems hijos que se generan mediante este tipo de modelos y mapas de modelos comparten las mismas características estructurales, se presume que tendrán características psicométricas similares; es decir, si se modificara algún aspecto del modelo (por ejemplo, el nivel de complejidad cognitiva), se verían afectadas las propiedades psicométricas de todos los ítems hijos que deriven de dicho modelo, como sucedería en el caso del nivel de dificultad.

El tercer proceso que se debe considerar en el diseño de un instrumento creado bajo los principios de la IT es la forma en que se elaborarán las plantillas y los ítems que de ellas deriven. Éstas deben incluir tres aspectos fundamentales. En primer lugar, se requiere un modelo representativo de los ítems, también llamado *plantilla de ítems* o *modelo de ítems* (Bejar et al., 2003; Haladyna, 2004), que pueda generar al menos dos ítems. En segundo término, es necesario contar con un sistema de evaluación, preferentemente automatizado, que organice las respuestas y les asigne un valor, de manera que se genere una clave de las respuestas. Por último, se debe tener un modelo de información que incluya a todos los componentes fijos y variables que serán manipulados, combinados y ensamblados, con la finalidad de construir ítems hijos. Si dicha información se genera mediante un algoritmo automatizado (informático), entonces se hablaría de generación automática de ítems.

El cuarto y último proceso propuesto por Luecht (2012) implica la forma en que se lleva a cabo la calibración psicométrica. Bajo los modelos de la teoría clásica de los tests (TCT) y la teoría de Respuesta al ítem (TRI), el diseño de un instrumento de evaluación implica elaborar los ítems, su pilotaje y por último la calibración para la obtención de sus propiedades psicométricas, con la finalidad de utilizar esa información, ajustarlos y así mejorarlos.

En la IT se establecen controles de calidad muy estrictos durante la fase de diseño de los instrumentos; es decir, a través de la cuidadosa elaboración de mapas de constructos, mapas de tareas, plantillas y modelos de ítems, se detalla a priori y con mucha precisión lo que se pretende obtener en términos de la dificultad y otras propiedades métricas de los ítems. De esta manera, la perspectiva psicométrica es confirmatoria, porque está basada en decisiones de diseño intencionales y sistemáticas, tomadas antes de la construcción de los ítems. Los procesos de análisis estadístico se enfocan en el MT como la unidad central, y se llevan a cabo con la finalidad de identificar y controlar la varianza observada entre los

ítems que pertenecen a un mismo modelo. El objetivo es reducir, en la medida de lo posible, dicha varianza.

Como se puede observar, el proceso de diseño de un instrumento de evaluación bajo las premisas de la IT es un paso fundamental que debe ser vigilado de manera estricta. Será en función de dicho diseño que se podrán elaborar modelos de ítems que contengan los elementos necesarios para generar los ítems que conformarán el instrumento final que se utilizará para evaluar los niveles de dominio de los estudiantes en los constructos delimitados. Asimismo, el uso y la aplicación de tecnología en el campo de la evaluación educativa ha permitido atender algunas de las demandas asociadas con la aplicación de evaluaciones a gran escala, y así, automatizar el proceso de generación de ítems.

2.3. Generación automática de ítems

2.3.1. Antecedentes

Recientemente, y bajo las premisas de la IT, se ha desarrollado una nueva forma de elaborar instrumentos de evaluación, la cual no se ajusta a los métodos tradicionales de diseño y elaboración, debido a que su objetivo no se centra en la construcción de ítems individuales, sino de grupos de ítems que evalúan un mismo contenido y a su vez tienen propiedades psicométricas similares. Se trata de la generación automática de ítems (GAI).

En este campo de trabajo ha surgido la necesidad de buscar distintos enfoques y modelos psicométricos para aproximarse al análisis de los ítems y la obtención de evidencias de validez y confiabilidad (Ferreyra, 2014; Pérez, 2013). Todo instrumento de gran escala y alto impacto que sea utilizado para evaluar el logro educativo de estudiantes, necesita contar con una amplia cantidad de ítems que cumplan con una serie de requerimientos de calidad técnica, y que garanticen su uso durante un tiempo razonable. Desde hace más de 10 años se están utilizando herramientas tecnológicas con la finalidad

de mejorar los procesos de elaboración de instrumentos y automatizarlos, inclusive administrarlos y calificarlos en línea (Gierl y Hollis, 2012).

Para algunos autores, la generación automática de ítems es una línea de investigación sustentada en los principios de las teorías psicométrica y cognitiva, cuyo objetivo es dirigir la producción de ítems, creándolos mediante el uso de algoritmos computacionales. Al mismo tiempo se trata de un proceso en el cual se utilizan especificaciones de ítems, llamadas modelos de ítems (MI), con la finalidad de generar ítems estadísticamente calibrados (Gierl y Hollis, 2012). Para otros, como Bezruczko (2014), no se trata de un procedimiento, sino de un objetivo, ya que mediante el uso de algoritmos computacionales se busca producir grandes cantidades de ítems sin llevar a cabo un pilotaje previo, pre calibración ni validación externa.

En 2005, Arendasy afirmó que el futuro de los instrumentos de evaluación de habilidades psicológicas, capaces de ser analizadas de manera estructurada, se encuentra dentro del campo de la generación automática de ítems. Recientemente la concepción de la GAI se ha transformado a la de una ciencia emergente (Haladyna y Rodríguez, 2013). Su objetivo actual es producir una gran cantidad de ítems con propiedades psicométricas preestablecidas, que ofrecen un ahorro económico sustancial al automatizar los procesos de elaboración de ítems y la oportunidad de crear, sobre la marcha y sin amenazas a la seguridad del examen, versiones paralelas del instrumento con experiencias evaluativas similares en los examinados (Bejar et al., 2003).

La generación de ítems de manera automática tiene sus orígenes en cinco *parteaguas* en la historia de la evaluación (Haladyna, 2012). El primero de ellos fue cuando Guttman creó en 1959 la *Teoría de facetas*, en la que la generación de ítems dependía de la cuidadosa elaboración de mapas de enunciados, divididos en diversas secciones que se mantenían fijas

y otras que podían variar (llamadas *facetas*) para modificar su contenido y convertirse en nuevos ítems. Uno de los problemas de esta teoría al llevarla a la práctica fue que los elaboradores de reactivos no siempre utilizaban el mismo criterio para determinar cómo elaborar los mapas, por lo cual se fueron desvaneciendo los esfuerzos.

El segundo acontecimiento se dio cuando Osburn (1968), seguido por Hively (1974), propuso lo que llamó *formas de ítem*, las cuales eran utilizadas para definir dominios de contenido. Tenían una estructura similar a la de los mapas de enunciados de Guttman, porque se trataba de estructuras sintácticas fijas en las que se podía reemplazar una o varias de sus secciones para cambiar los aspectos conceptuales que evaluaban y, con ello, generar una multiplicidad de reactivos distintos. Esta línea de trabajo fue adoptada durante la década de los años setenta y ochenta del siglo pasado, pero aunque ya no es utilizada bajo el mismo nombre sí se han aplicado algunos de sus principios en la GAI actual.

Posteriormente, surgió la GAI *basada en prosa*. Ésta contiene una propuesta realizada por Bormuth en 1970, quien detectó que la forma en que se estaban elaborando los reactivos de las pruebas era subjetiva e ineficiente, debido a la intervención del factor humano en la elaboración. El autor se refería a que, a pesar de que quienes elaboran los ítems se basan en una misma especificación y contenidos, pueden crear ítems muy distintos entre sí. Ante esta situación, propuso que se automatizara el proceso, para lo cual se centró en la elaboración de ítems cuya base sintáctica se pudiera transformar algorítmicamente de prosa a pregunta.

Aunque la propuesta teórica de Bormuth no tuvo buenos resultados, diversos investigadores la retomaron y la transformaron. Simplificaron sus algoritmos y la hicieron funcional (Finn, 1975; Roid y Finn, 1977; Roid y Haladyna, 1978). Esta línea de trabajo

desapareció debido a las dificultades que presentaba en la elección de los textos utilizados para elaborar los ítems, y al hecho de que evaluaban conocimientos memorísticos.

Un cuarto momento histórico se dio cuando surgieron los trabajos para generar ítems que midieran la formación de conceptos (Markle y Tiemann, 1970). Se tomaba un concepto y se identificaban sus principales componentes, atributos y elementos. El trabajo de los elaboradores de reactivos era crear distintas combinaciones de los atributos más importantes que conformaban dicho concepto, elaborar ejemplos y contraejemplos del mismo. El objetivo era que el examinado demostrara su habilidad para discriminar las respuestas correctas y descartara las incorrectas, al dar evidencias de un manejo adecuado del conocimiento evaluado.

El quinto evento importante se llevó a cabo cuando Irvine y Kyllonen publicaron un libro en el año 2002, bajo el título *Item generation for test development*, el cual compiló una serie de contribuciones realizadas por investigadores de distintos países en el campo de la GAI, particularmente de Estados Unidos y Gran Bretaña. Este libro fue producto de un seminario que se llevó a cabo en 1998 por el Educational Testing Service (ETS), cuyo tema central fue la GAI. En él se menciona la existencia de tres paradigmas de medición utilizados en la GAI: *los Modelos R*, que utilizan programas convencionales para medir el rendimiento escolar o las competencias profesionales; los *Modelos L* o de latencia, que son utilizados para determinar la velocidad de las respuestas de los examinados, y los *Modelos D* que implican una medición repetida debido a que son de tipo predictivo. En general, lo que se busca mediante estos modelos es el control de la dificultad de los ítems.

Como se observa, la evolución de la GAI ha seguido una variedad de líneas de trabajo y aunque existen puntos de convergencia también se observa desarticulada. Sin embargo, todos los esfuerzos realizados han tenido la misma necesidad subyacente: automatizar el

proceso para generar grandes cantidades de reactivos, que sean equivalentes en cuanto a conceptos y semejantes desde el punto de vista psicométrico. Actualmente los trabajos realizados dentro del campo de la GAI se han centrado en la elaboración de modelos de ítems, como el punto medular para la elaboración de reactivos con estructura y propiedades psicométricas similares, pero aún no existe consenso en la terminología a utilizar para referirse a ellos.

2.3.2. Modelos de ítems. Conceptualización

Un elemento central en la GAI es la noción de Modelo de ítem, que sustituye al término especificación de reactivos. Mientras que este se refiere a las características que debe tener un solo ítem, el primero hace referencia a las características de una familia de ítems que mide el mismo constructo de manera equivalente, condición que los hace intercambiables.

Los modelos de ítems fueron documentados por primera vez en 1968 por Osburn, quien los llamó *formas de ítems* (*item forms*); sin embargo fueron Hively, Patterson y Page (1968) quienes desarrollaron más el concepto. Ellos propusieron un sistema para generar una gran cantidad de problemas de matemáticas, en donde cada forma de ítem contenía un texto fijo con elementos intercambiables y reglas que servían para establecer las directrices a seguir para realizar dichos intercambios. Posteriormente, en 1974 Minsky llamó *marcos* (*frames*) a las estructuras mentales que una persona elabora para representar información y dar significado a lo que aprende en una situación dada. El conocimiento que se adquiere cuando se tiene una vivencia se convierte en un referente genérico de ésta y de nuevas experiencias. Así, cuando la persona se enfrenta a situaciones nuevas, pero que comparten algún aspecto de las vividas anteriormente, recurre a los marcos y los modifica o adapta cuando les agrega información de la nueva situación. De esta manera, los marcos

cuentan con dos niveles: el nivel más alto es el de la información fija y que representa lo que es verdadero de una situación, y el nivel más bajo es el de los espacios llamados *ranuras* (*slots*), que son rellenadas cuando la persona parte de la información fija, pero intenta dar significado a una nueva situación.

De manera paralela, en el proceso de elaboración de ítems cuando se trabaja con marcos, se cuenta con una estructura fija en la base del reactivo. Esta estructura contiene algunas secciones consideradas críticas para obtener los procesos de emisión de respuestas que se busca de los examinados. Dichas secciones son intercambiables, al generar nuevos ítems que se encuentran relacionados entre sí.

El término *modelo de ítems* fue introducido por primera vez en 1986 por LaDuca, Staples, Templeton y Holzman, quienes lo utilizaron dentro del contexto de la generación de ítems como una herramienta para elaborar ítems *isomorfos*; es decir, con contenidos y propiedades psicométricas similares (Bejar, 2002). En 1989 Haladyna y Shindoll añadieron el término *molde de ítem* (*ítem shell*) para nombrar a una técnica creada para responder a las necesidades de certificación de los profesionales de la salud, quienes requerían desarrollar un sistema de evaluación que produjera ítems de opción múltiple de alta calidad, y cuya construcción no implicara mucha inversión de tiempo. Denominaron *molde* a todo ítem "hueco" que contenía la estructura sintáctica y el contexto de un reactivo, más no su contenido específico.

La técnica implica seleccionar ítems específicos, pertenecientes a un instrumento de evaluación elaborado y administrado previamente y cuya evidencia empírica indica cuáles de ellos cuentan con las propiedades psicométricas necesarias para ser utilizados como moldes. Una vez seleccionados los ítems se toma su estructura sintáctica básica (base del reactivo) y se elimina su contenido específico, dejando como resultado un molde en el cual

se pueden incrustar distintos elementos que lo convierten en un nuevo ítem pero con estructura similar al original, pues se conservaron las propiedades psicométricas del reactivo inicial. De esta forma se cuenta con una estructura gramatical que servirá como punto de partida para que el elaborador de los ítems solamente elija entre los contenidos específicos que desea evaluar, permitiéndole la posibilidad de tener ítems similares en su estructura genérica, pero distintos en sus particularidades (Haladyna y Shindoll, 1989).

En el año 1991 Solano-Flores propuso una técnica centrada en la extracción de contenidos para elaborar reactivos de opción múltiple y construir instrumentos de evaluación mediante bancos de reactivos. Llamó a su propuesta *Diseño lógico de exámenes*, y se basó en el uso de algoritmos para la descripción y el análisis del contenido a evaluar. El objetivo primordial de su propuesta fue ayudar a solventar problemas prácticos en el diseño y la elaboración de exámenes, tales como la extensión, definición de contenidos y dificultad del instrumento. El modelo de su propuesta consta de tres fases para el diseño de exámenes: 1) formalización del procedimiento (*algoritmo*), 2) análisis de variables, y 3) generación y análisis lógico de reactivos.

En la primera fase se debe describir el algoritmo subyacente al dominio de cualquier conocimiento, de tal manera que se conozca a detalle cada paso que sigue una persona para alcanzar una meta, obtener un producto o solucionar una tarea. En esta fase se utilizan representaciones gráficas de los procedimientos (grafos y diagramas de flujo), las cuales ayudan a identificar la secuencia de pasos que se siguen para ejecutar una tarea. Según la complejidad del procedimiento se pueden hacer diagramas simples o sofisticados en los que se ilustren las secuencias de pasos, así como todas las relaciones que se dan entre los elementos que lo conforman.

En la segunda fase se realiza un análisis de las variables implicadas en las decisiones que se toman a lo largo del procedimiento. El objetivo es conocer todas las combinaciones y relaciones que se dan entre las variables para asignarles valores. Estas combinaciones permiten identificar los distintos problemas que pudiera enfrentar una persona cuando intenta solucionar un problema.

En la tercera fase y última, se utiliza la información recabada mediante el análisis de variables y se generan los distintos reactivos. Se toma como punto de partida la cantidad, el tipo y la viabilidad de las combinaciones identificadas. El elaborador de los reactivos podrá manipular los valores haciendo lo que Solano-Flores (1991) llamó *arreglos*, con la finalidad de generar más reactivos.

La técnica de Solano-Flores (1991) representa uno de los primeros esfuerzos en México dentro del campo de la generación de reactivos y la construcción de bancos de reactivos, como un procedimiento sistemático y objetivo para estimar a priori la dificultad de estos. Mediante el conocimiento a profundidad de los procesos subyacentes a la toma de decisiones, así como la identificación y el análisis de las variables implicadas en cada paso de un procedimiento, el elaborador del instrumento puede tomar decisiones respecto a los niveles de dificultad que tendrán los reactivos de un examen.

Originalmente los moldes de ítems fueron creados con la intención de ser utilizados para generar ítems equivalentes para exámenes de opción múltiple en formato lápiz-papel; sin embargo, con el paso del tiempo fueron utilizados exitosamente en distintos campos del conocimiento y con diversas finalidades, incluido el desarrollo de sistemas válidos y confiables para evaluar la ejecución de los estudiantes (Draaijer y Hartog, 2007; Enright, Morley y Sheehan, 2002; Haladyna, 1991; Liu y Haertel, 2011; Shea et al., 1992; Simon, 1989; Solano-Flores y Shavelson, 1997; Solano-Flores, Shavelson y Schneider, 2001).

Una de las ventajas que se pueden observar en el uso de los moldes de ítems es la posibilidad de elegir el nivel de complejidad cognitiva que se desea evaluar mediante la ejecución de los examinados. Este dependerá del tipo de estructura gramatical (redacción) de la cual se parta en la base del reactivo. Así, las preguntas cerradas o enunciados incompletos servirán para evaluar habilidades de pensamiento de primer orden, mientras que aquellas estructuras que se utilicen para evaluar conceptos de orden superior, presentarán al elaborador de los reactivos una serie de instigadores contextuales que le permitirán flexibilizar la estructura y contar con distintas rutas. De ellas podrá elegir la que más se adecúe al tipo de ejecución que desea evaluar.

Solano-Flores et al. (2001) han mencionado que los beneficios de utilizar moldes (término que tradujeron al español como *templete*) se debían a que son: a) herramientas para desarrollar pruebas de respuesta construida, b) documentos que formalizan las propiedades estructurales de los ejercicios; c) ambientes para la creación de ejercicios que permiten estandarizar y simplificar los formatos de respuesta para los estudiantes, y d) herramientas conceptuales que regulan el proceso de desarrollo de exámenes.

Los moldes de ítems permiten cuidar aspectos de redacción, gramática y ortografía que de otra forma podrían pasar inadvertidos cuando distintas personas participan en la elaboración del instrumento. Esto es posible ya que los ítems se construyen a partir de una estructura sintáctica preestablecida, y la función primordial del elaborador del ítem es concentrarse en los contenidos a evaluar y no en la redacción de los mismos. Siempre y cuando los moldes de ítems se desarrollen con cuidado, de manera que midan dominios específicos del aprendizaje, además de ser buenas herramientas para desarrollar instrumentos de evaluación, aportan evidencias de validez de contenido, ya que evalúan contenidos representativos del constructo en cuestión (Solano-Flores, Jovanovic, Shavelson y Bachman, 1999).

Por otro lado, como algunos autores han señalado, las limitaciones de los ítems isomorfos creados mediante un mismo molde: que son vulnerables al entrenamiento de los examinados (Arendasy y Sommer, 2012; Gierl, 2007), y que la similitud en las propiedades psicométricas de ítems del mismo molde no es señal de isomorfismo, sino de que el constructo no fue debidamente definido en el modelo de ítems (Gierl, Lai y Breithaupt, 2012). Otra desventaja que se ha encontrado al elaborar ítems mediante esta herramienta es la sobreexplotación o el "agotamiento" de aspectos específicos del campo o contenido que se mide, ya que se corre el riesgo de dejar de lado otros aspectos igualmente importantes de medir. Por ello, se recomienda el uso de distintos tipos de moldes para elaborar ítems dentro de un mismo instrumento (Haladyna y Shindol, 1989; Solano-Flores et al., 2001).

Sin embargo, la elaboración de moldes de ítems no es tarea fácil, ya que implica una serie de procedimientos en los que se definen las características que deberán tener los ítems para tomar la decisión respecto al tipo de molde que se utilizará. Este proceso conlleva complicaciones cuya magnitud dependerá de dos factores: el tipo y nivel taxonómico del campo de conocimiento que se desee evaluar, y el grado de complejidad cognitiva implicado en la tarea que se requerirá para resolver el ítem que se genere a partir de él.

En los años 1996 y 2002 el término *modelo de ítems* fue retomado por Bejar cuando trabajó dentro del campo de la elaboración de pruebas generadas con base en un modelo. Para Bejar, un modelo de ítems es una construcción de una forma de ítem, elaborada para evaluar una tarea compleja. Posteriormente Bejar et al. (2003) continuaron utilizando el término y coincidieron con LaDuca et al. (1986), ya que lo definieron como una aproximación centrada en el constructo, cuyo potencial no solamente descansa en mejorar

los procesos para obtener evidencias de validez, sino en reducir los costos, en comparación con los métodos tradicionales para la elaboración de ítems.

El trabajo de Bejar y sus colegas (2003) arrojó una serie de principios importantes para el campo de la GAI, entre los cuales destaca la importancia de realizar un análisis del constructo como punto de partida en el diseño de ítems. Es a partir de dicho análisis que se deben elaborar los modelos de tareas y de ítems, utilizando controles muy estrictos, los cuales servirán para evitar la fase de pilotaje y calibración de los mismos. Lo anterior, según Bejar (2002), solamente funcionará si se logra obtener evidencias de validez de los métodos y procedimientos utilizados en la elaboración de los modelos de ítems, ya que se trabaja bajo la premisa de que un buen diseño de modelo de ítems es suficiente para garantizar a priori que los ítems creados mediante la GAI son isomorfos (Haladyna, 2012).

En el año 2002 también se introdujeron otros dos términos para referirse a los modelos de ítems. Singley y Bennett (2002) los nombraron *esquema* (*schema*), mientras Embretson (2002) utilizó la palabra *plano* (*blueprint*). Un esquema, según los autores, es un término que surgió a partir de su propuesta teórica llamada *Teoría del esquema*, utilizada para la generación automática y variación de ítems, el análisis de soluciones múltiples en un problema, así como para dar estructura a las instrucciones. Para ellos, todos los problemas que en apariencia son distintos contienen una estructura subyacente (esquema), la cual comparten.

Por otro lado, el término *plano* se refiere a una especificación de ítems que consiste en una matriz o diagrama que representa la cantidad de ellos a incluir en la prueba. Este sirve para identificar los dominios y las habilidades a evaluar. Además, permite crear ítems que se pueden utilizar de manera inmediata con la finalidad de evaluar una fuente

específica de su dificultad y nivel cognitivo, sin la necesidad de llevar a cabo pruebas empíricas para cada ítem (Embretson, 2002).

El término *plantilla* (*template*) fue empleado por un grupo de investigación liderado por Rinconscente, Mislevy y Hamel en el año 2005, para referirse a una estructura que contiene el plano preliminar o esbozo de un instrumento de evaluación, de la cual se pretende generar múltiples planos para evaluar aspectos o dominios específicos. Aunque varían en algunos aspectos, las plantillas comparten un mismo marco referencial y gracias a que su sistema de diseño se apoya en el uso de herramientas informáticas, contienen atributos que pueden ser fijos o variables. Entre sus componentes, se incluyen las definiciones de las variables del modelo de evaluación y de las tareas evaluativas, sus relaciones jerárquicas, los productos generados, procedimientos de evaluación y demás elementos relacionadas con una familia de tareas evaluativas (Mislevy y Rinconscente, 2006).

2.3.3. GAI mediante teoría fuerte y teoría débil

De los distintos términos que se han empleado para denominar y describir la forma de elaborar ítems y manipular sus componentes, el que más se utiliza en el campo de la GAI es *modelo de ítems* (Bejar, 1996, 2002; Bejar et al., 2003; LaDuca et. al, 1986). Los modelos de ítems se pueden elaborar mediante dos aproximaciones: *teoría fuerte* y *teoría débil*. También existe un tercer método llamado *artesanal* (*from art*) que se basa en los principios teóricos y creencias del elaborador del instrumento. Este último, a pesar de que resulta atractivo es poco práctico para la GAI, pues aunque el propósito de su uso es producir la mayor cantidad posible de ítems no presta la atención necesaria a los niveles de dificultad, y se aleja de ser un método estructurado para la elaboración de instrumentos con principios teóricos robustos y probados (Drasgow, Luecht y Bennet, 2006).

Cuando se utiliza la *teoría fuerte* para elaborar modelos de ítems, los esfuerzos se centran en develar los mecanismos cognitivos subyacentes al proceso de solución de los ítems generados, así como en estipular cuáles son los elementos específicos que contienen y determinan su nivel de dificultad (Gitomer y Bennett, 2002). Se trata de un proceso mediante el cual se utiliza un modelo cognitivo para no solo identificar dichos elementos, sino utilizar el conocimiento teórico que se tiene tanto del contenido evaluado, como de las habilidades y conocimientos que utilizan los examinados para responder el ítem. La intención es manipular sus propiedades y, por ende, su nivel de dificultad (Gierl y Lai, 2012). Se trabaja bajo la premisa de que, al conocer la dificultad de las demandas cognitivas del contenido que evalúan los ítems, se pueden predecir los parámetros de un modelo de respuestas y controlar características psicométricas de los reactivos, tales como la homogeneidad y dificultad (Bejar, 1993).

Debido a lo anterior, uno de los grandes beneficios del uso de este método es que los ítems generados no requieren ser piloteados antes de su incorporación al instrumento de evaluación. Gracias al sustento teórico subyacente los ítems se pueden elaborar de manera sistemática, a la vez que, atender los niveles específicos de complejidad cognitiva con los que cuentan los estudiantes (Lai, Alves y Gierl, 2009). Sin embargo, entre sus limitaciones está el hecho de que no existen suficientes teorías cognitivas como para llevar a cabo estos principios a la práctica. Además, en ciertos instrumentos como los exámenes de admisión o de rendimiento académico en los que se evalúan varios campos del conocimiento no es práctico realizar un análisis cognitivo tan exhaustivo (Gitomer y Bennett, 2002). Por ello, la elaboración de instrumentos bajo estos principios se ha limitado a campos específicos del conocimiento en los que ya existen modelos cognitivos establecidos (Gierl et al., 2012).

Por otro lado, existe el método de elaboración de modelos de ítems mediante *teoría débil*, cuyo punto de partida radica en el uso de un *ítem padre* (Drasgow et al., 2006). Este

ítem se puede crear de varias maneras, ya sea mediante la revisión de los ítems de pruebas que ya han sido administradas, eligiéndolo de un inventario de ítems existentes, o simplemente elaborándolo (Gierl y Lai, 2012). El término *teoría débil* se atribuye al hecho de que al construir instrumentos mediante este método no es necesario un análisis exhaustivo para conocer y determinar cuáles son los procesos cognitivos detrás del dominio de contenido evaluado y de las respuestas de los examinados, como sucede cuando se utiliza el método de teoría fuerte, sino que se trabaja mediante una *Teoría de invariancia*, en la que se considera que el elaborador del instrumento debe detectar mediante su experiencia, intuición, conocimiento teórico o investigación, aquellas características del ítem padre que no afecten su operación; una vez que las identifica, las modifica, de tal forma que se generan variantes del ítem padre (Drasgow et al., 2006).

Cuando se utiliza teoría débil con la intención de producir ítems estadísticamente calibrados, a partir del ítem padre, la tarea del elaborador será manipular únicamente aquellos elementos del ítem que sirvan para producir ítems isomorfos. Pero cuando el objetivo es producir ítems sin calibración estadística, entonces se deberán manipular las características que ayuden a producir grandes cantidades de ítems, independientemente de las propiedades psicométricas que arrojen. A diferencia de los ítems isomorfos generados, este tipo de ítems sí requerirán ser piloteados (Gierl y Lai, 2012).

Los instrumentos que exploran dominios amplios del conocimiento, tales como los exámenes de admisión, generalmente utilizan la teoría débil como base para su elaboración, ya que los modelos de ítems construidos bajo esta modalidad muestran todas las propiedades, características y elementos que afectan o no los niveles de dificultad del ítem, y que a su vez permiten hacerle modificaciones para generar grandes cantidades de ítems similares. El objetivo de la elaboración de instrumentos mediante teoría débil es el mismo que el de teoría fuerte (generar grandes cantidades de ítems calibrados, de manera

automatizada); sin embargo, el procedimiento es distinto, ya que en la teoría débil se utilizan reglas o directrices para su diseño (Gitomer y Bennet, 2002), y no un análisis o mapeo de los procesos cognitivos subyacentes al contenido evaluado y a las respuestas emitidas por el estudiante.

Dentro de las limitaciones encontradas en el uso de teoría débil para la generación automática de ítems destaca que son pocos los elementos del ítem padre que se pueden manipular. Esto restringe las posibilidades de generar grandes cantidades de ítems; además, los que se elaboran terminan generalmente siendo muy parecidos entre sí. Asimismo, debido a que esta modalidad de elaboración de modelos de ítems no se basa en una teoría fuerte que respalde la toma de decisiones, sino en la experiencia, los juicios y el conocimiento de expertos en contenido, a menudo ocurre que sus predicciones respecto a los elementos que alterarán las propiedades psicométricas de los ítems generados no son del todo correctas (Gierl y Lai, 2012). Además, al no poder predecir adecuadamente su comportamiento psicométrico, los nuevos modelos deberán ser calibrados empíricamente, para que los ítems que se produzcan hereden tales calibraciones (Drasgow et al., 2006).

Los modelos de ítems contienen todas las variables que se incluirán en una tarea evaluativa y serán manipuladas para crear distintas versiones de los ítems. Están conformados por tres elementos: la base del reactivo, las opciones de respuesta y la información auxiliar. La base del reactivo contiene el contexto, contenido, ítem y/o la pregunta, que deberá responder el examinado. Las opciones de respuesta incluyen la respuesta correcta y al menos un distractor o respuesta incorrecta, de tal manera que el estudiante elija la que considere pertinente; la información auxiliar se refiere a todos los contenidos adicionales que complementan tanto a la base del reactivo como a las opciones de respuesta y puede presentarse en forma de texto, imágenes, tablas, diagramas, sonidos o videos (Gierl et al., 2012).

Tanto la base del reactivo como las opciones de respuesta se pueden dividir en elementos de dos tipos: aquellos que contienen información no numérica llamada *cadenas* (*strings*), y valores numéricos llamados *integrales* (*integers*). Mediante la manipulación sistemática de estos elementos se pueden generar grandes cantidades de ítems o *instancias*, las cuales se pueden clasificar en dos tipos: ítems *isomorfos* e ítems *variantes*.

En los ítems isomorfos se busca elaborar ítems con propiedades psicométricas similares. Esto se logra mediante la manipulación de los elementos superficiales de los ítems llamados *incidentales*, porque modifican su apariencia más no ejercen influencia significativa sobre su dificultad o propiedades psicométricas. En el caso de los ítems variantes lo que interesa es generar ítems con propiedades psicométricas distintas, para lo cual se manipularán otro tipo de elementos llamados *radicales.* Estos últimos son componentes de la estructura de los ítems que están relacionados con los aspectos teóricos que les subyacen, y que funcionan como variables cuasi independientes. Estas variables, al ser manipuladas causan modificaciones estadísticamente significativas en las dificultades de los ítems, lo cual se determina mediante la medición del índice de error y/o el tiempo de ejecución (Irvine, 2002).

A pesar de la importancia que reviste el proceso de elaboración de los modelos de ítems, los esfuerzos para documentar los principios y procedimientos relacionados con su elaboración es aún incipiente y existen pocos ejemplos en la literatura que han documentado este proceso. Uno de ellos es el estudio que realizaron Bejar et al. en el año 2003 con el objetivo de determinar la factibilidad del uso de modelos de ítems creados bajo teoría débil en pruebas adaptativas. Estos autores elaboraron modelos a partir de un grupo de ítems liberados del examen Graduate Record Examination (GRE), identificando aquellos que tenían mayor probabilidad de ser administrados. Documentaron el procedimiento que utilizaron para determinar las restricciones implementadas en la modificación de los

elementos cadena e integrales contenidos en la base del reactivo. Así como las opciones de respuesta. El resultado que obtuvieron fue la generación de ítems isomorfos, cuya estimación de los índices de confiabilidad, discriminación y otros parámetros se pueden aplicar a todas las instancias creadas mediante un mismo modelo de ítems, sin necesidad de probar empíricamente cada uno.

Sin embargo, también encontraron que existe deterioro en la precisión de la medición. Por ello, recomendaron la realización de un mayor número de estudios que permitan comprender los niveles de impacto de los modelos de ítems sobre la precisión de los puntajes (Bejar et al., 2003).

El estudio de Bejar et al. (2003) abrió el camino a la generación automática de ítems debido a que su procedimiento permitió examinar la arquitectura detrás de la elaboración de modelos de ítems creados bajo teoría débil. Una de las principales contribuciones de dicho trabajo fue la producción de ítems que cumplen dos características importantes: calidad psicométrica y contenido suficiente, como para incursionar en la generación de exámenes sumativos y formativos (Haladyna y Rodríguez, 2013).

Por su parte, Gierl, et al. (2008) propusieron una taxonomía para la elaboración de 10 distintos modelos de ítems bajo teoría débil, que utilizaron para generar ítems de diversas áreas del conocimiento. Para ello, diseñaron un sistema informático con la finalidad de crear instancias de cada modelo. Mediante la combinación de elementos de las distintas secciones de los modelos de ítems y el uso de tecnología, establecieron una aproximación integral al diseño de pruebas, y proporcionaron dos ejemplos detallados de cada uno de los 10 tipos de modelos propuestos en su taxonomía. Encontraron que el rol del experto en contenido es crítico durante la fase creativa de la elaboración de modelos de ítems útiles, pero que también lo es el papel que juega la tecnología, debido a la compleja tarea que

representa combinar mediante algoritmos todos los elementos contenidos en los modelos de ítems y almacenarlos en bancos. Concluyeron que al combinar el conocimiento de los expertos con la tecnología, los modelos de ítems se pueden utilizar para generar el contenido de todas las secciones que contempla un instrumento. De esta forma, consideraron que el uso de teoría débil para estos efectos es un procedimiento muy práctico para la generación de ítems de manera automática (Gierl et al., 2008).

El uso de modelos de ítems ha demostrado tener grandes ventajas en la generación automática de instrumentos de evaluación. Sin embargo, como se mencionó anteriormente, una de las grandes limitaciones que se han observado es que a diferencia de los métodos tradicionales para elaborar ítems (donde se cuenta con una definición clara del constructo a medir y los ítems elaborados pueden presentar gran variedad en su formato, contenidos y demanda cognitiva), todos los ítems generados mediante un mismo modelo son prácticamente iguales en cuanto a su apariencia. El uso de un sustento teórico fuerte, así como la incorporación de los principios de la psicología cognitiva, como base de un marco referencial que apoye la generación de ítems de manera automática, será el camino a seguir para resolver esta situación. Asimismo, cuando esto suceda, es probable que la GAI ayude a guiar los esfuerzos de los elaboradores de ítems en un primer momento, para posteriormente sustituirlos por completo (Haladyna y Rodríguez, 2013).

2.4. Validez de instrumentos de evaluación a gran escala

Uno de los documentos de consulta más importantes para el desarrollo y la evaluación de instrumentos de medición psicológica y educativa es el que se publicó en 1999 y recientemente en 2014, por un comité conjunto compuesto por la American Educational Research Association (AERA), la American Psychological Association (APA) y el National Council on Meassurement in Education (NCME). Se trata de *Standards for Educational and Psychological Testing* (*Estándares para elaborar pruebas educativas y psicológicas*). En ese

documento se establece un marco de referencia que contiene los lineamientos que todo organismo dedicado a la elaboración de instrumentos de evaluación debe seguir para el desarrollo, uso y la evaluación de los mismos.

Este documento hace hincapié en la importancia que tiene la aplicación del juicio profesional en la interpretación y el uso que se le dé a la información que se derive de dichos instrumentos (Linn, 2006). Aunque el documento menciona diversos aspectos que se deben contemplar en la construcción, el desarrollo, la evaluación, administración y documentación de una prueba, deja en claro que la validez es la consideración más importante que se debe tomar en cuenta al desarrollar y evaluar pruebas.

En este sentido, los Estándares definen la validez como:

> El grado en el que la evidencia y la teoría dan soporte a la interpretación que se hace de los puntajes que derivan del uso de las pruebas, según los propósitos con los que éstas son usadas. La validez es el aspecto fundamental que se debe considerar al desarrollar y evaluar instrumentos de evaluación. El proceso de validación implica acumular pruebas o evidencias relevantes, que proporcionen bases científicas para hacer las interpretaciones de los puntajes. Son las interpretaciones de las pruebas ante usos específicos, las que deben ser evaluadas y no la prueba misma. Cuando los resultados de una prueba son interpretados de distintas maneras (p. ej. para describir tanto los niveles de posesión del atributo medido, como para hacer predicciones de resultados futuros), cada una de esas interpretaciones debe ser validada (AERA, APA y NCME, 2014, p. 11, traducción libre).

De acuerdo con lo anterior, la atención se centra en la responsabilidad que tienen los desarrolladores de instrumentos con los usuarios de las pruebas, tanto los examinados como los actores involucrados en el proceso evaluativo. Serán ellos quienes harán inferencias respecto a los puntajes derivados de la aplicación de un instrumento; por ello,

el uso que se pretende hacer de las puntuaciones de las pruebas, debe estar debidamente justificado.

Históricamente, el concepto de *validez* ha tenido diversas connotaciones. Anastasi (1988) indica que la validez proporciona un control directo de la forma en que el instrumento de medición cumple su función. Por su parte Popham (1990) profundiza en la definición. Al hacer mención del uso que tienen los instrumentos de evaluación dentro del campo educativo establece que la diferencia entre una prueba válida y la validez de la información que de ella deriva, estriba en que una prueba es solamente una herramienta que proporciona datos que sirven para hacer inferencias, mientras que la interpretación que se hace de esos datos es lo que realmente debe considerarse válido o no.

Dentro del proceso instruccional la toma de decisiones está basada principalmente en inferencias. El uso creciente de pruebas dentro de dicho contexto exige que los instrumentos de evaluación que se utilizan proporcionen los elementos necesarios para poder hacer inferencias válidas. Por este motivo, Popham (1990) asegura que el concepto de validez, dentro del campo de la evaluación educativa, está estrechamente relacionado con la validez de las inferencias derivadas de los puntajes obtenidos de un instrumento de evaluación.

Por su parte, Messick (1993), en concordancia con lo que menciona Popham (1990), señala que la validez es un juicio evaluativo integral del grado en el que la evidencia empírica y la racionalidad teórica dan sustento a qué tan apropiadas y adecuadas son las inferencias y acciones basadas en los puntajes de una prueba o cualquier tipo de evaluación. Con esta definición coincide el Commiteee on the Foundations of Assessment (2001) del National Research Council, al hacer referencia a la validez como el grado en el que la evidencia y la teoría dan sustento a las interpretaciones derivadas de los puntajes de una evaluación. Por lo tanto, la validez considera y contrasta evidencias de tipo práctico y teórico con dos

aspectos derivados de la aplicación de un instrumento de evaluación: las interpretaciones que se hacen de las puntuaciones, y el uso que se da a dichas interpretaciones.

Kerlinger y Lee (2002) señalan que la definición más común de validez se orienta a responder: si el instrumento en cuestión mide aquello para lo cual fue diseñado y no otra cosa. Tanto Anastasi (1988), como Kerlinger y Lee sostienen que no existe un tipo único de validez, y que un instrumento o escala es válida en relación con el uso práctico o científico que el usuario haga de él. A esto se añade que existen tres tipos de validez: la de contenido, de criterio y de constructo.

Hace poco más de una década, Borsboom y Mellenbergh (2004) llevaron a cabo una revisión del concepto de validez. Ellos argumentaron que en la actualidad se requiere de una definición simple y útil que hasta ahora no se ha podido obtener. Su trabajo se centró en la realización de un análisis de las consideraciones que han llevado al estado actual del concepto y, a su vez, mostrar los aspectos que consideran irrelevantes, de tal forma que propusieron una alternativa clara, simple y práctica para abordar la validez.

Borsboom y Mellenbergh (2004) afirmaron que la validez no es algo complejo, con facetas dependientes de redes nomológicas o consecuencias sociales del uso de las pruebas, sino que se trata de un concepto elemental que se encuentra relacionado con el hecho de que una prueba es válida si mide lo que se supone debe medir. Argumentaron que una prueba es válida para medir un atributo, siempre y cuando cumpla con dos características: a) que el atributo efectivamente exista, y b) que las variaciones en el atributo produzcan variaciones en el resultado del procedimiento de medición. Basaron su idea principal en la Teoría de la medición causal (Borsboom, 2008; Borsboom, Mellenbegh y Van Heerden, 2004; Markus y Borsboom, 2013). Así, su planteamiento apunta a que la validez es una *propiedad* de las pruebas.

Como se ha visto, el tema de la validez de un instrumento ha sido ampliamente estudiado a través de los años y, conforme se han presentado avances en este campo, se ha otorgado mayor reconocimiento al hecho de que el desarrollo de un instrumento válido requiere de una multiplicidad de procedimientos que se deben realizar de manera secuenciada en las distintas etapas de la construcción del mismo (Anastasi, 1992).

Inicialmente se distinguían distintos tipos de validez, entre los cuales se mencionaban la de contenido, predictiva, convergente, discriminante y de constructo. Sin embargo, los Estándares actuales (AERA, APA y NCME, 2014) apuntan hacia un concepto unitario y los diferentes tipos de validez son considerados como distintas formas para obtener evidencia de un tipo único de validez, a la cual se le puede denominar *validez de constructo* (Martínez, Hernández y Hernández 2006).

Aunque el tema de la validez ha presentado algunas dificultades en su concepción, la mayoría de los autores coinciden en que se trata de una propiedad del instrumento de medida y que existen distintos tipos de validez. Sin embargo, la interpretación que se haga de esta información, se debe manejar con discreción, ya que pudiera causar confusiones o usos inapropiados del concepto. Al respecto, es de gran ayuda tener siempre presente que además de referirse al uso de las puntuaciones que derivan de una prueba, la validez es cuestión de grado y no de juicios absolutos (todo-nada), se basa en diferentes tipos de evidencia e implica un juicio evaluativo global en términos del soporte que brinde a sus interpretaciones (Popham, 2000).

La concepción de validez en el presente trabajo doctoral es la postulada en los Estándares publicados en 2014 por la AERA, APA y NCME, citada al inicio de esta sección. A continuación se revisa brevemente la clasificación de los distintos tipos de evidencia relacionada con la validez, y en el siguiente apartado se ahonda en la validez de contenido, debido a que es uno de los ejes sobre los cuales trata esta investigación.

2.4.1. Evidencias de validez relacionadas con el criterio

Este tipo de evidencias, se pueden obtener mediante la comparación de las puntuaciones de una prueba con uno o más criterios o variables externas, de las cuales se sabe con certeza que miden el atributo en cuestión (Kerlinger y Lee, 2002); es decir, se hace referencia a la correlación que existe entre las puntuaciones del instrumento y un criterio externo o elemento criterio.

Tradicionalmente existen dos tipos de evidencias de validez relacionadas con el criterio: la concurrente y la predictiva. La diferencia entre ambas estriba en la presencia de un intervalo de tiempo.

En los estudios de validez *predictiva* se involucra el uso de desempeños futuros del criterio que se pretende medir mediante el instrumento. Por ejemplo, para obtener este tipo de evidencia de validez en un instrumento diseñado para medir habilidades cuantitativas, lo que procede es administrar la prueba a un grupo de estudiantes al inicio del ciclo escolar, esperar a que concluyan el grado académico y comparar las calificaciones que obtuvieron en la asignatura de matemáticas con las del instrumento. Lo que se busca es el valor predictivo de la prueba, respecto a las calificaciones de los estudiantes al finalizar su grado escolar (desempeño futuro del criterio). Así, el criterio se fija en el futuro y, por ende, el proceso implicará una demanda de tiempo considerable.

En la validez concurrente no existe la presencia de la dimensión del tiempo, ya que los resultados del instrumento se correlacionan con el criterio en el mismo momento o de forma paralela. Este tipo de evidencias se utilizan generalmente cuando se desarrollan instrumentos nuevos, en comparación con instrumentos consolidados.

2.4.2. Evidencias de validez relacionadas con el constructo

La validez relacionada con el constructo quizá es la más importante, ya que se refiere al grado de coincidencia que existe entre los planteamientos teóricos y conceptuales propuestos, y los resultados obtenidos con la aplicación del instrumento. En 1955 Cronbach y Meehl presentaron una propuesta del concepto, la cual partía de los trabajos realizados por el Comité de Pruebas Psicológicas de la APA, durante los años 1950 a 1954. Su producto fue la elaboración en 1954 de las *Recomendaciones técnicas* (primeros estándares) donde se distinguían cuatro tipos de validez, siendo la de constructo el concepto más innovador.

Los esfuerzos de Cronbach y Meehl (1955) se centraron en explicar los conceptos y ahondar en las implicaciones de cada tipo de validez. Para ellos, la validez de constructo se debe investigar siempre que se quiera medir un atributo o cualidad que no se puede definir operacionalmente; es decir, que no exista un criterio o universo de contenido que lo expliquen a cabalidad o ayuden a definirlo. Afirmaban que es deseable que todo instrumento cuente con una definición de los constructos psicológicos que explican la ejecución en un test, ya que se llevarán a cabo mediciones y posteriormente interpretaciones a partir de los datos. Aunque consideraban que la validez de constructo es fundamental, aún no la concebían como un marco referencial y de organización general de la validez.

Fue en los trabajos posteriores de Cronbach, de 1971, que se mencionó la existencia de distintas aproximaciones a la validez de constructo, vislumbrándose la unificación de los tipos de evidencias derivadas de un test como un concepto unitario. Explicó que debido a que el fin de la validez es la comprensión y explicación de los constructos medidos, entonces todo tipo de validez es de constructo. Con esto enfatizó la necesidad de integrar distintos tipos de evidencias de validez para evaluar las interpretaciones y los usos propuestos para las puntuaciones de los tests.

Las distintas ediciones de los Estándares (AERA-APA-NCME, 1985, 1999 y 2014) han concebido la validez como un concepto unificado, pero que requiere distintos tipos de evidencias. Entre ellas se encuentra la de constructo, utilizada para brindar explicación de los resultados, a partir de la teoría que subyace al instrumento y los constructos que pretende medir. En este sentido, la validez de constructo se refiere a la precisión con la que el instrumento mide una característica o rasgo, e integra la evidencia que soporta la interpretación del sentido que poseen las puntuaciones del instrumento, explicando el modelo teórico y empírico que subyace a la variable de interés, llamada también *constructo* (Hernández, Fernández y Baptista, 2006).

2.4.3. Evidencias de validez relacionadas con el contenido

Bajo la consideración de que cualquier atributo psicológico o educativo posee un universo teórico de contenido, el cual consiste en todo lo que se puede observar, decir y opinar de dicho atributo (Kerlinger y Lee, 2002), la validez de contenido responde a preguntas como: ¿en qué medida los contenidos que evalúa la prueba cubren el atributo deseado?, ¿los contenidos de la prueba son representativos del universo de contenidos que rodean al atributo en cuestión?, y ¿cómo se puede lograr que el instrumento cuente con suficientes elementos como para considerarlo representativo del atributo que se pretende medir?

Para responder dichas interrogantes es importante destacar que la medición de atributos, propiedades o constructos no es tarea fácil, ya que está estrechamente relacionada con la representatividad de la teoría de la cual estos derivan. Por ello, la validez de contenido se basa en juicios profesionales de la relevancia que tiene el contenido de la prueba respecto al contenido de un dominio de interés específico; además, mediante este tipo de validez se busca juzgar la representatividad con la que el contenido de las tareas o los ítems que contiene una prueba cubren dicho dominio.

La validez de contenido no se preocupa por los resultados que derivan de la aplicación de un instrumento, por la estructura interna o externa de la prueba, las diferencias en la ejecución de los examinados ni por las consecuencias sociales que deriven de la interpretación de los resultados. La validez de contenido proporciona evidencias que dan soporte a la relevancia y representatividad del contenido de la prueba (Messick, 1993). Lo anterior sugiere que todo instrumento que se elabore con fines evaluativos, deberá incluir entre sus contenidos una muestra de elementos que sea suficientemente representativa como para conformar el dominio que se pretende medir.

Al paso de los años, se han presentado diversas definiciones del concepto de validez de contenido. Por ejemplo, se cuenta con las distintas ediciones de los *Standards for educational and psychological testing* (AERA-APA-NCME, 1985, 1999 y 2014), así como las definiciones propuestas por autores como Anastasi (1988), Suen (1990), Messick (1993), Nunnally y Bernstein (1994) y Walsh (1995). Aunque muestran diferencias dentro de las particularidades del concepto, comparten aspectos importantes que se resumen en que la validez de contenido es el grado o nivel en que los elementos de un instrumento de evaluación son relevantes y representativos del constructo que se mide con un propósito particular (Haynes, Richard y Kubany, 1995).

Al revisar este concepto, se puede afirmar que los *elementos* son todos aquellos aspectos del proceso evaluativo que afectan de manera directa o indirecta la obtención de datos, tales como la forma en que se encuentran redactados los ítems, los formatos en los que responderá el estudiante, y la claridad, especificidad y estandarización de las instrucciones que se proporcionen a los respondentes, entre otras cosas. El *grado* o *nivel* se refiere a que la validez de contenido se basa en juicios cuantitativos que proporcionan distintos valores, los cuales son sometidos a un proceso de análisis, con la finalidad de obtener conclusiones respecto a la pertinencia de las decisiones tomadas para la construcción del instrumento. El *constructo* implica la aproximación a la evaluación de todos

aquellos conceptos, atributos y variables que se pueden medir. El *propósito particular del instrumento* de evaluación determina los índices de relevancia y representatividad del mismo, así como la forma en la que estos pueden variar, de acuerdo con la función que tenga la evaluación. Son *relevantes* cuando sirven al propósito evaluativo, y *representativos* cuando son proporcionales a las facetas que conforma al constructo, llámense áreas, factores, dimensiones, secciones, etcétera.

Dentro de la Teoría clásica de los tests (TCT), el contenido de los ítems de un instrumento de evaluación es considerado válido, siempre y cuando el dominio que fue seleccionado para medir el constructo se pueda considerar representativo del mismo. Por ejemplo, un instrumento que pretende medir el constructo *habilidades verbales* debe contener una cantidad de ítems proporcional para cada una de las áreas que, según el referente teórico, conforman dicho constructo (Lynn, 1986; Nunnally y Bernstein, 1994; Suen y Ary, 1989).

De acuerdo con Anastasi (1988), la validez de contenido proporciona evidencias respecto a la validez de constructo de un instrumento de evaluación. Esto es, si la validez de constructo se refiere al grado en que un instrumento de evaluación mide el constructo en cuestión, y a su vez implica, según Messick (1993), todos los tipos de validez. Por tanto, desde esta perspectiva la validez de contenido es un componente indispensable para la obtención de la validez de constructo, ya que proporciona evidencias respecto al nivel y grado en que los elementos de un instrumento de evaluación son relevantes y representativos del constructo que se pretende medir. Debido a que la validez de contenido implica parte del proceso de obtención de evidencias de validez de un instrumento, sus procedimientos llevan, de manera natural, su mejoría y también una mejor definición del constructo que se pretende medir (Smith y McCarthy, 1995).

En la práctica, Messick (1993) señala que las evidencias de validez relacionadas con el contenido usualmente se obtienen mediante el juicio consensual de expertos, quienes emiten opiniones respecto a la relevancia del contenido de cada ítem de la prueba, en relación con el dominio que se supone representa. Dicho grupo deberá considerar todos aquellos aspectos del procedimiento evaluativo que afecten de manera significativa la ejecución de los examinados, entre los cuales se encuentra la especificación del referente del constructo en términos del contenido temático, así como los comportamientos y procesos subyacentes. También se requieren las especificaciones de la prueba, en las que se delinea el formato de presentación de los estímulos (ítems), las alternativas u opciones de respuesta, las condiciones de administración y los criterios que se seguirán para la obtención de los puntajes. Los expertos deberán tomar en consideración y evaluar el grado en el que el formato de cada ítem instiga un tipo de respuesta, y si ésta se apega al contenido específico que se supone debe estar evaluando.

Por su parte Lynn (1986) ha mencionado que para establecer si un instrumento cuenta con evidencias de validez de contenido, se requiere efectuar un procedimiento de dos fases con cinco etapas: desarrollo y jueceo-cuantificación (ver Figura 2.1).

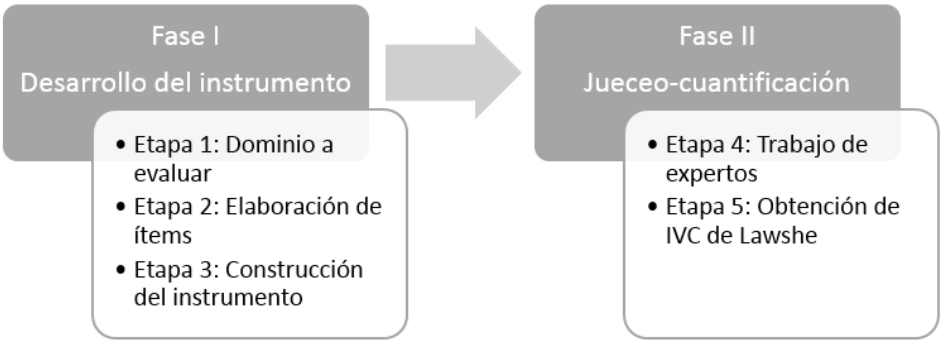


Figura 2.1. Modelo para obtención de evidencias de validez de contenido de Lynn (1986).

La primera fase es la de *desarrollo del instrumento*. Se lleva a cabo mediante una revisión exhaustiva de la literatura, así como la consulta a expertos en la materia. En ella se establecen tres etapas, las cuales varían según el objetivo que se persiga con la evaluación (Lynn 1986):

1) Identificación del dominio del contenido que interesa evaluar, lo que implica el uso de tablas de especificaciones o de mapas de los dominios de contenido.

2) Generación de ítems que conformarán una versión preliminar del instrumento, que se elaboran tomando secciones de los dominios de contenido identificados, de tal manera que se asegure que cada contenido mencionado en la tabla de especificaciones esté debidamente representado.

3) Construcción del instrumento, la cual consiste en tomar los ítems elaborados, para hacerles ajustes básicos y ordenarlos de manera específica para su presentación.

La segunda fase, llamada *jueceo y cuantificación de resultados,* incluye acciones de la cuarta y quinta etapa del procedimiento. En la etapa de jueceo (cuarta etapa) se involucra a un grupo de revisores expertos, quienes determinan la validez de contenido de cada uno de los ítems que conforman el instrumento. En esta etapa la cantidad de expertos que participan generalmente se establece de manera arbitraria, debido a que es muy difícil seleccionarlos en forma sistemática, ya que se eligen con base en criterios como: cuántos de ellos se pueden localizar y cuántos están dispuestos a colaborar en el proceso. Lynn (1986) propuso la participación de un mínimo de cinco expertos, excepto en áreas de contenido muy especializadas, en las cuales es difícil localizar a personas que cumplan el perfil, y en las que se recomienda incluir al menos tres. Asimismo, afirmó que, a pesar de no haber un límite establecido de participantes, es poco probable que supere la cantidad de 10.

La quinta y última etapa del procedimiento es lo que (Lynn 1986) llamó *cuantificación de resultados.* Esta etapa implica que los expertos evalúen el instrumento de manera global, para determinar si cuenta con evidencias de validez de contenido. Se debe calcular un valor llamado *índice de validez de contenido* (Lawshe, 1975), que se basa en el error estándar de la proporción de expertos que determinan la presencia de evidencias de validez de contenido en todo el instrumento. Es muy importante que los expertos sigan un procedimiento estructurado para emitir su dictamen, en el cual consideren aspectos de cobertura y relevancia (entre otros), calificando con base en una escala ordinal. El índice será el resultado de la proporción del total de ítems que fueron evaluados con validez de contenido.

Para Lynn (1986), esta forma de aproximación de la validez de contenido en dos etapas es fundamental para prácticamente todos los instrumentos de evaluación, ya que cuando se ha optado por obtener evidencias de validez de contenido mediante un solo procedimiento (desarrollo del instrumento o trabajo de expertos), ha habido cuestionamientos respecto a si realmente se trata de una forma de obtención de evidencias de validez o no (Jensen, 1980; Messick, 1981).

Por su parte Porter (2002) ha señalado la importancia de conocer el grado de congruencia o alineación entre los aprendizajes esperados, marcados en el currículum, y los exámenes creados para evaluar el dominio de dichos aprendizajes, por parte de los estudiantes. Propuso una forma de obtener evidencias de ello mediante la aplicación de encuestas a docentes. Utilizó una escala para asignar valores al tiempo que dedican a la enseñanza de los contenidos, así como al énfasis otorgado en el aula. Su propuesta sugiere la relevancia de revisar no solamente la alineación entre el examen y el currículum, sino entre lo que ocurre en el aula cuando el docente selecciona materiales y emplea su estilo instruccional, y el mismo examen.

Porter (2002) partió de la noción de que los docentes, por ser quienes interactúan directamente con los estudiantes, son los que determinan qué y cómo se enseña en el aula; además, toman decisiones respecto al énfasis que darán a los temas del programa, así como el tiempo que dedicarán para cubrirlos. Consideró de suma importancia revisar si existe alineación entre la forma de enseñanza del docente, el tiempo y el énfasis que da a los contenidos, así como los materiales educativos que utiliza para ello.

Para revisar si existe alineación entre los contenidos instruccionales y los materiales educativos que utilizan los docentes, sugirió la creación de un lenguaje común para describir los contenidos que se enseñan en el aula. Diseñó un cuestionario que aplicó a docentes, el cual incluye una tabla de doble entrada donde las columnas representan diversas categorías para describir la demanda cognitiva de cada contenido que enseñan, mientras que los renglones incluyen los temas o contenidos que enseñan. De esta manera, los docentes deben asignar un valor al grado de cobertura y énfasis que otorgan en el aula a cada demanda cognitiva de los contenidos revisados.

Los valores arrojan índices de alineación que ayudan a determinar la relación entre los resultados del aprovechamiento escolar y los contenidos instruccionales, o el currículum y los instrumentos de evaluación. Se responden preguntas como: cuando la alineación es baja, ¿qué es lo que la provoca?, ¿en qué áreas sí existe alineación? De esta manera se puede conocer objetiva y sistemáticamente el grado de alineación entre la instrucción, la evaluación y los contenidos que marca el currículum. Propuestas como la de Porter (2002) han aportado herramientas importantes en este campo de trabajo.

2.4.4. Aproximaciones a la medición de la validez de contenido

Para registrar de manera sistemática las opiniones de los expertos, existen procedimientos estadísticos que evalúan la relevancia de las opiniones y arrojan índices que sirven para calcular la congruencia que existe entre el ítem y la especificación. Algunas

técnicas como el análisis factorial o el escalamiento multidimensional de las calificaciones de los expertos pueden servir, tanto para documentar la naturaleza y grado de consenso al que se llega, como para develar diferencias en las opiniones y puntos de vista respecto a la importancia o relevancia que se juzgue sobre los dominios de contenido específicos o sobre el contenido de los ítems (Tucker, citado por Messick, 1993).

Por otro lado, la calidad técnica de los ítems también debe ser evaluada; es decir, todos aquellos aspectos que tengan que ver con la presentación de los ítems, tales como: claridad en la redacción; presencia de ambigüedades o aspectos irrelevantes en el ítem que no sean útiles para los objetivos de la evaluación; funcionamiento y utilidad de los distractores y de la respuesta correcta; demanda cognitiva y características de los formatos digitales, entre muchas otras cosas. Uno de los métodos propuestos durante la década de los años setenta para cuantificar la validez de contenido fue el que desarrolló Lawshe en 1975, en el que propone el empleo de paneles de expertos. Los integrantes de estos paneles proporcionan opiniones respecto a los ítems que conforman un instrumento de evaluación, y al instrumento mismo. Hasta la fecha este método es uno de los que más se utilizan cuando se aborda el proceso de validación del contenido de un instrumento.

Lawshe (1975) propuso una fórmula para obtener la razón de la validez de contenido (*content validity ratio* o CVR, por sus siglas en inglés), cuya fórmula es la siguiente:

$$\mathrm{CVR} = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

Donde:

- n_e es la cantidad de expertos que emitieron una respuesta de *ítem esencial.*
- N_2 es el total de expertos.

- CVR es una transformación lineal directa del porcentaje de expertos cuyas respuestas fueron *ítem esencial*.

Existen algunos principios propuestos por el autor, para determinar los valores del CVR:

1) Cuando menos de la mitad de los expertos emiten una opinión de *ítem esencial*, entonces el CVR es negativo,
2) Cuando la mitad de los expertos opina *ítem esencial*, mientras que la otra mitad opina lo contrario, el valor del CVR es cero.
3) Cuando todos los expertos opinan que el *ítem* es *esencial*, entonces el valor del CVR es 1.00 (por motivos de manejo estadístico, Lawshe propuso que se ajustara a 0.99).
4) Cuando más de la mitad y menos del total de los expertos opinan que el *ítem es esencial*, el valor del CVR oscila entre 0 y 0.99.

Esta fórmula se utiliza con la intención de identificar aquellos ítems que se deben rechazar o mantener en la prueba, y así mejorar el instrumento. Pero en ningún momento se puede decir que sustituye el uso de procedimientos analíticos tradicionales, como podría ser el índice de discriminación. La Tabla 2.1 indica los valores mínimos que se proponen para el CVR, con un nivel de significancia de 0.05, según la cantidad de expertos que conforman el grupo. Esto asegura que los acuerdos entre expertos no son sean al azar.

Tabla 2.1.
Valores mínimos de CVR y CVRt, de acuerdo con lo propuesto por Lawshe (1975)

Cantidad de expertos	Valor mínimo aceptado*
5	.99
6	.99
7	.99
8	.85
9	.78
10	.62
11	.59
12	.56
13	.54

14	.51
15	.49
20	.42
25	.37
30	.33
35	.31
40	.29

Nota: Adaptado de "A quantitative approach to content validity", por C. Lawshe, 1975, *Personnel Psychology*, 28, p. 568.
*****Prueba de una cola, p= .05

Lo que se puede observar es que entre mayor sea la cantidad de expertos dentro del grupo, menor será el porcentaje de acuerdo mínimo necesario entre ellos. Lawshe (1975) también propuso el uso del índice de validez de contenido (*content validity index* o CVI, por sus siglas en inglés). Este índice implica el uso de la media de los valores del CVR de todos los ítems que se conservarán en el instrumento. Se utiliza para todo el instrumento y no para los ítems en forma aislada.

Además del CVR y del CVI de Lawshe, existen otras aproximaciones a la obtención de evidencias de validez de contenido. Por ejemplo, Nunnally y Bernstein (1994) sugirieron que para medir la validez de contenido se requiere demostrar la presencia de consistencia interna, mediante el establecimiento de correlaciones entre los puntajes obtenidos en una evaluación y otro tipo de medida que evaluará el mismo constructo, para así poder mostrar diferencias entre los puntajes del postest y del pretest.

Una propuesta realizada por McGartland, Berg-Weger, Tebb, Lee y Rauch en el año 2003, definió la validez de contenido como el grado en que los ítems de un instrumento de medida evalúan un tipo de contenido, y cómo éste debe ser debidamente muestreado para obtener un instrumento representativo del dominio que se pretende medir. Los autores mencionaron que este tipo de validez se puede definir de dos maneras: *validez de faz o cara* y *validez lógica*. Así, consideraron dos aproximaciones para medir la validez de contenido, donde la validez de faz simplemente implica un juicio sobre qué tan válida “aparenta” ser la medida obtenida; mientras que la validez lógica implica un proceso más riguroso, como

el uso de paneles de expertos para la emisión de juicios evaluativos sobre la pertinencia, relevancia y otros aspectos del contenido que se pretende medir mediante el instrumento.

McGartland et al. (2003) coinciden con la propuesta de Lawshe (1975), al señalar que el uso de paneles de expertos proporciona una retroalimentación constructiva respecto a la calidad del instrumento, además de criterios objetivos mediante los cuales se puede evaluar cada uno de los ítems que lo conforman. Ellos añadieron que, aunque es costoso de inicio, el proceso representa una inversión y al mismo tiempo ahorro, debido a que evita procesos de pilotaje, calibración de ítems y revisiones posteriores que se vuelven costosas y redundantes.

El uso de expertos no se encuentra libre de limitaciones. Una de ellas se refiere a la subjetividad implícita en el tipo de retroalimentación que realizan, ya que pueden presentar prejuicios, los cuales surgen en el momento de las interacciones en grupo. Otra limitante es que la validez de contenido no elimina la necesidad de realizar estudios psicométricos adicionales. Además, mediante la validez de contenido no necesariamente se logra identificar aquellos contenidos que se dejaron fuera del instrumento en el proceso de selección de los mismos, tarea que no realiza el panel de expertos; sin embargo, esto se puede atender solicitando a los expertos la sugerencia de incluir otros contenidos y/o ítems en la prueba.

Sireci y Faulkner-Bond (2014) han mencionado que los aspectos más importantes para garantizar la calidad de un proceso de obtención de evidencias de validez de contenido de un instrumento radican en las características de los expertos. Es decir, ellos ubican la importancia de una selección cuidadosa, tomar en cuenta que posean amplia experiencia en el tema, que tengan pleno conocimiento de los contenidos que estarán evaluando, y que reciban una capacitación adecuada respecto al trabajo que realizarán, de tal manera que al revisar los ítems y llevar a cabo las tareas asignadas, lo hagan de manera minuciosa.

Por tanto, la selección de los expertos se debe efectuar mediante un escrutinio riguroso de sus antecedentes. McGartland et al. (2003) proponen conformar el grupo con expertos y personas no expertas, donde los primeros sean profesionistas y académicos que hayan publicado o trabajado en el área que evalúa el instrumento que será sometido a jueceo, y los no expertos sean personas para las cuales el tema es muy relevante, como es el caso de la población a la cual se aplicará el instrumento. Este segundo grupo se enfoca en la evaluación de aspectos como la redacción, la claridad de las instrucciones, y el uso de tecnicismos o palabras complejas, entre otras cosas.

Para saber cuántos expertos son necesarios para considerar que el panel está conformado por un número suficiente de personas como para garantizar un proceso objetivo, hay diversas recomendaciones. Para algunos el grupo debe estar integrado por un mínimo de tres expertos (Lynn, 1986). Para otros puede conformarse por un número entre 2 y 20 personas (Gable y Wolf, 1993; Walz, Strickland y Lenz, 1991), y algunos más señalan que debe ser de mínimo 3 y máximo 10 expertos, con la consideración de que habrá expertos y no expertos (McGartland et al., 2003). Sin embargo, todo dependerá del nivel de conocimientos y experiencia necesarios en los miembros del panel que se seleccionen para validar los contenidos del instrumento en cuestión.

2.4.5. Análisis de las opiniones de los expertos

Además de la razón del CVR y del CVI propuestos por Lawshe (1975), existen otros métodos para realizar el análisis de la información derivada de las opiniones de los paneles o grupos de expertos que evalúan los ítems. A continuación se describen dos de ellos.

1. Confiabilidad o acuerdo intergrupal (IRA). Denominado en inglés *interrater agreement*. Se trata de un método mediante el cual se determina el grado de confiabilidad de las puntuaciones que emiten los expertos. Se utiliza una escala dicotómica de cuatro puntos (es dicotómica porque se combinan los valores obtenidos en los puntos 1 y 2, así como

en 3 y 4). El hecho de dicotomizar la escala facilita identificar el nivel de acuerdo entre los expertos, respecto a si el ítem es o no representativo del contenido. La escala de cuatro puntos proporciona información adicional que servirá para tomar decisiones respecto a los ajustes que se requieran hacer al ítem, o a su eliminación definitiva del instrumento.

Además, no solamente se utiliza para evaluar ítems, sino para evaluar todo el instrumento en conjunto. Su limitante es que a medida que se cuente con un mayor número de expertos, menor será la probabilidad de consenso entre ellos. Este problema se consideró en la propuesta realizada por Lawshe en 1975, para establecer los valores mínimos del CVR, según la cantidad de personas que conforman el grupo.

2. Índice de validez factorial (FVI). Fue propuesto por McGartland et al. en el año 2003, como *Factorial validity index.* Se usa para determinar el grado en que los expertos asocian los ítems del instrumento con sus respectivos factores. Esto proporciona información preliminar de la validez factorial de la medición realizada.

   Para llevar a cabo el cálculo de este índice, en cada ítem se debe dividir el total de expertos que lo asociaron correctamente con su factor, entre el total de expertos que conforman el panel. Debido a que no existe un antecedente para poder determinar los niveles mínimos aceptables para este índice, estos investigadores proponen un FVI de por lo menos 0.80, tomando como referencia los valores del CVI.

A pesar de que el uso de paneles de expertos para obtener evidencias de validez de contenido es un método con limitaciones específicamente relacionadas con la subjetividad de los expertos, el uso de métodos analíticos, como los mencionados, puede proporcionar objetividad al proceso.

3. Evidencias de validez de contenido del Excoba: estructura del examen

Para llevar a cabo el estudio de validez de contenido del Excoba se adaptó el modelo propuesto por Lynn en 1986, que consta de dos fases, cada una con cinco etapas. En la primera fase, de desarrollo del instrumento, Lynn propuso tres etapas: 1) la definición del dominio a evaluar, 2) la descripción de la forma en que fueron elaborados los ítems, y 3) la descripción de cómo se construyó el instrumento. La segunda fase, de jueceo y cuantificación de los resultados, contiene las dos etapas restantes: 4) el trabajo con paneles de expertos, y 5) la obtención del CVI de Lawshe a partir de las evaluaciones realizadas.

El modelo metodológico propuesto para este estudio, se denomina a partir de aquí como: *Modelo para la obtención de evidencias de validez de contenido del Excoba* (MVCE), y contempla dos fases con cuatro etapas. Fue necesario adaptar el modelo de Lynn (1986), debido a que no se evaluaron los contenidos ni los ítems de un examen, sino los modelos de ítems, sus familias y los elementos cadena e integrales utilizados para generar los ítems hijos que producirán las distintas versiones del instrumento.

Es importante mencionar que el estudio contempló el análisis de los modelos de ítems de cuatro áreas del Excoba: Matemáticas, Historia, Química y Español, de nivel Secundaria, y no de la totalidad de los modelos de ítems que conforman al examen. Esto se debó a dos motivos. Por un lado la estructura conceptual del Excoba contempla la evaluación de competencias adquiridas durante la primaria y secundaria; ya que durante la primaria se adquieren las competencias necesarias para cursar el siguiente nivel educativo, y en la secundaria se adquieren nuevas competencias, con base en las de primaria. Por tanto, consideró que para efectos de esta investigación, las asignaturas de secundaria son lo suficientemente representativas de las competencias de ambos niveles.

Por otro lado, las asignaturas del nivel de secundaria se agrupan en cuatro grandes campos del conocimiento: matemáticas, ciencias sociales, ciencias naturales y lenguaje. Cada una de las cuatro asignaturas incluidas en este estudio son representativas de dichos campos.

La Figura 3.1 muestra las fases y etapas que conforman el MVCE. En la fase I que contempla una sola etapa se trabajó conceptualmente, con la finalidad de documentar el marco de referencia y el procedimiento que se siguió para la planeación, el diseño y la elaboración del Excoba.

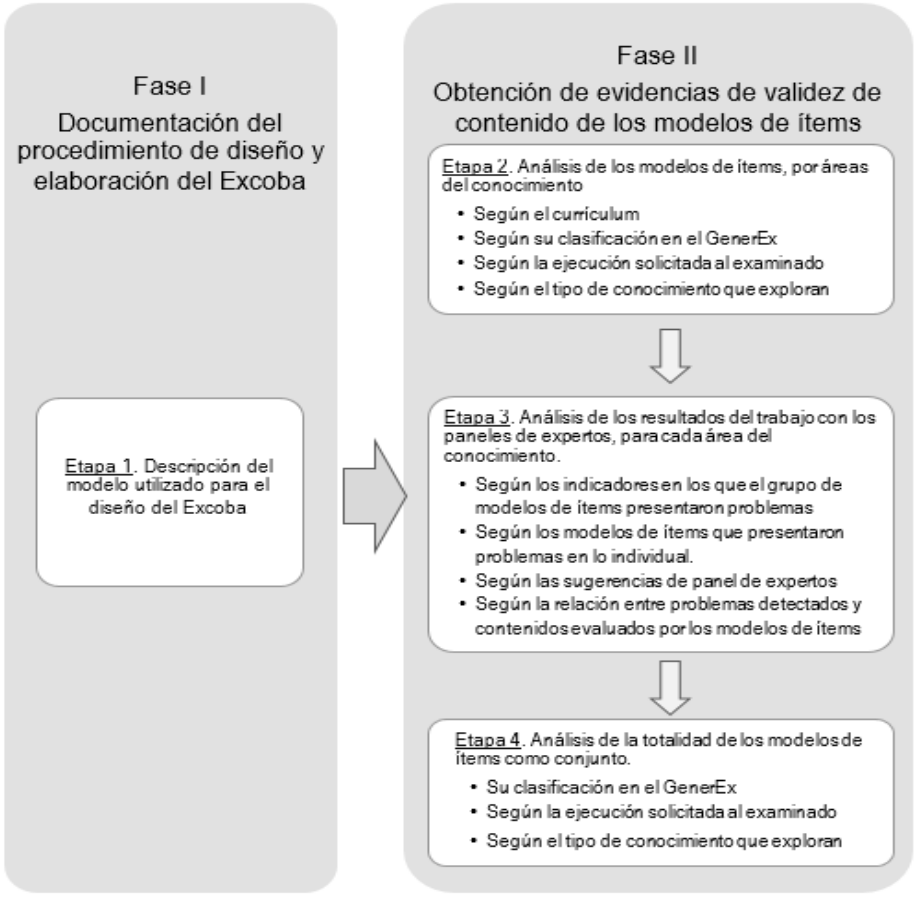


Figura 3.1. Modelo para la obtención de evidencias de validez de contenido de los modelos de ítems del Excoba.

La elaboración del Excoba fue respaldada por una secuencia metodológica que ha sido utilizada para el desarrollo de instrumentos a gran escala en el ámbito nacional (Backhoff, Sánchez, Peón, Monroy y Tanamachi, 2006), y que fue adaptada para su elaboración. No existe antecedente alguno en México para exámenes normativos a gran escala, elaborados mediante los principios de la GAI de teoría débil. Por ello, resultó imprescindible documentar dicha metodología mencionando las 7 fases y las 13 etapas que conforman el *Modelo de elaboración del Excoba* (MEE), así como los tipos de productos que fueron generados por los especialistas de cada una de ellas.

La fase II del MVCE se enfocó en la obtención de evidencias de validez de contenido de los modelos de ítems. Contempló las siguientes etapas: a) análisis de los modelos de ítems por área, según distintos criterios (currículum, clasificación en el sistema informático, tipo de ejecución y tipo de conocimiento); b) análisis de los resultados obtenidos del trabajo con los paneles de expertos de cada área, respecto a la revisión de los modelos de ítems, y c) análisis de los modelos de ítems considerándolos como una totalidad, con la finalidad de encontrar regularidades o patrones.

Para fines de una mejor organización de la información, se describen por separado las dos fases del MVCE. El resto del tercer capítulo se dedica a abordar los aspectos relacionados con la Fase I, llamada *Documentación del procedimiento de diseño y elaboración del Excoba*.

Primero se presenta el modelo que siguieron los autores para elaborar el Excoba, y se describe cada fase del mismo (planeación, estructuración, construcción, elaboración de la interfaz, administración, análisis de resultados, y recopilación de evidencias de validez). Enseguida se presentan los resultados de la Fase I, a manera de descripción del Excoba, su sustento, estructura conceptual, características, y los ejemplos de sus modelos de ítems y

reactivos. Es en este apartado donde se explica cómo se generan automáticamente los ítems hijos a partir de los modelos y el sistema informático GenerEx.

Finalmente, el capítulo cuatro se destinó a describir el procedimiento y los resultados obtenidos en la Fase II del MVCE, denominada *Obtención de evidencias de validez de contenido de los modelos de ítems del Excoba*. En esta se presentan, de manera separada, los resultados de cada una de las cuatro asignaturas analizadas, de acuerdo con la información obtenida mediante el proceso de evaluación realizado por los distintos paneles de expertos.

Fase I: Documentación del proceso de diseño y elaboración del Excoba

El objetivo de la primera fase del MVCE fue recopilar toda la información disponible sobre el diseño y la construcción del Excoba, así como de las actividades realizadas y los productos elaborados por los distintos especialistas que participaron en la construcción del instrumento. En la siguiente sección se describe el procedimiento que se siguió para tales efectos.

Procedimiento

Para explicar cuál fue el procedimiento que se siguió para documentar el punto de partida y las premisas en las que se basó el MEE, se revisaron los documentos históricos existentes en relación con el proceso de construcción del examen. Entre ellos, se consultaron: actas de reuniones del Consejos Técnico y del consultivo; ponencias, manuales técnicos, presentaciones, exposiciones, entrevistas, así como todo tipo de materiales que se utilizaron desde la planeación hasta la elaboración de la primera versión computarizada del Excoba.

Después de revisar dichos documentos, se pudieron llevar a cabo las siguientes acciones: a) explicar cómo y de dónde surgió el Excoba; b) conocer el tipo de decisiones que

fueron tomadas en torno a sus objetivos; c) describir el trabajo realizado con los participantes de los distintos comités de elaboración del Excoba; d) conocer el tipo de decisiones que tomó el Comité técnico a lo largo de su elaboración, y el razonamiento detrás de ellas; e) explicar los motivos por los cuales se llevó a cabo la transición de los reactivos de opción múltiple a reactivos de respuesta construida y semiconstruida, lo que hace que el Excoba sea un instrumento que se aproxima a la evaluación auténtica, y (f) describir la forma en la que se utilizaron herramientas informáticas para crear un instrumento de evaluación a gran escala bajo los principios de la GAI.

Como parte del proceso de elaboración de los antecedentes del Excoba, se llevó a cabo una revisión de los documentos generados en las distintas reuniones de Comité técnico y Consejo consultivo del EXHCOBA. La información generada fue relevante para el proceso de toma de decisiones respecto a transitar de un examen con reactivos de opción múltiple y respuesta única, a reactivos en los que el sustentante emitiera múltiples respuestas, generando puntajes parciales para cada reactivo que resolviera.

En marzo del año 2009 se llevó a cabo una reunión con los miembros del Comité técnico del EXHCOBA. Los miembros de este cuerpo colegiado son expertos en procesos de evaluación y diseño de instrumentos, estadística. También son conocedores de técnicas analíticas para el procesamiento de bases de datos, derivadas de aplicaciones a gran escala, planeación educativa, así como en las cuatro grandes disciplinas específicas que contempla el EXHCOBA: Español, Matemáticas, Ciencias Sociales y Ciencias Naturales.

En la agenda de dicha reunión, se presentó por primera vez una propuesta para evaluar campos cognitivos mediante reactivos que se aproximaran a estrategias de evaluación auténtica. Entre dichas propuestas, se presentaron diversos ejemplos que ilustraron la forma de operar la propuesta. En la Figura 3.2 se muestra el ejemplo de un tipo de reactivo al que se asignó en ese momento el nombre de *arrastre*.

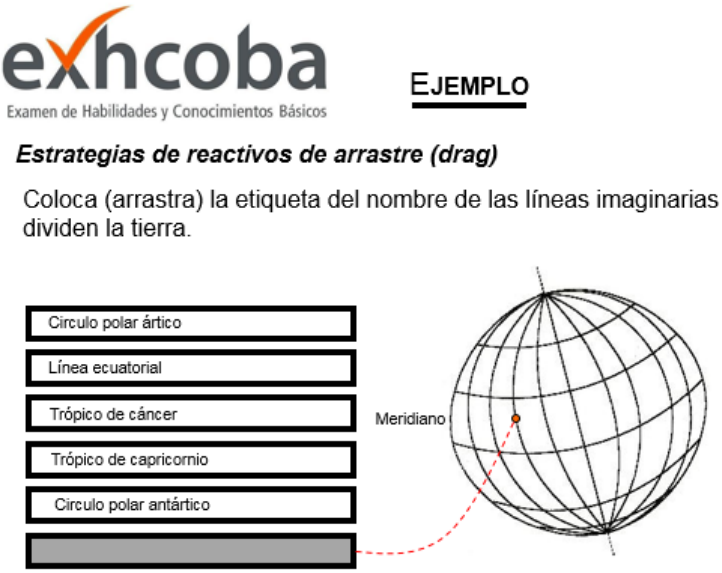


Figura 3.2. Ejemplo de un reactivo de *arrastre*.
Fuente: Segunda reunión de Consejo consultivo EXHCOBA, marzo 2009.

La propuesta en los reactivos de *arrastre* fue presentar una base de reactivo genérica, con diversas opciones de respuestas correctas e incorrectas, presentadas a manera de etiquetas. La intención era que el estudiante colocara sobre una imagen, mediante el uso del ratón, las etiquetas que considerara correctas. Lo innovador de esta propuesta fue que en un reactivo existía la posibilidad de seleccionar varias respuestas correctas y así obtener puntajes parciales.

Ante esta situación surgieron algunas discusiones que fueron registradas en la minuta de la reunión. Algunas de las discusiones fueron en torno a: cómo abordar distintos métodos analíticos para el procesamiento de datos, la elaboración de una nueva modalidad de especificaciones con características distintas a las actuales, el tipo de conocimientos que se deseaba abordar siguiendo una taxonomía específica; así como la necesidad de contar con equipos de especialistas que coadyuvaran en la elaboración de la nueva versión del instrumento.

En abril del año 2010 se llevó a cabo una reunión con el Consejo consultivo del EXHCOBA en la que estuvieron presentes los representantes de las instituciones educativas usuarias del mismo. Uno de los aspectos que se abordaron en esa reunión fue la transición del EXHCOBA al Excoba.

Se mostraron los nuevos reactivos, a los cuales se les asignó temporalmente el nombre de *Reactivos estructurales constructivos* (REESCO). La elaboración de estos tuvo como base dos principios fundamentales de la Teoría de los procesos cognoscitivos, de los que partía el EXHCOBA en su versión original: a) el conocimiento se desarrolla a partir de estructuras semánticas, y b) el proceso de apropiación de conocimientos se da a partir de que el sujeto cognoscente construye sus estructuras de conocimiento, las cuales le permiten el entendimiento. Posteriormente se les nombró *reactivos digitales*, porque además de los principios mencionados, son producto del aprovechamiento de la mediación computarizada (Tirado et al., 2014).

El proceso implicó una reestructuración en la manera de concebir las habilidades y los conocimientos básicos como ideas cristalizadas que implican nociones dentro de campos cognitivos en los que operan todos aquellos saberes, creencias, razonamientos y deducciones lógicas que utiliza el estudiante al solucionar cualquier problemática que se le presente. Se explicó que en ese proceso de pensamiento se formulan inferencias que a su vez amplían el saber, y la importancia que reviste contar con información objetiva que proporcione evidencias de que los estudiantes cuentan con los aprendizajes esperados para avanzar en su vida académica.

Fue entonces cuando se presentaron ejemplos de los nuevos reactivos en formato electrónico, en los que las respuestas se podían emitir mediante diversos tipos de ejecución: moviendo elementos y colocándolos en su lugar, escribiendo directamente la respuesta, o seleccionando segmentos de figuras, textos o gráficos. Se creó un sistema informático

llamado *Generador de Exámenes* (GenerEx), el cual forma parte de los dos desarrollos denominados Excoba/GenerEx, que trabajan de manera conjunta para poder generar automáticamente las distintas versiones del Excoba.

El sistema GenerEx se utiliza para administrar y calificar el Excoba. En él se depositaron todos los componentes de los modelos de ítems, con la finalidad de llevar a cabo su programación e implementar los algoritmos requeridos para que se seleccionaran de manera aleatoria todos los elementos necesarios para generar grandes cantidades de ítems hijos. Esto permitiría contar con una versión del examen para cada estudiante en cada grupo de aplicación, lo cual reduciría sustancialmente la posibilidad de que copiaran sus respuestas o las memorizaran.

Asimismo, se determinó que la evaluación de campos cognoscitivos implica la noción de diferentes tipos de conocimiento. Por eso, los modelos de ítems del Excoba responden a distintas exigencias en cuanto al nivel de dominio y la profundidad cognoscitiva que se solicita de los estudiantes. La taxonomía bajo la cual se clasificaron los modelos de ítems del Excoba se basó en la propuesta realizada por Ruiz-Primo en 2007. Se propuso una clasificación de cuatro tipos de conocimiento:

1) Conocimiento declarativo. Comprende desde un manejo de contenidos o detalles aislados como los datos y los hechos, hasta conocimientos más organizados como las definiciones, categorías y los principios teóricos. Si el estudiante tiene un manejo de este tipo de conocimientos, puede identificar los elementos implicados en la resolución de la tarea que se le pide ejecutar.
2) Conocimiento procedimental. Involucra el manejo de métodos y técnicas basadas en reglas establecidas. Si el estudiante maneja adecuadamente este tipo de conocimientos, sabe cómo realizar tareas ejecutando correctamente los pasos de un procedimiento.

3) Conocimiento esquemático. El estudiante organiza y maneja la información de manera sistemática, mediante el uso de modelos mentales y cuerpos de conocimiento. A través de ellos puede llegar a explicaciones y razonamientos que utiliza en el proceso de solución de problemas. El conocimiento esquemático permite al estudiante saber el motivo por el cual ocurren los fenómenos.
4) Conocimiento estratégico. Se fundamenta en los otros tipos de conocimiento, ya que mediante este se elaboran y planean estrategias en las que se revisan las distintas alternativas para llegar a la mejor solución de una situación dada. Es mediante el dominio de este tipo de conocimiento que el estudiante demuestra un manejo de cuándo, dónde y cómo se utiliza el conocimiento.

3.1. Etapa 1 del MVCE: Descripción del Modelo de Elaboración del Excoba (MEE)

El proceso que siguieron los desarrolladores del Excoba consistió en la implementación de un modelo de 7 fases y 13 etapas, llamado Modelo de Elaboración del Excoba (MEE). La Tabla 3.1 muestra las fases y sus etapas. Los productos que se obtuvieron en cada etapa sirvieron como insumos para la siguiente. Por ello, en el propio proceso de generación de este examen se contempló parte del proceso de obtención de evidencias de validez de contenido. La información detallada del MEE se puede revisar en el Apéndice 1.

Tabla 3.1.
Modelo de elaboración del Excoba (MEE)

Fases	Etapas
1. Planeación general	1. Diseño del plan general de evaluación
	2. Diseño y desarrollo del sistema informático GenerEx
	3. Diseño y elaboración de cuestionarios de contexto
2. Estructuración del Excoba	4. Diseño del Excoba
3. Construcción del Excoba	5. Elaboración de modelos de ítems con plantillas y generadores de ítems
4. Construcción de la interfaz	6. Programación del GenerEx
	7. Diseño gráfico de la interfaz
	8. Montaje en red del examen
5. Administración del Excoba	9. Piloteo

	10. Administración real del examen.
6. Análisis e interpretación de resultados del Excoba	11. Análisis psicométricos para comité técnico
	12. Análisis de resultados para el Consejo consultivo
7. Recopilación de evidencias de Validez del Excoba	13. Estudios de validación

El diseño del Excoba requirió del trabajo colegiado de expertos, y se fundamentó en un método sólido, y consistente con los propósitos que se perseguían (Contreras, 2000; Instituto Nacional para la Evaluación de la Educación [INEE], 2005; Nitko, 1994).

En el desarrollo del instrumento intervinieron diversos especialistas internos y externos, los cuales se organizaron en distintos grupos denominados comités: 1) Comité técnico, 2) Comités de área (Matemáticas, Español, Ciencias Sociales y Ciencias Naturales), 3) Comité elaborador de modelos de ítems. Cada uno de estos cuerpos colegiados cumplió una función específica y complementaria en el proceso de construcción, por lo que su trabajo se programó en forma escalonada y sus productos se convirtieron en insumos de las etapas subsecuentes. El MEE se describe a continuación.

Fase I. Planeación general del Excoba

El primer paso para construir del instrumento fue la fase de planeación general. En ella los miembros del Comité técnico, así como el personal técnico del Excoba, se dieron a la tarea de crear un plan de evaluación que incluyera las necesidades evaluativas de los usuarios del instrumento. Dado el formato automatizado del Excoba, especialistas en programación participaron en la planeación, realizando tareas de elaboración de los elementos gráficos que alimentaron el sistema informático creado exclusivamente para la aplicación, administración y calificación del Excoba.

Se realizaron reuniones de trabajo, seminarios y pruebas periódicas del sistema, con la finalidad de comunicar a los diseñadores del mismo y a los diseñadores gráficos, los criterios establecidos por el Comité técnico. De esta manera se buscaba empatar la interfaz gráfica, el lenguaje de programación y todos los detalles técnicos que surgieron durante el

proceso inicial. A su vez, expertos en evaluación del aprendizaje y diseño de cuestionarios de contexto trabajaron en la elaboración de un cuestionario, cuya función es captar información acerca de las condiciones socioeconómicas, académicas y contextuales de los estudiantes evaluados mediante el Excoba.

Fase II. Estructuración del Excoba

Una vez que se contó con un plan general de evaluación y se determinaron los criterios técnicos e informáticos para crear el sistema informático que permitiría realizar la aplicación del instrumento, el Comité técnico y el personal técnico del EXCOBA se dieron a la tarea de seleccionar, con base en criterios específicos relacionados con la trayectoria académica y profesional, a un grupo de docentes de escuelas primarias y secundarias públicas y privadas de Baja California, México. La finalidad fue conformar los Comités académicos de cada área, para seleccionar aquellos contenidos curriculares que se consideraran esenciales para ser incluidos dentro de un instrumento de esta naturaleza.

Los docentes trabajaron de manera grupal por área de conocimiento. Recibieron un curso de capacitación en el que se les explicó la naturaleza del instrumento, sus principios teóricos y metodológicos; así como el contexto de la evaluación auténtica y de la evaluación de campos cognitivos. Como insumos se les proporcionaron materiales que incluyeron libros de texto de las asignaturas y los grados escolares que se evalúan en el Excoba; retículas pertenecientes a la asignatura de su especialidad; planes y programas de estudio, y acceso a internet para la consulta de cualquier información pertinente a la labor encomendada. El objetivo fue realizar un análisis curricular y seleccionar una cantidad específica de contenidos que, con base en su conocimiento y experiencia, consideraran indispensables para incluir en el examen.

Fase III. Construcción del Excoba

Durante esta fase se conformaron los Comités elaboradores de modelos de ítems. El trabajo de los participantes de estos cuerpos colegiados fue guiado por los miembros del Comité técnico y el personal técnico del Excoba, con la finalidad de generar modelos de ítems que cumplieran con las características que se buscaba. Durante el proceso de capacitación se utilizaron diversos insumos que apoyaron el trabajo de los distintos comités, tales como: tablas de contenido elaboradas por los Comités académicos que trabajaron en la fase anterior, libros de texto oficiales (en el caso de primaria) y libros frecuentemente utilizados en escuelas públicas y privadas (en el caso de secundaria donde no existe un solo libro oficial).

Un aspecto muy importante del trabajo de estos comités fue la naturaleza iterativa del proceso. Esto significó para el Comité técnico del Excoba la revisión constante de los productos. Asimismo, implicó una inversión considerable de tiempo, ya que, a diferencia de una especificación en la que se redacta una base del reactivo, se elabora un ítem ejemplo y se construyen de tres a cinco opciones de respuesta, en los modelos de ítems el trabajo de los miembros de los comités implica la selección cuidadosa de una gran variedad de elementos de los libros de texto. La experiencia en la práctica docente, el conocimiento de estrategias pedagógicas, el dominio de la asignatura, y la creatividad fueron elementos clave para la realización de tal tarea.

De especial cuidado fue la estrategia que se emplearía para evaluar al estudiante, ya que el Excoba es un instrumento que aborda la evaluación de las competencias esperadas en los estudiantes, según el currículum de educación básica mexicano. Por ello, el objetivo fue utilizar mecanismos de evaluación que en la medida de lo posible se aproximaran a una evaluación auténtica. Esto implicó echar mano de ejemplos y contraejemplos de los libros de texto, de las clases impartidas en aula y también de las prácticas en laboratorio, como

referentes para la elaboración de los elementos cadena e integrales que serían utilizados para crear las familias de ítems y los hijos que de ellas derivarían.

Como producto de esta fase se obtuvieron 120 modelos de ítems y, con ello, se dio estructura al Excoba. Dichos modelos integraron la información necesaria para asegurar que al generar los ítems se evaluara el aprendizaje esperado del estudiante. Entre los componentes del modelo de ítems se incluyeron: a) los datos de identificación del contenido a evaluar, los cuales sirven para identificar su ubicación en el currículum; b) las características relacionadas con su importancia curricular, en términos de los motivos por los cuales se consideró importante incluirlo dentro del instrumento, y c) las especificaciones, los elementos conceptuales, las reglas y restricciones para generar los ítems.

Fase IV. Construcción de la interfaz

Una vez que se contó con los modelos de ítems, en los que se describió de manera detallada la forma en que los ítems debían aparecer al examinado, el Comité técnico del Excoba se dio a la tarea de trabajar con un grupo de especialistas en bases de datos, en sistemas de información, así como con diseñadores gráficos. El trabajo de estos grupos de profesionales generó distintas versiones de un sistema creado ex profeso, llamado GenerEx (Generador de Exámenes). Este sistema se encuentra actualmente en su versión número 3.31, lo que refleja cómo han sido detectados y atendidos los detalles que han surgido mediante su uso y aplicación. En él se alojan las 22 clasificaciones distintas de los modelos de ítems, de acuerdo con el tipo de programación que requirieron para ser funcionales y generar los ítems hijos según las reglas y restricciones descritas en los modelos de ítems.

Fase V. Administración del Excoba

Durante esta fase se llevaron a cabo dos importantes acciones para el desarrollo del Excoba: el piloteo y la administración del examen en un escenario real. Durante la etapa de

piloteo del examen se obtuvo información empírica que ayudó a conocer el comportamiento de los reactivos, realizar ajustes a los modelos de ítems cuyos ítems hijos obtuvieran índices de calidad psicométrica por debajo de lo esperado. Para realizar lo anterior se generaron dos versiones paralelas de los ítems hijos, de cada uno de los 120 modelos. Estos 120 ítems se administraron a una muestra de estudiantes de bachillerato.

Mediante la Teoría clásica de tests (TCT) se obtuvo información acerca de las propiedades psicométricas de los ítems. Desde la Teoría de respuesta al ítem (TRI) se analizó su unidimensionalidad (modelamiento Rasch-Masters). También se utilizó el Modelamiento de ecuaciones estructurales (SEM), a través del Análisis factorial confirmatorio (AFC), con la finalidad de confirmar la estructura teórica del examen.

Los resultados se analizaron mediante métodos empíricos. Y además de determinar un buen comportamiento estadístico de los ítems hijos, se obtuvo información que ayudó a detectar errores y hacer los ajustes necesarios a los modelos de ítems (base del reactivo, estrategia evaluativa, elementos cadena e integrales, entre otros aspectos). Posteriormente se llevó a cabo el montaje final del examen en la plataforma del GenerEx, a través de la captura de todos los modelos de ítems, para hacer pruebas de su funcionamiento.

En la etapa de administración del examen se involucró tanto personal técnico como investigadores del Excoba, especialistas en tecnologías y sistemas, así como personal técnico de las distintas instituciones usuarias del Excoba. Mediante acciones de naturaleza logística y técnica, se llevó a cabo la instalación del sistema GenerEx en los centros informáticos de cada institución. Se realizaron las pruebas y posteriormente la aplicación, de la cual se generaron bases de datos con las respuestas de los examinados.

Fase VI. Análisis e interpretación de los resultados del Excoba

Una vez que se contó con las bases de datos de aplicaciones en escenarios reales, es decir, a estudiantes que se encontraban en proceso de selección para ingreso al bachillerato, inició la etapa de análisis psicométricos de la información. Como producto de diversas reuniones con los miembros del Comité técnico del Excoba y de asesores expertos en medición, se elaboraron los informes técnicos sobre el comportamiento psicométrico del examen. Asimismo, se trabajó en los informes de caracterización de los estudiantes, de acuerdo con las diversas variables contextuales captadas en la información arrojada por el cuestionario de contexto que se administró a cada estudiante, durante el proceso de aplicación del Excoba.

Fase VII. Recopilación de evidencias de validez del Excoba

La última fase del MEE aborda la recopilación de evidencias de validez de distinta índole. Hasta el momento se cuenta con dos estudios de validez. El primero, que culminó en una tesis doctoral, abordó la obtención de evidencias de validez cognitiva y fue realizado por Pérez en el año 2013, con la finalidad de conocer las características de los procesos cognitivos empleados por los estudiantes evaluados al responder a los ítems hijos del Excoba. El segundo, que también se consolidó como una tesis doctoral, fue realizado por Ferreyra en el año 2014; se trata de un estudio de validez de la estructura interna del instrumento, y representa una propuesta de aproximación para obtener evidencias de validez de exámenes no adaptativos, elaborados mediante la GAI de teoría débil. Actualmente y mediante el presente estudio, se aborda la obtención de evidencias de validez de contenido de los modelos de ítems del GAI Excoba.

3.2. Resultados

Generador Automático de Ítems Excoba

El desarrollo del Generador automático de ítems del Excoba se basó en la teoría débil de la GAI. La razón de ello es que los aprendizajes escolares que evalúa se sustentan en el currículum mexicano de la educación básica y los contenidos esenciales que marca: las habilidades, los aprendizajes esperados y las competencias que los estudiantes deben adquirir a lo largo de su formación académica. El GAI Excoba no se basa en un modelo de los procesos cognitivos subyacentes a las respuestas de los examinados.

El Excoba se diseñó con la finalidad de seleccionar estudiantes para ingresar a nivel medio superior y superior. Debido a su estructura conceptual, se puede aplicar según las necesidades de cada institución, una versión simplificada para seleccionar estudiantes de nuevo ingreso al bachillerato, o la versión completa para seleccionar estudiantes que desean ingresar a estudios de nivel superior. Lo que se busca evaluar mediante el Excoba son las competencias básicas que el estudiante ha adquirido durante su experiencia escolar y que se supone son necesarias para que alcance aprendizajes significativos de mayor nivel, y así cursar con éxito los siguientes niveles escolares.

Se parte del supuesto central de que la comprensión de los conocimientos básicos y la adquisición de las competencias que marca el currículum resultan indispensable para entender cabalmente una disciplina y llevarla a la práctica. Por lo tanto, es más importante evaluar la comprensión y aplicación de las competencias académicas esenciales, bajo las cuales el estudiante determina el significado y la aplicación de sus aprendizajes, que aquellos conocimientos sustentados básicamente en la capacidad memorística del estudiante, y que no son esenciales en la estructura de una disciplina.

La estructura conceptual del Excoba contempla la evaluación de 300 contenidos y competencias del currículum de educación básica mexicano, considerados como esenciales para detectar si los estudiantes cuentan con las competencias y los aprendizajes esperados. Su diseño permite realizar distintos ensamblajes del instrumento, según los objetivos evaluativos que se persigan, de tal manera que en la versión simplificada que se utiliza para el bachillerato se evalúan únicamente los 120 contenidos y competencias de primaria y secundaria; mientras que en la versión completa que se utiliza como instrumento de selección a la universidad, se agregan 60 contenidos de bachillerato correspondientes a tres de nueve bloques que se eligen según el área de conocimiento a la cual pertenezca el programa de licenciatura al que aspire ingresar el estudiante.

Para brindar estructura y delinear la forma en la que se evaluarían todos los contenidos que conforman el Excoba, los autores del instrumento se dieron a la tarea de diseñar y elaborar los modelos de ítems que contienen la información necesaria para generar automáticamente una gran cantidad de ítems de cada uno de ellos. Estos se agrupan conceptualmente en tres secciones, según el nivel educativo al que correspondan los contenidos curriculares que evalúan:

1) Primaria. Contiene 40 modelos de ítems, los cuales abordan aspectos relacionados con el uso del lenguaje y las matemáticas, tales como la comprensión de textos y la solución de problemas aritméticos.
2) Secundaria. Contiene 80 modelos de ítems, distribuidos en cuatro áreas del conocimiento: Español, Matemáticas, Ciencias Sociales y Ciencias Naturales. Ciencias Sociales se subdivide en Historia, Geografía y Formación cívica y ética. El área de Ciencias Naturales se subdivide en Biología, Física y Química.
3) Bachillerato. Consta de 180 modelos de ítems, distribuidos en nueve áreas de especialidad: matemáticas para la estadística, matemáticas para cálculo, física, química, biología, lenguaje, ciencias sociales, humanidades y ciencias

económico-administrativas. La Figura 3.3 muestra la estructura conceptual y la distribución de los contenidos en cada nivel educativo y área del conocimiento.

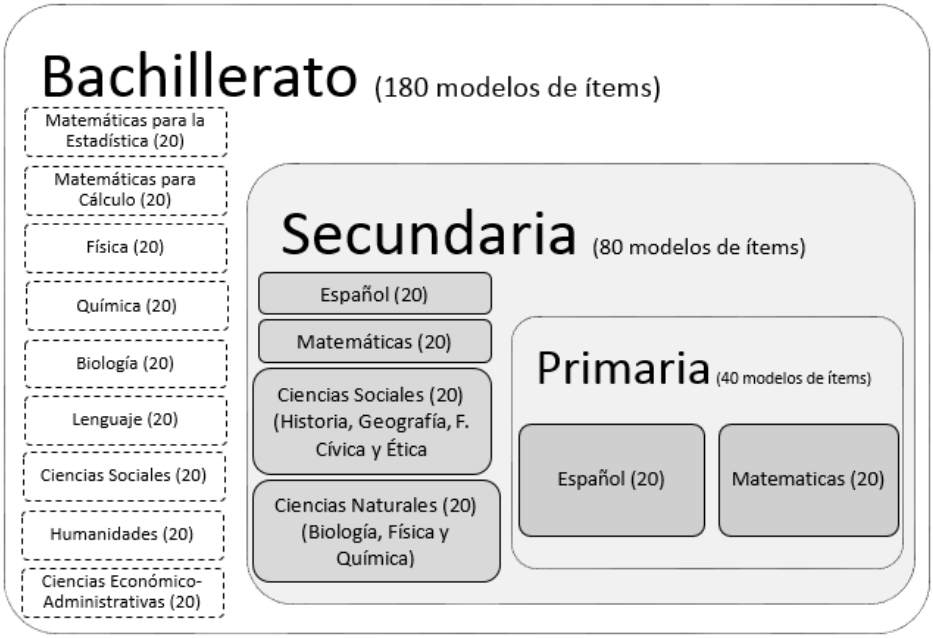


Figura 3.3. Estructura conceptual del Excoba. Distribución de los modelos de ítems por nivel educativo y área del conocimiento, así como dos formas de ensamblar el examen, según se utilice como instrumento de selección de estudiantes al bachillerato o a la universidad.

Se puede observar que los recuadros de las nueve áreas del conocimiento que conforman la sección de bachillerato tienen líneas punteadas. Esto significa que, aunque esta sección es parte de la estructura conceptual, aún se encuentra en desarrollo. Actualmente el Excoba se utiliza únicamente en su versión simplificada, es decir, para seleccionar estudiantes de nuevo ingreso al bachillerato. Se estima que al final del año 2015 se utilice como examen de selección a la universidad.

Como resultado de la amplitud de posibilidades que ofrece la estructura de los modelos de ítems dentro de la GAI, así como del uso de las herramientas informáticas, se

logró transitar del formato único de generación de ítems de opción múltiple, a la generación de este tipo de ítems, más los que Tirado (2010) llamó *reactivos estructurales constructivos* (REESCO), posteriormente nombrados *reactivos digitales* (Tirado et al., 2014).

Los modelos de ítems del Excoba incluyen generadores que funcionan a manera de plantillas o moldes, y que contienen los elementos genéricos que se requieren para la redacción de una base del reactivo, así como los complementos que se necesitan para su presentación al estudiante (textos auxiliares, gráficos, entre otros). Para poder "rellenar" la plantilla, los elaboradores del instrumento insertan bajo una serie de reglas previamente establecidas, segmentos de contenido que están directamente asociados al campo conceptual que se desea medir (ver Figura 3.4).

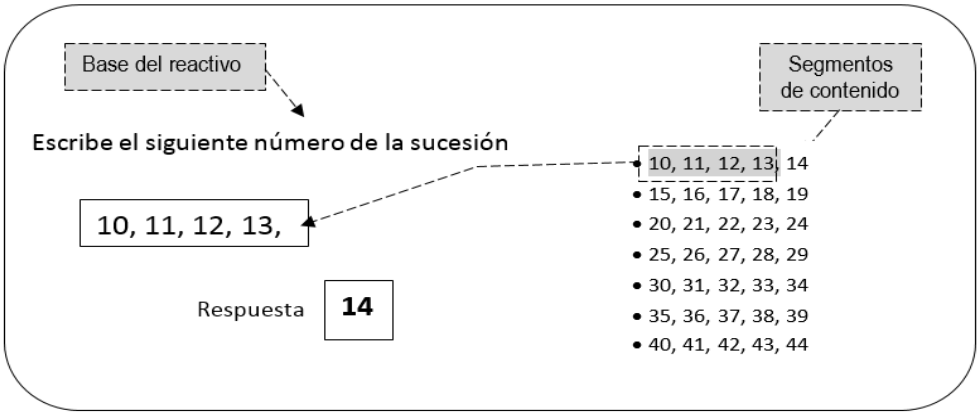


Figura 3.4. Ejemplo de una plantilla del Excoba.

Las plantillas permiten superar el formato tradicional de elaboración de ítems para exámenes de gran escala, aún empleado en muchos de los instrumentos desarrollados bajo los principios de la GAI, en los que se utilizan ítems de opción múltiple para que el estudiante elija la respuesta correcta. Para evolucionar a un formato de respuesta construida y semiconstruida, el cual permite la generación de múltiples versiones de ítems de un mismo modelo.

Una aportación muy importante que el Excoba hace al campo de la GAI es la forma en que se diseñan los modelos de ítems. Debido a que el modelo utilizado para su diseño y construcción se basa en la evaluación auténtica, la información que contienen los modelos de ítems no solo permite generar múltiples versiones de ellos, sino que también permiten utilizar un gran cúmulo de opciones asociadas a un campo de conocimiento, y hacer una selección de ellas al azar, para construir ítems en los que el estudiante pueda emitir más de una respuestas.

También permite elaborar reactivos con distintas modalidades de respuesta, tales como respuesta construida y semiconstruida. Esta última incluye selección de elementos, ordenamiento de textos, identificación de elementos, asociaciones, entre otras. De esta manera, el diseño del Excoba no solo evalúa el nivel de habilidades y conocimientos con los que cuenta el estudiante, sino el dominio de campos cognitivos que se requieren tanto para la ejecución de una tarea, como para mostrar evidencias del nivel de desempeño en las habilidades y los conocimientos considerados básicos y esperados en todo estudiante.

Para poder generar de manera automática los ítems y operarlos siguiendo las reglas especificadas en los modelos de ítems, fue necesario crear el sistema informático GenerEx. Se trata de una plataforma digital que contiene distintos módulos, en los cuales se captura toda la información contenida en los modelos de ítems. La Figura 3.5 muestra dichos módulos. Se puede apreciar un conjunto de “pestañas” en las que se puede capturar información de distinta naturaleza, tal como la definición del contenido evaluado, la estrategia evaluativa a utilizarse, un ejemplo de ítem hijo, etcétera.

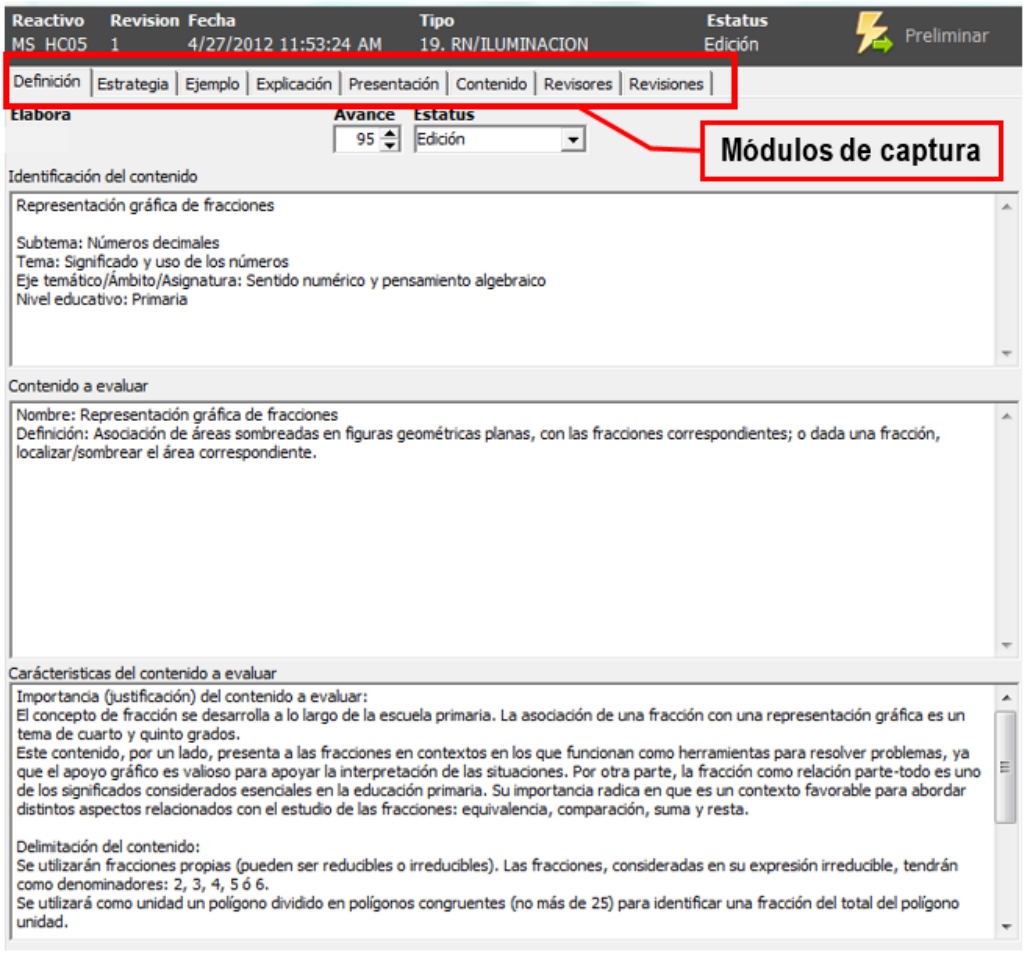


Figura 3.5. Módulos de captura del GenerEx.

Para poder generar automáticamente los ítems, el GenerEx contiene una base de programación que incluye algoritmos computacionales que atienden las necesidades especificadas en los modelos de ítems. Este sistema aplica la información y las reglas capturadas, y se apega a las restricciones establecidas para combinar los distintos elementos cadena e integrales, que permitirán presentar los ítems hijos en la interfaz que se presentará al examinado.

Para atender las características específicas de cada uno de los modelos de ítems que conforman la estructura conceptual del Excoba, los desarrolladores del sistema crearon 22

clasificaciones diferentes de los modelos de ítems, según sus características programáticas. La Tabla 3.2 muestra dicha clasificación y una descripción del tipo de algoritmos y reglas que utiliza el sistema informático para generar los ítems, así como una breve descripción del funcionamiento de los ítems hijos que se generan mediante dichas reglas.

Tabla 3.2.
Clasificación de los modelos de ítems en el sistema GenerEx del Excoba

Tipo	Descripción
1. Opción múltiple	La programación de este tipo de ítems no requiere de códigos de programación sofisticados, ya que se trata del formato tradicional de ítems en los que se cuenta con una base del reactivo y cinco opciones de respuesta.
2. Opción periodo	El sistema contiene una serie de textos asociados a elementos, divididos en grupos que tienen un orden preestablecido, el cual se definió en las reglas y restricciones del modelo de ítems. La programación exige el reconocimiento de las respuestas de los examinados dentro de dicho orden, de tal manera que asigne el puntaje adecuado.
3. Elemento imagen	El sistema requiere la alimentación de un banco de datos de elementos, asociados a una imagen en la cual se delimitan sectores en los que se ubicarán dichos elementos. El número de sectores será mayor al de los elementos, de tal manera que al responder, el estudiante colocará su respuesta donde lo desee.
4. Selección elementos	Estos ítems requieren que el sistema presente textos o fórmulas en los que se observen espacios vacíos o segmentos marcados (como podría ser una palabra, frase o enunciado), de tal manera que al hacer clic sobre cada espacio marcado, se despliegue una ventana con distintas opciones. De ellas, el estudiante deberá elegir una para completar correctamente el segmento.
5. Elemento categoría	La programación exige clasificar la información en una serie de categorías a las cuales se asocian elementos específicos que se seleccionan al azar por el sistema. La interfaz detecta el uso del ratón cuando se seleccionan y mueven correcta o incorrectamente los elementos a las categorías.
6. Orden oraciones	En este tipo de ítems, la interfaz presenta una serie de párrafos u oraciones en desorden para que el estudiante las ordene y forme un párrafo coherente.
7. RN/fórmulas *	La interfaz utiliza reglas o fórmulas que sirven para calcular el valor de una o más variables contenidas en la base del reactivo o en figuras auxiliares, sustituyéndolas de manera aleatoria en cada ítem hijo que se genere.
8. RN/ecuaciones	La programación exige la selección aleatoria de una ecuación lineal o cuadrática que se muestra al estudiante. El sistema debe permitir escribir una o dos respuestas en forma numérica, según sea el caso.
9. RN/triángulos	Utiliza los mismos principios que los ítems de tipo RN/fórmulas, con la adición de una figura en la que se encuentran datos que el sistema genera de manera aleatoria, mediante las reglas y restricciones indicadas.
10. RN/Pendiente	Mediante el uso de fórmulas, el sistema selecciona aleatoriamente los valores que se requieren para generar una gráfica de función, la cual se muestra como parte del ítem, en un plano cartesiano.
11. RN/rangos	Utiliza los mismos principios que los ítems de tipo RN/fórmulas, con la diferencia de que para cada respuesta emitida por el estudiante, se puede determinar un rango diferente de respuestas correctas.
12. RN y selección	Utiliza los mismos principios que los ítems de tipo RN/fórmulas, pero al aparecer la ventana de opciones, se muestra una lista de elementos más extensa.
13. RN/sucesiones	La programación de este tipo de reactivos requiere que se capturen distintas series de valores numéricos consecutivos, bajo una serie de reglas preestablecidas, de las cuales el sistema selecciona aleatoriamente una, así como un grupo de valores contenidos en ella.
14. Orden números	El sistema presenta una serie de elementos que el estudiante debe ordenar. Dichos elementos son números.

15. RN/etiquetas	El sistema requiere alimentar un banco de datos de figuras o imágenes, de las cuales selecciona una y sustituye una o varias de sus medidas (lados, ángulos, etcétera). Cuando el estudiante escribe su respuesta, ésta es comparada con la correcta y se le asigna un puntaje.
16. Selección frases	El sistema genera ítems en los que se muestra un texto en el que se marcan segmentos (palabras, frases o enunciados), los cuales se deben seleccionar.
17. Orden elemento múltiple	La programación de este tipo de ítems exige que se presenten imágenes para ordenarlas de acuerdo con un criterio. El sistema elige, según el orden preestablecido, una cantidad de imágenes que el estudiante debe ubicar en el orden correcto.
18. Frase imagen	El sistema elige un texto del banco de elementos, en el que se marcan segmentos (palabras, frases o enunciados) que se deben ubicar dentro de una imagen con espacios en blanco.
19. RN/iluminación	El sistema elige una de las dos formas posibles de presentar el ítem hijo: 1) al mostrar una figura del banco de elementos, la cual se encuentra dividida en partes iguales, de tal manera que el estudiante ilumine (mediante la acción de clic del ratón) segmentos de la misma; o 2) al mostrar una figura con segmentos iluminados por el sistema, con la cual se debe escribir una respuesta numérica.
20. RN/gráficas	La forma en que está programado este tipo de reactivos implica el uso de gráficas dinámicas que se modifican cuando el sistema sustituye aleatoriamente los valores de algunas variables.
21. RN/R. algebraica	En este tipo de reactivos, la programación permite hacer uso combinado de cualquiera de los tipos RN/ fórmulas y R. Algebraica, generando ítems que transforman la escritura de respuestas numéricas a respuestas algebraicas, y que a su vez permiten seleccionar imágenes asociadas a dichas respuestas, mediante el contraste de éstas con una serie de reglas o fórmulas preestablecidas.
22. R. algebraica	El tipo de programación requerida para su funcionamiento implica el uso de un editor de ecuaciones que permite que el estudiante escriba una ecuación utilizando algunos caracteres del teclado, que el sistema automáticamente transforma, de tal manera que se pueda visualizar como ecuación.

*R= Respuesta numérica, R= Respuesta

3.3. Modelos de ítems del Excoba

Un modelo de ítems es similar a lo que comúnmente se conoce como especificación de ítems cuando se habla acerca de instrumentos con reactivos de opción múltiple. La diferencia entre estos y los modelos de ítems del Excoba radica en que, además de la definición, delimitación del contenido a evaluar y la descripción de la estrategia a utilizar, se incorpora un banco de información con contenidos curriculares, junto con una serie de instrucciones detalladas y muy bien delimitadas. Estas instrucciones permiten generar una gran cantidad de ítems a partir de reglas que se declaran de manera explícita, con la finalidad de tomar elementos del banco de información y realizar con ellos diferentes combinaciones para producir los ítems que se presentarán al estudiante. A esto se le denomina *generador de ítems*.

El modelo de ítems del Excoba incluye toda la información necesaria para asegurar que al construir un ítem se evalúe el aprendizaje que se espera que el estudiante domine en un contenido particular. El concepto de contenido curricular se define no solo por la ejecución del estudiante al demostrar su nivel de dominio de la información y de los conceptos del currículum, sino principalmente por las habilidades y competencias que posee, y la forma en la que su ejecución proporciona evidencias de los aprendizajes que el currículum señala que debe haber aprendido. De esta manera, lo que evalúa el Excoba no se limita a contenidos informativos, sino a los procesos de razonamiento y manejo práctico de información, asociados al aprendizaje esperado de un estudiante con nivel académico de secundaria o bachillerato.

El modelo de ítems de este instrumento se divide en tres grandes secciones: la primera de ellas incluye todos los datos que identifican el contenido a evaluar, la segunda describe las características del contenido curricular evaluado, y la tercera incluye la plantilla con los elementos indispensables para generar los ítems (ver Figura 3.6).

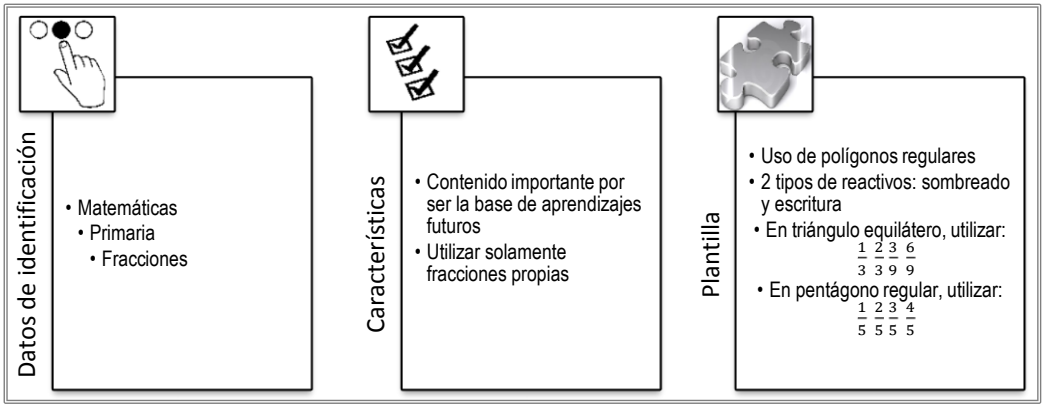


Figura 3.6. Secciones que conforman el Modelo de ítems del Excoba.

La primera sección contiene todos los datos de identificación del contenido que se evalúa en el examen, tales como:

a) La clave que se le asignó dentro del esquema de organización del Excoba, que sirve para identificarlo y ubicarlo jerárquicamente en relación con el nivel educativo al que pertenece, para conocer el lugar que guarda dentro del currículum.
b) El nombre de la asignatura o campo disciplinario al que pertenece.
c) El nivel educativo en el que se debe dominar dicho contenido
d) El ámbito al que pertenece dentro del currículum mexicano, entendiéndose éste como un estándar curricular que expresa lo que los estudiantes deben saber y ser capaces de hacer al concluir el nivel académico correspondiente.
e) El nombre del contenido y la competencia a evaluar, así como su definición en términos de los aprendizajes esperados en el estudiante (ver Figura 3.7).

Asignatura o Área		**Nivel educativo**	
Matemáticas		Primaria	
Clave	**Eje temático/Ámbito/Asignatura**	**Tema**	**Subtema**
HC05	Sentido numérico y pensamiento algebraico	Significado y uso de los números	Números decimales
Contenido	**Nombre**	Representación gráfica de fracciones	
	Definición	Asociación de áreas sombreadas en figuras geométricas planas, con las fracciones correspondientes; o dada una fracción, localizar/sombrear el área correspondiente.	

Figura 3.7. Ejemplo de modelo de ítems. Sección de datos de identificación del contenido a evaluar.

En el ejemplo de la Figura 3.7 se aborda la descripción del contenido *representación gráfica de fracciones.* Esta descripción permite saber que fue seleccionado del plan de estudios de matemáticas de primaria, específicamente del eje temático *sentido numérico y pensamiento algebraico,* en donde se encuentra el tema "significado y uso de los números", así como el subtema "números decimales". De esta manera, quien desee profundizar sobre el origen de este contenido curricular y de los ítems que se generen para evaluarlo, podrá ubicarlo con precisión y obtener información detallada sobre sus características.

La segunda sección contiene las características del contenido curricular evaluado, tales como los motivos por los cuales se consideró importante incluirlo dentro del instrumento de evaluación, su delimitación conceptual y los conocimientos y habilidades que requiere el estudiante para poder demostrar su dominio en él (Figura 3.8).

Importancia (justificación) del contenido a evaluar
El concepto de fracción se desarrolla a lo largo de la escuela primaria. La asociación de una fracción con una representación gráfica es un tema de cuarto y quinto grados. Este contenido, por un lado, presenta a las fracciones en contextos en los que funcionan como herramientas para resolver problemas, ya que el apoyo gráfico es valioso para apoyar la interpretación de las situaciones. Por otra parte, la fracción como relación parte–todo es uno de los significados considerados esenciales en la educación primaria. Su importancia radica en que es un contexto favorable para abordar distintos aspectos relacionados con el estudio de las fracciones: equivalencia, comparación, suma y resta.
Delimitación del contenido
Se utilizarán fracciones propias (pueden ser reducibles o irreducibles). Las fracciones, consideradas en su expresión irreducible, tendrán como denominadores: 2, 3, 4, 5 ó 6. Se utilizará como unidad un polígono dividido en polígonos congruentes (no más de 25) para identificar una fracción del total del polígono unidad. La fracción a identificar se corresponderá con un número entero de los polígonos que dividen a la unidad. No se presentará el caso en que el denominador de la fracción a graficar coincida con la cantidad de polígonos en que se dividió la unidad (por considerarse elemental).
Conocimientos y habilidades involucrados en la solución correcta del reactivo
Concepto de fracción: numerador, denominador. Conceptos de fracciones propias y fracciones equivalentes. Representación gráfica de las partes iguales de un todo. Este reactivo requiere de los niveles cognitivos de comprensión y aplicación.

Figura 3.8. Ejemplo de modelo de ítems. Sección de características del contenido a evaluar.

En la Figura 3.8 se puede observar que para los expertos que elaboraron este modelo de ítems, evaluar el dominio del concepto de *fracción* resultó importante, porque se utiliza a lo largo de toda la educación primaria, particularmente las fracciones que se puedan o no reducir a otras expresiones. En este caso, la decisión de evaluarlo se justifica con un argumento sobre la importancia de utilizar figuras geométricas o polígonos divididos en segmentos o fracciones. Además de justificar la inclusión de ciertos dominios de contenido en el instrumento de evaluación, el modelo de ítems incluye un apartado en el cual se describen los tipos de conocimientos curriculares y habilidades cognitivas que requieren los estudiantes para resolver correctamente los ítems generados.

El ejemplo indica que para responder correctamente los ítems que evalúan este dominio, el estudiante deberá comprender y saber aplicar los conceptos de *fracción*, *fracción propia y equivalente*, así como la representación gráfica de las partes iguales de una totalidad. Hasta esta sección el modelo de ítems funciona a manera de mapa, mediante el cual se ubican las características generales del contenido curricular evaluado, y los motivos que justifican su selección del currículum y su relevancia.

La tercera sección del modelo de ítem es la que contiene las especificaciones de este, para generar diversos ítems. Es ahí donde se precisan los algoritmos y las reglas o restricciones que se establecieron para combinar los elementos conceptuales (cadena e integrales) y producir ítems. Debido a que la información debe ser lo suficientemente detallada como para que de ella se elaboren grandes cantidades de reactivos, es indispensable que en esta sección se expliciten con el mayor detalle posible, los siguientes elementos: a) la estrategia de evaluación que se seguirá para poder evaluar lo que se desea; b) las bases del reactivo que servirán como instrucciones para el examinado, las cuales deberán ser lo más claras y concisas posible; c) los elementos que se utilizarán para generar los ítems (cadenas e integrales), así como los algoritmos (procedimientos, instrucciones y restricciones) que se deberán seguir para construir los ítems, con la finalidad de establecer criterios o reglas a seguir, sin excepción, por parte de los elaboradores del examen y del GenerEx (ver Figura 3.9).

Estrategia de evaluación

Se presentan al sustentante dos opciones:

1. Una fracción propia y se pide sombrear su fracción equivalente en una figura plana reticulada.
2. Una figura plana reticulada coloreada representando una fracción y el sustentante tendrá que escribir la fracción que representa.

Nota: Al sustentante sólo se le presenta una de dos opciones.

Base del reactivo

Opción 1. Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar haz clic nuevamente sobre ellas.

Opción 2. Observa la figura y escribe la fracción que representan las partes en blanco.

Datos para el programador

1. Se presenta al sustentante como base del reactivo uno de los dos enunciados siguientes (elegidos al azar):
   - **A.** Selecciona las partes de la figura que indica la fracción. Haz clic sobre las partes que elijas, y si deseas desmarcar haz clic nuevamente sobre ellas.
   - **B.** Observa la figura y escribe la fracción que representan las partes en blanco.
2. Para el caso del inciso "**A.**" Ver la tabla que aparece al final:
   a. De la tabla del banco de información, se elige al azar una fracción $\frac{a}{b}$ (segunda columna) de los renglones I, II, III ó IV, con su respectivo polígono
   b. Para responder, el sustentante pintará las partes que representen a la fracción.
   c. Las partes pintadas podrán estar en cualquier posición de la figura.
   d. La respuesta correcta se muestra en la cuarta columna de la tabla (Respuesta); es decir, el número de partes que deben aparecer pintadas.
3. Para el caso del inciso "**B.**" Ver la tabla que aparece al final:
   a. Se elige al azar una fracción (segunda columna) de los renglones I, II, III ó IV, junto con su respectivo polígono.
   b. Se muestra al sustentante la figura con las partes pintadas que representan a la fracción elegida.
   c. La cantidad de partes que deben aparecer pintadas se muestran en la columna 4 de la tabla.
   d. Las partes pintadas pueden aparecer en orden aleatorio, compartiendo al menos uno de sus lados.
   e. La respuesta será correcta si coincide con la fracción $\frac{a}{b}$ seleccionada.
   f. Si no coincide, se deberá verificar la equivalencia, $\frac{a}{b}$ respuesta del programador y $\frac{c}{d}$ respuesta del sustentante; si: $a*d = b*c$ se toma como respuesta correcta.

Figura 3.9. Ejemplo de modelo de ítems. Generador con algoritmos y reglas para generar ítems.

En la Figura 3.9 se puede observar que la estrategia de evaluación del modelo de ítems del ejemplo incluye algoritmos (procedimientos) para dos aproximaciones en la generación de los ítems que evaluarán el dominio del concepto de fracciones. En la primera (tipo 1) se construirán ítems en los que se presentará una figura geométrica segmentada y una fracción expresada en números. En la base del reactivo se solicitará al estudiante que sombree el segmento o los segmentos de la figura, que considere que representan la fracción numérica presentada en el ítem. En la segunda aproximación (tipo 2) se construirán ítems de manera inversa; es decir, en los que se presente una figura geométrica con segmentos sombreados, así como un espacio donde el estudiante deberá escribir la fracción

que están representando. En lo que respecta a la base del reactivo, el ejemplo muestra cómo se pueden generar distintas instrucciones para medir el mismo tipo de contenido curricular. En este caso, la redacción puede variar según se le solicite al estudiante, que escriba su respuesta o que la indique mediante el sombreado de áreas en la figura geométrica.

En el segmento denominado *Datos para el programador* se encuentra el generador de ítems que explica cómo construir ítems para los dos tipos de ejecución: sombrear secciones de una figura o escribir la fracción correspondiente. Las reglas contenidas en los algoritmos indican que para construir ítems en los que el estudiante sombree segmentos de un polígono, se deberá seleccionar al azar (de un banco de elementos conceptuales) una fracción numérica, junto con el polígono al cual se encuentre asociada. De esta manera, el estudiante puede visualizar ambas cosas y seleccionar o marcar cualquier segmento o segmentos del polígono, que en conjunto representen la fracción numérica que se le presentó, sin importar si comparten sus lados o no.

Por otro lado, si se desea construir ítems en los que el estudiante escriba la fracción que corresponde a los segmentos que se le presentan sombreados dentro de un polígono, el sistema deberá utilizar el algoritmo para seleccionar un polígono al azar (del banco de elementos), junto con una fracción asociada. Se presentará la figura con los segmentos sombreados, tomando en cuenta la restricción de que siempre deberán compartir al menos uno de sus lados. Cuando el estudiante escriba su respuesta, esta deberá coincidir exactamente con la fracción que fue seleccionada para presentarle. De no ser así, se deberá aplicar una regla de equivalencia para calificar, mediante la cual se comparará la respuesta del estudiante con la respuesta correcta original. En caso de mostrar equivalencia se calificará como correcta.

La sección más importante del modelo de ítems es la que contiene todos los elementos conceptuales que forman parte del banco de información curricular, ya que es el lugar de donde se seleccionan los elementos cadena e integrales que se combinarán de distintas maneras, generando múltiples ítems. Con el mismo ejemplo del modelo de ítem de matemáticas, en el cual se evalúa el concepto de *fracción* mediante figuras geométricas en las que se sombrea (tipo 1) o se escribe (tipo 2) la respuesta, en la Figura 3.10 se muestran los elementos que se pueden utilizar para construir reactivos de cualquiera de los dos tipos de ejecución.

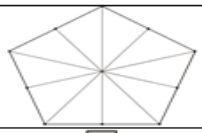
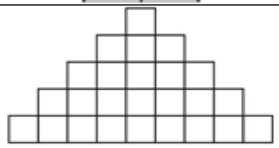
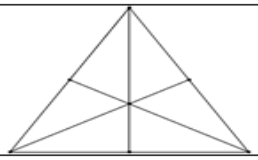
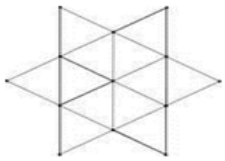
	Fracciones propias $(\frac{a}{b})$	Polígono	Respuesta $(\frac{a}{b} \times n)$ n = número de particiones del polígono
I	$\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ $\frac{3}{15}, \frac{6}{15}, \frac{9}{15}, \frac{12}{15}$		Sombrear $\frac{a}{b}$ x cualesquiera 10 segmentos
II	$\frac{1}{5}, \frac{2}{5}, \frac{3}{5}, \frac{4}{5}$ $\frac{2}{10}, \frac{4}{10}, \frac{6}{10}, \frac{8}{10}$ $\frac{3}{15}, \frac{6}{15}, \frac{9}{15}, \frac{12}{15}$		Sombrear $\frac{a}{b}$ x cualesquiera 25 segmentos
III	$\frac{1}{3}, \frac{2}{3}$ $\frac{3}{9}, \frac{6}{9}$ $\frac{2}{12}, \frac{3}{12}, \frac{4}{12}, \frac{6}{12}, \frac{8}{12}, \frac{10}{12}$		Sombrear $\frac{a}{b}$ x cualesquiera 6 segmentos
IV	$\frac{1}{3}, \frac{2}{3}$ $\frac{1}{4}, \frac{2}{4}, \frac{3}{4}$ $\frac{1}{6}, \frac{2}{6}, \frac{3}{6}, \frac{4}{6}, \frac{5}{6}$ $\frac{2}{8}, \frac{4}{8}, \frac{6}{8}$ $\frac{3}{9}, \frac{6}{9}$		Sombrear $\frac{a}{b}$ x cualesquiera 12 segmentos

Figura 3.10. Ejemplo de modelo de ítems. Banco de información curricular que se utiliza para generar ítems que evalúan el manejo del concepto de *fracciones*.

En la columna titulada *Fracciones propias* de la Figura 3.10, se observa que el formato en el que se presenta el numerador y el denominador de la fracción es muy específico (el primero arriba y el segundo abajo), ya que debe estar alineado al formato que utilizan los

libros de texto, sus ejemplos y ejercicios. También se observa que cada grupo de fracciones está asociado a un polígono, por lo cual no se deberán mezclar fracciones de un polígono con otro. Así, en el caso del pentágono, se cuenta con un banco de ocho elementos integrales (fracciones numéricas), en el de los cuadros apilados se tienen 12, en el triángulo hay 10 y en la estrella 15.

Si se construyen ítems del tipo 1, en los que el estudiante debe sombrear los segmentos del polígono que corresponden a la fracción numérica que se le muestra, con el banco de elementos conceptuales que se cuenta en este ejemplo, se tiene la posibilidad de generar 45 ítems distintos. En el caso de los ítems tipo 2, en los que el estudiante escribe la fracción correspondiente al área sombreada que aparece dentro del polígono, se cuenta con información para generar otros 45 ítems distintos, sin considerar que pueden aparecer diferentes áreas sombreadas para la misma fracción. Esto implica que con el generador de ítems contenido en este modelo de ítems, se puede tener al menos 90 reactivos distintos que evalúan el mismo contenido curricular.

En el ejemplo anterior se muestra la forma en la que se puede diseñar una gran cantidad de ítems para evaluar una habilidad. En este caso la habilidad matemática implica que el estudiante demuestre que se domina la competencia para el manejo de fracciones y su representación en figuras geométricas. Se presentan dos tipos de bases de reactivo que solicitan al estudiante diferentes tipos de ejecución, pero con la misma finalidad evaluativa.

En la Figura 3.11 se muestra la imagen de las dos versiones de un reactivo que fue generado mediante el modelo de ítems explicado en los párrafos anteriores. El estudiante tendrá únicamente uno de ellos en su examen. La figura del lado izquierdo corresponde al ítem del tipo 1 (sombreado) y la figura de la derecha al reactivo de tipo 2 (escritura).

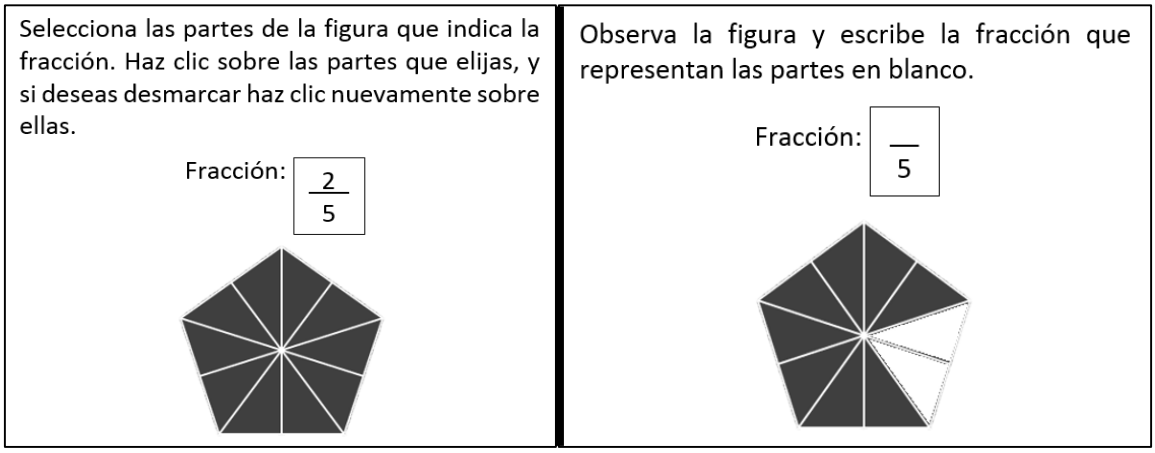


Figura 3.11. Ítems hijos de matemáticas, provenientes de un modelo de ítems con dos familias: sombreado y escritura.

3.4. Familias de ítems, ítems padre e ítems hijos

Del modelo de ítems, se obtiene un mapa del cual se extraen tareas específicas que deberá realizar el estudiante, ordenadas jerárquicamente en tres niveles, que producirán uno o más grupos de reactivos: las *familias de ítems*, *ítems padre* e *ítems hijos*. Estos tres niveles sirven para construir una gran variedad de ítems isomorfos que evalúan el mismo contenido curricular, habilidad o aprendizaje esperado del estudiante. En general, muchos modelos de ítems solamente producen una familia de reactivos, pero existen casos en los que los procedimientos de construcción permiten la existencia de dos o más familias. Tal es el caso del ejemplo anterior, en el que el modelo de ítems busca evaluar el manejo del concepto de fracción mediante el uso de figuras geométricas.

De acuerdo con el mapa de clasificación de los niveles jerárquicos del modelo de ítems del ejemplo anterior, si se tienen dos tipos de reactivos: aquellos que solicitan sombrear y aquellos que solicitan escribir la respuesta, se diría que existen dos familias de reactivos. También se cuenta con cuatro polígonos distintos: pentágono, cuadros apilados, triángulo

y estrella. Al seguir la clasificación, cada uno se llamaría ítem padre, ya que de ellos "nacen" los ítems hijos. La figura 3.12 muestra la organización de las familias, padres e hijos de este ejemplo.

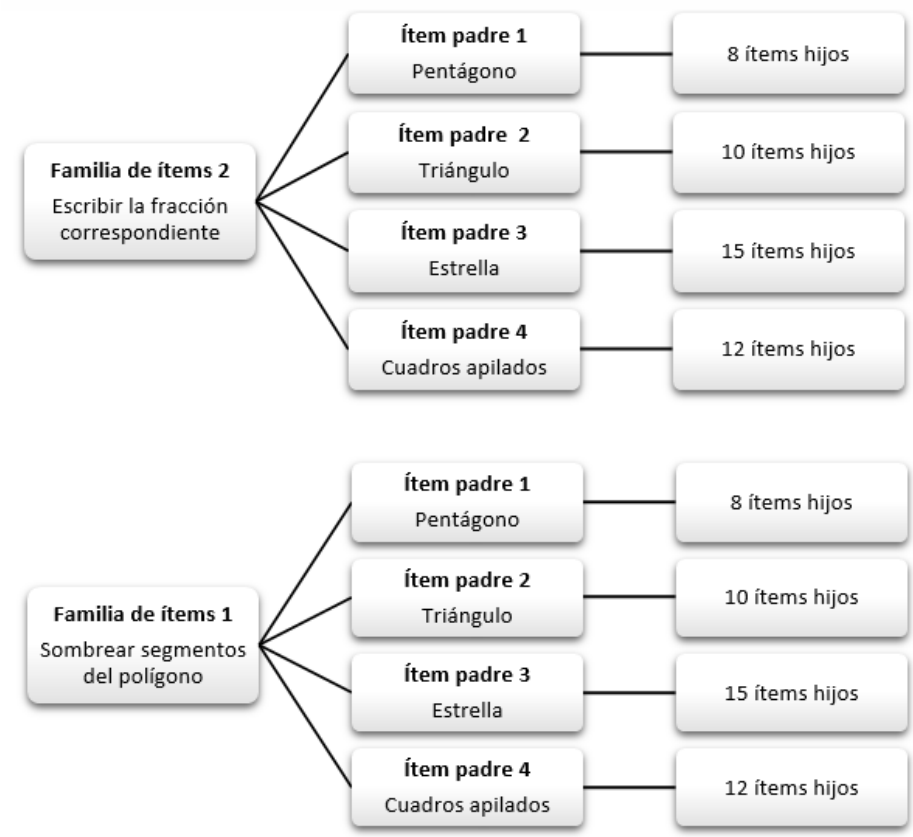


Figura 3.12. Familias de ítems, ítems padre e ítems hijos en un modelo de ítems para evaluar el manejo del concepto de fracciones.

En este caso, las familias de ítems se clasificaron de acuerdo con las distintas formas de ejecución que se solicitan del estudiante; sin embargo, se puede hacer con una lógica distinta. Un modelo de ítems puede contener una o más formas de clasificar las familias, ya sea por medio de los conceptos que maneja, la cantidad de textos que se tienen en el banco de información, los tipos de tareas que se solicitan al estudiante, entre otras. Las familias

deberán incluir una cantidad suficiente de elementos como para seleccionar aleatoriamente unos cuantos (más no la totalidad) de ellos, según las reglas y restricciones establecidas ya que así se podrán hacer combinaciones y construir muchos ítems hijos. Los tres niveles jerárquicos no se encontrarán en todos los modelos. En muchos casos los ítems hijos se generarán directamente de los ítems padre, sin la presencia de familias de ítems; sin embargo, en todos los modelos de ítems habrá *n* cantidad de ítems hijos y al menos un ítem padre que los generará.

4. Evidencias de validez de contenido del Excoba: modelos de ítems

4.1. Fase II del MVCE. Obtención de evidencias de validez de contenido de los modelos de ítems del Excoba

La segunda fase del MVCE implicó diversas acciones cuyo propósito fue:

a) Conocer las características de los modelos de ítems, de acuerdo con el tipo de contenido según el currículum oficial, la clasificación que tienen en el sistema informático que administra el examen, la ejecución que debe realizar el estudiante al responder, y el tipo de conocimiento que evalúa.

b) Determinar qué tan representativos son los contenidos seleccionados para conformar la estructura del Excoba, con el universo de contenidos que integran el currículum de educación básica, específicamente de las asignaturas de Matemáticas, Historia, Química y Español de educación secundaria.

c) Determinar la proporción en la que los modelos de ítems se apegan a la evaluación de los contenidos que se pretende evaluar del currículum de las asignaturas de Matemáticas, Español, Química e Historia de educación secundaria.

Para llevar a cabo la clasificación de los modelos de ítems, se revisaron y contrastaron con la información contenida en el currículum oficial de la educación básica mexicana. Se clasificaron de acuerdo con su registro en el sistema GenerEx. Se revisaron las estrategias evaluativas contenidas en cada modelo de ítems, para determinar qué tipo de ejecución se solicita del estudiante, y clasificarlos en: arrastre, selección, escritura o mixta. Y se determinó el tipo de conocimiento que evalúa cada modelo de ítems, utilizando como

referente la taxonomía propuesta por Ruiz-Primo en el año 2007: conocimiento de tipo declarativo, procedimental, esquemático y estratégico.

En cuanto al proceso de obtención de evidencias de validez de contenido de los modelos de ítems, se utilizó el método de validación mediante el trabajo de paneles de expertos, quienes proporcionaron opiniones acerca de: 1) la representatividad de los contenidos seleccionados del currículum, respecto al universo del currículum de educación básica, específicamente las asignaturas de Matemáticas, Historia, Química y Español de secundaria; 2) la estructura de los modelos de ítems y su claridad, en relación con las reglas que contienen para la generación de reactivos; 3) los elementos conceptuales (cadena e integrales) relacionados con el contenido curricular evaluado, y que sirven como insumo para generar grandes cantidades de ítems y versiones del examen; 4) la claridad de las instrucciones de las bases de los reactivos, y 5) el grado de correspondencia que guarda la ejecución solicitada al estudiante en la resolución del examen, con los aprendizajes esperados.

Participantes

Se seleccionó y capacitó a un panel de dos expertos por cada área del examen incluida en este estudio: Matemáticas, Español, Química e Historia, de secundaria. Los integrantes de cada panel trabajaron de manera colegiada durante todo el proceso. La determinación de la cantidad de especialistas se hizo bajo el criterio de accesibilidad a los mismos, así como su disposición para colaborar en esta investigación.

En total, intervinieron ocho especialistas que evaluaron los aspectos mencionados con anterioridad. La selección de los expertos se hizo con base en su perfil académico, así como su experiencia profesional. Fueron profesores de escuelas secundarias públicas y privadas, en ejercicio docente. Entre las características que se determinaron para la selección de los integrantes de los distintos paneles, se encuentran las siguientes:

- Tener experiencia mínima de cinco años frente a grupo.
- Tener especialidad en la asignatura y grado en el que colaboraron como expertos.
- Tener un amplio conocimiento del currículum de la asignatura en la que colaboraron, así como de la operación del mismo dentro del aula.
- Ser un buen docente y contar con prestigio como tal ante sus colegas y estudiantes. Para corroborar esta característica es de particular importancia la opinión de los directivos del plantel o los planteles donde ejerce su práctica docente, así como la recomendación de sus colegas.
- Conocer, mediante la experiencia docente y la constante actualización en cuanto a los programas de estudio, las habilidades y conocimientos que se supone deben adquirir los estudiantes a lo largo de su educación básica (primaria y secundaria).
- Estar interesados en el proceso de validación del instrumento.
- Contar con disponibilidad de tiempo para realizar la labor dentro de los tiempos asignados y bajo las condiciones que marcó el programa de trabajo.
- Fue deseable, mas no determinante, que los docentes participantes conocieran y tuvieran experiencia o capacitación en el campo de la evaluación educativa, así como en el diseño de instrumentos de evaluación del aprendizaje, particularmente de gran escala.

Materiales

Para registrar el proceso de toma de decisiones, así como los resultados individuales y grupales del proceso de evaluación, se entregó a cada miembro de los distintos paneles un conjunto de materiales de trabajo, que a continuación se describen:

- Manual de trabajo. Contiene toda la información revisada durante el proceso de capacitación, de forma condensada y secuencial. Su función fue como material de consulta permanente y de apoyo para los evaluadores, durante la realización del trabajo encomendado.

- Formato de currículum vitae. Contiene la información del especialista: datos de contacto, formación académica, lugar de trabajo, experiencia laboral dentro del ámbito académico, actividades académicas y profesionales que actualmente realiza (ver Apéndice 2).
- Protocolo para la evaluación de los modelos de ítems. Registra las decisiones de los expertos. Los elementos que evalúa son: a) la representatividad de los contenidos seleccionados, respecto al universo de contenidos del currículum de primaria y secundaria; b) la importancia que tienen los contenidos seleccionados, respecto a las habilidades y los conocimientos esperados del estudiante; c) su alineación curricular; d) la claridad y precisión al definir los contenidos, así como al delimitarlos; e) la plantilla del reactivo, y f) el generador de reactivos que contiene los elementos cadena e integrales, las reglas, restricciones y la información que se utilizará para generar las múltiples versiones del examen (ver Apéndice 3).
- Formato de compromiso de confidencialidad. Detalla las normas de confidencialidad que deben seguir los expertos que participan en el trabajo de validación de los contenidos de los modelos de ítems del Excoba (ver Apéndice 4).
- Material multimedia. Se refiere a todo el material utilizado con la finalidad de presentar información a los miembros de cada panel durante el proceso de capacitación, así como de evaluación de los modelos de ítems e ítems hijos. Está conformado por diferentes conjuntos de diapositivas digitales, gráficos, figuras, imágenes y texto.
- Materiales auxiliares. Incluye los libros de texto, así como los planes y programas de estudio de cada asignatura que contempla el Excoba.

Procedimiento

La fase de obtención de evidencias de validez de contenido de los modelos de ítems que conforman el Excoba tuvo como propósito revisar y verificar la congruencia entre estos y los ítems hijos. En esta labor participaron diversos paneles de expertos en las asignaturas

de Matemáticas, Español, Química e Historia, de educación secundaria. Debido a que la cantidad de modelos de ítems a evaluar oscilaba entre 8 y 20, según la asignatura de la que se tratara, el procedimiento se llevó a cabo en tres modalidades: trabajo individual de cada experto, trabajo colegiado y trabajo mixto (ver Figura 4.1).

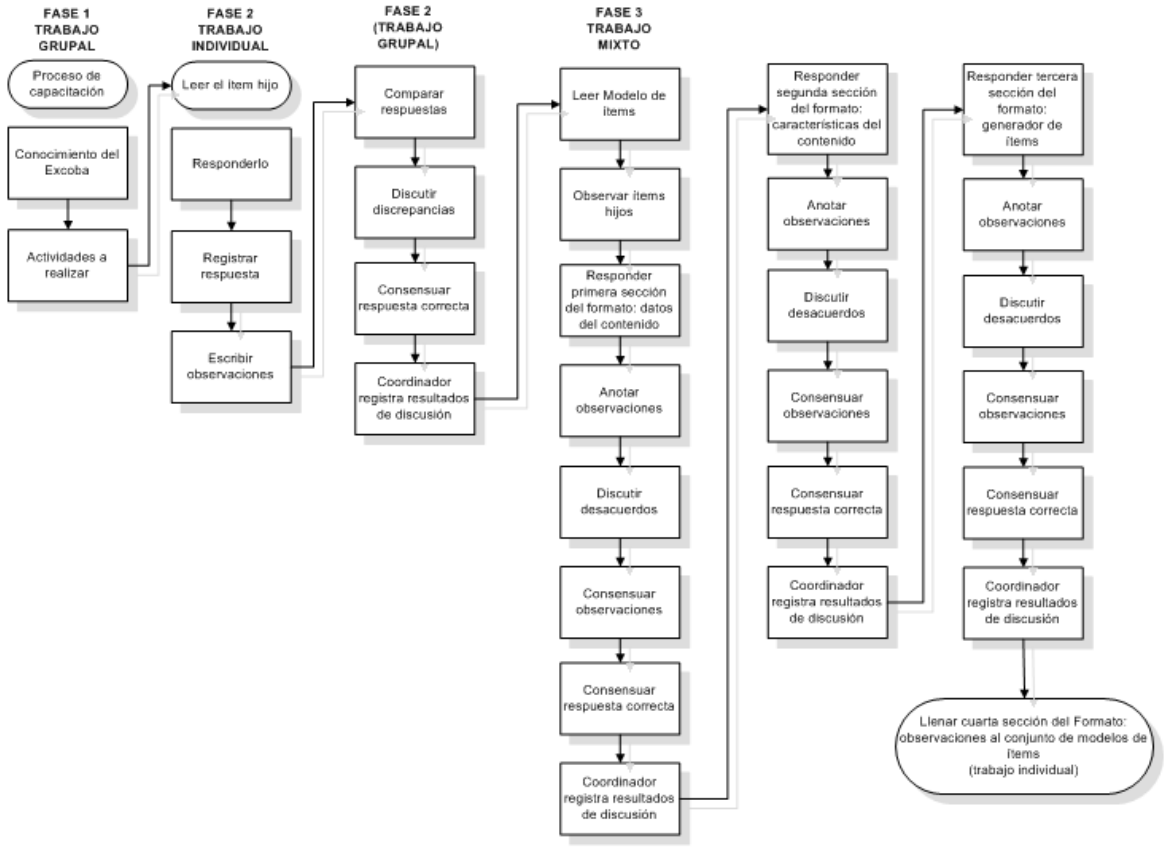


Figura 4.1. Proceso de obtención de evidencias de validez de contenido de los modelos de ítems del Excoba. Trabajo con paneles de expertos.

Durante el proceso, los expertos atendieron puntualmente la evaluación de tres aspectos: a) el contenido seleccionado del currículum, mencionado y delimitado en cada modelo de ítems; b) la estrategia evaluativa utilizada para generar reactivos, junto con el generador de ítems y los elementos cadena e integrales, y c) un ítem hijo generado a partir del modelo de ítems. La Figura 4.1 muestra cómo el trabajo con los paneles de expertos se

llevó a cabo en tres fases: capacitación, realización del examen y evaluación de los modelos de ítems.

A continuación se describe cada uno de los componentes de las tres fases.

Fase 1. En la primera sesión el trabajo consistió en un proceso de capacitación. Los participantes recibieron información respecto al marco teórico-conceptual del modelo y la metodología utilizada para la elaboración del Excoba, así como el trabajo a realizar durante el proceso de validación de los modelos de ítems del mismo. Se brindó información de carácter contextual, con la finalidad de que conocieran los siguientes aspectos:

- El Modelo de Elaboración del Excoba (MEE), el proceso de planeación y diseño del instrumento, la metodología de trabajo que siguieron los distintos paneles, así como la importancia de que éstos estuvieran conformados por diferentes especialistas en cada fase de la elaboración del examen.
- La estructura, las características y particularidades que tiene el Excoba, cómo se encuentra conformado por áreas, campos del conocimiento y contenidos, y por qué es un generador de exámenes.
- La propuesta de innovación, mediante la transición del uso de reactivos de opción múltiple al uso de modelos de ítems para generar grandes cantidades de ítems que se aproximen a una evaluación auténtica de los aprendizajes esperados en los estudiantes.
- La importancia del trabajo de obtención de evidencias de validez del Excoba (particularmente de contenido) y sus modelos de ítems, debido a su naturaleza normativa y uso para la selección de estudiantes a ingresar al bachillerato.
- La formación de grupos de expertos por área temática o asignatura, así como el trabajo individual y grupal de cada experto.
- El estricto apego al manual de procedimientos y la importancia de que las opiniones de los expertos sean consistentes.

- El acuerdo de confidencialidad y resguardo de la información recibida, así como de los productos generados durante el proceso.

Fase 2. En un segundo momento se llevó a cabo la administración de cada uno de los ítems hijos que fueron seleccionados, para presentarlos a los expertos. Cada uno de ellos ejerció el papel de examinado, con la finalidad de que se sensibilizaran respecto a la de experiencia de los estudiantes al responder el Excoba.

Una vez que resolvieron el ítem se compararon las respuestas de los evaluadores y se inició la discusión. En los casos en los que hubo diferencia en las respuestas, se definió la naturaleza de las mismas y se eligió la correcta por consenso. Este paso fue muy importante, ya que además de sensibilizarlos, permitió a cada experto tener la oportunidad de evaluar, desde otra perspectiva, la funcionalidad de cada uno de los reactivos. Asimismo, este ejercicio les facilitó opinar sobre otros aspectos y aportar comentarios que posteriormente complementaron los formatos de evaluación y ayudaron a mejorar el instrumento. En este sentido, la discusión generada una vez que concluyó la aplicación de cada uno de los ítems del examen fue de gran relevancia.

Fase 3. La última fase fue destinada al trabajo de evaluación colegiada de los expertos, en relación con el modelo de ítem (contenidos curriculares, estrategia evaluativa, generadores de reactivos, elementos cadena e integrales e ítems hijos). Se llevó a cabo en distintas sesiones de trabajo, organizadas por asignatura. Esta fase, ocurrió inmediatamente después de que se presentaron cada uno de los ítems hijos. Se empleó un proyector multimedia para mostrar las distintas secciones del modelo de ítems ante los miembros de cada grupo de expertos, con la finalidad de que llevaran a cabo una evaluación individual de cada sección.

La tarea de los participantes consistió en leer con detenimiento cada sección e identificar todos los problemas, errores u omisiones contenidos en los modelos de ítems.

Para ello, trabajaron en el siguiente orden: a) revisaron la sección de los datos de identificación del contenido, b) analizaron las características del contenido a evaluar, y c) revisaron la sección que contiene el generador de ítems, elementos cadena e integrales y las reglas y restricciones para la combinación de dichos elementos.

Los expertos llevaron a cabo la evaluación de manera individual. Posteriormente discutieron grupalmente los hallazgos, haciendo pausas para el trabajo colegiado de cada sección del modelo de ítems, hasta llegar a un consenso respecto a los aspectos evaluados. Todas las opiniones se registraron en un formato de evaluación (ver Apéndice 3), en el que también agregaron comentarios sobre contenidos curriculares que no se incluyeron en el examen, y otros aspectos que consideraron relevante mencionar.

Entre los aspectos que evaluaron se encuentran los siguientes:

- Alineación y correspondencia del contenido evaluado con el nivel y área temática correspondiente en el currículum.
- La pertinencia del contenido y su carácter de esencial para ser incluido en el examen.
- La congruencia de las distintas secciones que conforman el modelo de ítems (por ejemplo: estrategia evaluativa utilizada, generador de reactivos y elementos cadena e integrales).
- El uso de un lenguaje apropiado a la edad y al nivel educativo del estudiante en la instrucción, en la base del reactivo, y en los elementos utilizados para construir los ítems hijos;
- Aspectos de ortografía y redacción.
- Características de los textos auxiliares, gráficas, tablas y/o imágenes utilizados en los modelos de ítems.
- Contenidos curriculares omitidos en el Excoba, considerados esenciales por el panel de expertos.

De esta forma, la tarea del panel de evaluadores fue conocer y valorar cada uno de los elementos que conforman los modelos de ítems de los distintos contenidos que conforman el Excoba, y que fueron incluidos en este estudio.

4.2. Resultados de la Fase II del MVCE

A continuación se presentan los resultados obtenidos mediante la aplicación de la segunda fase del MVCE. Esta implicó las etapas 2, 3 y 4 del mismo, que son: análisis y clasificación de los modelos de ítems por áreas del conocimiento (etapa 2); trabajo de paneles de expertos en la revisión detallada de los modelos de ítems de la asignatura correspondiente (etapa 3), y el análisis de los modelos de ítems como conjunto (etapa 4). Los resultados se muestran de manera separada para cada una de las cuatro áreas que fueron seleccionadas para esta investigación: Matemáticas, Historia, Química y Español.

Siguiendo la secuencia del MVCE, en la etapa 2 se presenta una descripción de los modelos de ítems, de acuerdo con cuatro clasificaciones: 1) el tipo de contenido curricular que evalúa cada uno de ellos, según el eje temático al que pertenecen dentro del programa de estudios correspondiente; 2) la clasificación que tienen los tipos de modelos de ítems en el sistema informático que administra el examen (GenerEx); 3) el tipo de ejecución que debe realizar el estudiante para emitir su respuesta al utilizar la interfaz gráfica del examen (arrastre, selección, escritura o mixta), y 4) el tipo de conocimiento que exploran, según los procesos intelectuales que se espera que lleve a cabo el estudiante al aprender los contenidos curriculares dentro de la asignatura (declarativo, procedimental, esquemático o estratégico).

Los resultados de la etapa 3 muestran el análisis de cada área temática. La información fue obtenida mediante la participación de los paneles de expertos. El análisis se encuentra organizado en dos niveles: 1) como conjunto de modelos de ítems, de acuerdo

con los indicadores en los que los evaluadores señalaron la presencia de problemas, y 2) como modelos de ítems individuales, de acuerdo con los problemas que se presentaron en los indicadores bajo los cuales fueron evaluados. Al final de este apartado, se presentan las sugerencias de los expertos, así como una síntesis de los resultados, con la finalidad de identificar las posibles causas de los problemas detectados.

En un último apartado se presenta un análisis de los resultados encontrados en las cuatro áreas como conjunto. La finalidad de esto es buscar la existencia de regularidades en tres aspectos: 1) la clasificación de los modelos de ítems en el sistema informático GenerEx; 2) la ejecución que debe realizar el estudiante para emitir su respuesta (arrastre, selección, escritura o mixta), y 3) el tipo de conocimiento que explora cada uno de los modelos de ítems (declarativo, procedimental, esquemático o estratégico).

En el siguiente apartado se describe la forma en la que se obtuvo la información emitida por los docentes que fungieron como expertos, y la manera en que se analizó la información resultante.

Proceso de jueceo y calificación

La evaluación de cada uno de los modelos de ítems se realizó mediante el llenado del formato que se utilizó para dicho fin (ver Apéndice 3). Durante este proceso se obtuvo información concerniente a distintos aspectos de los modelos de ítems, tales como: cobertura, representatividad y pertinencia de los contenidos seleccionados del currículum; la funcionalidad de la estrategia evaluativa utilizada, y los problemas detectados en los ítems hijos que se generan a partir de dichos modelos.

El panorama global de la apreciación de los expertos, respecto a los modelos de ítems de cada asignatura se obtuvo mediante el registro de las opiniones consensuadas por los miembros de los paneles. Ellos las expresaron en forma verbal y las registraron de manera

escrita y electrónica en el formato de evaluación. Dicha evaluación constó de 27 afirmaciones, ante las cuales cada evaluador emitió una opinión de *acuerdo* o *desacuerdo*, además de comentarios específicos y sugerencias para mejorar cada aspecto evaluado.

Con base en esta información, se obtuvieron los porcentajes de indicadores que fueron validados (acuerdo entre expertos) y que no lo fueron (desacuerdo entre expertos). Dichos resultados representan las opiniones consensuadas por el panel, y en ningún momento se deben interpretar como evaluaciones individuales.

Para efectos de considerar que un modelo de ítems o un indicador del formato de evaluación presentan problemas significativos, se estableció como criterio un porcentaje igual o mayor a 20% de los indicadores que no fueron validados por los expertos.

De la información general, se desprendió un análisis de los problemas detectados por el panel de expertos en cada uno de los modelos de ítems de las asignaturas evaluadas, tales como: a) claridad, alineación curricular y congruencia de los datos de identificación del contenido, b) características, correspondencia y pertinencia del contenido evaluado, y c) características del generador de reactivos y de los ítems hijos, como su diseño, funcionalidad y su correspondiente banco de elementos cadena e integrales, entre otros.

Por último, se presentan las sugerencias realizadas por parte del panel de expertos, con la finalidad de mejorar cada uno de los modelos de ítems. Dichas observaciones señalan aspectos específicos que los evaluadores consideraron que deben ser atendidos para fortalecer el examen.

Al final de este capítulo, se presenta un análisis de las cuatro áreas evaluadas, en el cual se muestra la clasificación de los modelos de ítems de cada una, según la forma en que están clasificados en el sistema informático, la ejecución que solicitan del estudiante, y el tipo de conocimiento que exploran. Lo anterior se realizó con la finalidad de buscar

regularidades entre los problemas que presentan los modelos y su clasificación en dichas categorías.

4.2.1. Modelos de ítems del área de Matemáticas

Este apartado, correspondiente a la Fase II del MVCE está dividido en dos secciones referente a las etapas 2 y 3 del MVCE. Siguiendo este orden, en los resultados de la etapa 2 se describen las características de los modelos de ítems que conforman el área, de acuerdo con el tipo de contenido curricular que evalúan, la clasificación que tienen en el sistema informático que administra el examen, la ejecución que debe realizar el estudiante al responder, y el tipo de conocimiento que exploran.

En los resultados de la etapa 3 se presenta el análisis, cuya información se obtuvo mediante el proceso de validación de los modelos de ítem. En este proceso hubo una serie de opiniones emitidas por los expertos, acerca de los indicadores contenidos en el formato de evaluación.

Etapa 2: Características de los modelos de ítems de Matemáticas

Los tipos de características de los ítems que se analizaron fueron cuatro: 1) los contenidos curriculares, clasificados de acuerdo con el eje temático al que pertenecen dentro del programa de estudios 2006, 2) las características informáticas de los modelos de ítems, 3) la ejecución que los modelos de ítems solicitan al estudiante para responder al reactivo, y 4) los conocimientos evaluados, de acuerdo con la taxonomía propuesta por Ruiz-Primo en 2007.

Tipo de contenidos curriculares

Para la elaboración del GAI Excoba se utilizó el programa de estudios de Matemáticas, de educación secundaria, del año 2006. En él, todos los contenidos de la asignatura se organizan en tres ejes: 1) sentido numérico y pensamiento algebraico, 2) forma, espacio y

medida, y 3) manejo de la información (SEP, 2006c).[2] La Tabla 4.1 muestra la distribución de los 20 modelos de ítems siguiendo esta clasificación.

Tabla 4.1.
Distribución de modelos de ítems de Matemáticas por eje temático curricular

Modelos de ítems	Contenido	Eje temático	Distribución (porcentaje)
MAT01	Cálculo potencia (enteros +)	Sentido numérico y pensamiento algebraico	60
MAT02	Cálculo potencia (enteros + y/o -)		
MAT03	Sucesiones aritméticas		
MAT04	Operaciones c/polinomios		
MAT05	Ecuaciones primer grado		
MAT06	Sistemas de ecuaciones		
MAT07	Ecuaciones cuadráticas		
MAT08	Representación de recta		
MAT10	Teorema Pitágoras		
MAT13	Teorema Tales		
MAT16	Conteo		
MAT19	Pendiente recta		
MAT09	Simetría axial	Forma, espacio y medida	20
MAT11	Ángulos entre paralelas y secante		
MAT12	Desarrollo plano de cuerpos geométricos		
MAT14	Trigonometría		
MAT15	Proporcionalidad inversa	Manejo de la información	20
MAT17	Interpretación de gráficas		
MAT18	Medidas tendencia central y dispersión		
MAT20	Gráfica de una parábola		

[2] **Sentido numérico y pensamiento algebraico**. Se refiere a aquellos contenidos que permiten que el estudiante dé sentido al lenguaje matemático en sus formas oral y escrita, de tal manera que pueda establecer una relación entre la aritmética y el álgebra, más allá de las bases que adquirió durante sus estudios primarios, ya que durante la educación secundaria se profundiza en este tipo de contenidos, buscando su consolidación.

Forma, espacio y medida. Incluye los conocimientos relacionados con la geometría y medición, en los cuales se revisan los aspectos de forma y espacio de manera detallada, particularmente lo relacionado con el trazo, la construcción, propiedades y medición de las formas.

Manejo de la información. Se refiere a aquellos contenidos curriculares cuya naturaleza enseña al estudiante a identificar y analizar el origen de la información contenida en diversas fuentes, tales como gráficas o tablas; de tal manera que pueda seleccionar y organizar aquella que es relevante para interpretar correctamente los datos que en ella se encuentran representados.

Se puede apreciar que 12 (60%) de los modelos de ítems evalúan competencias de *sentido numérico* y *pensamiento algebraico*, 4 de ellos (20%) evalúan habilidades relacionadas con *Forma, espacio y medida*, y otros 4(20%) evalúan contenidos de *Manejo de la información*.

Características del sistema informático GenerEx

El sistema informático creado para la captura de datos, administración, aplicación y calificación de los exámenes que se generan mediante la información contenida en los modelos de ítems, permite clasificar 22 tipos diferentes de modelos de ítems. Esto es posible, según la programación que se requiera para su correcto funcionamiento y visualización en la interfaz gráfica del GAI Excoba. La descripción detallada de cada uno de ellos se encuentra en el capítulo 3 de esta tesis, en el apartado en que se describe el Modelo de Elaboración del Excoba.

En el caso del área de matemáticas de secundaria, se utilizaron nueve diferentes tipos de modelos de ítems diferentes, según esta clasificación (ver Tabla 4.2).

Tabla 4.2.
Clasificación y distribución de modelos de ítems de Matemáticas en el GenerEx

	Total	Porcentaje
R. Algebraica	5	25
RN/ ecuaciones	2	10
RN R. Algebraica	2	10
Elemento categoría	2	10
RN/ etiquetas	3	15
RN/ fórmulas	3	15
RN/ sucesiones	1	5
RN/ gráficas	1	5
RN/ pendiente	1	5

La distribución que se presentó en los 20 modelos de ítems, muestra que 25% (5) implica una programación que genera ítems hijos que permiten registrar respuestas escritas

en forma algebraica; 10% (2) de los modelos requieren que la interfaz genere ecuaciones para crear los ítems hijos; 10% (2) implica dos modalidades de reactivos según la selección aleatoria de elementos que haga el sistema: aquellos que requieren respuestas escritas en forma de números, o aquellos que requieran que la respuesta sea de tipo algebraica; 10% (2) de los modelos de ítems implica una interfaz que permite realizar la clasificación de elementos dentro de categorías; 15% (3) genera ítems hijos cuya base del reactivo o los elementos gráficos auxiliares (p. ej. figuras geométricas) intercambian una o más variables de manera aleatoria; 15% (3) corresponde a ítems hijos que requieren del uso de fórmulas, cuyas variables se intercambian directamente en la base del reactivo, según las reglas establecidas en la estrategia evaluativa del modelo del ítem; 5% (1) de los modelos de ítems requieren de una programación que arroje sucesiones numéricas con cantidades fijas de elementos; 5% (1) de los modelos de ítems incluye el uso de gráficas dinámicas, en las que se obtiene el valor de cada variable contenida en la base del reactivo, mediante la aplicación de una fórmula, y 5% (1) requiere del uso de planos cartesianos dinámicos, en donde se programa el sistema con rangos de máximos y mínimos para cada ordenada dentro del plano.

Tipo de ejecución solicitada

Los modelos de ítems que fueron validados por los expertos implicaron estrategias evaluativas que generan distintos tipos de ejecución por parte del estudiante cuando emite sus respuestas en los ítems hijos. En la Tabla 4.3 se presenta la distribución de los mismos, según fueron de arrastre de elementos, selección de elementos, escritura libre, o una combinación de selección de elementos con escritura libre.

Tabla 4.3.
Distribución de modelos de ítems de Matemáticas, según el tipo de ejecución que demandan del estudiante

Tipo de ejecución	Modelo de ítems	
	Total	Porcentaje
Arrastre	2	10
Selección	-----	-----
Escritura	16	80
Mixta (selección y escritura)	2	10

Como se puede observar, en el área de matemáticas, la gran mayoría de los modelos de ítems, es decir 80% (16), genera ítems hijos en donde la ejecución implica que el estudiante escriba libremente y construya su respuesta. Un 20% se encuentra distribuido de manera equitativa entre los modelos que generan ítems de arrastre y de ejecución mixta. Los modelos de ítems cuyos ítems hijos requieren ejecución de arrastre, es decir, donde el estudiante requiere del uso del ratón para el movimiento y colocación de elementos dentro de categorías o secciones de la interfaz, representan 10% (2). Los de ejecución mixta, donde se requiere la selección de elementos, así como la escritura libre de la respuesta, son también 10% (2).

De los modelos que generan ítems hijos que requieren una ejecución en la que el estudiante seleccione elementos y los clasifique en categorías, no se presentó ningún caso.

Tipo de conocimientos evaluados

La distribución de los modelos de ítems, según el tipo de conocimiento que exploran los ítems hijos generados, es la que se muestra en la Tabla 4.4.

Tabla 4.4.

Tipo de conocimiento que evalúan los modelos de ítems de Matemáticas

Conocimiento evaluado	Modelos de ítems	
	Total	Porcentaje
Declarativo	2	10
Procedimental	18	90
Esquemático	-----	-----
Estratégico	-----	-----

Se puede observar que 90% (18) de los modelos de ítems evalúan principalmente conocimientos de tipo procedimental. En ellos se requiere que el estudiante utilice su conocimiento acerca de métodos y procedimientos, lo que implican que sepa seguir reglas o pasos para obtener un resultado determinado.

Por ejemplo, en el modelo de ítems MAT01, que genera ítems hijos que evalúan el conocimiento que tiene el estudiante respecto a la forma en que se calcula una potencia, para resolverlos correctamente se necesita conocer los pasos implicados en el procedimiento para aplicar cada una de las reglas de los exponentes, específicamente las de suma, multiplicación y resta. De esta manera, el hecho de que el estudiante tenga un manejo adecuado de los procedimientos implicados en dichas reglas será determinante para que responda correctamente a los ítems hijos que exploran este contenido curricular.

Por otro lado, se observa que 10% (2) de los modelos de ítems evalúan conocimientos de tipo declarativo. En ellos se requiere un manejo organizado de la información previamente aprendida, mediante la clasificación de datos y elementos conceptuales. De esta manera, las respuestas dadas expresan los principios teóricos y conceptuales que el estudiante aprendió en la escuela. Así, por ejemplo, en el modelo de ítems MAT09, que evalúa el manejo del concepto de simetría axial, el estudiante debe contar con un manejo conceptual y teórico de los elementos que conforman la definición de simetría axial, para

responder correctamente al ítem hijo, ya que deberá discriminar entre una serie de figuras, aquellas que sean simétricas respecto a su eje y aquellas que no lo sean.

En el caso de los conocimientos de tipo esquemático y estratégico, no hubo modelos de ítems en el área de Matemáticas que generaran ítems hijos en estos niveles taxonómicos.

Etapa 3: Análisis de los resultados del trabajo con el panel de expertos

Resultados generales

En total se validaron 20 modelos de ítems, para lo cual se utilizó un reactivo hijo, como muestra de cada modelo de ítems, previamente seleccionado y generado mediante el sistema informático del Generador Automático de Ítems (GenerEx). Dichos reactivos corresponden a una versión fija del Excoba, que fue utilizada para su pilotaje, en un estudio realizado para la obtención de las propiedades psicométricas y la obtención de evidencias de validez de constructo (Ferreyra, 2014).

El proceso de evaluación que realizaron los miembros del panel de expertos proporcionó información que permitió identificar los problemas que presentan los 20 modelos de ítems que conforman el área de Matemáticas de secundaria del Excoba. De igual manera, ayudó a detectar los aspectos específicos que requerían atenderse para su mejoramiento. Los expertos evaluaron 27 indicadores de calidad en cada uno de los 20 modelos de ítems, consensuando acuerdo o desacuerdo en cada uno de ellos. Se valoró la claridad, alineación curricular, pertenencia y congruencia de los datos de identificación del contenido expresado en el modelo de ítems; las características, la correspondencia y pertinencia del contenido, así como las características del generador de reactivos y de los ítems hijos, tales como su diseño, funcionalidad y su correspondiente banco de elementos cadena e integrales.

En la Tabla 4.5 se muestran los resultados generales. Dicha tabla proporciona, de manera condensada, la información que se recopiló de las opiniones de los expertos respecto a factores de calidad y pertinencia de los contenidos seleccionados del currículum para ser evaluados. El símbolo de *paloma* (✓) señala que hubo acuerdo al valorar los indicadores (que resultaron válidos); mientras que la *tacha* (✗) indica desacuerdo (que los indicadores no fueron válidos).

Tabla 4.5.

Resultados globales del proceso de evaluación de panel de expertos del área de Matemáticas del Excoba

	Indicador	MAT01 Cálculo potencia (enteros +)	MAT02 Cálculo potencia (enteros + y/o -)	MAT03 Sucesiones aritméticas	MAT04 Operaciones c/polinomios	MAT05 Ecuaciones 1er grado	MAT06 Sistemas de ecuaciones	MAT07 Ecuaciones cuadráticas	MAT08 Representación de recta	MAT09 Simetría axial	MAT10 Teorema Pitágoras	MAT 11 Ángulos entre paralelas y secante	MAT12 Desarrollo plano cuerpos geométricos	MAT13 Teorema Tales	MAT14 Trigonometría	MAT15 Proporcionalidad inversa	MAT16 Conteo	MAT17 Interpretación de gráficas	MAT18 Medidas tendencia central y dispersión	MAT19 Pendiente recta	MAT20 Gráfica de parábola	% Validado (✓)	% No validado (×)
I1	Definición de contenido clara y precisa	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95	5
I2	Definición de contenido congruente con nombre	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95	5
I3	Definición de contenido alineada al currículum de asignatura	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	95	5
I4	Contenido coherente con lo que se enseña en aula	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	95	5
I5	Dominio de contenido es básico para asignatura	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	90	10
I6	Dominio de contenido es esperado del promedio de estudiantes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	95	5
I7	Aprendizaje de contenido es importante p/dominio de asignatura	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	90	10
I8	Delimitación del contenido alineada y derivada de definición	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I9	Habilidades y contenidos delimitados representan lo esencial	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95	5
I10	Estrategia ev. adecuada p/evaluar contenidos delimitados	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	90	10
I11	Estrategia ev. adecuada p/evaluar aprendizajes esperados	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	90	10
I12	Estrategia evaluativa semejante a como se enseña en aula	✓	✓	✓	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	85	15
I13	Ítems hijos reflejan uso de conocimiento adquirido	×	×	×	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	75	25
I14	Base del reactivo clara y suficiente para emitir respuesta	×	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	85	15
I15	Instrucciones adicionales claras y suficientes p/emitir respuesta	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---
I16	Reglas p/generar ítems hijos responden a estrategia evaluativa	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	85	15
I17	Textos auxiliares apropiados	---	✓	✓	✓	✓	---	---	✓	---	---	---	---	---	---	---	---	---	---	×	✓	85.71	14.29
I18	Gráficos e imágenes apropiados	---	---	---	---	---	---	---	✓	✓	✓	✓	✓	✓	✓	---	---	✓	---	✓	✓	100	0
I19	Banco de información corresponde al contenido seleccionado	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95	5
I20	Tipo de ejecución simple y facilita evaluación del contenido	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I21	Ítems hijos representan contenido delimitado	✓	✓	×	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	×	80	20
I22	Ítems hijos sin errores de redacción	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	90	10
I23	Ítems hijos redactados con palabras de uso común de estudiantes	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I24	Ítems hijos sin pistas	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I25	Ítems hijos con nivel de dificultad apropiado al grado escolar	✓	✓	×	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	×	75	25
I26	Ítems hijos sin sesgo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I27	Banco de información e Ítems hijos libres de otro tipo de errores	✓	✓	✓	×	✓	×	✓	×	×	×	✓	✓	×	✓	×	×	✓	×	×	×	45	55
	% Validado	91.67	96	60	84	88	79.17	100	92.31	84	96	96	100	96	100	95.83	66.67	100	95.83	92.31	76.92		
	% No validado	8.33	4	40	16	12	20.83	0	7.69	16	4	4	0	4	0	4.17	33.33	0	4.17	7.69	23.08		

Nota: Los símbolos en las columnas representan los siguientes aspectos: (✓) = Indicador validado, (×) = Indicador no validado, (---) = No aplica.

El formato de evaluación contempló aspectos comunes a la mayoría de los modelos de ítems seleccionados del Excoba. Sin embargo, hubo excepciones en las que algunos indicadores no fueron tomados en cuenta para su evaluación, debido a la naturaleza de los modelos de ítems. Cuando se presentó esta situación, la ausencia de respuesta se registró con una marca de tres guiones ortográficos (---), para señalar que la evaluación de dicho indicador no aplica. Como ejemplo se puede mencionar el caso del indicador número 15, que explora si las instrucciones adicionales a la base del reactivo son claras y suficientes para que el estudiante emita su respuesta. En este caso no fue necesario evaluar, ya que a diferencia de las otras áreas del Excoba (Español, Química e Historia), los ítems hijos de Matemáticas no utilizan instrucciones adicionales a la base del reactivo.

En las columnas de la Tabla 4.5 se observan las opiniones de los expertos en cada uno de los 20 modelos de ítems. En los renglones se registraron las opiniones respecto a los 27 indicadores evaluados. Cabe destacar que en la parte inferior se muestran, en términos de porcentajes, los indicadores que fueron y no fueron validados por los expertos en cada uno de los modelos de ítems, y en la columna del extremo derecho, los porcentajes de ítems validados en cada uno de los 27 indicadores.

Análisis de resultados

Este apartado, como el de las demás asignaturas, está dividido en tres secciones. En la primera de ellas se analiza el conjunto de modelos de ítems que conforman el área de Matemáticas. Se describen las opiniones de los expertos en cada uno de los 27 indicadores que integran el formato de evaluación. En la segunda, se describen de manera individual, las características de los modelos de ítems que presentaron cuando menos 20% de los indicadores no validados. El apartado concluye con la tercera sección, donde se presentan

una serie de observaciones y sugerencias realizadas por los expertos con la finalidad de mejorar el área evaluada.

Resultados por problemas detectados en el conjunto de modelos de ítems

La Figura 4.2 muestra, de manera gráfica, una síntesis de los resultados, según el porcentaje de los modelos de ítems que no fueron validados por los expertos, en los distintos indicadores evaluados.

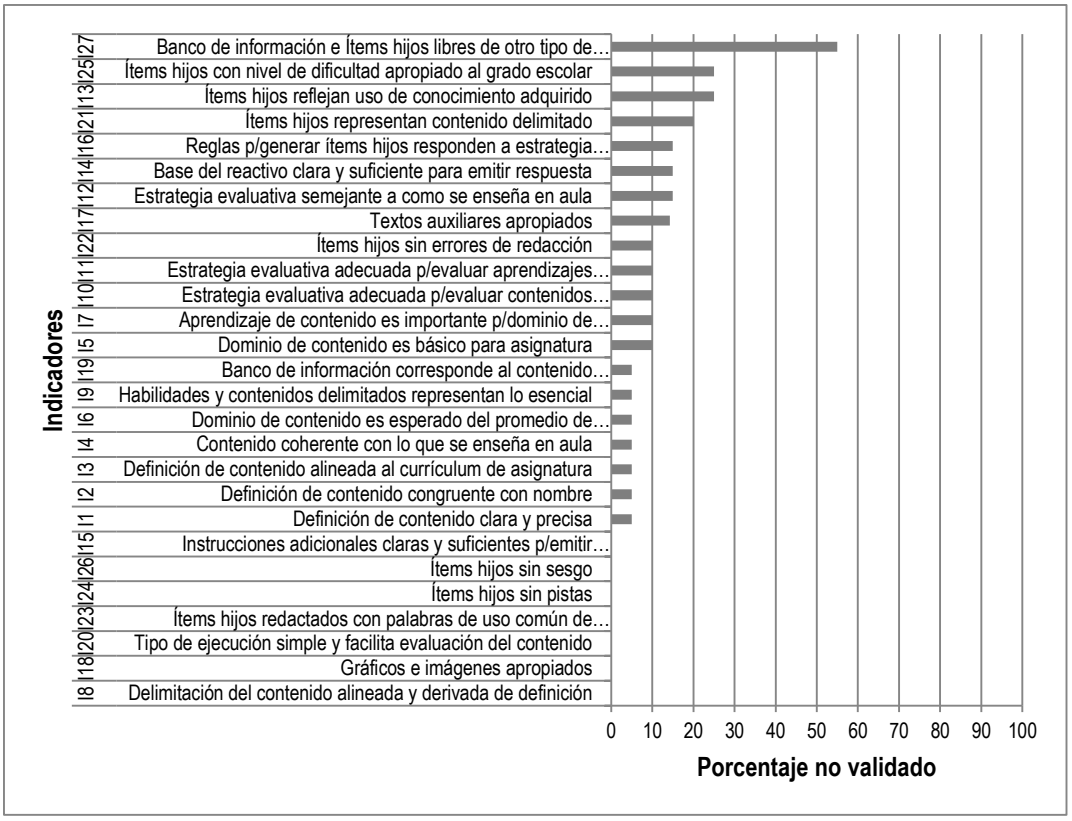


Figura 4.2. Porcentaje de modelos de ítems de Matemáticas no validados por los expertos en los indicadores

Se observa que hubo cuatro indicadores que presentaron un porcentaje de modelos de ítems no validados, mayor a 20%, lo que revela los problemas del área de Matemáticas que se presentan con mayor frecuencia.

En primer lugar se encuentra el indicador 27, cuyo criterio no fue cubierto por 55% de los modelos de ítems, y que evalúa aspectos relacionados con los ítems hijos y los bancos de información. Esto significa que en 11 de los 20 modelos de ítems, los expertos identificaron problemas específicos, tales como: elementos dentro del banco de información no viables para la producción de ítems, errores en las respuestas etiquetadas como correctas, inconsistencias en los niveles de dificultad entre familias de ítems pertenecientes a un mismo modelo de ítems, y falta de claridad en el tipo de respuesta que se le solicita al estudiante (fracciones o decimales, cantidad de decimales que puede escribir, entre otros).

En segundo término se encuentran los indicadores 13 y 25, que valoran la dificultad y pertinencia de los ítems hijos. El indicador 13 evalúa si los ítems hijos requieren que, al responder, el estudiante demuestre que sabe utilizar el conocimiento que se le enseñó y adquirió a lo largo del proceso de enseñanza-aprendizaje, por el que ha pasado. En este indicador, los expertos expresaron que 25% de los modelos de ítems no cumplían el criterio. Ante ello, mencionaron que esto se debe a que generan reactivos que exploran principalmente la memorización de procedimientos y reglas, más que su razonamiento y aplicación, lo cual no necesariamente implica que las respuestas del estudiante demuestran que sabe utilizar el conocimiento adquirido.

Por otro lado el indicador 25 en el que los expertos evaluaron si los ítems hijos cuentan con un nivel de dificultad apropiado para el grado escolar en el que se encuentran los estudiantes, el resultado fue que un 25% de modelos de ítems no fueron validados, es decir,

cinco modelos presentaron alguna de las siguientes situaciones: producen reactivos que son más difíciles de lo que el estudiante puede resolver; evalúan el contenido de una manera muy simple, y requieren que se eleve el grado de dificultad para que se adapten al grado académico de los examinados; generan ítems hijos con niveles de dificultad variable, lo que resulta en una situación de desventaja para algunos examinados.

Por último, el indicador 21, que evalúa si los ítems hijos representan adecuadamente el segmento del contenido que fue seleccionado del currículum de Matemáticas y delimitado con la finalidad de evaluar los aspectos más esenciales del mismo, no fue validado en 20% de los modelos de ítems. Esto indica que en cuatro modelos los expertos detectaron que no existe correspondencia o hay inconsistencias entre el segmento de contenido que fue elegido para ser evaluado, y los ítems hijos generados mediante el GAI Excoba.

Asimismo, se aprecia otro grupo de cuatro indicadores que señalan que de 14% a 15% de los modelos de ítems presentan problemas con la estrategia evaluativa y la base del reactivo. En el caso de los indicadores 12, 14 y 16, no fueron validados en 15% de los modelos de ítems, señalando dos tipos de problemas: el primero de ellos implica que existen modelos de ítems en los que la estrategia evaluativa utilizada para generar los ítems hijos, se aleja de la forma en que se enseña en el aula o no cumple las reglas establecidas para la generación de los reactivos. El segundo problema es la insuficiencia de claridad en la base del reactivo.

Por otro lado, el indicador 17 refleja que entre los modelos de ítems que generan reactivos en los que se utilizan textos auxiliares, 14% no obtuvieron la validación por falta de pertinencia. También se puede observar un grupo de cinco indicadores (5, 7, 10, 11 y 22) en los que los expertos determinaron que 10% de los modelos de ítems no cumplen con la

característica mencionada en ellos. Los problemas detectados en este bloque se relacionan con el hecho de que algunos contenidos no son indispensables para dominar la asignatura, por lo cual evaluarlos no resulta del todo relevante.

Estos indicadores también señalan que la estrategia evaluativa utilizada en estos modelos de ítems no es la más adecuada para evaluar los aprendizajes esperados del estudiante ni los contenidos que fueron seleccionados del currículum y delimitados para tal efecto. De igual manera, los expertos señalaron que los modelos de ítems de este último grupo contienen errores de redacción o hacen un uso inadecuado de algunos términos, lo cual podría confundir a los examinados. En la misma gráfica se puede ver la presencia de un grupo de siete indicadores (1, 2, 3, 4, 6, 9 y 19), que no fueron validados por los expertos en 5% de los modelos de ítems. Dichos indicadores reflejan problemas en los contenidos y en el banco de información utilizado para generar los ítems hijos.

En cuanto a los contenidos, los expertos señalaron una falta de claridad en su definición, así como errores u omisiones de alineación curricular, delimitación y correspondencia entre las diferentes secciones de los modelos. Por otra parte, en lo relacionado al banco de elementos cadena e integrales, los expertos detectaron la inclusión de algunos elementos que generan ítems en los que se evalúan contenidos distintos a los que se especifican en el modelo de ítems.

Hubo seis indicadores (8, 18, 20, 23, 24 y 26) en los que los expertos consideraron que 100% de los modelos de ítem los cubrían. Señalaron, con ello, que se cumplieron satisfactoriamente los criterios de calidad evaluados. Dichos indicadores exploraron aspectos como la alineación de la delimitación del contenido con su definición (indicador 8); la pertinencia y características de los gráficos e imágenes utilizados en los ítems, tales como su diseño, nitidez, uso de colores, nivel de complejidad, entre otras (indicador 18); la

simplicidad de las acciones de arrastre, selección y/o escritura que el estudiante debe realizar para responder los reactivos hijos (indicador 20); el uso apropiado del lenguaje y la familiaridad que el estudiante tiene con el estilo de redacción de los ítems hijos (indicador 23); la ausencia de pistas en la redacción de los ítems hijos, así como en sus elementos auxiliares (indicador 24), y la ausencia de sesgo en los ítems hijos por razones de edad, sexo, región, clase social o raza (indicador 26).

Por último, los expertos determinaron que el indicador 15 no era aplicable a ninguno de los modelos de ítems, ya que evalúa las características de las instrucciones adicionales a la base del reactivo y tal como se mencionó anteriormente, ninguno de los modelos de ítems del área de Matemáticas genera reactivos que requieran de ellas. Por lo tanto no se puede hablar de un indicador no validado, sino que uno que fue omitido por los expertos.

Resultados por problemas detectados en los modelos de ítems en lo individual

La Figura 4.3 muestra, de manera gráfica, una síntesis de los resultados según el porcentaje de indicadores que no resultaron válidos en cada uno de los 20 modelos de ítems evaluados. Es de especial interés analizar los modelos de ítems que presentaron mayor cantidad de indicadores no validados (problemas), por lo que se examinan a detalle, a continuación, aquellos que obtuvieron un 20% o más de indicadores no validados: MAT03, MAT16, MAT20 y MAT06.

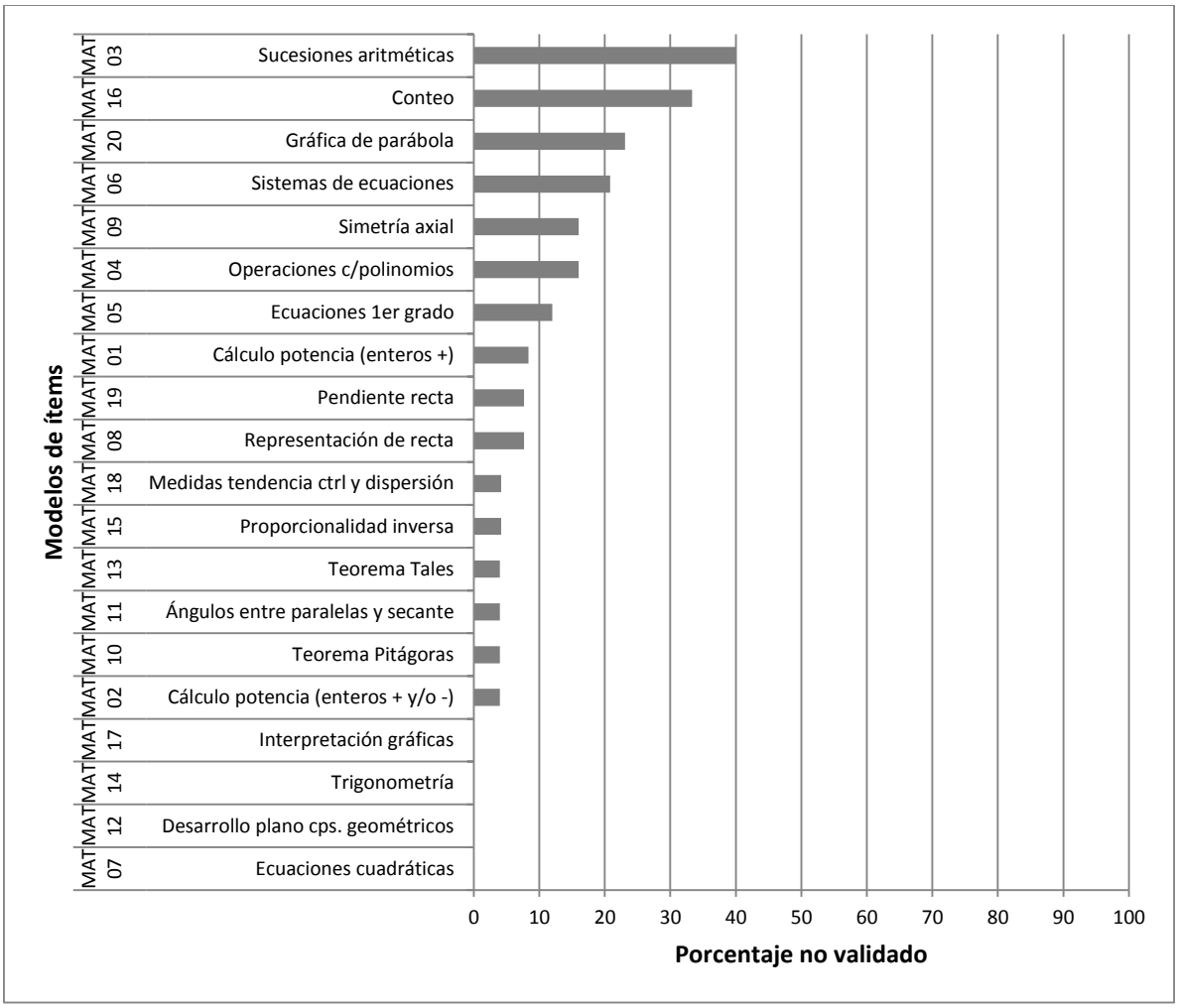


Figura 4.3. Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Matemáticas.

El modelo de ítems que presentó mayor cantidad de problemas fue MAT03 (40% de indicadores no validados), que evalúa si el estudiante es capaz de identificar las reglas generales que se necesitan para realizar cálculos dentro de una sucesión numérica, utilizando lenguaje algebraico. En segundo término se encuentra MAT16 (33.33% de indicadores no validados), que evalúa si el estudiante cuenta con conocimientos de las estrategias de ordenamiento de elementos que pertenecen a un conjunto, de tal manera

que pueda solucionar problemas de conteo. En tercer lugar se ubica MAT20 (23.08% de indicadores no validados), que explora si el estudiante es capaz de identificar la forma que tiene la gráfica de una parábola, y asociarla con su ecuación. Por último se encuentra el modelo de ítems MAT06 (20.83% de indicadores no validados), cuyos ítems hijos evalúan si el estudiante cuenta con las nociones algebraicas necesarias para solucionar un sistema de dos ecuaciones lineales. A continuación se describe con detalle el tipo de problemas que presentó cada uno de los cuatro modelos de ítems mencionados.

MAT03. Sucesiones aritméticas

Este modelo de ítems genera ítems hijos que exploran el conocimiento del tema *sucesiones aritméticas*. Solicita al estudiante que determine y escriba en su forma algebraica, la regla que subyace al cálculo de los términos de una sucesión numérica dada. En la Figura 4.4 se muestra un ítem hijo similar a los que se generan en el Excoba, en el cual se presenta una sucesión de números, y se solicita al estudiante que escriba cuál es la regla que se utiliza para generarla.

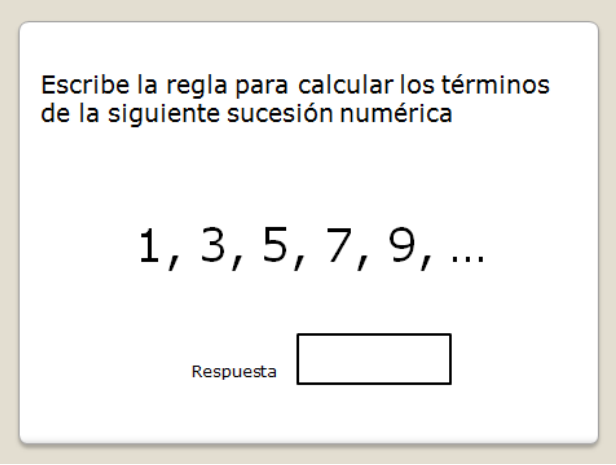


Figura 4.4. Ejemplo de ítem hijo del modelo MAT03: sucesiones aritméticas.

Para resolver este ítem, el estudiante debe descubrir que existe un orden en el conjunto de números, y que éste responde a un patrón. En el ejemplo, dicho patrón corresponde a que la diferencia entre cualquier término y el anterior es siempre 2, de tal manera que si el estudiante está familiarizado con la fórmula básica de las sucesiones: a*n* + b, donde a y b son constantes, y n es el número del término deseado, llegará a la respuesta correcta sustituyendo los valores en la fórmula.

Siguiendo el ejemplo, si la diferencia entre términos es 2, el término general sería 2n + b. Para encontrar el valor de b, se puede utilizar el primer término de la sucesión, donde n = 1. Así 2(1)+b=1 por ende, b= (-1). De esta manera, el estudiante obtendrá y escribirá la respuesta: 2n-1.

Respecta al proceso de validación, la Figura 4.5 muestra los indicadores en los que determinaron los expertos que el modelo de ítems MAT03 tenía problemas. Esto fue realizado mediante una discusión grupal en la que el dictamen final se emitió mediante la opinión consensuada de los presentes. Para ello, hicieron uso del formato de validación con 27 indicadores de calidad técnica.

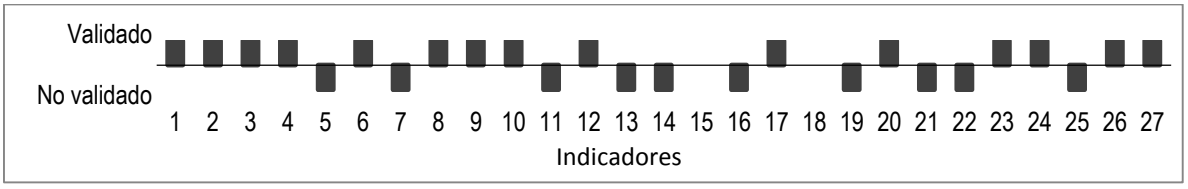


Figura 4.5. Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT03.

Entre los problemas que presenta este modelo de ítems, los expertos mencionaron que la evaluación del tema *sucesiones aritméticas* no es un contenido que se considere esencial para dominar la asignatura de Matemáticas, ya que esto depende de que el estudiante tenga conocimientos y buen manejo del lenguaje algebraico.

En opinión de los expertos (indicadores 5 y 7), es mucho más importante evaluar si el estudiante cuenta con conocimientos de este tipo de lenguaje, ya que es básico para el dominio y la adquisición de la mayoría de los contenidos de la asignatura, incluyendo el tema de sucesiones aritméticas, mas no viceversa.

Respecto a la estrategia evaluativa utilizada (indicadores 11, 13, 16, 19, 21 y 25), los expertos encontraron dos problemas, ya que consideraron que no se están cumpliendo satisfactoriamente las reglas estipuladas para generar los ítems hijos. El primero de ellos es que el modelo de ítems planteó desde el inicio una regla en la que únicamente se utilizarían números positivos para generar las sucesiones, pero al no cumplirse se generan ítems hijos que utilizan números negativos. Esto provoca que requieran la ejecución de una mayor cantidad de pasos, por ende, estrategias de razonamiento más complejo para su solución.

El segundo problema detectado por los expertos fue el incumplimiento de otra regla en la generación de los ítems hijos, la cual establece los parámetros para sustituir la variable *n* dentro de la fórmula a*n* + b utilizada para producirlos. A pesar de que cuando se delimitó el contenido que se evaluaría del tema se estipuló claramente el rango de los valores de la variable, el generador produce ítems hijos con valores que difieren de dicha delimitación, por lo tanto, presentan distintos grados de dificultad para los estudiantes.

Por otro lado, en lo que respecta a la redacción de la base del reactivo (indicadores 14 y 22), los expertos encontraron dos tipos de problemas: el primero de ellos es que la presencia de tecnicismos mal empleados, lo cual podría confundir a los estudiantes, generando respuestas erróneas que no se deban a la falta de conocimiento del tema, sino a la confusión originada por la redacción de la base del reactivo. Además consideraron que la redacción se aleja de la forma en que los estudiantes aprenden en los libros de texto, ya que los ejercicios que revisan en ellos utilizan un lenguaje coloquial y sencillo, mientras que

los ítems hijos que produce el modelo de ítems, contienen un lenguaje más formal y no tan común para los estudiantes.

MAT16. Conteo

En este modelo de ítems, los ítems hijos generados evalúan si el estudiante cuenta con conocimientos de las estrategias de ordenamiento de elementos que pertenecen a un conjunto, de tal manera que pueda solucionar problemas de conteo. Un ejemplo de este tipo de ítems se muestra en la Figura 4.6, en donde se plantea un problema en el que se solicita al estudiante que calcule y escriba la cantidad de arreglos distintos que se pueden hacer con una serie de elementos dados.

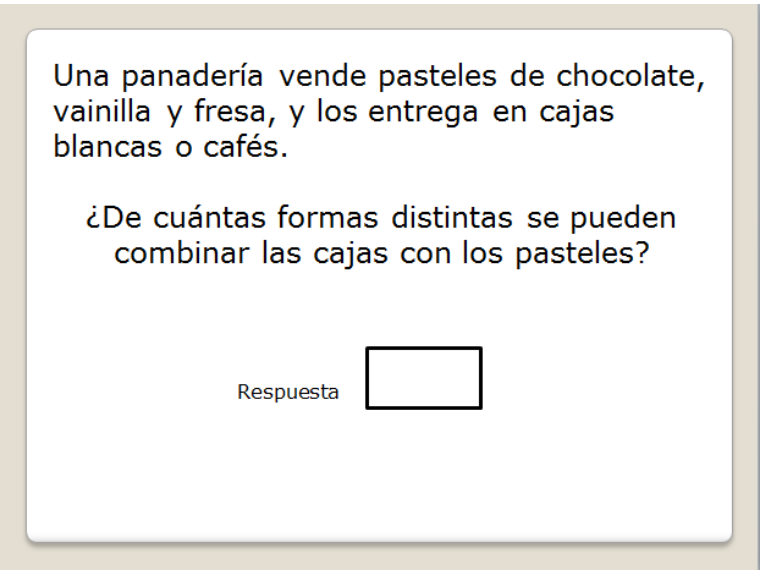


Figura 4.6. Ejemplo de ítem hijo del modelo MAT16: conteo.

El estudiante debe determinar un procedimiento sistemático de enumeración y llegar a la solución. En el caso que se ejemplifica, debe mostrar todas las posibilidades que tiene para combinar los sabores de los pasteles con el color de las cajas. Así, si tiene dos colores de cajas y para cada uno tiene tres posibilidades con los pasteles, puede llegar al resultado

multiplicando la cantidad de posibilidades de la primera característica (color de caja), por la cantidad de posibilidades de la segunda (sabor del pastel); de tal forma que tendrá seis combinaciones distintas: caja blanca con pastel de chocolate (1), vainilla (2) o fresa (3), y caja café con pastel de chocolate (4), vainilla (5) o fresa (6).

La figura 4.7 muestra las respuestas consensuadas por los expertos, a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

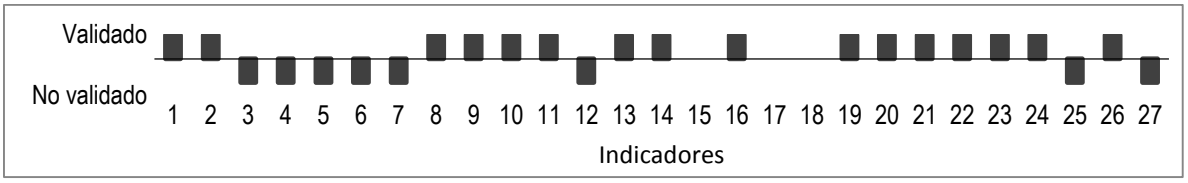


Figura 4.7. Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT16.

Dentro de los problemas que presenta este modelo de ítems, destacan aquellos relacionados con la alineación y pertinencia curricular del contenido que evalúa (indicadores 3 al 7). Se trata de un tema que los expertos no consideran como un pilar para la adquisición de otros conocimientos y competencias en el bachillerato. Mencionaron que aunque es un contenido que se enseña en secundaria, debido a las recientes modificaciones que se han realizado en el plan de estudios, ha perdido importancia y actualmente se ha convertido en un contenido aislado dentro del currículum.

En lo concerniente a la estrategia evaluativa utilizada para generar los ítems hijos, así como a las características de estos (indicadores 12, 25 y 27), los expertos mencionaron que el grado de dificultad de la mayoría de los ítems hijos que utiliza el Excoba es mayor al de los ejercicios que se presentan al estudiante en los libros de texto. En contraparte, existen algunos elementos en el banco de información que, por su bajo nivel de dificultad, producen ítems hijos que ponen en ventaja a algunos examinados, respecto a otros.

MAT20. Gráfica de una parábola

En este modelo de ítems, se explora si el estudiante es capaz de identificar la forma que tiene la gráfica de una parábola, y si puede asociarla con su ecuación. Contiene dos formas de generar ítems hijos: una en la que se muestra la gráfica de la parábola, y el estudiante determina y escribe la ecuación que le corresponde; otra donde se muestra la ecuación, y el estudiante debe escribir las coordenadas de ubicación de la parábola en el plano cartesiano y elegir la gráfica que la representa.

En la Figura 4.8 se observa un ejemplo de ambos tipos de ítems hijos. En la imagen de la izquierda, lo que debe hacer el estudiante es emplear la fórmula de la ecuación de una parábola para sustituir los valores de cada coeficiente. Para ello, debe utilizar las coordenadas dadas por el plano cartesiano (2, 7), de tal manera que al sustituir los coeficientes en la fórmula, pueda obtener la respuesta $y= 2x^2 – 8x + 1$.

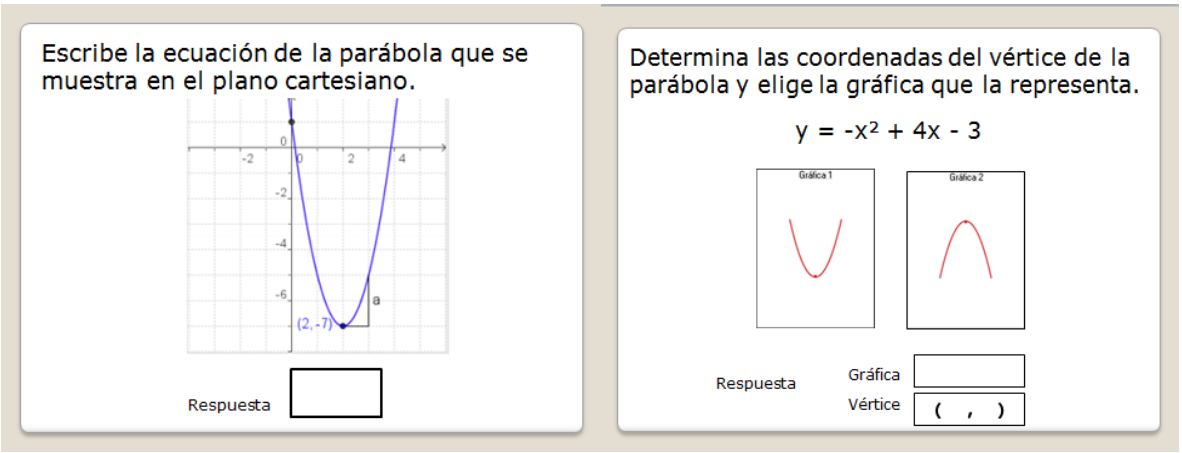


Figura 4.8. Ejemplo de ítem hijo del modelo MAT20: gráfica de una parábola. La imagen muestra ambos tipos de ítems que se pueden generar.

En la imagen de la derecha, se muestra al estudiante una ecuación y la imagen de dos parábolas. El estudiante debe determinar las coordenadas del vértice y los puntos de corte

con el eje x. Para elegir la gráfica, basta con que conozca que cuando un coeficiente es negativo, la parábola tendrá su abertura hacia abajo.

La figura 4.9 muestra las respuestas consensuadas por el panel de expertos por cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

Figura 4.9. Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT20.

Dentro de los problemas que presenta este modelo de ítems, mencionaron que la estrategia evaluativa no es congruente con lo que se propuso evaluar al delimitar el contenido, ya que en ella se menciona el uso de funciones que no se utilizan para generar los ítems hijos (indicadores 10, 11 y 16). En cuanto a la representación que hacen los ítems hijos del contenido delimitado (indicadores 21 y 27), encontraron que no existe correspondencia alguna y sugieren modificar la delimitación para que sea congruente con la estrategia evaluativa y los ítems hijos que de ella deriven. Por último, encontraron que hay algunos ítems hijos que tienen un grado de dificultad mayor al estipulado en la estrategia evaluativa (indicador 25).

MAT06. Sistemas de ecuaciones

En este modelo de ítems se explora si el estudiante cuenta con las nociones algebraicas necesarias para solucionar un sistema de dos ecuaciones lineales. En la Figura 4.10 se muestra un ejemplo en el que se solicita al estudiante que realice un cálculo para obtener el valor de la incógnita *x*. Para resolverlo debe saber que la forma para encontrar

dicho valor implica la eliminación de incógnitas, despeje de ecuaciones y correcto uso de signos.

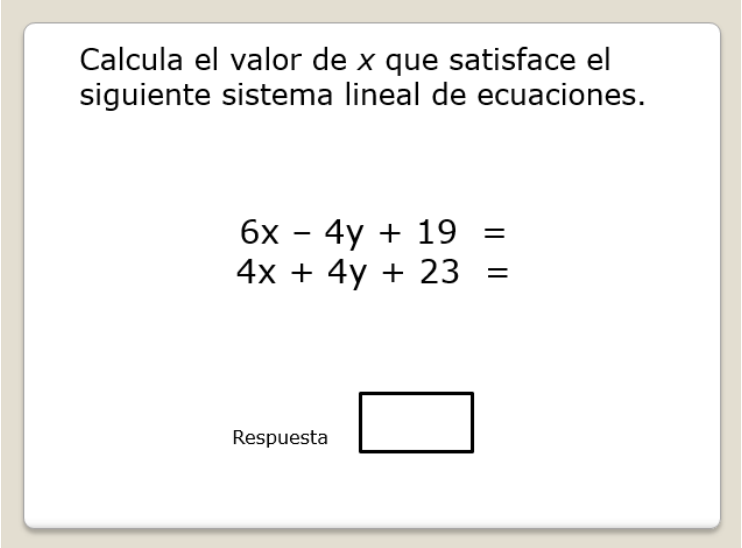


Figura 4.10. Ejemplo de ítem hijo del modelo MAT06: Sistemas de ecuaciones.

En el ejemplo mostrado en la Figura 4.10 se deben eliminar las incógnitas -4y y +4y debido a sus signos, permaneciendo las ecuaciones 6x+19 y 4x+23. Enseguida se suman los valores numéricos, para posteriormente despejar la incógnita x. Este proceso indica que el resultado final es que *x* tiene un valor de -4.2.

La Figura 4.11 muestra las respuestas consensuadas por los expertos, a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems, del modelo de ítems MAT06.

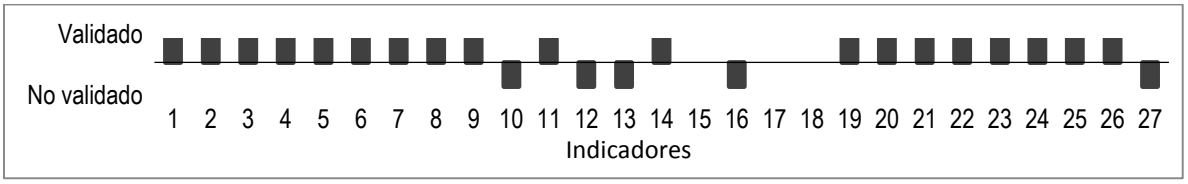


Figura 4.11. Opinión de los expertos en los 27 indicadores validados del modelo de ítems MAT06.

Dentro de los problemas que presenta este modelo de ítems en lo relacionado a la estrategia evaluativa (indicadores 10, 12 y 13), los expertos mencionaron que no es la más adecuada para el propósito que se persigue. Esta solicita al estudiante que proporcione el valor de una de dos incógnitas, mientras que en los libros de texto se enseña a los estudiantes a encontrar ambas.

Por otro lado, también indicaron que si lo que se busca evaluar es si el estudiante tiene la capacidad de demostrar que sabe utilizar el conocimiento adquirido en el aula, se deberá replantear la estrategia, ya que actualmente los ítems hijos exploran la memorización de un método que lleva a la solución.

Finalmente, en cuanto al banco de elementos que se utiliza para generar los ítems hijos (indicadores 16 y 27), los evaluadores encontraron distintos niveles de dificultad, así como falta de correspondencia con los valores que se establecieron en la estrategia evaluativa.

Resultados conjuntos

Además de los análisis realizados en los apartados anteriores, se contrastó la información derivada del trabajo con el panel de expertos, de tal manera que se tomaron los cuatro modelos de ítems que tuvieron por lo menos 20% de los indicadores de calidad no validados, y los 4 indicadores que no fueron validados por los expertos en al menos 20% de los modelos de ítems de toda la asignatura. De esta manera, se detectó la presencia de regularidades en el tipo de problemáticas que presentan. La Tabla 4.6 muestra esta información.

Tabla 4.6.
Modelos de ítems de Matemáticas y regularidades que presentaron

No.	Indicador	Modelos de ítems				Frecuencia relativa
		MAT03 Sucesiones aritméticas	MAT06 Sistemas de ecuaciones	MAT16 Conteo	MAT20 Gráfica de una parábola	
25	Ítems hijos con nivel de dificultad adecuado, y apropiados para el grado escolar del contenido evaluado	×	---	×	×	3 de 4
27	Banco de información y/o ítems hijos libres de otro tipo de errores	---	×	×	×	
13	Ítems hijos requieren que al responder, el estudiante demuestre que sabe utilizar el conocimiento adquirido dentro del aula	×	×	---	---	2 de 4
21	Los ítems hijos representan adecuadamente el contenido delimitado	×	---	---	×	

Nota: Se muestra únicamente la información de los modelos de ítems que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Matemáticas. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

Se observa que existen coincidencias entre modelos de ítems, respecto a los tipos de problemas presentes en ellos. Por ejemplo, los tres modelos de ítems, MAT 03 (sucesiones aritméticas), MAT16 (conteo), y MAT20 (gráfica de una parábola), además de tener al menos 20% de los indicadores de calidad no validados, coincidieron en que tienen problemas con la forma en que se generan sus ítems hijos en cuanto a sus niveles de dificultad y correspondencia con el grado escolar al que pertenece el contenido evaluado (indicador 25).

De igual manera, los modelos de ítems MAT06 (sistemas de ecuaciones), MAT16 (conteo) y MAT20 (gráfica de una parábola), no fueron validados en el indicador 27, el cual explora problemas específicos en los bancos de elementos y en los ítems hijos. Esto sugiere

que los elementos cadena e integrales que son utilizados para generar los ítems hijos, requieren ser revisados a detalle para determinar su pertinencia, grado de dificultad, redacción y diseño, entre otras cosas.

Así mismo, se detectó que en su construcción, dos modelos: MAT03 (sucesiones aritméticas) y MAT06 (sistemas de ecuaciones), coincidieron en que no fue validado el indicador 13, el cual sugiere que al responder no necesariamente se utilizan los conocimientos adquiridos en el aula.

Por último, también se observa que en los modelos de ítems MAT03 (sucesiones aritméticas) y MAT20 (gráfica de una parábola), no se validó el indicador 21, el cual menciona que en el caso de estos dos modelos, los ítems hijos que generan no representan de manera adecuada el contenido que fue delimitado del currículum.

Resultados según las sugerencias del panel de expertos para el mejoramiento del área

Adicionalmente a los problemas que fueron detectados por los expertos, hubo una serie de sugerencias que ellos aportaron para mejorar el área de Matemáticas del Excoba.

Entre ellas, se destaca su apreciación general respecto a la pertinencia de los contenidos seleccionados del currículum. Estimaron que 90% (18) de ellos representan lo que el estudiante debe dominar de la asignatura, pero que es necesario hacer una revisión de los contenidos que no consideran esenciales en la evaluación que realiza este instrumento, y así valorar la conveniencia de su sustitución por otros más esenciales. Los contenidos señalados corresponden a los modelos de ítems MAT03 (sucesiones aritméticas) y MAT16 (conteo).

Por otro lado, estimaron que aproximadamente 70% de los estudiantes responderían correctamente a los 20 reactivos hijos que produzca el Excoba en el área de Matemáticas,

ya que se debe considerar que existen variables que pueden afectar su aprendizaje y rendimiento en el examen, las cuales no son atribuibles al proceso de evaluación. Entre ellas mencionaron: la capacidad del estudiante para la retención de información, su falta de interés en los temas evaluados, y la falta de madurez emocional y cognitiva, entre otros.

Resultados según la relación entre los problemas detectados y los contenidos evaluados

Una vez que se identificó el eje temático curricular al que pertenece cada modelo de ítems, se procedió a realizar un análisis para detectar si existe algún tipo de relación entre el contenido curricular que evalúan los modelos de ítems que presentaron mayor cantidad de indicadores no validados ($\geq 20\%$) y el tipo de problema expresado en los indicadores que no fueron validados en 20% o más de los modelos de ítems. La Tabla 4.7 resume los hallazgos.

Tabla 4.7.
Modelos de ítems, ejes temáticos e indicadores no validados con mayor frecuencia en Matemáticas

Modelos de ítems	Contenido	Eje temático	No. de indicador			
			13	21	25	27
MAT03	Sucesiones aritméticas	Sentido numérico y pensamiento algebraico	×	×	×	--
MAT06	Sistemas de ecuaciones	Sentido numérico y pensamiento algebraico	×	--	--	×
MAT16	Conteo	Sentido numérico y pensamiento algebraico	--	--	×	×
MAT20	Gráfica de una parábola	Manejo de la información	--	×	×	×

Nota: Se muestra únicamente la información de los modelos de ítem que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Matemáticas. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

De los cuatro modelos de ítems que cumplieron con el criterio de 20% o más indicadores no validados por los expertos, se puede observar que tres corresponden al eje temático Sentido numérico y pensamiento algebraico (MAT03, MAT06 y MAT16); mientras el cuarto, al eje temático *Manejo de la información*. Esto indica que la mayoría de los modelos de ítems con un alto número de indicadores no validados (y por ende problemas),

generan ítems hijos en los que se requiere que el estudiante conozca y tenga un buen manejo del lenguaje aritmético y algebraico, de manera que pueda establecer una relación entre ellos y así resolver adecuadamente problemas matemáticos. También coinciden en que se requiere la habilidad para identificar, analizar y utilizar correctamente la información contenida en fuentes, como gráficas o tablas, con la finalidad de hacer una correcta interpretación de los datos.

Este tipo de problemática se relaciona con la dificultad, representatividad y pertinencia de los ítems hijos. Los expertos estimaron que no representan adecuadamente el contenido delimitado, no son del todo apropiados para el grado escolar de los estudiantes, y tampoco reflejan adecuadamente si el estudiante sabe utilizar el conocimiento adquirido. Expresaron que debido a la naturaleza de los elementos integrales utilizados para generar los ítems hijos, los niveles de dificultad varían. Esto pone en desventaja a algunos estudiantes cuyo dominio del conocimiento es menor al resto de los evaluados.

Lo anterior sugiere que sí existe relación entre el tipo de contenido al que pertenecen los modelos de ítems en el currículum y la naturaleza de la problemática señalada por los expertos. Esta información indica la necesidad de realizar una revisión detallada de estos modelos de ítems, en los siguientes aspectos: la delimitación del contenido y su congruencia con los elementos del generador de ítems; la cercanía conceptual de los elementos integrales en cuanto a sus niveles de dificultad, y la representatividad en cuanto al tipo y nivel de profundidad del conocimiento que están evaluando.

4.2.2. Modelos de ítems del área de Historia

Al igual que los resultados del área de Matemáticas, este apartado está dividido en dos secciones, correspondientes a las etapas 2 y 3 del MVCE. La primera sección (etapa 2)

describe las características de los modelos de ítems que conforman el área, de acuerdo con cuatro clasificaciones: 1) el tipo de contenido curricular que evalúa cada uno de ellos según el grado escolar, eje curricular y bloque temático al que pertenecen dentro del programa de estudios correspondiente; 2) la clasificación que tienen los tipos de modelos de ítems en el sistema informático GenerEx, que administra el examen; 3) el tipo de ejecución que debe realizar el estudiante para emitir su respuesta al utilizar la interfaz gráfica del examen (arrastre, selección, escritura o mixta), y 4) el tipo de conocimiento que exploran según los procesos intelectuales que se espera que utilice el estudiante al aprender los contenidos curriculares dentro de la asignatura (declarativo, procedimental, esquemático o estratégico).

La segunda sección (etapa 3 del MVCE) muestra el análisis de los resultados obtenidos del proceso de validación de los modelos de ítem, según las opiniones de los expertos en los indicadores contenidos en el formato de evaluación. En esta sección se presentan los resultados en dos niveles: 1) como conjunto de modelos de ítems; es decir, de acuerdo con los indicadores en donde se presentaron los problemas, y 2) como modelos de ítems individuales; es decir, de acuerdo con los problemas que presentaron los modelos en dichos indicadores. De manera complementaria, se presentan las sugerencias realizadas por parte de los expertos, con la finalidad de mejorar cada uno de los modelos de ítems en aquellos indicadores de calidad donde se encontraron deficiencias o áreas de mejora.

Al final, se hace una síntesis de los resultados, con la finalidad de identificar las posibles causas de los problemas detectados, de acuerdo con el tipo de contenido que evalúa cada modelo de ítems, según su grado escolar, eje curricular y bloque temático al que pertenecen dentro del programa de estudios.

Etapa 2: Características de los modelos de ítems de Historia

Siguiendo la clasificación propuesta en este estudio, a continuación se presenta la distribución de los ocho modelos de ítems que conforman el área de Historia del GAI Excoba, según: 1) los contenidos curriculares, clasificados de acuerdo con el bloque temático al que pertenecen dentro del programa de estudios 2006; 2) las características informáticas de los modelos de ítems; 3) el tipo de ejecución que los modelos de ítems solicitan al estudiante que realice para responder al reactivo, y 4) los conocimientos evaluados, de acuerdo con la taxonomía propuesta por Ruiz-Primo en 2007.

Tipo de contenidos curriculares

Para la elaboración del GAI Excoba se utilizó el programa de estudios de Historia de educación secundaria del año 2006, el cual se imparte durante los primeros dos grados escolares (SEP, 2008). La Figura 4.12 muestra su estructura.

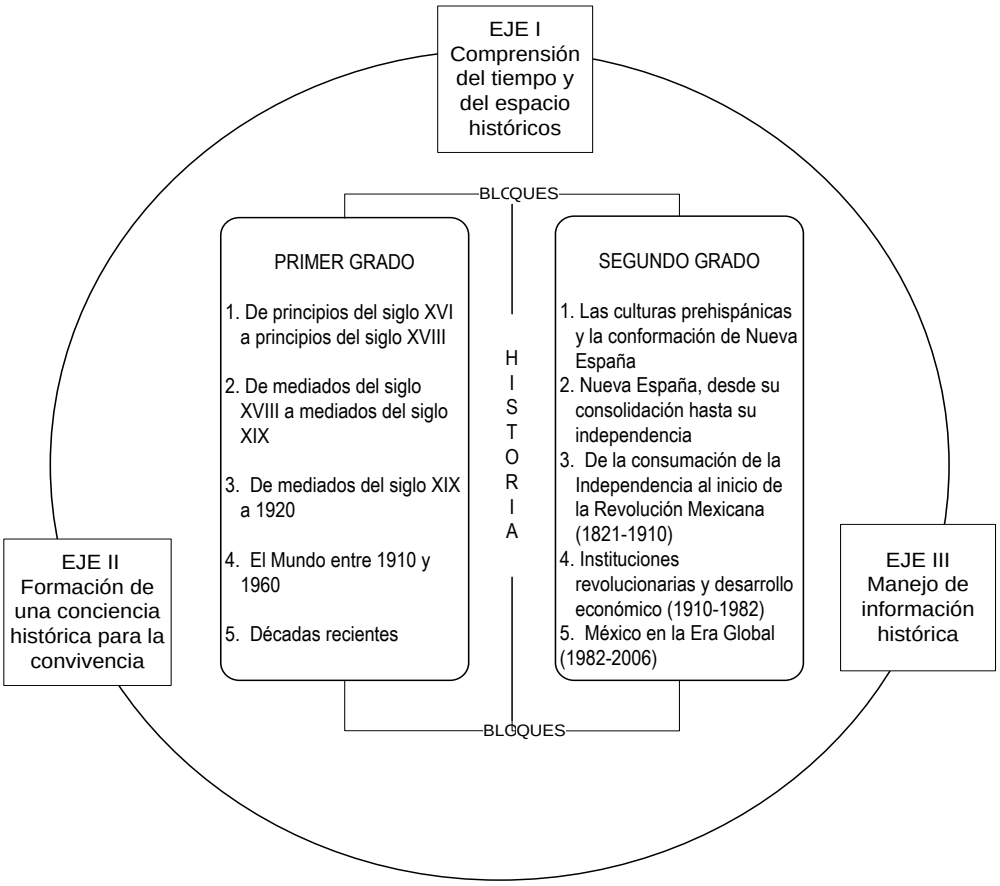


Figura 4.12. Estructura del programa de estudios de la asignatura de Historia.

Puede observarse que la manera en que se organiza el programa de estudios, obedece a tres grandes ejes curriculares, que orientan e integran de manera transversal los aprendizajes esperados del estudiante: 1) *comprensión del tiempo y espacio históricos*, 2) *manejo de información básica*, y 3) *formación de una conciencia histórica para la convivencia*. A su vez, todos los contenidos de la asignatura se organizan de manera cronológica en cinco bloques temáticos, que varían según se trate de la asignatura del primero o segundo año escolar.

La distribución de los ocho modelos de ítems aparece en la Tabla 4.8, en la que se observa que ninguno de los contenidos pertenece al mismo bloque temático. Esto se debe a que los contenidos se encuentran organizados cronológicamente e interrelacionados de manera secuencial, donde el aprendizaje de uno implica el conocimiento y manejo sistemático del que lo precede, de tal manera que el estudiante analice los cambios que se suscitaron en la historia y, con ello, comprenda y explique su sociedad. En cuanto a la distribución por grados escolares, de la totalidad de modelos de ítems, cinco (62.5%) pertenecen a la asignatura de Historia de primer grado, mientras que los tres restantes (32.5%) corresponden a Historia II.

Tabla 4.8.
Distribución de modelos de ítems de Historia por bloque temático

Modelos de ítems	Contenido	Bloque temático	Porcentaje
		HISTORIA I	
HIS06	Viajes de exploración, hegemonía europea y colonización	De principios del siglo XVI a principios del siglo XVIII	
HIS07	Revolución industrial y las Revoluciones Atlánticas	De mediados del siglo XVIII a mediados del siglo XIX	
HIS08	Países imperialistas	De mediados del siglo XIX a 1920	62.5
HIS09	Las grandes guerras en el mundo	El Mundo entre 1910 y 1960	
HIS10	Conflictos sociales, políticos, culturales y religiosos	Décadas recientes	
		HISTORIA II	
HIS11	Culturas prehispánicas	Culturas prehispánicas y conformación de Nueva España	
HIS12	Hechos y personajes. Conquista, Colonia e Indep.	Nueva España. Consolidación hasta independencia	32.5
HIS13	Hechos y personajes de la Revolución Mexicana	Instituciones revolucionarias y des. económico (1910-1982)	

Características del sistema informático GenerEx

En el área de Historia se utilizaron 2 de los 22 tipos diferentes de reactivos que conforman la estructura del sistema informático del GAI. La distribución que se presentó en los ocho modelos de ítems, se puede observar en la Tabla 4.9, la cual muestra que 87.5% (7) son del tipo *Elemento categoría*, mientras que 12.5% (1) corresponden al tipo *Elemento imagen*.

Tabla 4.9.
Clasificación y distribución de modelos de ítems de Historia en el sistema editor de reactivos

	Total	Porcentaje
Elemento categoría	7	87.5
Elemento imagen	1	12.5

El primer tipo de reactivos requiere una forma de programación en la que el sistema construye ítems hijos mediante la selección aleatoria de un número de categorías, junto con una cantidad preestablecida de elementos asociados a ellas, los cuales se muestran en la interfaz gráfica con la que interactúa el estudiante. Así, cuando el examinado responde, el sistema califica haciendo una comparación entre el lugar en donde fueron colocados los elementos, y las respuestas que se encuentran alimentadas en él.

La Figura 4.13 presenta un ejemplo de este tipo de reactivos. En la parte superior izquierda se observa que la interfaz muestra un ítem hijo generado aleatoriamente, que aún no ha sido respondido. En él se plantea la tarea de clasificar en tres categorías una serie de cinco elementos conceptuales, tomados aleatoriamente del banco de información, y que previamente fueron alimentados en el sistema informático. Las categorías se refieren a regiones geográficas en donde ocurrieron algunos acontecimientos históricos, y los elementos se refieren a los nombres de tales acontecimientos.

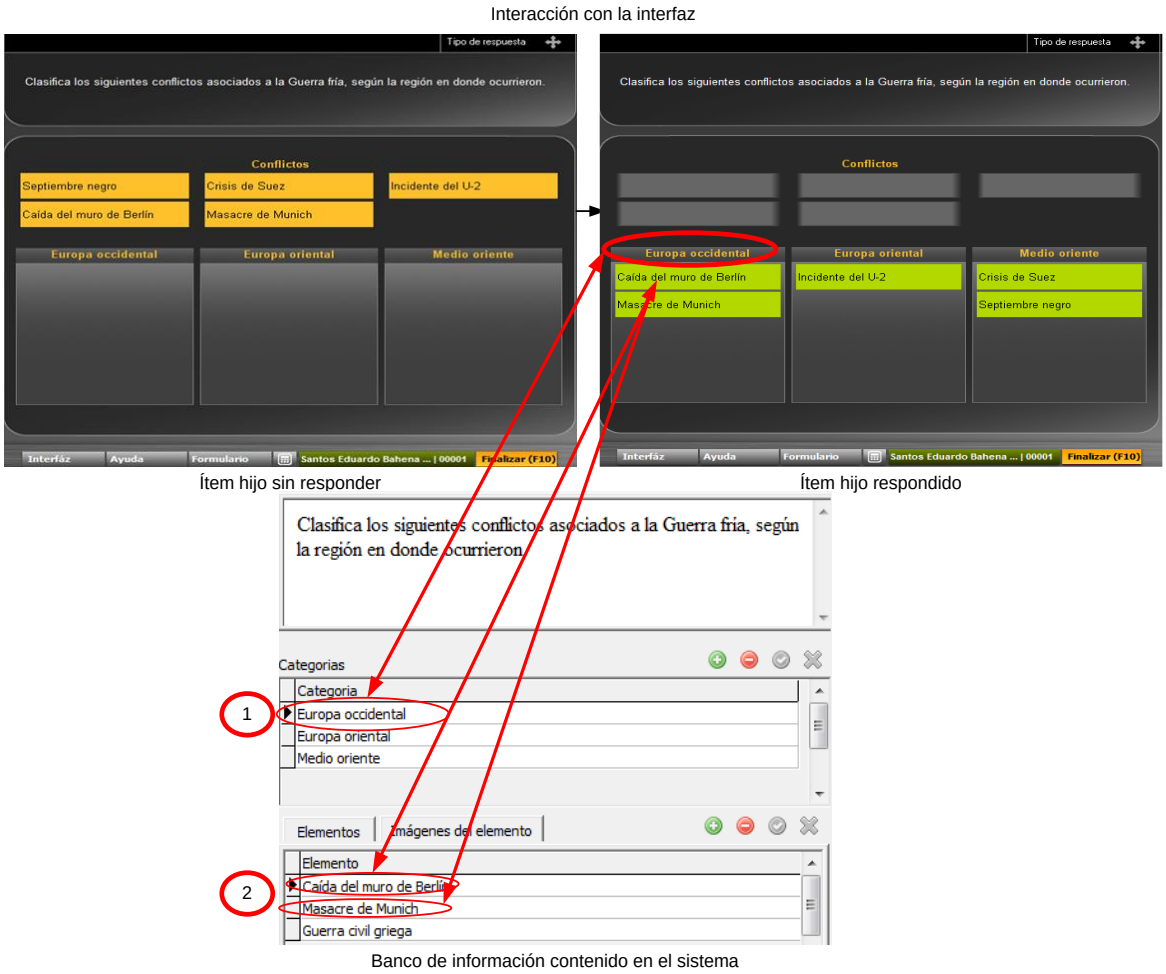


Figura 4.13. Ejemplo del procedimiento de calificación del sistema informático en reactivos de tipo *Elemento categoría*.

La sección superior derecha de la Figura 4.3 muestra el reactivo resuelto por el estudiante. Se observa que los cinco eventos históricos fueron *depositados* por el estudiante en las distintas categorías. En la parte inferior de la figura, es de especial importancia observar que tanto las categorías, como los elementos correspondientes a cada una de ellas son seleccionadas de un banco de información que previamente fue

alimentado al sistema. Este se programa para mostrar en la interfaz una cantidad determinada de elementos y categorías, así como la forma en la que se presentarán.

En este ejemplo, el sistema mediante códigos programáticos preestablecidos realizará dos acciones: 1) elegirá en forma aleatoria tantas categorías como se haya establecido, y 2) de cada categoría, tomará aleatoriamente una cantidad determinada de elementos, los cuales mostrará en la interfaz, y una vez que el estudiante emita su respuesta, los cotejará con lo que se encuentra capturado en el banco de información.

El segundo tipo de ítems que se utilizan para evaluar el área de Historia, corresponden al tipo llamado *Elemento imagen*, que representan 12.5% (uno) del total de los modelos de ítems de dicha sección del instrumento. Estos ítems, debido a sus características programáticas, utilizan una imagen auxiliar en la interfaz (p. ej. una línea de tiempo), en la cual se delimitan sectores mediante el establecimiento de coordenadas, con la finalidad de que se determinen los límites gráficos de la posición de las respuestas correctas e incorrectas. Al calificar, el sistema utilizará como criterio la ubicación y el orden de las respuestas del estudiante.

Al igual que en el tipo de ítem anterior, en la Figura 4.14 se observa una imagen dividida en tres secciones. En la parte superior izquierda se muestra un ítem hijo sin responder, que fue generado para que muestre en la interfaz una serie de cinco elementos conceptuales, así como la imagen de una línea de tiempo. La sección superior derecha de la Figura muestra el ítem respondido, y la sección inferior indica la forma en que el sistema califica las respuestas del estudiante.

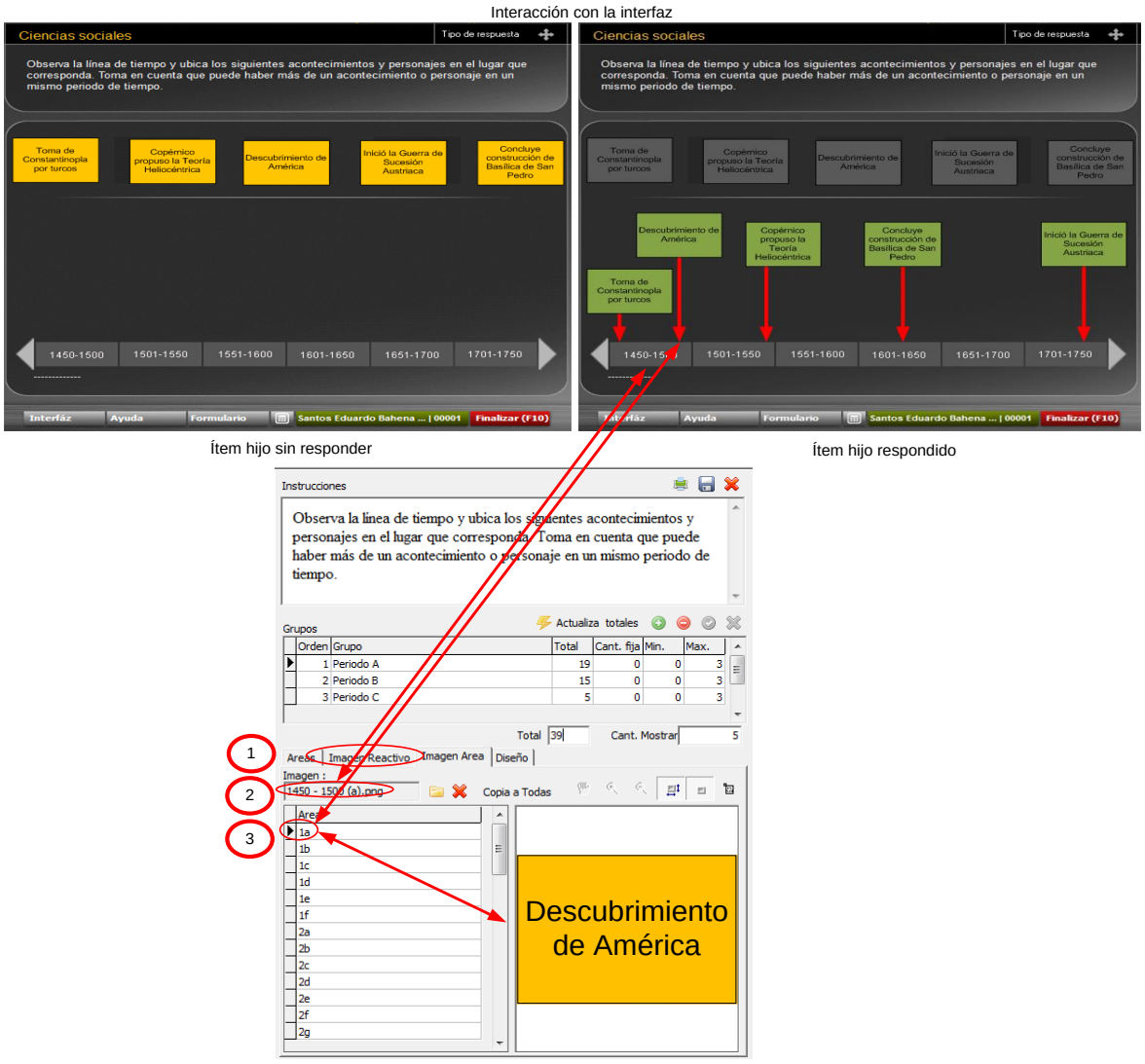


Figura 4.14. Ejemplo del procedimiento de calificación del sistema informático en reactivos de tipo *Elemento imagen*.

El procedimiento para generar el ítem hijo y su calificación implican tres aspectos, como se muestra en la sección inferior de la Figura 4.14:

1) Se selecciona aleatoriamente la imagen, que en el ejemplo es una línea de tiempo.

2) Se establecen los sectores del gráfico para dividirlo en secciones y se asocian elementos a cada uno de ellos, los cuales se seleccionan aleatoriamente según la cantidad que se haya establecido. En el ejemplo se dividió cada sector de la imagen en periodos de 50 años y se incluyó una lista de distintos acontecimientos históricos, de los cuales se seleccionaron aleatoriamente cinco.
3) Cuando el estudiante responde colocando los eventos en algún lugar de la línea de tiempo, el sistema coteja la respuesta con lo que se encuentra alimentado en el banco de información, de tal manera que pueda asignar un puntaje.

Tipo de ejecución solicitada

Los modelos de ítems que evalúan los contenidos de la asignatura de Historia, contemplaron el uso de estrategias evaluativas que generan un solo tipo de ejecución por parte del estudiante, cuando emite sus respuestas en los ítems hijos por *arrastre*. Esto implica que el estudiante haga uso del ratón (*mouse*) con la finalidad de seleccionar un elemento entre las opciones que se le presentan, y lo coloque en el lugar que considere adecuado.

Dicho lugar puede referirse a una categoría (p. ej., nombres de conflictos bélicos, causa o consecuencia de un evento histórico, etcétera) o a una sección específica de una imagen, tal como sucede cuando se colocan las opciones de respuesta en orden cronológico dentro de una línea de tiempo. La Figura 4.15 muestra dos tipos de ítems en los que el estudiante debe mover elementos y colocarlos dentro de diferentes categorías.

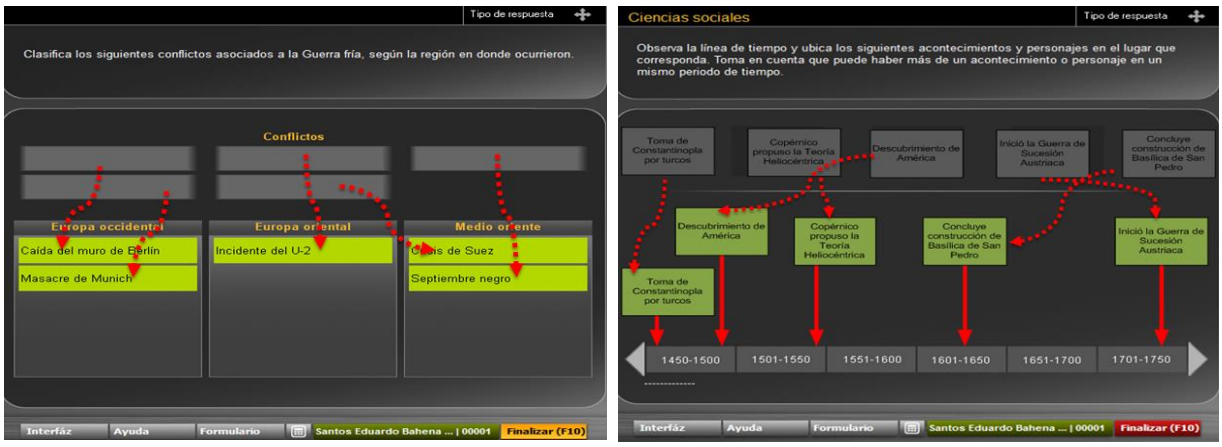


Figura 4.15. Ejemplos de dos distintos tipos de ítems hijos en los que se utiliza ejecución de *Arrastre* para emitir las respuestas.

Del lado izquierdo se muestra un ítem en el cual se clasificaron cinco eventos históricos distintos en tres categorías, que se refieren a las zonas geográficas en las que estos ocurrieron. Con las flechas rojas punteadas se puede observar la trayectoria que siguieron los movimientos del ratón cuando el estudiante lo utilizó para colocar cada elemento dentro de las tres categorías disponibles (Europa occidental, Europa oriental, y Medio oriente).

Del lado derecho de la Figura 4.15 se muestra un ítem en el que se presentan cinco acontecimientos históricos que deben moverse y colocarse, según su periodicidad, en una línea de tiempo dividida en cortes cronológicos con periodos de 50 años. Al igual que en el reactivo anterior, las flechas rojas punteadas indican cómo el estudiante utilizó el ratón para seleccionar cada elemento y moverlo, hasta colocarlo sobre la fecha en la que consideró que sucedió tal acontecimiento.

La ejecución que realiza el estudiante en ambos tipos de ítems es la misma e implica el uso del ratón para la selección, movimiento y colocación de elementos de la interfaz gráfica, dentro de repositorios llamados categorías.

Tipo de conocimientos evaluados

Debe recordarse que, para efectos de este estudio, se empleó una clasificación de cuatro tipos de conocimiento: declarativo, procedimental, esquemático y estratégico. En el área de Historia, 100% de los modelos de ítems exploran conocimientos de tipo esquemático, debido a que evalúan aspectos vinculados con la organización cronológica de acontecimientos históricos, y con la identificación y el establecimiento de relaciones entre elementos, como: lugares, personajes y sucesos históricos relevantes en la historia de México y el mundo.

Lo anterior implica que para responder a cualquier ítem hijo de esta área, el estudiante debe manejar la información de manera ordenada y sistemática, y utilizar modelos o esquemas mentales que le permitan acceder a los conocimientos aprendidos durante su formación académica. De esta manera, sus respuestas deben reflejar que conoce los motivos por los cuales ocurrieron los hechos históricos en un orden específico o bajo determinadas circunstancias.

Por ejemplo, en el modelo de ítems HIS06, que genera ítems hijos para evaluar el conocimiento del estudiante respecto a los acontecimientos más relevantes entre el siglo XV y el XVIII, se utiliza una línea de tiempo en la cual se deben colocar en orden cronológico los elementos que se presentan como opciones de respuesta, los cuales incluyen nombres de personajes y eventos históricos de gran relevancia.

Para resolver correctamente este tipo de ítems, el estudiante debe contar con un esquema mental muy claro respecto a la cronología de hechos. Debe asociarlos a aspectos

sociales, económicos, políticos, científicos y culturales, de tal manera que tengan un hilo conductor cuyo eje rector sea el tiempo. Al ordenar correctamente sus respuestas y colocarlas en una línea de tiempo, mostrará evidencias de que cuenta con las competencias básicas en esta área del examen.

Etapa 3: Análisis de los resultados del trabajo con el panel de expertos

Resultados generales

En total se validaron ocho modelos de ítems, para lo cual se utilizó un ítem hijo como muestra de cada modelo. Al igual que en el área de Matemáticas, los ítems que fueron seleccionados para presentar a los expertos corresponden a una versión fija del Excoba que fue utilizada para su pilotaje y el análisis de su estructura interna.

El proceso de evaluación que realizaron los miembros del panel de expertos proporcionó información útil en la identificación de los problemas que presentan los ocho modelos de ítems que conforman el área de Historia de secundaria del Excoba. También ayudó a detectar los aspectos específicos que requieren atenderse y modificarse para fortalecer el instrumento.

En el proceso de evaluación los expertos valoraron 27 indicadores de calidad en cada uno de los ocho modelos de ítems que conforman el área de Historia del Excoba. Los indicadores en los que se enfocaron fueron los relacionados con la estructura y correspondencia de las distintas secciones del modelo de ítems, las características del banco de información curricular, así como aspectos de diseño, funcionalidad y operatividad de los reactivos hijos que se generan mediante el GAI. En cada indicador consensuaron acuerdo o desacuerdo, y registraron sus observaciones.

En la Tabla 4.10 se muestra la información general obtenida como producto del proceso de evaluación de los modelos de ítems del área de Historia. Dicha tabla proporciona, de manera condensada, la información recopilada mediante las opiniones que los expertos emitieron respecto a factores de calidad y pertinencia de los contenidos que fueron seleccionados del currículum para ser evaluados y conformar dicha área.

Tabla 4.10.

Resultados globales del proceso de evaluación de panel de expertos del área de Historia del Excoba

	Indicador	HIS06	HIS07	HIS08	HIS09	HIS10	HIS11	HIS12	HIS13	% Validado (✓)	% No validado (×)
		Viajes de exploración, hegemonía europea y colonización	Revolución industrial y las Revoluciones Atlánticas	Países imperialistas	Las grandes guerras en el mundo	Conflictos sociales, políticos, culturales y religiosos	Culturas prehispánicas	Hechos y personajes de la Conquista, Colonia e Independencia	Hechos y personajes de la Revolución Mexicana		
I1	Definición de contenido clara y precisa	✓	✓	×	✓	✓	✓	✓	×	75	25
I2	Definición de contenido congruente con nombre	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I3	Definición de contenido alineada al currículum de asignatura	×	✓	✓	✓	✓	✓	✓	✓	87.5	12.5
I4	Contenido coherente con lo que se enseña en aula	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I5	Dominio de contenido es básico para asignatura	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I6	Dominio de contenido es esperado del promedio de estudiantes	✓	✓	✓	✓	×	✓	✓	✓	87.5	12.5
I7	Aprendizaje de contenido es importante p/dominio de asignatura	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I8	Delimitación del contenido alineada y derivada de definición	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I9	Habilidades y contenidos delimitados representan lo esencial	✓	×	✓	✓	✓	✓	✓	✓	87.5	12.5
I10	Estrategia ev. adecuada p/evaluar contenidos delimitados	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I11	Estrategia ev. adecuada p/evaluar aprendizajes esperados	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I12	Estrategia evaluativa semejante a cómo se enseña en aula	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I13	Ítems hijos reflejan uso de conocimiento adquirido	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I14	Base del reactivo clara y suficiente para emitir respuesta	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I15	Instrucciones adicionales claras y suficientes p/emitir respuesta	---	---	---	---	---	---	---	---	---	---
I16	Reglas p/generar ítems hijos responden a estrategia evaluativa	✓	✓	✓	✓	✓	✓	×	×	75	25
I17	Textos auxiliares apropiados	---	---	---	---	---	---	---	---	---	---
I18	Gráficos e imágenes apropiados	×	---	---	---	---	---	---	---	0	100
I19	Banco de información corresponde al contenido seleccionado	×	×	×	×	×	×	×	×	0	100
I20	Tipo de ejecución simple y facilita evaluación del contenido	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I21	Ítems hijos representan contenido delimitado	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I22	Ítems hijos sin errores de redacción	×	✓	✓	✓	✓	✓	✓	✓	87.5	12.5
I23	Ítems hijos redactados con palabras de uso común de estudiantes	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I24	Ítems hijos sin pistas	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I25	Ítems hijos con nivel de dificultad apropiado al grado escolar	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I26	Ítems hijos sin sesgo	✓	✓	✓	✓	✓	✓	✓	✓	100	0
I27	Banco de información e Ítems hijos libres de otro tipo de errores	×	✓	✓	✓	✓	✓	✓	✓	87.5	12.5
	% Validado	80	91.67	91.67	95.83	91.67	95.83	91.67	87.5		
	% No validado	20	8.33	8.33	4.17	8.33	4.17	8.33	12.5		

Nota: Los símbolos en las columnas representan los siguientes aspectos: (✓) = Indicador validado, (×) = Indicador no validado, (---) = No aplica.

Como en la codificación utilizada en el área de matemáticas, los símbolos de *paloma* (✓) señalan que hubo acuerdo entre los expertos al valorar los indicadores; es decir, que los validaron. Las *tachas* (✗) indican desacuerdo o que los indicadores no fueron validados.

De igual manera, hubo excepciones en las que algunos indicadores no fueron tomados en cuenta para su evaluación, por lo cual, la ausencia de respuesta se registró con una marca de tres guiones ortográficos (---). Esto indica que la evaluación de dicho indicador no aplicaba para tales casos. Como ejemplo, se puede mencionar el caso de los indicadores 15 y 17, que exploran si las instrucciones adicionales a la base del reactivo y/o los textos auxiliares son apropiados, claros y suficientes para que el estudiante emita su respuesta, pero que no fue necesario evaluar debido a que los ítems hijos de Historia no utilizan instrucciones adicionales a la base del reactivo ni textos auxiliares.

En las columnas de la Tabla 4.10 se observan las opiniones de los expertos a cada uno de los ocho modelos de ítems, mientras que en los renglones se registraron las opiniones a cada uno de los 27 indicadores evaluados. En la parte inferior se muestran, en términos de porcentajes, los indicadores que fueron y no fueron validados por los expertos en cada uno de los modelos de ítems. En la columna del extremo derecho aparecen los porcentajes de ítems validados en cada uno de los 27 indicadores.

Análisis de resultados

En este apartado, se presenta la información en tres secciones. En la primera se analiza el conjunto de modelos de ítems que conforman el área de Historia y se describen las opiniones de los expertos en cada uno de los 27 indicadores que conforman el formato de evaluación. En la segunda se describen de manera individual las características de los modelos de ítems que presentaron cuando menos 20% de indicadores no validados. En la tercera sección se presentan

las observaciones y sugerencias realizadas por los expertos con la finalidad de mejorar el área evaluada.

Resultados por problemas detectados en el conjunto de modelos de ítems

La Figura 4.16 muestra, de manera gráfica, una síntesis de los resultados del área de Historia, de acuerdo con el porcentaje de los modelos de ítems que no fueron validados por los expertos, en los distintos indicadores evaluados.

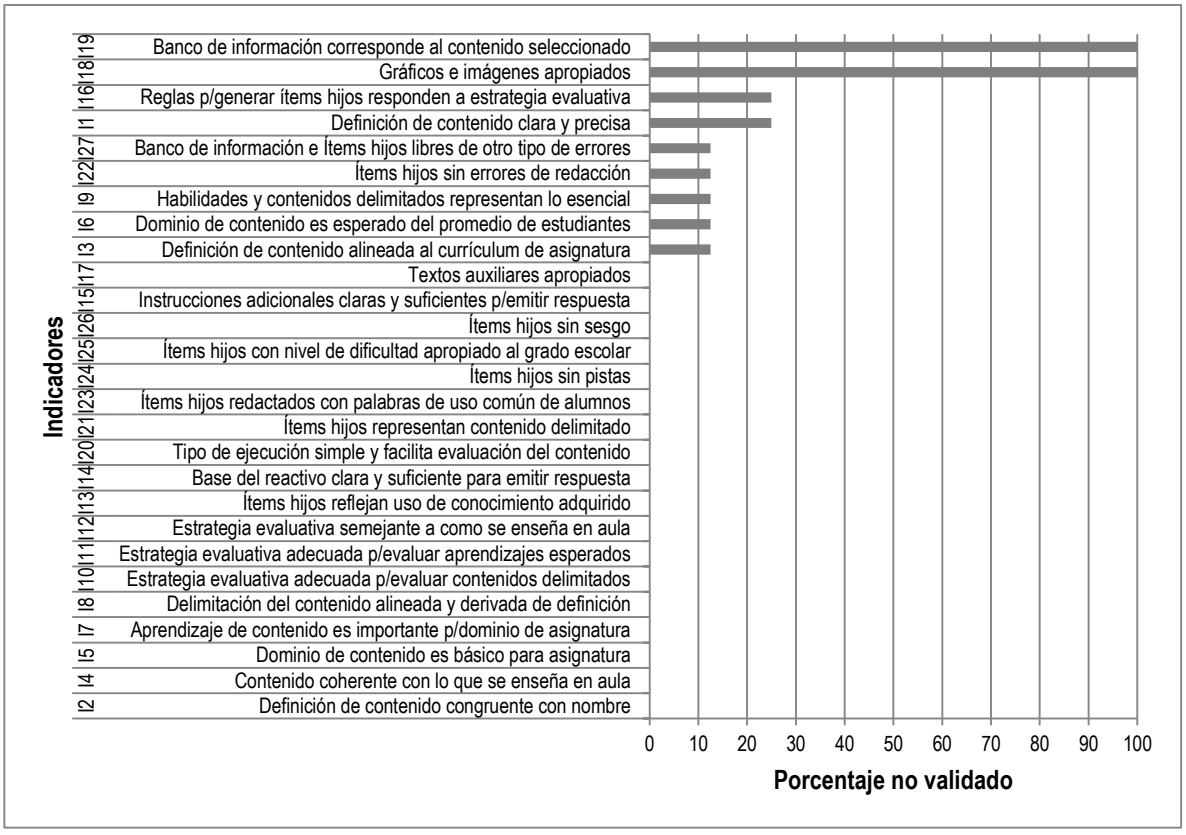


Figura 4.16. Porcentaje de modelos de ítems de Historia no validados por los expertos en los indicadores

En la Figura 4.16 se observa que hubo cuatro indicadores que presentaron un porcentaje de modelos de ítems no validados por parte de los expertos, mayor a 20% y por ende, que revelaron los problemas del área de Historia que se presentan con mayor frecuencia. En primer

lugar se encuentran los indicadores 18 y 19, que no fueron validados en 100% de los modelos de ítems en los que fueron evaluados.

El indicador 18 presentó una situación especial, ya que únicamente se requirió evaluar en un modelo de ítems. Esto se debió a que explora aspectos de calidad en el diseño y nivel de complejidad de los gráficos o imágenes utilizadas por el generador de reactivos, y solamente uno de los ocho modelos de ítems utiliza un gráfico. Como resultado, no fue validado debido a que los expertos señalaron problemas con los cortes establecidos en la línea de tiempo, que fueron utilizados para determinar los periodos en los que esta se dividió.

El indicador 19, que sí fue evaluado en los ocho modelos de ítems, pero resultó que en 100% de ellos no fue validado, explora aspectos relacionados con los ítems hijos y los elementos cadena que se utilizan para generarlos. Esto significa que en todos los modelos de ítems los expertos identificaron problemas específicos, tales como: presencia de elementos dentro del banco de información poco relevantes o alejados de la intención evaluativa y/o de los contenidos de la asignatura de Historia; problemas de redacción que generan confusión en los estudiantes o se alejan de la forma en que se enseñan en el aula, y elementos marcados como respuestas correctas cuando no lo son, entre otros.

En segundo término se encuentran los indicadores 1 y 16, en los cuales se identificó un problema de falta de correspondencia de la información contenida en las distintas secciones del modelo de ítems. Ambos indicadores no fueron validados en 25% de los modelos de ítems, es decir en dos de ellos. Por un lado, el indicador 1 aborda la claridad y precisión en la forma en la que fue definido el contenido a evaluar, y por otro, los expertos detectaron que en los modelos de ítems en los que se presentó este problema, para definir y describir el contenido a evaluar se utilizan términos y conceptos que usualmente no se encuentran asociados a él o lo hacen de

manera indirecta, ya que pertenecen a un periodo histórico distinto al que evalúa el modelo de ítems.

Por otro lado, en el indicador 16 que evalúa si las reglas establecidas para generar los reactivos hijos responden a la estrategia evaluativa, los expertos encontraron que esto no sucede en uno de los modelos de ítems, ya que los ítems hijos utilizan una línea de tiempo para que el estudiante ubique acontecimientos y personajes de la historia, mientras que las reglas establecen que se clasificarán las opciones de respuesta en categorías.

Hubo un grupo de cinco indicadores (3, 6, 9, 22 y 27) en los que los expertos determinaron que 12.5% de los modelos de ítems no cumplen con los criterios de calidad evaluados. Los problemas detectados en este bloque están relacionados con aspectos de definición, relevancia y correspondencia del contenido, ya que se detectó un caso en el que el programa de estudios marca que pertenece a un tema curricular que se ubica dentro de un periodo histórico específico, mientras que el modelo de ítems menciona otro.

Asimismo, los expertos mencionaron que hubo un modelo de ítems en el cual se evalúa un contenido que no es un aprendizaje esperado del promedio de los estudiantes, debido a su alto nivel de complejidad. Esto conlleva problemas tanto para su enseñanza, como para su aprendizaje, por lo cual su evaluación no se considera indispensable. También se detectaron problemas de pertinencia en algunos de los elementos curriculares contenidos en el banco de ítems, ya que los expertos argumentaron que no son temas regularmente abordados en clase.

Hubo dos indicadores (15 y 17) que no fueron evaluados por los expertos en ninguno de los modelos de ítems. Los aspectos que abordan están relacionados con el uso de instrucciones adicionales a la base del reactivo, así como textos auxiliares en los ítems hijos. Asimismo, debido a que ningún modelo de ítems de la asignatura de Historia presenta estas características, no fue necesaria su evaluación.

Finalmente, en la Figura 4.16 se observa que 16 indicadores fueron validados en 100% de los modelos de ítems. Esto significa que se cumplieron satisfactoriamente los aspectos de calidad relacionados con la forma en la que el contenido fue definido (indicadores 2 y 8), la correspondencia entre lo que se planteó en el modelo de ítems con lo que se enseña dentro del aula (indicador 4), así como la importancia que éste tiene en lo que respecta al dominio de toda la asignatura por parte del estudiante (indicadores 5 y 7).

También se puede observar que en este grupo de indicadores los expertos juzgaron que la estrategia que se utilizó para evaluar los contenidos y aprendizajes esperados, y por ende generar los ítems hijos de todos los modelos de ítems, no solamente es adecuada, sino que es semejante a la forma en que se enseña en el aula (indicadores 10, 11 y 12).

En cuanto a los ítems hijos, los expertos consideraron que representan adecuadamente el contenido que se desea evaluar, y sirven para evaluar si el estudiante utiliza el conocimiento que adquirió en el aula (indicadores 13 y 21). Asimismo, determinaron que tanto su nivel de dificultad como el tipo de ejecución que solicitan que el estudiante realice para emitir su respuesta, son apropiados para el grado escolar de los examinados (indicadores 20 y 25).

En lo que respecta a aspectos técnicos de redacción y sesgo, los expertos indicaron que los ítems hijos se encuentran libres de problemas, ya que el lenguaje es simple y de uso común a todos los estudiantes (indicadores 23, 24, 25 y 26).

Resultados por problemas detectados en los modelos de ítems en lo individual

La figura 4.17 muestra de manera gráfica una síntesis de los resultados de los modelos de ítems de Historia, de acuerdo con el porcentaje de indicadores que no fueron validados por los expertos en cada uno de dichos modelos. Siguiendo el tipo de estructura propuesta con anterioridad para presentar y analizar los resultados, a continuación se examinan a detalle aquellos modelos de ítems que obtuvieron un 20% o más de indicadores no validados.

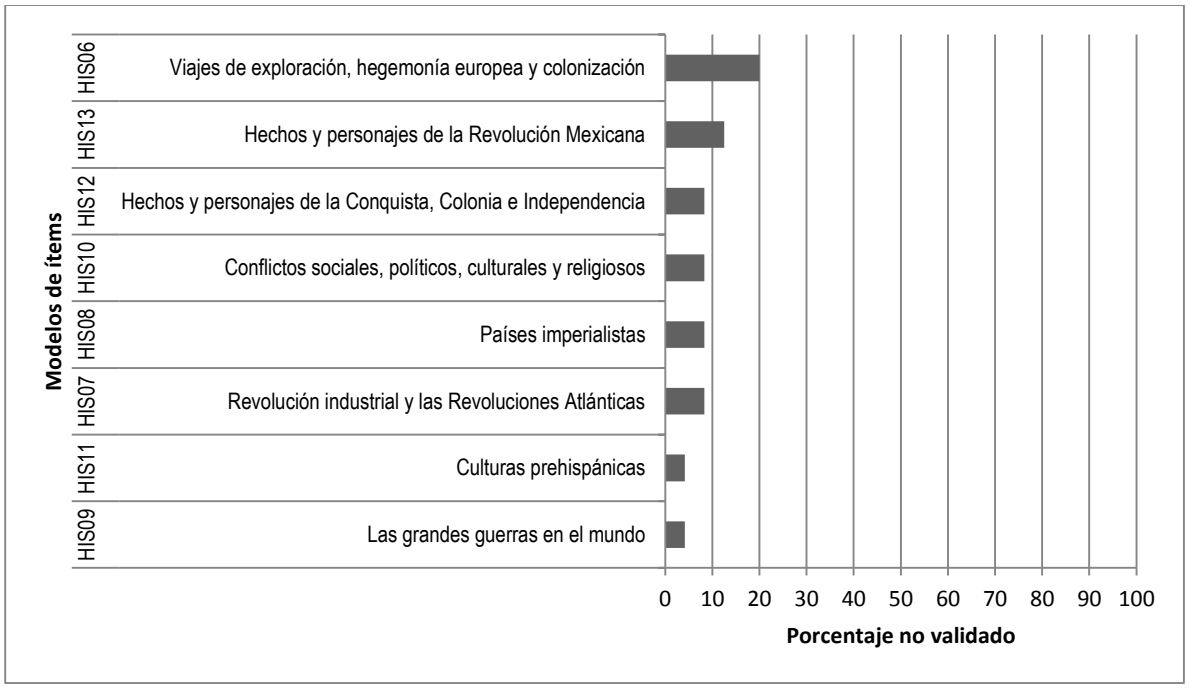


Figura 4.17. Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Historia.

Como se observa en la Figura 4.17, solamente hubo un modelo de ítems que fue identificado por los expertos con problemas en 20% de los indicadores de calidad. Dicho modelo es HIS06, el cual evalúa si el estudiante cuenta con una visión histórica ordenada, y si ubica espacial y cronológicamente los acontecimientos históricos asociados a los viajes de exploración, hegemonía europea y colonización.

El resto de los modelos de ítems se distribuyen en tres grupos, según la cantidad de indicadores que no fueron validados por los expertos: el primero, que incluye un solo modelo de ítems (HIS13), el cual presentó problemas en 12.5% (3) de los indicadores; un segundo grupo que incluye cuatro modelos de ítems (HIS07, HIS08, HIS10, HIS12) que presentaron 8.33% de indicadores no validados, y por último, un grupo con dos modelos de ítems (HIS09 e HIS11) que resultaron con 4.17% (1) de los indicadores no validados.

HIS06. Viajes de exploración, hegemonía europea y colonización

Este modelo de ítems genera ítems hijos que solicitan al estudiante que identifique y ubique en una línea de tiempo los sucesos históricos, personajes y avances científicos y tecnológicos más relevantes del mundo antiguo y Edad Media, asociados al surgimiento de la economía mundial y la historia moderna. En la figura 4.18, se muestra el ejemplo de un ítem en el cual se presentan diversos acontecimientos históricos del periodo mencionado, los cuales deben organizarse cronológicamente y colocar en una línea de tiempo.

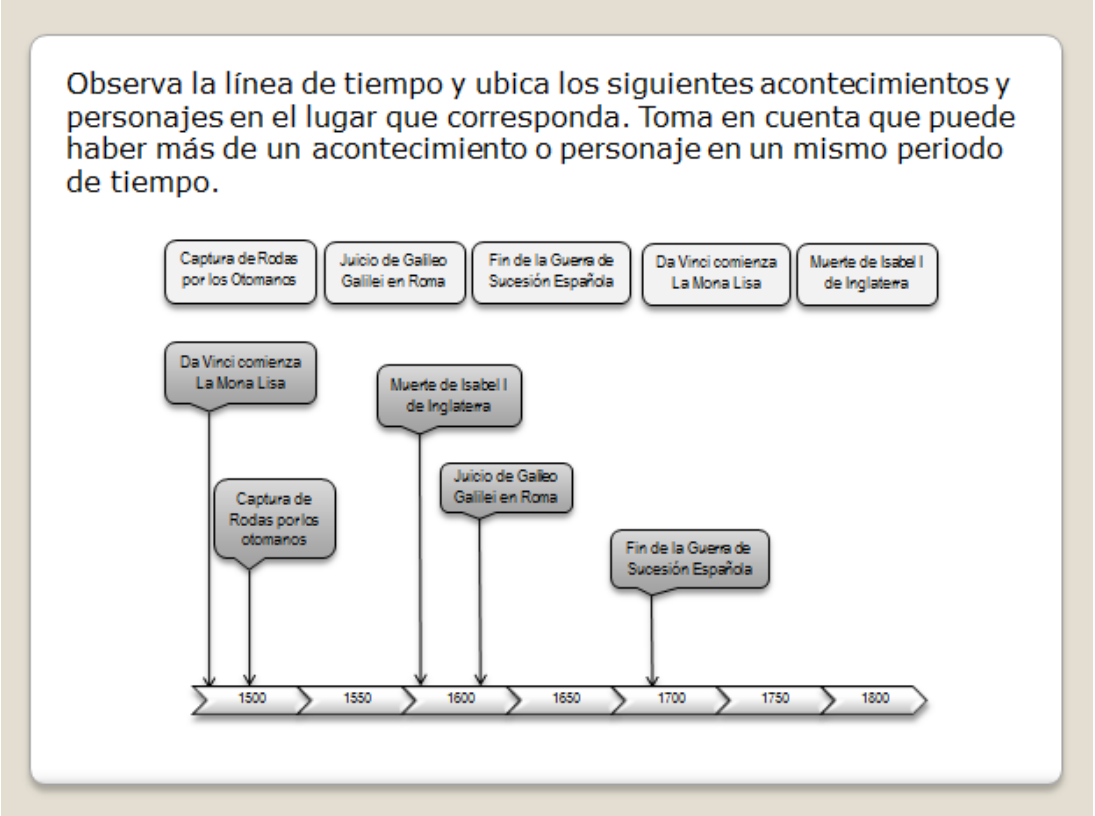


Figura 4.18. Ejemplo de ítem hijo del modelo HIS06: Viajes de exploración, hegemonía europea y colonización.

Para resolver este reactivo, el estudiante debe conocer la cronología de los hechos ocurridos en la Antigüedad y la Edad Media, de tal manera que demuestre su dominio y comprensión de las implicaciones que tuvo la expansión europea de los siglos XVI al XVIII en el mundo moderno. Asimismo, debe demostrar que conoce los cambios asociados a la hegemonía europea y las transformaciones que se dieron, como resultado del contacto entre diversas civilizaciones. Para emitir sus respuestas, el estudiante debe seleccionar una por una las opciones y colocarlas sobre la imagen de la línea de tiempo, dentro de los espacios en los que considere que sucedió el evento.

Respecto al proceso de validación, la Figura 4.19 muestra para este modelo de ítems, las respuestas consensuadas por los expertos a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de los modelos de ítems.

Figura 4.19. Opinión de los expertos en los 27 indicadores validados del modelo de ítems HIS06.

Se puede observar que hubo cinco indicadores (3, 18, 19, 22 y 27) que no fueron validados por los expertos. Entre los problemas que presenta este modelo de ítems señalaron discrepancia entre la alineación curricular que debe tener el contenido evaluado y lo que marca el programa de estudios (indicadores 3 y 27). Esto se debe a que el modelo de ítems menciona que el contenido se ubica dentro del periodo comprendido entre principios del siglo XV y principios del siglo XVIII, mientras que el programa de estudios de la asignatura estipula que dicho periodo es de principios del siglo XVI a principios del siglo XVIII. Para solventar esta situación, los expertos

sugirieron hacer la modificación necesaria al modelo de ítems para que exista correspondencia entre éste y el programa de estudios.

Por otro lado, este modelo de ítems indica que al generar los distintos ítems hijos se debe utilizar la imagen de una línea de tiempo, con la finalidad de que el estudiante coloque diversos eventos históricos en orden cronológico. Sin embargo, los expertos encontraron (indicador 18) que la manera en la que están representados los cortes en el tiempo, dificulta la emisión de la respuesta, ya que son muy amplios (50 años), mientras que en el aula usualmente se emplean líneas de tiempo cuyos cortes son por periodos de 10 a 25 años. Por este motivo, sugirieron modificar la imagen y adaptarla a la forma en que se utiliza en el aula.

Asimismo, se detectaron varios elementos que conforman el banco de información, que los expertos consideran poco relevantes o descontextualizados, ya que: a) forman parte de la información que se revisa en otras asignaturas como Geografía y Formación cívica y ética; b) no corresponden al periodo evaluado en este contenido; c) no se revisan en el aula por alejarse de los propósitos de enseñanza de la asignatura, como serían las aportaciones sociales o históricas que hicieron algunos personajes, y no sus datos biográficos (indicador 19). Ante esto, la sugerencia fue que se eliminaran los elementos que señalaron.

Por último, otro aspecto que detectaron los expertos fue que una de las opciones de respuesta contenidas en el banco de elementos contiene un error ortográfico que requiere ser corregido (indicador 22).

Resultados conjuntos

El propósito de este apartado es buscar si existen coincidencias en los problemas detectados en los modelos de ítems que conforman la sección de Historia del instrumento, que presentan problemas en al menos 20% de los indicadores, con los indicadores que no fueron validados por los expertos en al menos 20% de los modelos de ítems. Sin embargo, se encontró

que únicamente uno de ocho modelos de ítems presentó ambas características, lo cual es insuficiente para establecer una comparación.

Resultados según las sugerencias del panel de expertos para el mejoramiento del área

Además de los problemas que detectaron los expertos, hubo una serie de sugerencias que realizaron para el mejoramiento y fortalecimiento del instrumento en el área de Historia. Entre las más notables, mencionaron que al revisar los ocho modelos de ítems como conjunto, estimaron que 100% de ellos representan lo que el estudiante debe dominar de la asignatura de Historia debido a que las competencias que se favorecen y señalan en el programa de estudios, y que deben fomentarse en los estudiantes de secundaria se abordan con los distintos tipos de ítems que se incluyen en el examen, así como en los contenidos curriculares que éstos evalúan.

También mencionaron que mediante los contenidos y tipos de reactivos que se incluyeron en el examen se fomentan las competencias de comprensión del tiempo histórico y del espacio histórico, así como las de manejo de información histórica, las cuales son fundamentales y se mencionan en el programa de estudios en el que se basó el área de Historia del GAI Excoba.

Consideraron que hay otros contenidos que deberían incluirse en el examen; por ejemplo, opinaron que es esencial evaluar el tema *En busca de un sistema político*, debido a que en él se sientan las bases del sistema político mexicano actual como un *parte-aguas* en la historia del país.

En lo que respecta a la estimación que hicieron del porcentaje de estudiantes que responderían correctamente a estos ocho reactivos, consideran que sería aproximadamente 75% de ellos, ya que de acuerdo con su experiencia, algunos estudiantes tienen dificultades para desarrollar habilidades de análisis y deducción de respuestas, lo cual se traduciría en problemas de evocación y anclaje del conocimiento previamente adquirido, con el contenido evaluado en algunos de los reactivos del examen.

Resultados según la relación entre los problemas detectados y los contenidos evaluados

El propósito de este apartado es presentar un análisis que muestre la existencia de relaciones entre los tipos de contenidos curriculares que evalúan los modelos de ítems del área de Historia que presentan una mayor cantidad de problemas en cuanto a los indicadores de calidad (al menos 20% no validados), y el tipo de problema que detectan los indicadores que no fueron validados en al menos 20% de los modelos de ítems del área.

Hubo varias razones por las cuales no se pudo realizar este análisis. La primera y más obvia de ellas implicó el hecho de que solamente hubo un modelo de ítems que cumplió con el criterio de 20% de indicadores no validados, siendo inviable una comparación. La segunda razón está relacionada con la naturaleza y estructura del programa de estudios; esto es, los ocho modelos de ítems del área de Historia generan reactivos que evalúan contenidos de ocho distintos bloques temáticos debido a que el programa está estructurado de manera secuencial y en orden cronológico, más no conceptual.

Finalmente, otro motivo importante fue que aunque el programa cuenta con tres ejes curriculares que permean de manera transversal a todos los contenidos de la asignatura (incluyendo los ocho seleccionados para formar parte de este instrumento), éstos no pueden ser separados ni analizados desde el nivel pretendido, debido a que están relacionados con dos aspectos universales que orientan e integran los aprendizajes esperados: valores y competencias.

En la dimensión valoral se considera la formación de una conciencia histórica para la convivencia, y en la de competencias se atiende la comprensión del tiempo y espacio históricos, así como el manejo de información básica. Esto, de igual manera, hace inviable una comparación entre modelos de ítems desde la perspectiva del tipo de contenido curricular, ya que además de no ser el objeto del análisis, los ocho persiguen el desarrollo de estos dos aspectos.

Por otro lado, la forma en que se consideró la asignatura de Historia para ser analizada y extraer de ella los contenidos considerados indispensables de ser evaluados en el Excoba, tiene una característica especial que la distingue de otras del nivel educativo de secundaria y se refiere al hecho de que curricularmente, está dividida en dos asignaturas que se imparten en distintos grados académicos: Historia I e Historia II.

En este sentido, si se buscan coincidencias entre modelos de ítems y grado escolar al que pertenecen, dejando de lado los criterio básicos de que deben formar parte del grupo que presentan 20% o más indicadores no validados y el tipo de contenido curricular, se observó que cinco modelos de ítems pertenecen a la asignatura Historia I (HIS06, HIS07, HIS08, HIS09, HIS10), mientras que los otros tres corresponden a Historia II (HIS11, HIS12, HIS13).

Siguiendo este análisis, se encontró que un modelo de ítems de Historia I (HIS08) y uno de Historia II (HIS13) coinciden en que presentan problemas de precisión en cuanto a la definición que se hizo del contenido y/o los elementos específicos que se utilizaron para describirlo con detalle, tales como los ejemplos. Asimismo, dos modelos de ítems de Historia II (HIS12, HIS13) coincidieron en que además de presentar problemas en el banco de información (lo cual sucedió en la totalidad de modelos de ítems), tuvieron señalamientos en cuanto a la forma en que fue operada la estrategia evaluativa.

4.2.3. **Modelos de ítems del área de Química**

Este apartado, como en los dos casos anteriores, corresponde a la Fase II del MVCE y describe las etapas 2 y 3 de dicha fase. Primero se presentan los resultados que describen las características de los modelos de ítems que conforman el área de acuerdo con el tipo de contenido curricular que evalúa, la clasificación que tienen en el sistema informático que administra el examen, la ejecución que debe realizar el estudiante al responder, y el tipo de

conocimiento que explora. Segundo, se presenta el análisis derivado del trabajo con el panel de expertos.

Etapa 2: Características de los modelos de ítems de Química

Los resultados de los modelos de ítems de química se agrupan en cuatro niveles de análisis: 1) según los contenidos curriculares de los que derivan los modelos de ítems; 2) por su clasificación en el GenerEx; 3) por la ejecución que el estudiante debe realizar para responder al ítem hijo, y 4) por el tipo de conocimiento evaluado (declarativo, procedimental, esquemático o estratégico.

Tipo de contenidos curriculares

Para la elaboración del GAI Excoba se utilizó el programa de estudios de Química de educación secundaria del año 2006. En él, todos los contenidos de la asignatura se organizan en cinco bloques: 1) las características de los materiales, 2) la diversidad de propiedades de los materiales y su clasificación química, 3) la transformación de los materiales: la reacción química, 4) la formación de nuevos materiales, y 5) química y tecnología[3] (SEP, 2006a). La Tabla 4.11 muestra la distribución de los ocho modelos de ítems siguiendo esta clasificación.

[3] **Las características de los materiales**. Se refiere a aquellos contenidos que buscan que el estudiante identifique las características fundamentales del conocimiento científico y tecnológico: experimentación e interpretación, y abstracción y generalización. En estos contenidos los modelos y sus características son fundamentales (abstracción o generalización, lenguaje matemático, precisión, brevedad, alcances y limitaciones). Se busca que el estudiante inicie el estudio de los materiales y los primeros sistemas de clasificación de sustancias, de tal manera que identifique los fundamentos básicos de las técnicas que acompañan a la investigación científica.

La diversidad de propiedades de los materiales y su clasificación química. Incluye los conocimientos relacionados con los materiales y la clasificación de sustancias, así como las características de los materiales. También se busca que mediante el manejo de estos contenidos, el estudiante identifique características macroscópicas de los materiales metálicos, así como el sistema de clasificación de la tabla periódica.

Tabla 4.11.
Distribución de modelos de ítems de Química por bloque temático curricular

Modelos de ítems	Contenido	Bloque temático	Distribución (porcentaje)
QUI13 QUI14	Propiedades intensivas y extensivas de la materia Mezclas homogéneas y heterogéneas	I Las características de los materiales	25
QUI15 QUI16	Clasificación de las sustancias Modelos de enlace: covalente, iónico y metálico	II La diversidad de propiedades de los materiales y su clasificación química	25
QUI17	El enlace químico y la valencia	III La transformación de materiales: reacción química	12.5
QUI18 QUI19	Aplicaciones de ácidos y bases Las reacciones Redox	IV La formación de nuevos materiales	25
QUI20	Características del método científico	Otro no contemplado en el programa	12.5

Se puede apreciar que dos (25%) de los modelos de ítems evalúan competencias del Bloque I: *Las características de los materiales*, dos más (25%) evalúan habilidades relacionadas con el segundo bloque: *La diversidad de propiedades de los materiales y su clasificación química*, dos (25%) abordan contenidos del bloque IV: *La formación de nuevos materiales,* uno (12.5%) evalúa conocimientos del bloque III: *La transformación de materiales: reacción química,* y uno más (12.5%) evalúa un contenido que se encuentra fuera del programa de ciencias de educación Secundaria. Esto es debido a que los desarrolladores del Excoba consideraron muy importante incluir la evaluación de este contenido por su relevancia en etapas escolares posteriores, a pesar de no encontrarse de manera explícita entre los contenidos del programa; sin embargo, como se

La transformación de los materiales: la reacción química. Se refiere a aquellos contenidos curriculares en los que se introducen conceptos, se presentan modelos y formas de representar el enlace químico.

La formación de nuevos materiales. Mediante los contenidos de este bloque se busca que el estudiante adquiera la capacidad de predecir a nivel básico los productos de las reacciones químicas ácido-base y óxido-reducción.

Química y tecnología. Los contenidos de este bloque incluyen las características del conocimiento científico y su interacción con la tecnología. También se busca que el estudiante aprenda la forma en que los resultados deben ser comunicados.

verá más adelante es uno de los modelos de ítems que fue señalado por los expertos con mayor cantidad de problemas de diversa naturaleza.

Características del sistema informático GenerEx

En el área de Química se utilizaron cuatro de los 22 diferentes tipos de modelos de ítems que conforman la estructura del sistema informático del GAI. La distribución que se presentó en los ocho modelos se puede observar en la Tabla 4.12, la cual muestra que 62.5% (cinco) son del tipo *Elemento categoría*, mientras que 12.5% (uno) corresponde al tipo *Orden elementos múltiple*, 12.5% (uno) a la clasificación S*elección de elementos*, y 12.5% (uno) al tipo *Frase imagen*.

Tabla 4.12.
Clasificación y distribución de modelos de ítems de Química en el sistema GenerEx

	Total	Porcentaje
Elemento categoría	5	62.5
Orden elementos múltiple	1	12.5
Selección de elementos	1	12.5
Frase imagen	1	12.5

El primer tipo de modelos de ítems (*Elemento categoría*), requiere una forma de programación en la que el sistema construye ítems hijos mediante la selección aleatoria de un número de categorías, junto con una cantidad preestablecida de elementos asociados a ellas, los cuales se muestran en la interfaz gráfica con la que interactúa el estudiante. Así, cuando el examinado responde, el sistema califica haciendo una comparación entre el lugar donde fueron colocados los elementos, y las respuestas que se encuentran alimentadas en él. En el apartado de resultados del área de Historia se presentó un ejemplo de este tipo de reactivos.

En lo que respecta al tipo *Orden elementos múltiple*, el sistema selecciona al azar y presenta un grupo de imágenes que representan objetos con propiedades (masa, peso, volumen, etcétera), las cuales deben ordenarse según un criterio establecido (menor a mayor, pequeño a grande,

etcétera). Al elegirlas, el sistema muestra las imágenes en la interfaz y el estudiante debe conocer en qué media se encuentra presente la propiedad para seleccionarlas, moverlas al lugar en que considere que muestran el orden correcto. La forma de calificar del sistema exige que se considere la ubicación de cada imagen dentro de un continuo ordenado.

En los modelos de ítems del tipo *Selección de elementos*, el sistema presenta ecuaciones químicas en las que se observan espacios vacíos o segmentos marcados en los que se debe seleccionar un valor. Al hacer clic sobre cada espacio marcado, se despliega una ventana con distintas opciones numéricas para que el estudiante elija la que balancee la ecuación.

Por último, para operar los modelos de ítems de tipo *Frase imagen*, el sistema debe contener un banco de información con distintos tipos de textos, de los cuales seleccionará uno al azar. Cada texto deberá tener segmentos (palabras, frases o enunciados) que funcionarán como elementos cadena que se podrán intercambiar y presentar en distinto orden. Siguiendo las reglas establecidas en el modelo de ítems, se generará el ítem hijo y se mostrará al estudiante en la interfaz, donde deberá seleccionar, mover mediante el uso del ratón y depositar en el espacio que considere refleja el orden correcto de los segmentos que conforman el texto.

Tipo de ejecución solicitada

Los modelos de ítems que fueron validados por los expertos implicaron el uso de estrategias evaluativas que solicitan distintos tipos de ejecución por parte del estudiante cuando emite sus respuestas en los ítems hijos. En la Tabla 4.13 se presenta la distribución de los mismos, según impliquen ejecuciones de arrastre de elementos, selección de elementos, escritura libre, o una combinación de selección de elementos con escritura libre.

Tabla 4.13.
Distribución de modelos de ítems de Química, según el tipo de ejecución que demandan del estudiante

Tipo de ejecución	Modelo de ítems	
	Total	Porcentaje
Arrastre	7	87.5
Selección	1	12.5

Como puede observarse, en el área de Química, la gran mayoría de los modelos de ítems, es decir 87.5% (7) genera ítems hijos en los que la ejecución implica que el estudiante realice la acción de mover elementos, es decir, que requiere del uso del ratón para el movimiento y colocación de elementos dentro de secciones de la interfaz. El modelo de ítems restante representa 25%, y el tipo de ejecución que se requiere realizar para responder es de selección de elementos, en donde el estudiante debe utilizar el ratón para dar clic y hacer que aparezca una ventana en la cual se presentan una serie de opciones, entre las cuales seleccionará la que consideres es la respuesta correcta.

En el caso de los otros dos tipos de ejecución (escritura y mixta), no hubo modelos de ítems en el área de Química que cumplieran con estas características.

Tipo de conocimientos evaluados

La distribución de los modelos de ítems según el tipo de conocimiento que exploran los ítems hijos generados es la que se muestra en la Tabla 4.14.

Tabla 4.14.
Tipo de conocimiento que evalúan los modelos de ítems de Química

Conocimiento evaluado	Modelos de ítems	
	Total	Porcentaje
Declarativo	4	50
Procedimental	3	37.5
Esquemático	1	12.5
Estratégico	-----	-----

Se puede observar que 50% (4) de los modelos de ítems evalúan principalmente conocimientos de tipo declarativo, en los que se requiere un manejo organizado de la información previamente aprendida, mediante la clasificación de datos y elementos conceptuales, de tal forma que las respuestas dadas expresen los principios teóricos y conceptuales que el estudiante aprendió en la escuela.

Así, por ejemplo, en el modelo de ítems QUI13, que evalúa el manejo del concepto de propiedades intensivas y extensivas de la materia, el estudiante deberá tener un manejo conceptual y teórico de las características de los materiales y sus propiedades, con la finalidad de que discrimine el grado en que cada una de las sustancias presentadas posee la propiedad en cuestión.

Por otro lado, se observa que 37.5% (3) de los modelos de ítems evalúan conocimientos de tipo procedimental, en los que se requiere que el estudiante utilice su conocimiento de métodos y procedimientos que implican el seguimiento de reglas o pasos para obtener un resultado determinado. Por ejemplo, en el modelo de ítems QUI14, que genera ítems hijos que evalúan el conocimiento que tiene el estudiante respecto a las mezclas homogéneas y heterogéneas, así como sus métodos de separación, se necesita conocer de manera muy específica cuáles son las características de cada método, así como el tipo de mezclas para los cuales son útiles. De esta manera, el hecho de que el estudiante tenga un manejo adecuado de los procedimientos implicados en la separación de sustancias será determinante para que responda correctamente a los ítems hijos que exploren este contenido curricular.

En el caso de los conocimientos de tipo esquemático, 12.5% (1) de los modelos de ítems evalúan aspectos vinculados con el enlace químico y la valencia, particularmente la relación de modelos de compuestos y sus fórmulas químicas, para balancear ecuaciones. Esto implica que para responder a cualquier ítem hijo de esta área, el estudiante debe manejar la información de

manera ordenada y sistemática, y utilizar modelos o esquemas mentales que le permitan acceder a los conocimientos adquiridos durante su aprendizaje, de tal manera que sus respuestas reflejen que conoce la forma en que se utilizan las valencias para balancear ecuaciones.

Por ejemplo, en el modelo de ítems QUI17, que genera ítems hijos que evalúan el conocimiento que tiene el estudiante respecto al enlace químico y la valencia, se utilizan reacciones químicas, representadas mediante ecuaciones, las cuales deberá balancear eligiendo los números que expresan las valencias correctas. Para resolver correctamente este tipo de ítems, el estudiante debe contar con un esquema mental muy claro respecto a los modelos detrás de las reacciones químicas y asociarlas a las valencias requeridas para lograr un correcto balanceo.

Por último, la evaluación del conocimiento de tipo estratégico, en el que se requiere la elaboración de un plan detallado y la presentación de alternativas de solución a un problema dado, no fue abordada por los modelos de ítems del área de Química.

Etapa 3: Análisis de los resultados del trabajo con el panel de expertos

Resultados generales

En total se validaron ocho modelos de ítems, para lo cual se utilizó un ítem hijo como muestra de cada modelo. Al igual que las áreas de Matemáticas e Historia, los ítems que fueron seleccionados para presentar a los expertos, corresponden a una versión fija del Excoba que fue utilizada para su pilotaje y análisis de estructura interna.

El proceso de evaluación que realizaron los miembros del panel de expertos proporcionó información útil en la identificación de los problemas que presentan los ocho modelos de ítems que conforman el área de Química de secundaria del Excoba, así como en la detección de los aspectos específicos que requieren ser atendidos y modificados para fortalecer el instrumento.

Recordando el proceso de evaluación que se siguió en esta etapa, los expertos valoraron 27 indicadores de calidad en cada uno de los ocho modelos de ítems que conforman el área de Química del Excoba, relacionados con la estructura y correspondencia de las distintas secciones del modelo, las características del banco de información curricular y los elementos cadena e integrales que lo conforman, así como aspectos de diseño, funcionalidad y operatividad de los ítems hijos generados. En cada indicador consensuaron acuerdo o desacuerdo, registrando sus observaciones.

En la Tabla 4.15 se muestra la información general obtenida como producto del proceso de evaluación de los modelos de ítems del área de Química. Dicha tabla proporciona de manera condensada, la información recopilada mediante las opiniones que los expertos emitieron respecto a factores de calidad y pertinencia de los contenidos que fueron seleccionados del currículum para ser evaluados y conformar dicha área. Retomando la codificación utilizada en las otras áreas analizadas, los símbolos de *paloma* (✓) señalan que hubo acuerdo entre los expertos al valorar los indicadores; es decir, que los validaron. Las *tachas* (✗) indican desacuerdo o que los indicadores no fueron validados.

De igual manera, hubo excepciones en las que algunos indicadores no fueron tomados en cuenta para su evaluación, por lo cual, la ausencia de respuesta se registró con una marca de tres guiones ortográficos (---), indicando que la evaluación de dicho indicador no aplicaba para tales casos.

Tabla 4.15.

Resultados globales del proceso de evaluación de panel de expertos del área de Química del Excoba

		QUI13	QUI14	QUI15	QUI16	QUI17	QUI18	QUI19	QUI20		
	Indicador	Propiedades intensivas y extensivas de la materia	Mezclas homogéneas y heterogéneas	Clasificación de las sustancias	Modelos de enlace: covalente, iónico y metálico	El enlace químico y la valencia	Aplicaciones de ácidos y bases	Las reacciones Redox	Características del método científico	% Validado (✓)	% No validado (×)
I1	Definición de contenido clara y precisa	✓	✓	×	×	×	✓	✓	✓	62.5	37.5
I2	Definición de contenido congruente con nombre	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I3	Definición de contenido alineada al currículum de asignatura	✓	✓	✓	✓	✓	✓	✓	×	87.5	12.5
I4	Contenido coherente con lo que se enseña en aula	×	✓	✓	✓	✓	✓	✓	×	75.0	25.0
I5	Dominio de contenido es básico para asignatura	✓	✓	✓	✓	✓	✓	✓	×	87.5	12.5
I6	Dominio de contenido es esperado del promedio de estudiantes	✓	✓	✓	✓	✓	✓	✓	×	87.5	12.5
I7	Aprendizaje de contenido es importante p/dominio de asignatura	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I8	Delimitación del contenido alineada y derivada de definición	×	×	✓	✓	×	✓	✓	✓	62.5	37.5
I9	Habilidades y contenidos delimitados representan lo esencial	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I10	Estrategia evaluativa adecuada p/evaluar contenidos delimitados	×	✓	✓	✓	✓	✓	✓	✓	87.5	12.5
I11	Estrategia evaluativa adecuada p/evaluar aprendizajes esperados	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I12	Estrategia evaluativa semejante a como se enseña en aula	✓	✓	✓	✓	✓	×	✓	×	75.0	25.0
I13	Ítems hijos reflejan uso de conocimiento adquirido	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I14	Base del reactivo clara y suficiente para emitir respuesta	✓	×	✓	✓	×	×	✓	✓	62.5	37.5
I15	Instrucciones adicionales claras y suficientes p/emitir respuesta	---	---	---	---	---	---	---	---	---	---
I16	Reglas p/generar ítems hijos responden a estrategia evaluativa	✓	×	✓	✓	✓	✓	✓	✓	87.5	12.5
I17	Textos auxiliares apropiados	---	---	---	---	---	---	---	✓	100.0	0.0
I18	Gráficos e imágenes apropiados	×	---	---	---	---	---	---	---	0.0	100.0
I19	Banco de información corresponde al contenido seleccionado	×	×	×	✓	×	×	✓	×	25.0	75.0
I20	Tipo de ejecución simple y facilita evaluación del contenido	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I21	Ítems hijos representan contenido delimitado	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I22	Ítems hijos sin errores de redacción	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I23	Ítems hijos redactados con palabras de uso común de alumnos	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I24	Ítems hijos sin pistas	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I25	Ítems hijos con nivel de dificultad apropiado al grado escolar	×	✓	✓	✓	×	✓	✓	✓	75.0	25.0
I26	Ítems hijos sin sesgo	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I27	Banco de información e Ítems hijos libres de otro tipo de errores	✓	×	✓	✓	✓	✓	✓	✓	87.5	12.5
	% Validado	76.0	79.2	91.7	95.8	79.2	87.5	100.0	76.0		
	% No validado	24.0	20.8	8.3	4.2	20.8	12.5	0.0	24.0		

Nota: Los símbolos en las columnas representan los siguientes aspectos: (✓) = Indicador validado, (×) = Indicador no validado, (---) = No aplica.

En las columnas de la Tabla 4.15 se observan las opiniones de los expertos a cada uno de los ocho modelos de ítems, mientras que en los renglones se registraron las opiniones a cada uno de los 27 indicadores evaluados. En la parte inferior se muestran, en términos de porcentajes, los indicadores que fueron y no fueron validados por los expertos en cada uno de los modelos de ítems, y en la columna del extremo derecho, los porcentajes de ítems validados en cada uno de los 27 indicadores

Análisis de resultados

En este apartado, se analiza la información desde tres perspectivas. En la primera se analiza el conjunto de modelos de ítems que conforman el área de Química y se describen las opiniones de los expertos en cada uno de los 27 indicadores que conforman el formato de evaluación. En la segunda se describen de manera individual, las características de los modelos de ítems que presentaron cuando menos 20% de indicadores no validados. En la tercera sección se presentan las observaciones y sugerencias realizadas por los expertos con la finalidad de mejorar el área evaluada.

Resultados por problemas detectados en el conjunto de modelos de ítems

La Figura 4.20 muestra de manera gráfica una síntesis de los resultados de acuerdo con el porcentaje de los modelos de ítems que no fueron validados por los expertos, en los distintos indicadores evaluados.

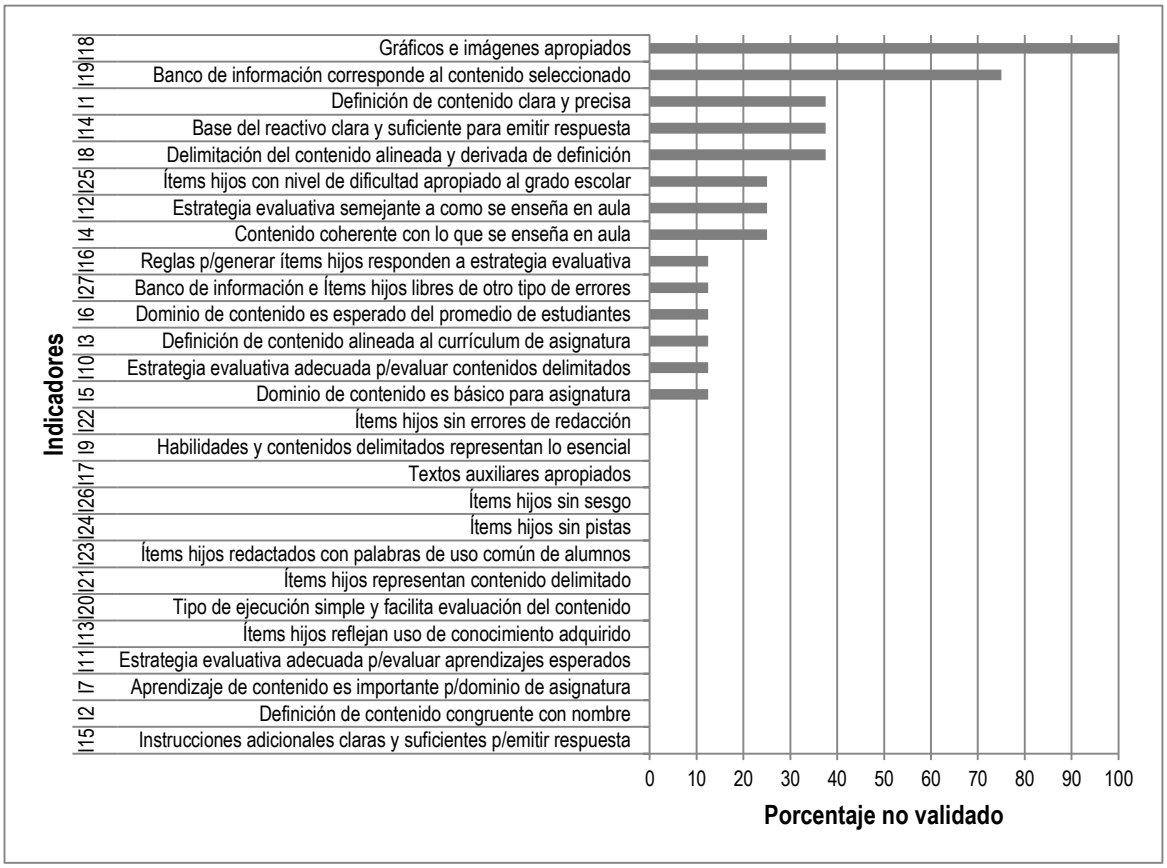


Figura 4.20. Porcentaje de modelos de ítems de Química no validados por los expertos en los indicadores

Se observa que hubo ocho indicadores que presentaron un porcentaje de modelos de ítems no validados por los expertos, mayor a 20% y por ende que revelaron los problemas del área de Química que se presentan con mayor frecuencia. En primer lugar se encuentra el indicador 18, que no fue validado en 100% de los modelos de ítems en los que fue evaluado. Al igual que lo sucedido en el área de Historia, este indicador fue evaluado únicamente en un modelo de ítems debido a que explora aspectos de calidad en el diseño de las imágenes utilizadas por el generador, y solamente uno de los ocho modelos utiliza imágenes para generar los ítems. Los expertos consideraron que casi todas las imágenes utilizadas son pertinentes y de buena calidad, pero señalaron que una de ellas no cumple con este indicador debido al color empleado en ella. Como se trata de una sustancia que el estudiante debe identificar visualmente, argumentaron que es

de vital importancia que se presente tal como la aprendieron en el aula, recomendando que se modifique el color.

En segundo término, el indicador 19 no fue validado en 75% de los modelos de ítems. Este indicador explora aspectos relacionados con los ítems hijos y los elementos cadena e integrales que se utilizan para generarlos; esto significa que en seis de ocho modelos de ítems los expertos identificaron problemas específicos, tales como: la presencia de elementos cadena dentro de los bancos de información que pueden generar confusión en los estudiantes debido a su cercanía conceptual, haciendo muy difícil la distinción entre correctos e incorrectos; elementos cadena registrados como correctos cuando no lo son; falta de homogeneidad en los niveles de dificultad de los elementos cadena, la cual provoca distinta demanda cognitiva del estudiante; elementos cadena que no corresponden al grado escolar del estudiante o a contenidos que marca el programa, y uso de términos poco familiares para los alumnos. Los expertos sugirieron una revisión detallada y modificación de los elementos cadena de los seis modelos de ítems que presentaron problemas en este indicador.

En tercer lugar se encuentran los indicadores 1, 8 y 14, los cuales no fueron validados en 37.5% de los modelos de ítems. Los expertos señalaron diversos problemas con la claridad, precisión, alineación curricular y delimitación del contenido, así como con la estructura de las bases de los reactivos. Por un lado, el indicador 1 aborda la claridad y precisión en la forma en que fue definido el contenido a evaluar, y los expertos detectaron que en los tres modelos de ítems en los cuales se presentó este problema, se utilizan términos y conceptos que pueden causar confusión o que no forman parte del contenido evaluado.

Por otro lado, en el indicador 8 se explora si la delimitación que se hizo del contenido es pertinente. Los expertos señalaron que hay tres modelos de ítems que no están delimitados adecuadamente ya que excluyen segmentos muy importantes de la definición, que deben ser

evaluados, o no son claros respecto a su inclusión de ellos. Sugirieron la revisión de este aspecto en los tres modelos de ítems.

Finalmente, respecto al indicador 14 que explora la claridad y suficiencia de la base de los reactivos, los expertos mencionaron que los tres modelos de ítems con este problema requieren revisión, ya que las instrucciones incluidas no contienen una explicación lo suficientemente clara para que el estudiante comprenda lo que debe realizar para emitir sus respuestas. También señalaron que en uno de los modelos de ítems la instrucción dada contiene el uso de un término que no corresponde a las enseñanzas de los estudiantes. Las sugerencias giraron en torno a modificaciones en la redacción.

Hubo un grupo de tres indicadores (4, 12 y 25) en los que los expertos determinaron que 25% de los modelos de ítems no cumplen con los criterios de calidad evaluados. El indicador 4 aborda la correspondencia entre lo que se planteó en el modelo de ítems con lo que se enseña dentro del aula. Los expertos mencionaron que uno de los dos modelos de ítems que presenta este problema requiere que se le agregue un segmento conceptual que se enseña en el aula y fue excluido del modelo. También mencionaron que el segundo modelo de ítems que presentó este problema fue señalado por que no pertenece al programa de estudios vigente.

En cuanto al indicador 12, que evalúa la relación de la estrategia evaluativa con la forma en que se enseña en el aula, se encontró que en uno de los dos modelos de ítems que presentaron este problema excluye un aspecto importante de la forma en que se enseña el contenido, sugiriendo que se incluya a futuro. El segundo modelo de ítems fue señalado por los expertos con este problema debido a que no se encuentra en el plan de estudios.

Hubo un grupo de seis indicadores que no fueron validados en 12.5% de los modelos de ítems (3, 5, 6, 10, 16 y 27). Estos corresponden a distintos aspectos como la importancia de la inclusión de los contenidos en el examen, el tipo de estrategia utilizada para evaluar si el

estudiante domina o no los contenidos, y el cumplimiento de las reglas y restricciones para la generación de ítems, marcadas en los modelos. Particularmente se señaló que uno de los modelos de ítems coincide en el incumplimiento de tres indicadores debido a que no está contemplado en el programa de estudios y aunque lo consideran esencial en el aprendizaje de la Química y de las Ciencias en general debido a que aborda las etapas del método científico, no se enseña en el aula y evaluarlo penalizaría injustamente a los estudiantes.

Finalmente, se observa en la gráfica que 12 indicadores fueron validados en 100% de los modelos de ítems y uno no fue evaluado debido a que el criterio no era aplicable a ninguno de los ocho modelos de ítems. Lo anterior significa que se cumplieron satisfactoriamente los aspectos de calidad relacionados con: a) la forma en que el contenido fue definido, delimitado y es representativo de los conocimientos y habilidades que deben dominar los estudiantes (indicadores 2, 7 y 9); b) la pertinencia de la estrategia evaluativa utilizada, el uso de textos auxiliares para evaluar los aprendizajes esperados de los estudiantes y la forma en que los ítems hijos reflejan que saben utilizar el conocimiento adquirido en el aula (indicadores 11, 13 y 17), y c) las características de los ítems hijos en cuanto a la representatividad del contenido, sencillez en el uso del lenguaje y tipo de ejecución solicitada del estudiante, así como ausencia de sesgo, errores de redacción o elementos que den pistas para su solución (indicadores 17, 20, 21, 22, 23, 24 y 26).

Resultados por problemas detectados en los modelos de ítems en lo individual

La Figura 4.21 muestra de manera gráfica una síntesis de los resultados de los modelos de ítems de Química, de acuerdo con el porcentaje de indicadores que no fueron validados por los expertos en cada uno de los ocho modelos evaluados. Siguiendo el tipo de estructura propuesta con anterioridad para presentar y analizar los resultados de Matemáticas e Historia, a continuación se examinarán a detalle aquellos modelos de ítems que obtuvieron un 20% o más de indicadores no validados.

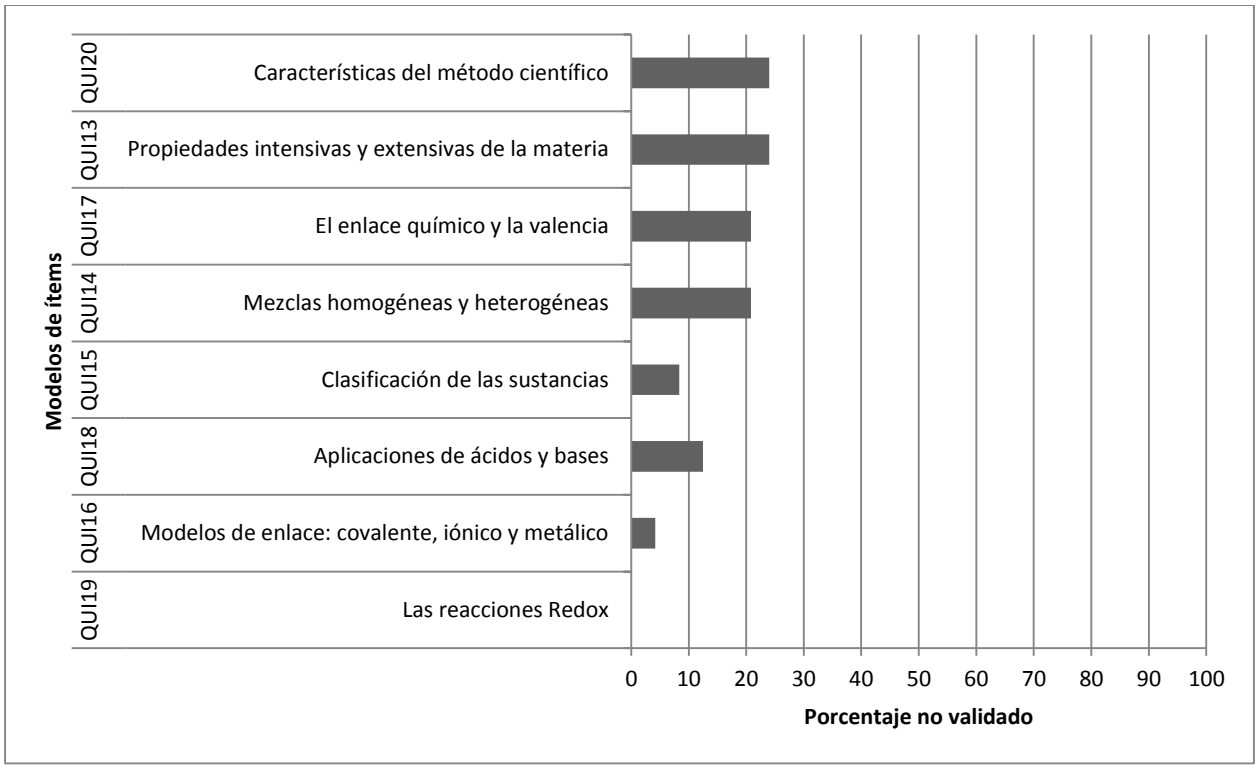


Figura 4.21. Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Química

Como puede observarse, hubo cuatro modelos de ítems que fueron identificados por los expertos con problemas en al menos 20% de los indicadores de calidad, los cuales se distribuyen en pares. En el primer par se encuentran los modelos QUI13 y QUI20, que presentaron problemas en 24% de los indicadores de calidad. El segundo par incluye a los modelos QUI14 y QUI17, con problemas en 20.8% de los indicadores.

El resto de los modelos de ítems se distribuyeron según la cantidad de indicadores que no fueron validados por los expertos: QUI18 presentó problemas en 12.5% (3) de los indicadores; QUI15 obtuvo 8.33% de indicadores no validados; QUI16 presentó problemas en 4.17% (1) de los indicadores, y QUI19 fue señalado por los expertos como libre de problemas. A continuación se presentan los detalles de los cuatro modelos de ítems que presentaron problemas en al menos 20% de los indicadores de calidad.

QUI13. Propiedades intensivas y extensivas de la materia

Este modelo de ítems genera ítems hijos que solicitan al estudiante que identifique y ubique en una escala un grupo de sustancias, según el grado en el que posean una propiedad intensiva (como la temperatura de fusión o densidad) o extensiva (como la masa o volumen). En la Figura 4.22, se muestra el ejemplo de un ítem en el cual se presentan diversas sustancias que deben ordenarse según posean una propiedad en menor o mayor grado. El estudiante deberá dar clic con el ratón sobre la imagen de la sustancia, moverla y depositarla en la ubicación de la escala que considere refleja el grado en el que posee la propiedad en cuestión.

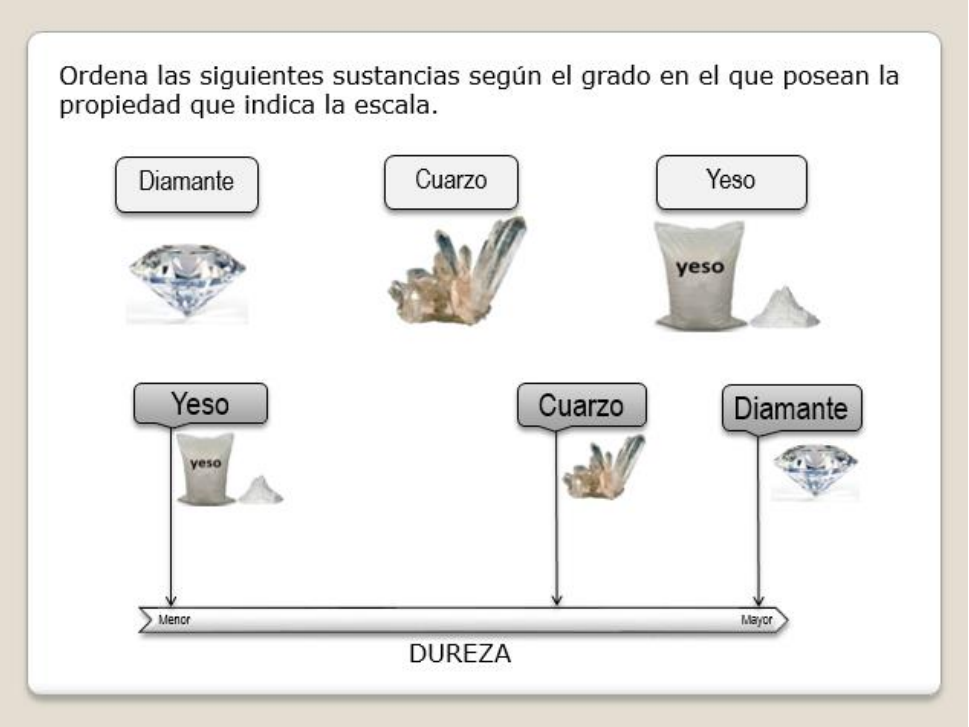


Figura 4.22. Ejemplo de ítem hijo del modelo QUI13: Propiedades intensivas y extensivas de la materia.

Para resolver este reactivo, el estudiante debe conocer cuáles son las propiedades intensivas y extensivas de la materia, de tal manera que pueda distinguir entre ellas y así caracterizar distintas sustancias mediante la identificación del grado en el que poseen cada

propiedad. Para emitir sus respuestas, debe seleccionar la imagen de cada una de las sustancias que se le presenten y colocarlas sobre un espacio de una escala, indicando por su ubicación si tienen menor o mayor presencia de una propiedad.

En lo que respecta al proceso de validación, la Figura 4.23 muestra para este modelo de ítems, las respuestas consensuadas por los expertos a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de los modelos de ítems.

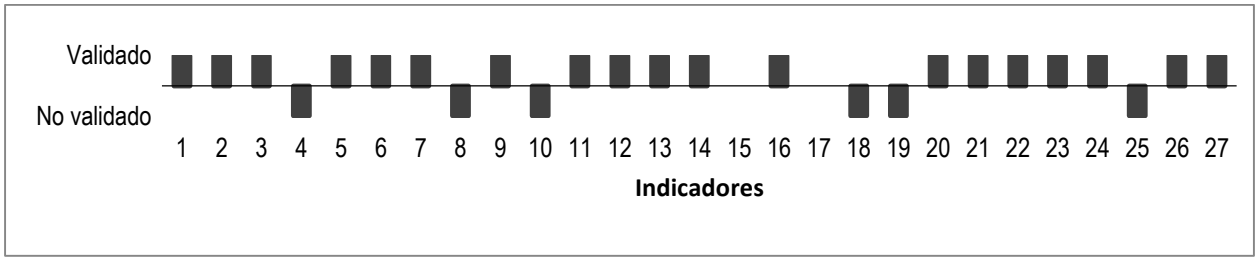


Figura 4.23. Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI13.

Se puede observar que hubo seis indicadores (4, 8, 10, 18, 19 y 25) que no fueron validados por los expertos. Entre los problemas que detectaron los expertos en este modelo de ítems, se encuentran aspectos abordados por los indicadores 4, 8 y 10. Al respecto mencionaron que aunque el contenido evaluado sí tiene relación directa con lo que se enseña en el aula, existe otra dimensión del mismo que no está siendo evaluada, ya que en el salón de clases también se solicita al alumno que realice estimaciones acerca de algunas propiedades de objetos y sustancias, y que utilice instrumentos de medición para la realización de cálculos y estimaciones precisas. En cuanto a la delimitación que se hizo del contenido, los expertos expresaron que esta no explicita cuáles propiedades de la materia serán evaluadas, señalando que en la definición del contenido se mencionan ocho, mientras que los ítems hijos generados abordan la evaluación de cuatro de ellas. Esto conlleva a que la estrategia evaluativa también esté planteada de manera errónea al mencionar la evaluación de ocho propiedades. En este sentido, los expertos sugirieron que se modifique la información contenida en la estrategia evaluativa, eliminando algunas de las

categorías, debido a que sería muy difícil evaluarlas correctamente sin proporcionar pistas al estudiante.

Por otro lado, este modelo de ítems indica que al generar los distintos ítems hijos se debe utilizar una serie de imágenes auxiliares para ejemplificar cada elemento del banco de elementos cadena, las cuales fueron revisadas por los expertos. Expresaron que consideran que todas son apropiadas a excepción de una (glicerina) en la cual se utiliza erróneamente un color grisáceo para mostrar la sustancia. Sugirieron que se modifique el color a uno que refleje transparencia, ya que la forma en que se enseña en el aula es indicando que se trata de una sustancia incolora (indicador 18).

Otro problema encontrado mediante la evaluación del indicador 19 fue que existe la presencia de dos tipos de problemas con los elementos cadena que fueron seleccionados del currículum para conformar el banco de información: nombres e imágenes de sustancias que pueden provocar confusión en el estudiante, y un ordenamiento incorrecto en el banco de elementos del modelo de ítems.

En cuanto al primer tipo de problemas, mencionaron que en la propiedad *densidad* el alumno debe tener un mayor dominio conceptual que en el resto de las propiedades de la materia, ya que necesita conocer con cierto grado de precisión los valores que cada sustancia tiene en este rubro para así poder ordenarlas correctamente. El resto de las propiedades de la materia están libres de esta situación. Para corregir lo anterior, los expertos sugirieron utilizar elementos cadena que contengan imágenes en las que se pueda identificar con mayor facilidad la densidad, tales como objetos sólidos disímiles.

Respecto al segundo tipo de problemas encontrados mediante la evaluación del indicador mencionado, señalaron la necesidad de reordenar los elementos de una de las categorías (densidad) por contener un error, y propusieron un nuevo orden. Finalmente agregaron que la

propiedad *viscosidad* usualmente no se incluye en las enseñanzas que marca el currículum, y sugirieron que se mantuviera en el banco de elementos, pero que se eliminaran algunas de las sustancias que lo conforman debido a que su proximidad en la gradiente puede resultar en confusión por parte de los estudiantes, situación que también surgió con un elemento asociado a la propiedad *masa,* en donde sugirieron que se sustituyera por otra sustancia.

Por último, la modificación de algunos de los elementos cadena mencionados en los párrafos anteriores, atenderá cabalmente la diferencia encontrada por los expertos en los niveles de dificultad de los ítems hijos, situación que fue detectada al revisar el indicador de calidad 25.

QUI20. Características del método científico

En este modelo de ítems, los ítems hijos generados evalúan si el estudiante conoce cuáles son los cuatro pasos básicos del método científico y los identifica por sus características. Un ejemplo de este tipo de ítems se muestra en la Figura 4.24.

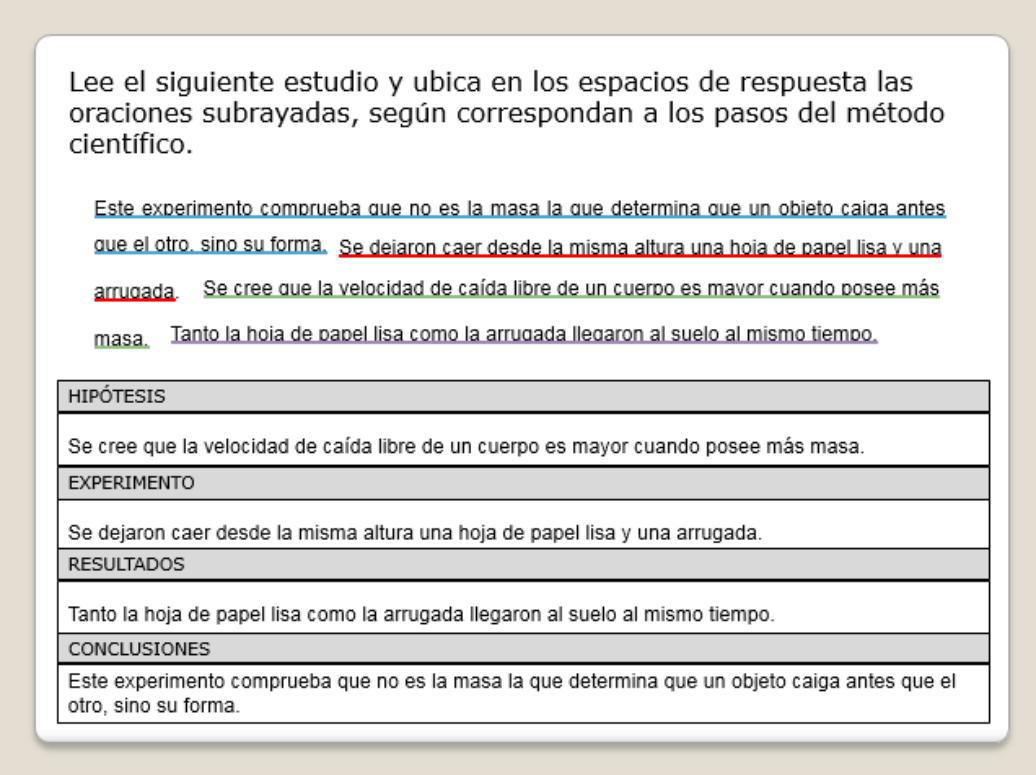


Figura 4.24. Ejemplo de ítem hijo del modelo QUI20: Características del método científico.

En la Figura 4.24 se presenta un texto dividido en distintos fragmentos, que describen los pasos de un experimento científico. Lo que se solicita al estudiante es que ordene los fragmentos pasos, clasificándolos según se trate de hipótesis, experimento, resultados o conclusiones. El estudiante deberá utilizar el ratón para seleccionar cada uno de los enunciados y moverlos hasta depositarlos en el lugar correspondiente a cada paso del método científico.

La Figura 4.25 muestra las respuestas consensuadas por los expertos, a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

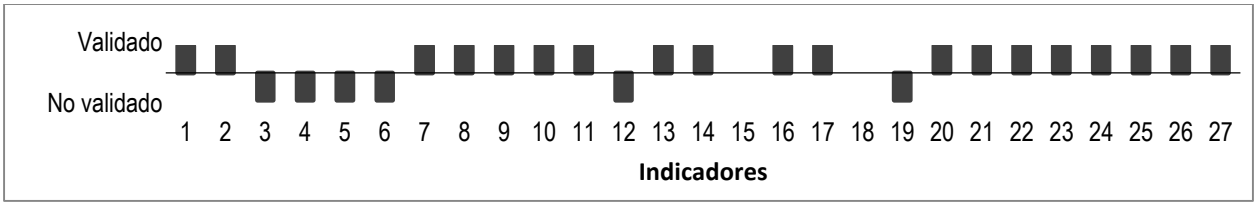


Figura 4.25. Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI20.

Hubo seis indicadores que no fueron validados por los jueces en este modelo de ítems, que señalan diversos problemas, entre los cuales destacan aquellos relacionados con la alineación y pertinencia curricular del contenido que evalúa, así como su importancia para la asignatura de Química (indicadores 3 al 6). Mencionaron que se trata de un contenido que no se encuentra integrado como tal en el programa de estudios de Química, puntualizando que sí forma parte de los contenidos del bloque de Ciencias, pero de la asignatura de Bilogía que se imparte durante el primer grado de secundaria. Señalaron que evaluarlo como parte de los contenidos de la asignatura de Química puede presentar ventaja de unos estudiantes sobre otros debido al interés particular de sus maestros por impartir el tema, aun sin ser oficial u obligatorio. Debido a lo anterior, agregaron que el aprendizaje de este contenido no es algo que se espera del promedio de los estudiantes.

Como consecuencia a lo mencionado en el párrafo anterior, la estrategia evaluativa utilizada para generar ítems hijos no presenta similitud con el proceso de enseñanza aprendizaje llevado a cabo en el aula (indicador 12). Adicionalmente, los expertos mencionaron que los elementos que fueron seleccionados del currículum para elaborar el banco de reactivos no son adecuados (indicador 19) debido a que no se contempla la enseñanza de este tema en el currículum actual; sin embargo enfatizaron que si el contenido se volviera a incluir en el programa, sí los consideran que adecuados y pertinentes.

QUI14. Mezclas homogéneas y heterogéneas

Este modelo de ítems genera ítems hijos que exploran el conocimiento de los diferentes tipos de mezclas y sus métodos de separación. El estudiante debe saber diferenciar entre las mezclas homogéneas y heterogéneas a partir del uso de diversos criterios de clasificación, entre los cuales se encuentran sus propiedades físicas y los métodos de separación. Así, los ítems hijos generados mediante este modelo de ítems solicitan al estudiante que clasifique mezclas según el método más apropiado para su separación. En la Figura 4.26 se muestra un ítem hijo similar a los que se generan en el Excoba, en el cual se presentan distintos nombres de mezclas que el estudiante debe clasificar según el criterio mencionado en la base del reactivo.

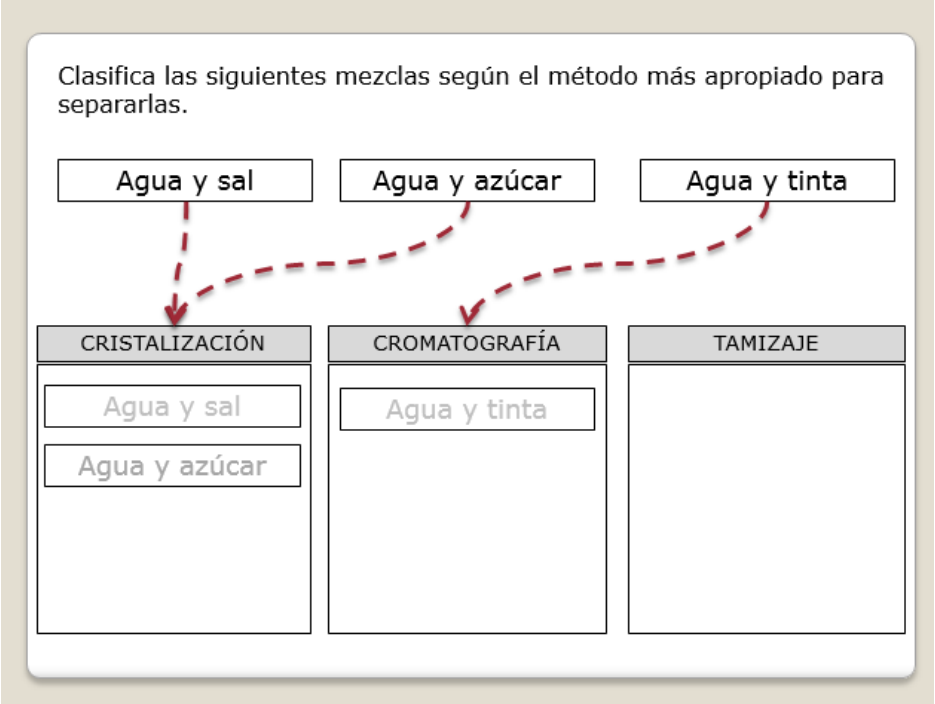


Figura 4.26. Ejemplo de ítem hijo del modelo QUI14: Mezclas homogéneas y heterogéneas.

Para resolver este ítem, el estudiante debe seleccionar cada mezcla con el ratón, mover el recuadro y depositarlo en el método de separación más apropiado para dicha mezcla.

En lo que respecta al proceso de validación, la Figura 4.27 muestra los indicadores en los que determinaron los expertos que el modelo de ítems QUI14 tenía problemas. Esto fue realizado mediante una discusión grupal en la que el dictamen final se emitió mediante la opinión consensuada de los presentes. Para ello hicieron uso del formato de validación en el cual se encontraban 27 indicadores de calidad técnica.

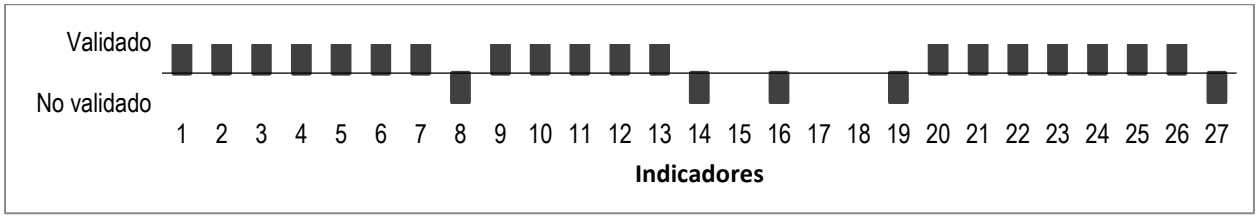


Figura 4.27. Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI14.

Se observa que hubo cinco indicadores que no fueron validados por los expertos, los cuales atienden aspectos relacionados con la alineación curricular del contenido, la calidad de la base del reactivo, la congruencia entre las reglas para generar los ítems y la estrategia evaluativa, así como la calidad de los elementos contenidos en el banco de información.

En lo concerniente a la alineación existente entre la definición del contenido curricular y el segmento de contenido que fue delimitado para ser evaluado (indicador 8), los jueces realizaron dos recomendaciones: 1) que se incluya en la delimitación del contenido y por ende en la estrategia evaluativa el método de separación que tradicionalmente es utilizado para enseñar la separación de mezclas homogéneas (cromatografía), y 2) que se excluya uno de los métodos que sí se consideraron en la delimitación del contenido, ya que puede generar confusión en los estudiantes (decantación).

También mencionaron que la base del reactivo no es lo suficientemente clara para que el estudiante emita su respuesta (indicador 14), ya que consideraron que es necesario incluir en su

redacción la explicación de que puede haber categorías en las cuales no se coloque ningún elemento.

Por otro lado, en cuanto a las reglas establecidas para generar ítems hijos (indicador 16), los expertos mencionaron que es muy importante evitar que se presenten algunos de los métodos de separación en el mismo ítem, para lo cual sugirieron que se agregara esta restricción al generador de ítems. También mencionaron que los elementos cadena que conforman el banco de información presentan dos tipos de problemas (indicador 19): 1) algunos de ellos se encuentran mal clasificados, para lo cual presentaron una propuesta, y 2) otros no son utilizados como ejemplos en el aula, abriendo la posibilidad de confusión en los estudiantes debido a su elevado nivel de dificultad; la sugerencia fue su eliminación del banco de elementos y presentaron una propuesta para sustituirlos.

Por último, señalaron la presencia de un método de separación en el banco de elementos que puede generar confusión en los estudiantes debido a que su proceso implica el uso de métodos de separación adicionales y de diversa naturaleza. La sugerencia fue eliminarla debido a que representa un nivel de complejidad muy distinto al del resto de los métodos.

QUI17. El enlace químico y la valencia

Este modelo de ítems genera ítems hijos que exploran expresiones de ecuaciones químicas en las que se utiliza el principio de la conservación de la masa y valencia. El estudiante debe mostrar su dominio del modelo de enlace químico y la transferencia de electrones, con la finalidad de balancear ecuaciones. En la Figura 4.28 se muestra un ítem hijo similar a los que se generan en el Excoba, en el cual se presenta una ecuación química a balancear.

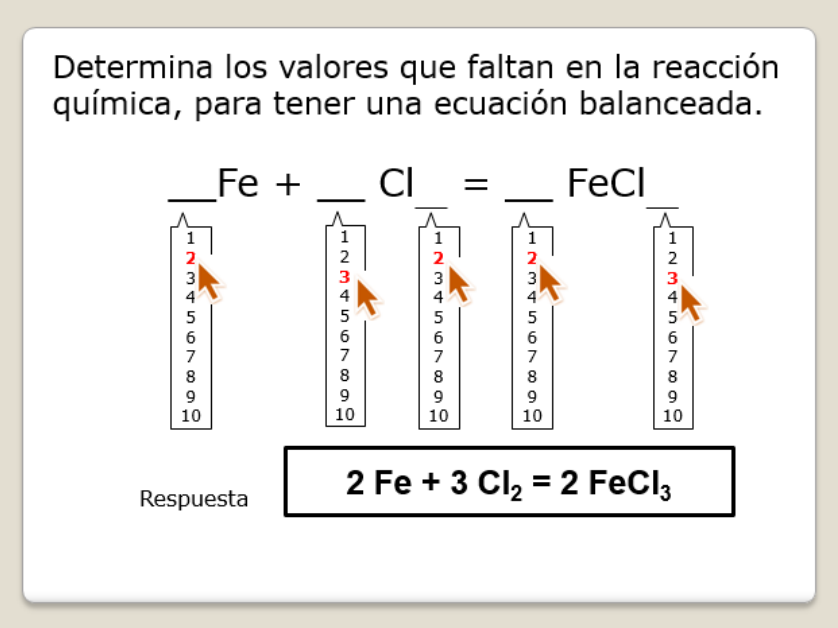


Figura 4.28. Ejemplo de ítem hijo del modelo QUI17: El enlace químico y la valencia.

Para resolver este ítem, el estudiante debe utilizar un método para identificar cuándo se ganan o pierden electrones; por lo tanto, deberá conocer la carga de los átomos dentro de las moléculas, siguiendo reglas específicas asociadas al método. Para responder al reactivo, debido a que es de tipo selección de elementos, el estudiante deberá ubicar el puntero del ratón y dar clic en cada uno de los espacios donde hace falta el número de la valencia. Conforme selecciona las valencias, la respuesta irá apareciendo en la interfaz.

En lo que respecta al proceso de validación, la Figura 4.29 muestra las respuestas consensuadas por los expertos a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de los modelos de ítems.

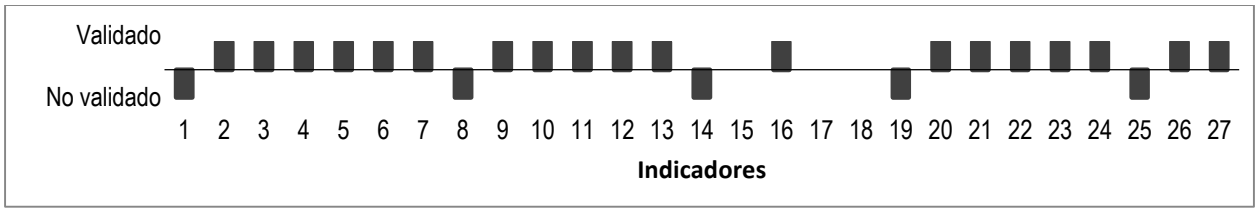


Figura 4.29. Opinión de los expertos en los 27 indicadores validados del modelo de ítems QUI17.

Se puede observar que hubo cinco indicadores que no fueron validados por los expertos (1, 8, 14, 19 y 25). Entre los problemas que presenta este modelo de ítems, mencionaron que la definición del contenido no se presenta de manera clara ni precisa (indicador 1), ya que incluye la descripción de elementos que no son evaluados mediante el modelo de ítems, tales como la relación entre un modelo químico y su fórmula; ante esto sugirieron que para resolver dicha situación, se requiere eliminar el primer segmento de la definición.

También señalaron que aunque la delimitación del contenido se encuentra alineada y deriva de su definición (indicador 8), requiere mayor precisión en cuanto al tipo de procedimiento que se espera utilice el estudiante al responder a los ítems hijos. El estudiante usualmente balancea ecuaciones utilizando el método de tanteo y no por procedimientos en los que se utilicen valencias, mencionando que es debido a que este procedimiento es mucho más complicado. La sugerencia de los expertos fue incluir en la delimitación del contenido esta precisión.

Otro aspecto encontrado fue que la forma en que se encuentra redactada la base del reactivo es incorrecta (indicador 14), en el sentido de que al enseñar este contenido se utiliza el término *coeficientes* y no *valores,* como se indica en la base. Los expertos sugirieron que se sustituyan los términos para evitar confusiones en el estudiante.

En cuanto a la correspondencia de los elementos que conforman el banco de información (indicador 19), los expertos señalaron que uno de ellos no es pertinente, ya que evalúa un

contenido de mayor nivel de complejidad que rebasa los aprendizajes esperados en los estudios de la educación secundaria. De igual manera, identificaron que existe otro elemento (ecuación) cuyo nivel de dificultad es inapropiado para el grado escolar de los examinados, ya que requiere mayor tiempo y esfuerzo para balancearla (indicador 25). Opinaron que cuando los estudiantes se enfrenten a un ítem hijo que contenga estos elementos no aplicarán su conocimiento, sino que resolverán el ítem mediante el método de tanteo, sustituyendo uno por uno los numerales hasta llegar al correcto. Sugirieron la sustitución de estos elementos y presentaron una propuesta para ello.

Resultados conjuntos

Además de los análisis realizados en los apartados anteriores, se contrastó la información derivada del trabajo con el panel de expertos, de tal manera que se tomaron los cuatro modelos de ítems que tuvieron por lo menos 20% de los indicadores de calidad no validados, y los ocho indicadores que no fueron validados por los expertos en al menos 20% de los modelos de ítems de toda la asignatura. Se detectó la presencia de regularidades en el tipo de problemáticas que presentan. La Tabla 4.16 muestra esta información.

Tabla 4.16.
Modelos de ítems de Química y regularidades que presentaron

No.	Indicador	Modelos de ítems				Frecuencia relativa
		QUI13 Propiedades intensivas y extensivas de la materia	QUI14 Mezclas homogéneas y heterogéneas	QUI17 El enlace químico y la valencia	QUI20 Características del método científico	
19	Banco de información corresponde al contenido seleccionado	×	×	×	×	4 de 4
8	Delimitación del contenido alineada y derivada de definición	×	×	×	----	3 de 4
14	Base del reactivo clara y suficiente para emitir respuesta	----	×	×	----	
25	Ítems hijos con nivel de dificultad adecuado y apropiados para el grado escolar del contenido evaluado	×	----	×	----	2 de 4
4	Contenido coherente con lo que se enseña en aula	×	----	----	×	
1	Definición de contenido clara y precisa	----	----	×	----	
12	Estrategia evaluativa semejante a como se enseña en aula	----	----	----	×	1 de 4
18	Gráficos e imágenes apropiados	×	----	----	----	

Nota: Se muestra únicamente la información de los modelos de ítems que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Química. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

Se observa que existen coincidencias entre varios modelos de ítems, respecto a los tipos de problemas presentes en ellos. Por ejemplo, los cuatro modelos de ítems: QUI13 (Propiedades intensivas y extensivas de la materia), QUI14 (Mezclas homogéneas y heterogéneas), QUI17 (El enlace químico y la valencia) y QUI20 (Características del método científico), además de tener al menos 20% de los indicadores de calidad no validados, coincidieron en que tienen problemas con la correspondencia del banco de información con el contenido evaluado (indicador 19).

De igual manera, los modelos de ítems QUI13 (Propiedades intensivas y extensivas de la materia), QUI14 (Mezclas homogéneas y heterogéneas) y QUI17 (El enlace químico y la valencia), no fueron validados en el indicador 8, el cual explora problemas específicos de correspondencia y alineación entre el contenido seleccionado del currículum para ser evaluado en el examen y el segmento delimitado para generar ítems hijos.

Así mismo, se detectó que los modelos de ítems QUI14 (Mezclas homogéneas y heterogéneas) y QUI17 (El enlace químico y la valencia), coincidieron en que no fue validado el indicador 14, el cual señala problemas de falta de claridad y suficiencia en la base del reactivo.

Otra coincidencia que surgió de este análisis fue que los modelos de ítems QUI13 (Propiedades intensivas y extensivas de la materia) y QUI17 (El enlace químico y la valencia), presentaron problemas en cuanto a los niveles de dificultad y correspondencia al grado escolar de los contenidos que se enseñan en secundaria (indicador 25). Lo anterior implica que los ítems hijos que se generan a partir de estos modelos contienen elementos cadena o integrales que producirán ítems con distintos niveles de dificultad o que no son apropiados para el nivel educativo debido a que no se incluyen en las enseñanzas de la asignatura.

Por último, también se observa que en los modelos de ítems QUI13 (Propiedades intensivas y extensivas de la materia) y QUI20 (Características del método científico), no se validó el indicador 4, lo cual significa que en el caso de estos dos modelos, los expertos consideran que el contenido evaluado no es coherente con los contenidos que se enseñan en el aula.

En lo que respecta a los indicadores 1, 12 y 18, aunque no fueron validados por los expertos en más de 20% de los modelos de ítems del área de Química, en el caso de estos cuatro modelos no hubo coincidencias.

Resultados según las sugerencias del panel de expertos para el mejoramiento del área

Adicionalmente a los problemas que fueron detectados por los expertos, hubo una serie de sugerencias que estos realizaron para el mejoramiento del área de Química del Excoba.

Entre ellas, se destaca su apreciación general respecto a la pertinencia de los contenidos seleccionados del currículum. Estimaron que en conjunto, 75% (6) de los modelos de ítems de la asignatura de Química que evaluaron representan las habilidades y conocimientos básicos que deben dominar los estudiantes al término de la secundaria. Esto fue debido a que el modelo de ítems QUI17, elaborado para evaluar el contenido Enlace químico y valencia, tiene un nivel de dificultad mayor a lo que el estudiante usualmente resuelve; por otro lado el modelo de ítems QUI20, que evalúa el contenido *Características del método científico*, no está incluido en el programa de estudios, y puntualizaron que sí consideran muy importante tanto su enseñanza como evaluación.

Por otro lado, consideraron que existe un contenido muy importante que no contempla el examen y debería incluirse debido a su trascendencia para el aprendizaje de la asignatura. Argumentaron que el Programa de estudios vigente (2011) incluye en el Bloque II, llamado *Propiedades de los materiales y su clasificación*, el Tema *Tabla periódica*, con el Subtema *Regularidades en la tabla periódica de los elementos químicos representativos*. En este sentido, comentaron que se debe elaborar un modelo de ítems que evalúe dicho contenido y sugirieron que los ítems hijos que se generen soliciten al estudiante que clasifique los elementos químicos en metales y no metales.

Finalmente, al solicitar a los expertos la estimación de qué porcentaje de estudiantes que concluyen la secundaria consideran que responderían correctamente a los ítems hijos generados mediante los ocho modelos de ítems que conforman la asignatura de Química, mencionaron que si se trata de estudiantes provenientes de escuelas públicas estiman que de 50 a 60% lo harían,

pero que este porcentaje ascendería a 70 u 80% si provinieran de escuelas privadas. Esta estimación fue hecha considerando que entre los aprendizajes esperados de los estudiantes, marcados en el currículum, se encuentran solamente siete de los ocho contenidos evaluados por el examen.

Resultados según la relación entre los problemas detectados y los contenidos evaluados

Una vez que se identificó el bloque temático curricular al que pertenece cada modelo de ítems, se procedió a realizar un análisis para detectar si existe algún tipo de relación entre el contenido curricular que evalúan los modelos de ítems que presentaron mayor cantidad de indicadores no validados ($\geq 20\%$) y el tipo de problema expresado en los indicadores que no fueron validados en 20% o más de los modelos de ítems. La Tabla 4.17 resume los hallazgos.

Tabla 4.17.
Modelos de ítems, bloques temáticos e indicadores no validados con mayor frecuencia en Química

Modelos de ítems	Contenido	Bloque temático	No. de indicador							
			1	4	8	12	14	18	19	25
QUI13	Propiedades intensivas y extensivas de la materia	I Las características de los materiales	--	×	×	--	--	×	×	×
QUI14	Mezclas homogéneas y heterogéneas	I Las características de los materiales	--	--	×	--	×	--	×	--
QUI17	El enlace químico y la valencia	III La transformación de los materiales: la reacción química	×	--	×	--	×	--	×	×
QUI20	Características del método científico	Otro no contemplado en el programa de la asignatura	--	×	--	×	--	--	×	--

Nota: Se muestra únicamente la información de los modelos de ítem que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Química. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

De los cuatro modelos de ítems que cumplieron con el criterio de 20% o más indicadores no validados por los expertos, se puede observar que dos corresponden al bloque temático I *Las características de los materiales* (QUI13 y QUI14), uno pertenece al bloque III *La transformación*

de los materiales: la reacción química, mientras que el restante proviene de un bloque temático que no está contemplado en el programa de la asignatura de Química.

En el primer caso (QUI13 y QUI14), la mitad de los modelos de ítems con un alto número de indicadores no validados (y por ende problemas) generan ítems hijos en los que se requiere que el estudiante identifique las características fundamentales del conocimiento científico y tecnológico, tales como la experimentación e interpretación, abstracción y generalización. En estos modelos de ítems, los contenidos evaluados requieren un buen manejo del conocimiento de las características de algunos modelos químicos, entre las cuales destacan la abstracción o generalización, lenguaje matemático, precisión, brevedad, así como sus alcances y limitaciones. Particularmente se busca que el estudiante inicie el estudio de los materiales y los primeros sistemas de clasificación de sustancias, de tal manera que logre identificar los fundamentos básicos de las técnicas que acompañan a la investigación científica.

El tipo de problemática detectada por los expertos en estos dos modelos de ítems está relacionada con la coherencia, representatividad y pertinencia tanto de los contenidos delimitados y evaluados por estos dos modelos de ítems, como de los elementos que fueron seleccionados del currículum como cadena e integrales y que son utilizados para generar los ítems hijos. Los expertos estimaron que en ambos modelos, la delimitación del contenido no representa adecuadamente lo que debe evaluarse, ya que no es lo suficientemente explícita; mencionaron que tampoco refleja adecuadamente los componentes conceptuales que deben conformarla y por ende, ser incluidos en la evaluación. En cuanto a los elementos cadena e integrales que conforman el banco de información curricular, puntualizaron la presencia de errores y elementos descontextualizados, ya que no son utilizados durante el proceso de enseñanza-aprendizaje, comentando que es importante atender esta situación, ya que de no hacerlo habría confusión en los estudiantes y los resultados de la evaluación no serían del todo confiables.

Lo anterior sugiere que al menos en los dos modelos de ítems que pertenecen a este bloque de análisis, sí existe relación entre el tipo de contenido al que corresponden en el currículum y la naturaleza de la problemática señalada por los expertos. Esta información indica la necesidad de realizar una revisión detallada de estos modelos de ítems en los siguientes aspectos: la delimitación del contenido y su congruencia con los elementos del generador de ítems, la correspondencia del banco de elementos cadena e integrales con el contenido evaluado, la coherencia entre los contenidos evaluados y la forma en que se enseña en el aula, y el balanceo y correspondencia de los niveles de dificultad de los ítems hijos con el grado escolar de los estudiantes examinados. En el caso de los otros dos modelos de ítems no hubo coincidencia en la organización curricular y el tipo de problemas encontrados por los expertos.

4.2.4. **Modelos de ítems del área de Español**

Al igual que en las tres áreas del Excoba que fueron analizadas en los apartados anteriores, los resultados presentados para el área de Español inician con la sección que corresponde a la Fase II del MVCE y describe las etapas 2 y 3 del mismo. Primero se presentan los resultados que describen las características de los modelos de ítems que conforman el área, de acuerdo con cuatro características: tipo de contenido curricular, la clasificación que tienen en el sistema informático que administra el examen, la ejecución que debe realizar el estudiante al responder, y el tipo de conocimiento que explora. Segundo, se presenta el análisis derivado del trabajo con el panel de expertos.

Etapa 2: Características de los modelos de ítems de Español

Los resultados se agrupan en cuatro niveles de análisis: 1) según los contenidos curriculares de los que se derivan los modelos de ítems; 2) por la clasificación en el sistema informático GenerEx; 3) por la ejecución que el estudiante debe realizar para responder al ítem hijo, y 4) por el tipo de conocimiento evaluado (declarativo, procedimental, esquemático o estratégico.

Tipo de contenidos curriculares

Para la elaboración de esta sección del GAI Excoba se utilizó el programa de estudios de Español de educación secundaria del año 2006. El eje central en la definición de los contenidos de la asignatura es la conceptualización de las prácticas sociales del lenguaje como modos de interacción que enmarcan la producción e interpretación de textos. Se considera que es mediante el aprendizaje del uso, producción e interpretación de textos e intercambios orales, los individuos aprenden a interactuar. Estas prácticas sociales se encuentran agrupadas en tres ámbitos que se encuentran presentes en los tres grados escolares en que se enseña la asignatura de Español: Estudio, Literatura y Participación ciudadana[4] (SEP, 2006b). La Figura 4.30 muestra las prácticas del lenguaje que se trabajan a lo largo de los tres grados escolares, su organización por ámbitos, así como las prácticas generales y específicas que se exigen en cada uno de ellos.

[4] **Estudio**. Implica una actitud atenta y reflexiva respecto al contenido de los textos y sus modos de expresión, un intercambio oral formal y un dominio preciso de la expresión escrita. Se espera que el estudiante aprenda a expresarse de manera oral y escrita en un lenguaje formal y académico, que se apropie del discurso en que se expresan las áreas de las ciencias y humanidades, y que mediante la producción de textos logre desarrollar habilidades para preparar información y exponerla, expresar ideas con claridad, organizar un texto coherente y utilizar vocabulario técnico y especializado. Se trabajan con mayor énfasis los temas relacionados con la estructura sintáctica y semántica de los textos, su organización gráfica y la puntuación.

Literatura. El trabajo de la asignatura en este ámbito busca fomentar una actitud libre y creativa en la que se descubra el poder creador de la palabra y se experimente el goce estético que la variedad de formas y ficción literaria pueden producir. Se enfatiza la intención creativa e imaginativa del lenguaje mediante prácticas de lectura en las que se siguen temáticas o movimientos literarios, con la finalidad de que el estudiante compare los patrones del lenguaje.

Participación ciudadana. En este ámbito se busca que los estudiantes reflexionen sobre la dimensión ideológica y legal de la palabra, el poder de las leyes y regulaciones sociales, así como la participación mediada por el diálogo. Incluye lectura y uso de documentos administrativos y legales, así como aquellos que impliquen la expresión y defensa de la opinión personal y el desarrollo de una actitud crítica ante la información que se recibe de los medios de comunicación. En este ámbito se busca que los estudiantes aprendan a interpretar textos e imágenes en circulación, identificar valores y formas de vida que los medios de comunicación difunden, y a descubrir su postura ideológica, entre otras cosas.

A su vez, los contenidos están agrupados en cinco bloques que especifican de manera fina las actividades que se deben realizar, así como los temas en que el estudiante debe reflexionar para lograr los aprendizajes esperados en cada uno de ellos.

ESTUDIO

1.- Obtener y organizar información.

2.- Revisar y reescribir textos producidos en distintas áreas de estudio.

3.- Participar en eventos comunicativos formales.

Prácticas generales	Prácticas específicas por grado escolar		
	1°	2°	3°
1	• Buscar, seleccionar y registrar información de distintos textos. • Escribir resúmenes como apoyo al estudio o al trabajo de investigación.	• Seleccionar, comparar y registrar información de distintos textos. • Utilizar la entrevista como medio para obtener información.	• Leer y comparar diferentes tratamientos de un mismo tema. • Comunicar información obtenida mediante entrevistas.
2	• Escribir un texto que integre información de resúmenes y notas. • Revisar informes sobre observaciones de procesos.	• Escribir la biografía de un personaje.	• Revisar y reescribir informes sobre experimentos.
3	• Exponer los resultados de una investigación.	• Organizar mesas redondas sobre temas investigados previamente.	• Participar en debates sobre temas investigados previamente.

PARTICIPACIÓN CIUDADANA

1.- Leer y utilizar distintos documentos administrativos y legales.

2.- Investigar y debatir sobre la diversidad lingüística.

3.- Analizar y valorar críticamente a los medios de comunicación.

Prácticas generales	Prácticas específicas por grado escolar		
	1°	2°	3°
1	• Escribir cartas para hacer aclaraciones o presentar reclamos. • Explorar, leer y participar en la elaboración de reglamentos.	• Explorar los documentos que acreditan la propiedad de bienes o validez de transacciones comerciales. • Analizar el papel de diversos documentos nacionales e internacionales para garantizar derechos y obligaciones de los ciudadanos.	• Utilizar documentos para presentar solicitudes. • Participar en la solución de problemas de la escuela o comunidad.
2	• Investigar sobre la diversidad lingüística y cultural de los pueblos indígenas de México.	• Investigar la diversidad lingüística y cultural de los pueblos hispanohablantes.	• Investigar la diversidad lingüística y cultural de los pueblos del mundo.
3	• Hacer encuestar sobre el uso de los medios de comunicación. • Dar seguimiento y comentar programas televisivos de divulgación de las ciencias, cultura y artes.	• Realizar seguimiento de noticias en medios de comunicación y hacer análisis comparativo. • Grabar programa en audio o video. • Leer y escribir reportajes.	• Realizar encuestas sobre la influencia de la publicidad. • Analizar mensajes publicitarios de diversos medios. • Leer y escribir artículos de opinión.

LITERATURA

1.- Leer y escribir para compartir la interpretación de textos literarios.

2.- Hacer el seguimiento de algún subgénero, temática o movimiento.

3.- Leer para conocer otros pueblos.

4.- Escribir textos con propósitos expresivos y estéticos.

5.- Participar en experiencias teatrales.

Prácticas generales	Prácticas específicas por grado escolar		
	1°	2°	3°
1	• Compartir poemas de la lírica tradicional.	• Reseñar una novela.	• Elaborar y prologar antologías
2	• Hacer el seguimiento de un subgénero narrativo: cuento de terror, de ciencia ficción, policiaco o algún otro.	• Hacer el seguimiento de una temática en textos literarios.	• Hacer el seguimiento de un periodo o movimiento poético.
3	• Investigar sobre relatos míticos y leyendas de distintos pueblos.	• Leer cuentos de la narrativa latinoamericana de los siglos XIX y XX.	• Leer una obra del español medieval o renacentista.
4	• Escribir poemas tomando como referente los movimientos de vanguardia del Siglo XX.	• Escribir cuentos.	• Escribir autobiografía.
5	• Leer obras dramáticas contemporáneas breves. • Escribir una obra corta para ser representada.	• Escribir un guion de teatro a partir de un texto narrativo.	• Leer en atril una obra de teatro del Siglo de Oro.

Figura 4.30. Distribución de los contenidos de la asignatura Español. En los óvalos se encuentran los ámbitos y las prácticas generales que se realizan durante el proceso de enseñanza-aprendizaje. Las flechas y números rojos indican cómo cada práctica general establece de manera transversal y para cada ámbito, las directrices para operar prácticas específicas en cada uno de los tres grados escolares en que se imparte la asignatura.

Para fines de análisis, los modelos de ítems del Excoba se agruparon en los tres ámbitos en que se encuentra organizado el currículum, ya que estos permean de manera transversal a todos los contenidos y aprendizajes esperados de los estudiantes en cuanto al dominio de la asignatura Español de secundaria. El análisis no se abordó siguiendo la organización del currículum en prácticas generales, específicas o bloques, ya que no permitieron encontrar coincidencias entre los distintos tipos de problemas con la calidad técnica de los modelos de ítems y su organización curricular. La Tabla 4.18 muestra la distribución de los veinte modelos de ítems siguiendo esta clasificación.

Tabla 4.18.
Distribución de modelos de ítems de Español por ámbito de estudio

Modelos de ítems	Contenido	Ámbito	Distribución por ámbito (porcentaje)
ESP01	Preguntas de acuerdo con propósitos específicos	Estudio	45
ESP02	Estructura del párrafo		
ESP03	Gráficos, esquemas y diagramas en textos		
ESP04	Evaluación de información sobre un mismo tema, de diversas fuentes		
ESP05	Argumentación		
ESP06	Recursos lingüísticos: nexos y adverbios		
ESP07	Recursos lingüísticos: vocabulario técnico, voz pasiva, tiempos verbales y forma impersonal		
ESP08	Discurso directo e indirecto		
ESP09	Acento diacrítico y enfático		
ESP10	Géneros literarios	Literatura	20
ESP11	Recursos de cohesión		
ESP12	Poemas		
ESP13	Predicción del contenido de una obra		
ESP14	Contraste de mensajes en las noticias	Participación ciudadana	35
ESP15	Reglamento o instructivo		
ESP16	Propósitos y argumentos de la carta formal		
ESP17	Documentos que establecen derechos y obligaciones		
ESP18	Propósitos comunicativos y puntos de vista		
ESP19	Llenado de formatos		
ESP20	Recursos lingüísticos y visuales en los mensajes publicitarios		

Se puede apreciar que nueve (45%) de los modelos de ítems evalúan contenidos del ámbito *Estudio*, cuatro (20%) abordan la evaluación de contenidos relacionados con el ámbito *Literatura*, y siete (35%) corresponden al ámbito *Participación ciudadana*.

Características del sistema informático GenerEx

Como se revisó en los apartados anteriores, el sistema informático GenerEx, creado para la captura de datos, administración, aplicación y calificación de los exámenes que se generan mediante la información contenida en los modelos de ítems, cuenta con la posibilidad de clasificar 22 diferentes tipos de modelos de ítems, según la programación que se requiera para su correcto funcionamiento y visualización en la interfaz gráfica del GAI Excoba.

En el caso del área de Español del Excoba, se utilizaron cuatro diferentes tipos de reactivos según esta clasificación (ver Tabla 4.19).

Tabla 4.19.
Clasificación y distribución de modelos de ítems de Español en el GenerEx

	Total	Porcentaje
Elemento categoría	11	55
Selección elementos	6	30
Frase imagen	2	10
Selección frases	1	5

La distribución que se presentó en los 20 modelos de ítems, muestra que 55% (11) implica una programación que genera ítems hijos que permiten realizar la clasificación de elementos dentro de categorías; 30% (6) de los modelos

El tipo de modelos de ítems que destaca en este grupo es el denominado *Elemento categoría*, ya que 55% (11) de los modelos de ítems de Español requieren una forma de programación en la que el sistema construye ítems hijos mediante la selección aleatoria de un número de categorías, siguiendo la reglas establecidas por los diseñadores del examen. También

selecciona una cantidad preestablecida de elementos asociados a cada categoría, de tal manera que cuando el estudiante interactúa con la interfaz responderá colocando elementos dentro de categorías. El sistema califica haciendo una comparación entre el lugar en el que fueron colocados los elementos, y las respuestas que se encuentran alimentadas en él y registradas como correctas.

En los modelos de ítems del tipo *Selección de elementos*, se encontró que 30% (6) corresponden a esta categoría. En ellos se presentan textos (párrafos, enunciados, oraciones, palabras) que contienen segmentos marcados en los que se debe seleccionar una o varias palabras. Al hacer clic sobre cada espacio marcado, el sistema despliega una ventana con distintas opciones para que el estudiante elija la que sustituya o complete el texto con la información correcta.

Hubo 10% (2) de modelos de ítems clasificados como de tipo *Frase imagen*. Este tipo de modelos implican que el sistema contiene un banco de información con distintos tipos de textos, de los cuales seleccionará uno al azar. Cada texto cuenta con segmentos (palabras, frases o enunciados) que funcionan como elementos cadena que se podrán intercambiar y presentar en distinto orden. Siguiendo las reglas establecidas en el modelo de ítems, se genera el ítem hijo y se muestra al estudiante en la interfaz, donde deberá seleccionar, mover mediante el uso del ratón y depositar en el espacio que considere refleja el orden o lugar correcto de los segmentos que conforman el texto.

Por último, se puede observar que 5% (1) de los modelos de ítems fue de tipo *Selección frases*. Para operar los modelos de ítems de este tipo el sistema debe contener un banco de textos, del cual seleccionará uno para mostrar en la interfaz. Todos los segmentos en los cuales está dividido el texto (palabras, frases o enunciados) deben estar marcados, de tal manera que el estudiante seleccione únicamente el que responde a lo que se plantea en la base del reactivo.

Tipo de ejecución solicitada

Los modelos de ítems que fueron validados por los expertos implicaron estrategias evaluativas que generan distintos tipos de ejecución por parte del estudiante cuando emite sus respuestas en los ítems hijos. En la Tabla 4.20 se presenta la distribución de los mismos, según fueron de arrastre de elementos, selección de elementos, escritura libre, o una combinación de selección de elementos con escritura libre.

Tabla 4.20.
Distribución de modelos de ítems de Español, según el tipo de ejecución que demandan del estudiante

Tipo de ejecución	Modelo de ítems	
	Total	Porcentaje
Arrastre	13	65
Selección	7	35
Escritura	-----	-----
Mixta (selección y escritura)	-----	-----

Como puede observarse, en el área de Español la mayoría de los modelos de ítems, es decir, 65% (13), genera ítems hijos en los que la ejecución implica que el estudiante haga uso del ratón para el movimiento y colocación de elementos dentro de categorías o secciones de la interfaz. El resto de los modelos de ítems representan 35% (7) y requieren que el estudiante seleccione segmentos de la interfaz, también mediante el uso del ratón.

En el caso de los modelos de ítems que requieren una ejecución el estudiante escriba su respuesta o realice una combinación de acciones, no se presentó ningún caso.

Tipo de conocimientos evaluados

La distribución de los modelos de ítems según el tipo de conocimiento que exploran los ítems hijos generados, es la que se muestra en la Tabla 4.21.

Tabla 4.21.
Tipo de conocimiento que evalúan los modelos de ítems de Español

Conocimiento evaluado	Modelos de ítems	
	Total	Porcentaje
Declarativo	13	65
Procedimental	1	5
Esquemático	6	30
Estratégico	-----	-----

Se puede observar que 65% (13) de los modelos de ítems evalúan principalmente conocimientos de tipo declarativo, en los que se requiere un manejo organizado de la información previamente aprendida y recuperación de detalles aislados, mediante la clasificación de datos y elementos conceptuales, de tal forma que las respuestas dadas expresen los principios teóricos y conceptuales que el estudiante aprendió en la escuela y manifieste que puede identificar los elementos implicados en la resolución de los ítems.

Así por ejemplo, en el modelo de ítems ESP01 que aborda la evaluación de la identificación y discriminación de los elementos esenciales para la búsqueda de información de un tema específico, el estudiante debe contar con un manejo conceptual y teórico de los elementos que le ayudarán a responder al ítem hijo, de tal manera que discrimine entre aquellos que le sirven y los que no.

En el caso de los conocimientos de tipo esquemático, 30% (6) de los modelos de ítems evalúan aspectos en los que el estudiante debe manejar la información de manera ordenada y sistemática y utilizar cuerpos de conocimiento que le permitan llegar a explicaciones y razonamientos, de tal manera que sus respuestas reflejen que conoce los motivos por los cuales ocurren los fenómenos. Por ejemplo, en el modelo de ítems ESP02 que genera ítems hijos que evalúan el conocimiento que tiene el estudiante respecto a las oraciones temáticas dentro de un párrafo, se utilizan textos breves en los cuales deberá identificar y señalar la oración que considere sea representativa de la idea temática del texto. Para resolver correctamente este tipo

de ítems, el estudiante debe contar con un cuerpo de conocimiento suficiente sobre las características de las oraciones temáticas dentro de un párrafo, y que a su vez le permita saber el motivo por el cual su elección fue la correcta.

Por otro lado, se observa que 5% (1) de los modelos de ítems evalúa conocimientos de tipo procedimental, en los que se requiere que el estudiante utilice su conocimiento de métodos y procedimientos que implican el seguimiento de reglas o pasos para obtener un resultado determinado. Por ejemplo, en el modelo de ítems ESP19, que genera ítems hijos que evalúan el conocimiento que tiene el estudiante respecto a cómo llenar formatos, se necesita conocer de manera muy específica cuáles son las características de dicho documento, así como el tipo y orden de los datos que se requieren para llenarlo correctamente. De esta manera, el hecho de que el estudiante conozca el orden en que deben encontrarse los distintos elementos de un formato de solicitud será determinante para que responda correctamente a los ítems hijos que exploren este contenido curricular.

Por último, la evaluación del conocimiento de tipo estratégico, en el que se requiere la elaboración de un plan detallado y la presentación de alternativas de solución a un problema dado, no fue abordada por los modelos de ítems del área de Español.

Etapa 3: Análisis de los resultados del trabajo con el panel de expertos

Resultados generales

En total se validaron 20 modelos de ítems, para lo cual se utilizó un ítem hijo como muestra de cada modelo. Al igual que las tres áreas previamente analizadas, los ítems que fueron seleccionados para presentar a los expertos corresponden a una versión fija del Excoba que fue utilizada para su pilotaje y análisis de estructura interna. El trabajo realizado por los expertos se registró en los formatos de evaluación, de los cuales se obtuvo una tabla en la que se concentró

toda la información. En la Tabla 4.22 se muestra la información general de los acuerdos y desacuerdos de los expertos al valorar cada modelo de ítems.

Tabla 4.22.

Resultados globales del proceso de evaluación de panel de expertos del área de Español del Excoba

	Indicador	ESP01 Preguntas con propósitos específicos	ESP02 Estructura del párrafo	ESP03 Gráficos, esquemas y diagramas en textos	ESP04 Evaluación de info. de un tema, diversas fuentes	ESP05 Argumen-tación	ESP06 Recursos lingüísticos: nexos y adverbios	ESP07 Recursos lingüísticos: v. técnico, voz pasiva, t. verbales y f. impersonal	ESP08 Discurso directo e indirecto	ESP09 Acento diacrítico y enfático	ESP10 Géneros literarios	ESP11 Recursos de cohesión	ESP12 Poemas	ESP13 Predicción del contenido de una obra	ESP14 Contraste de mensajes en noticias	ESP15 Reglamento o instructivo	ESP16 Propósitos y argumentos de la carta formal	ESP17 Documentos que establecen derechos y obligaciones	ESP18 Propósitos comunicativos y puntos de vista	ESP19 Llenado de formatos	ESP20 Recursos lingüísticos y visuales en los mensajes publicitarios	% Validado (✓)	% No validado (×)
I1	Definición de contenido clara y precisa	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I2	Definición de contenido congruente con nombre	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I3	Definición de contenido alineada al currículum de asignatura	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	90.0	10.0
I4	Contenido coherente con lo que se enseña en aula	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	✓	✓	✓	×	×	✓	✓	✓	✓	80.0	20.0
I5	Dominio de contenido es básico para asignatura	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	×	✓	×	×	75.0	25.0
I6	Dominio de contenido es esperado del promedio de estudiantes	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	90.0	10.0
I7	Aprendizaje de contenido es importante p/dominio de asignatura	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	×	✓	85.0	15.0
I8	Delimitación del contenido alineada y derivada de definición	✓	×	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	90.0	10.0
I9	Habilidades y contenidos delimitados representan lo esencial	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I10	Estrategia ev. adecuada p/evaluar contenidos delimitados	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I11	Estrategia ev. adecuada p/evaluar aprendizajes esperados	×	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	90.0	10.0
I12	Estrategia evaluativa semejante a como se enseña en aula	×	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	80.0	20.0
I13	Ítems hijos reflejan uso de conocimiento adquirido	×	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	85.0	15.0
I14	Base del reactivo clara y suficiente para emitir respuesta	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	×	×	×	×	✓	✓	✓	✓	✓	75.0	25.0
I15	Instrucciones adicionales claras y suficientes p/emitir respuesta	---	✓	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	0.0	0.0
I16	Reglas p/generar ítems hijos responden a estrategia evaluativa	×	✓	×	×	✓	✓	✓	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	75.0	25.0
I17	Textos auxiliares apropiados	---	✓	---	---	---	---	---	---	---	---	---	×	✓	✓	---	✓	✓	✓	✓	---	87.5	12.5
I18	Gráficos e imágenes apropiados	---	---	×	---	---	---	---	---	---	---	---	---	---	---	---	---	---	---	✓	✓	66.7	33.3
I19	Banco de información corresponde al contenido seleccionado	×	✓	✓	×	✓	✓	✓	✓	✓	×	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	80.0	20.0
I20	Tipo de ejecución simple y facilita evaluación del contenido	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I21	Ítems hijos representan contenido delimitado	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	90.0	10.0
I22	Ítems hijos sin errores de redacción	✓	✓	×	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	90.0	10.0
I23	Ítems hijos redactados con palabras de uso común de estudiantes	✓	✓	×	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	95.0	5.0
I24	Ítems hijos sin pistas	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I25	Ítems hijos con nivel de dificultad apropiado al grado escolar	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	✓	×	✓	✓	×	✓	✓	✓	✓	✓	85.0	15.0
I26	Ítems hijos sin sesgo	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	100.0	0.0
I27	Banco de información e Ítems hijos libres de otro tipo de errores	✓	×	✓	✓	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	✓	✓	✓	✓	×	75.0	25.0
	% Validado	79.17	92.31	72.00	87.50	100.00	95.83	100.00	95.83	100.00	50.00	95.83	84.00	92.00	92.00	66.67	92.00	96.00	100.00	88.46	92.00		
	% No validado	20.83	7.69	28.00	12.50	0.00	4.17	0.00	4.17	0.00	50.00	4.17	16.00	8.00	8.00	33.33	8.00	4.00	0.00	11.54	8.00		

Nota: Los símbolos en las columnas representan los siguientes aspectos: (✓) = Indicador validado, (×) = Indicador no validado, (---) = No aplica.

En las columnas de la Tabla 4.22 se observan las opiniones de los expertos a cada uno de los 20 modelos de ítems, mientras que en los renglones se registraron las opiniones a cada uno de los 27 indicadores evaluados. En la parte inferior se muestran, en términos de porcentajes, los indicadores que fueron y no fueron validados por los expertos en cada uno de los modelos de ítems, y en la columna del extremo derecho, los porcentajes de ítems validados en cada uno de los 27 indicadores.

Análisis de resultados

En este apartado, se analiza la información en tres vertientes. En la primera se analiza el conjunto de modelos de ítems que conforman el área de Español y se describen las opiniones de los expertos en cada uno de los 27 indicadores que conforman el formato de evaluación. En la segunda se describen de manera individual, las características de los modelos de ítems que presentaron cuando menos 20% de indicadores no validados. En la tercera sección se presentan las observaciones y sugerencias realizadas por los expertos con la finalidad de mejorar el área evaluada.

Resultados por problemas detectados en el conjunto de modelos de ítems

La Figura 4.31 muestra de manera gráfica una síntesis de los resultados del área de Español, de acuerdo con el porcentaje de los modelos de ítems que no fueron validados por los expertos en los distintos indicadores evaluados.

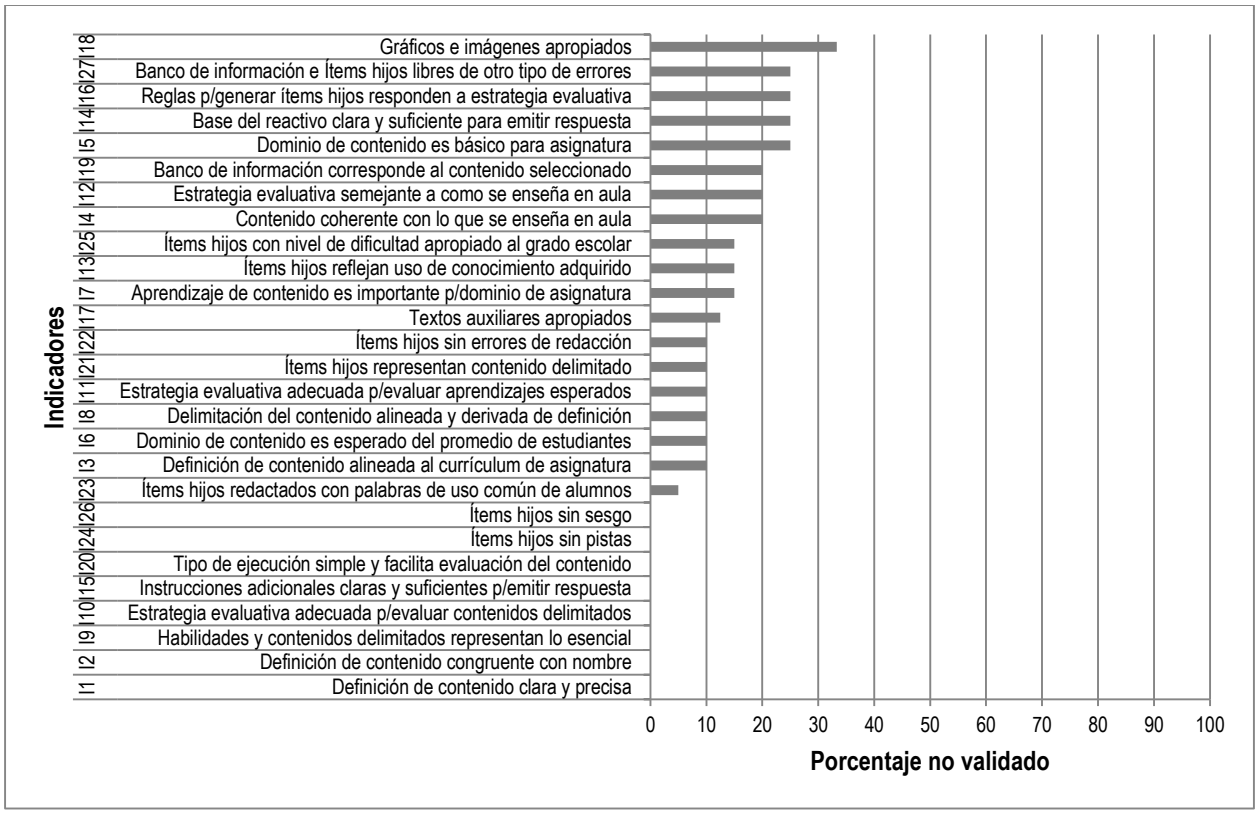


Figura 4.31. Porcentaje de modelos de ítems de Español no validados por los expertos en los indicadores

Se observa que hubo ocho indicadores (1, 4, 8, 12, 14, 18, 19 y 25) que presentaron porcentajes de modelos de ítems no validados mayores a 20% y por ende que revelaron los problemas del área de Español que se presentan con mayor frecuencia. En primer lugar se encuentra el indicador 18, que no fue validado en 33% de los modelos de ítems en los que fue evaluado y que presentó una situación especial ya que únicamente se requirió evaluar en tres modelos de ítems, lo cual fue debido a que explora aspectos de calidad en el diseño y nivel de complejidad de los gráficos o imágenes utilizadas por el generador de reactivos, y solamente tres de los 20 modelos de ítems utilizan imágenes, de los cuales dos sí cumplieron con el criterio pero uno no lo hizo. Los expertos señalaron problemas con la nitidez del texto contenido en las imágenes, ya que se encuentra resaltado en negrillas y consideraron que eso provoca pérdida de nitidez.

Enseguida se observa en la gráfica que hay un grupo de cuatro indicadores que no fueron validados en 25% de los modelos de ítems: 5, 14, 16 y 27. Los problemas detectados en este bloque están relacionados con aspectos de qué tan relevante consideraron que es el dominio de estos contenidos por parte de los estudiantes, la claridad de la base de los reactivos y la correspondencia de las reglas especificadas en los modelos de ítems para generar los ítems hijos. Mediante la evaluación del indicador 5 los expertos señalaron que los temas *Géneros literarios, Reglamento o instructivo, Documentos que establecen derechos y obligaciones, llenado de formatos, y Recursos lingüísticos y visuales en los mensajes publicitarios,* aunque sí son importantes no son considerados esenciales para la asignatura debido a que se relacionan más con aspectos de participación social que con la estructura, funciones y uso del lenguaje.

En cuanto a la claridad y suficiencia de la base del reactivo (indicador 14), los expertos indicaron que los cinco modelos de ítems con este problema requieren revisión, ya que las instrucciones incluidas no contienen una explicación lo suficientemente clara para que el estudiante comprenda lo que debe realizar para emitir sus respuestas. También señalaron que en tres de los modelos de ítems la instrucción dada contiene problemas de redacción y uso de términos que no le corresponden, son incorrectos o resultan confusos. Las sugerencias giraron en torno a modificaciones en la redacción.

Por otro lado, en el indicador 16 que evalúa si las reglas establecidas para generar los reactivos hijos responden a la estrategia evaluativa, los expertos encontraron que no existe correspondencia porque en cinco modelos de ítems hay contradicciones según se revisen sus distintas secciones. Sugirieron que se revise este aspecto con la finalidad de homologar todas las secciones del modelo de ítems.

El último indicador perteneciente a este primer grupo es el 27 que aborda otro tipo de problemáticas no detectadas mediante el resto de los indicadores, pero que consideran

importantes en el proceso de evaluación de los modelos de ítems. Los expertos detectaron que uno de los modelos de ítems (ESP08) requiere mayor revisión en cuanto a su redacción, ya que detectaron problemas con los tiempos verbales de los textos que se utilizan para generar los ítems hijos. También recomendaron la revisión exhaustiva o reestructuración del modelo ESP10 debido a que consideran que tanto la definición como la estrategia evaluativa y el banco de elementos cadena que lo integran, no son suficientes para evaluar el domino del contenido que marca. Por otro lado, señalaron que el modelo ESP14 incluye distractores que pueden causar confusión en los estudiantes al momento de intentar responder a los ítems hijos.

Hubo otro grupo de indicadores (4, 12 y 19) que no fueron validados en 20% de los modelos de ítems y abordan aspectos relacionados con la coherencia entre los contenidos evaluados y el proceso de enseñanza-aprendizaje, así como la correspondencia entre el banco de elementos cadena y los contenidos seleccionados del currículum. Señalaron que en todos los modelos de ítems hay presencia de elementos dentro del banco de información poco relevantes o alejados de la intención evaluativa, de la forma de enseñanza en el aula y/o de los contenidos de la asignatura de Español; además mencionaron que sí hay elementos pertinentes porque fueron seleccionados del currículum, pero no sirven para evaluar lo pretendido.

Hubo un grupo de tres indicadores (7, 13 y 25) en donde los expertos determinaron que 15% de los modelos de ítems no cumplen con los criterios de calidad evaluados, los cuales abordan aspectos de la importancia para el dominio de la asignatura, las niveles de dificultad de los ítems hijos y su capacidad para captar si los estudiantes saben utilizar los conocimientos adquiridos.

También se presentó el caso del indicador 17, que fue evaluado en los ocho modelos de ítems que emplean textos auxiliares, pero no fue validado en 13% (1) de ellos. Otro grupo de indicadores (3, 6, 8, 11, 21 y 22) no fue validado en 10% de los modelos de ítems, los cuales están

relacionados con aspectos de alineación y delimitación del contenido al currículum y a los aprendizajes esperados de los estudiantes, así como de la representatividad y calidad de los ítems hijos.

Finalmente se puede observar en la gráfica que hubo un indicador que no fue validado en 5% de los modelos de ítems, lo cual significa que los expertos detectaron que en un modelo se presentaron problemas de redacción en los ítems hijos.

Resultados por problemas detectados en los modelos de ítems en lo individual

La Figura 4.32 muestra de manera gráfica una síntesis de los resultados de acuerdo con el porcentaje de indicadores que no resultaron válidos en cada uno de los 20 modelos de ítems evaluados. Se examinarán a detalle aquellos que obtuvieron un 20% o más de indicadores no validados: ESP01, ESP03, ESP15 y ESP10.

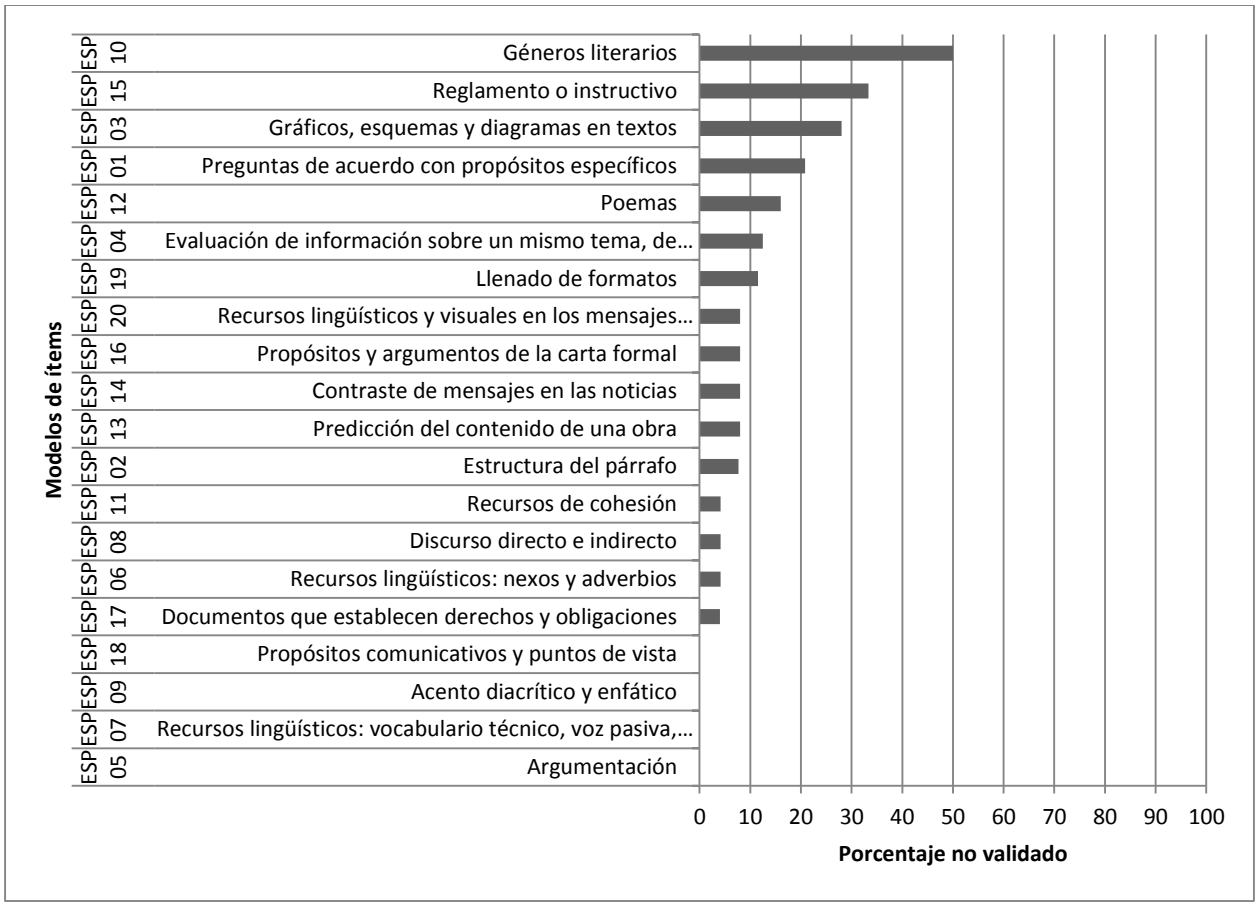


Figura 4.32. Porcentaje de indicadores no validados por los expertos en los modelos de ítems de Español.

El modelo de ítems que presentó mayor cantidad de problemas fue ESP10 (50% de indicadores no validados), que evalúa si el estudiante es capaz de identificar los distintos géneros literarios a partir de los títulos de diversas obras. En segundo término se encuentra ESP15 (33.33% de indicadores no validados), que evalúa si el estudiante cuenta con conocimientos para interpretar adecuadamente documentos que regulen la convivencia o sirvan para reglamentar situaciones de su entorno. En tercer lugar se ubica ESP03 (28% de indicadores no validados), que explora si el estudiante cuenta con la habilidad para comprender e interpretar el mensaje en los textos, a través de otros recursos como las gráficas, esquemas y tablas, entre otros. Por último se encuentra el modelo de ítems ESP01 (20.83% de indicadores no validados), cuyos ítems hijos

evalúan si el estudiante es capaz de identificar y discriminar los elementos que son esenciales en la búsqueda de información de un tema específico. A continuación se describe con detalle el tipo de problemas que presentó cada uno de los cuatro modelos de ítems mencionados, ordenados de acuerdo con el porcentaje de indicadores que no fueron validados.

ESP10. Géneros literarios

Este modelo de ítems genera ítems hijos que exploran el conocimiento del tema *Géneros literarios*, solicitando al estudiante que clasifique títulos de obras literarias según el género al que correspondan. En la Figura 4.33, se muestra un ítem hijo similar a los que se generan en el Excoba, en el cual se presenta una sucesión de números, solicitando al estudiante que escriba cuál es la regla que se utiliza para generarla.

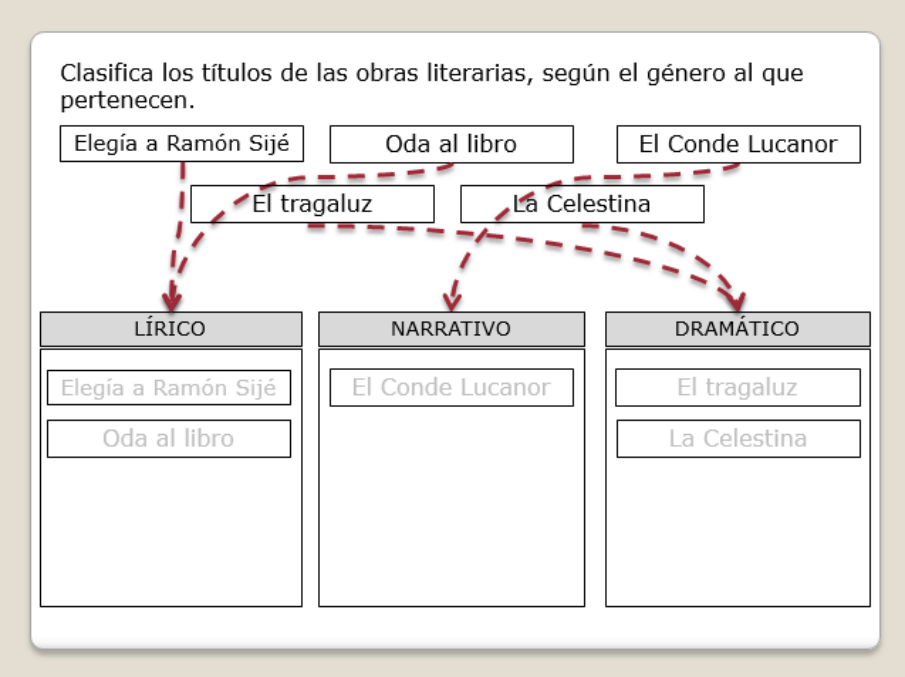


Figura 4.33. Ejemplo de ítem hijo del modelo ESP10: Géneros literarios.

Para resolver este tipo de ítems, el estudiante debe saber cómo identificar los aspectos estructurales y temáticos relevantes a cada género literario. En el ejemplo, se encuentran tres tipos de géneros dispuestos como categorías: lírico, narrativo y dramático. El estudiante debe depositar en cada categoría los recuadros que contienen los nombres de distintas obras literarias, según considere su correspondencia. Para ello debe hacer uso del ratón para seleccionar y mover los recuadros.

En lo que respecta al proceso de validación, la Figura 4.34 muestra los indicadores en que los expertos determinaron que el modelo de ítems ESP10 tuvo problemas.

Figura 4.34. Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP10.

Entre los problemas que presenta este modelo de ítems, los expertos mencionaron que para responder a los ítems hijos derivados de este modelo, el estudiante debe haber leído al menos un fragmento de las obras mencionadas para saber el género literario al que corresponden. Continuaron señalando que la definición del contenido no está alineada al programa de estudios ni es coherente con lo que se enseña en el aula (indicadores 3 y 4) debido a que se centra en la evaluación del reconocimiento y clasificación de títulos de obras, y no en el conocimiento de las características y estructura de cada género literario, lo cual sí está contemplado en el currículum de la asignatura y se enseña en el aula.

También mencionaron que el programa de la asignatura se enfoca en las prácticas sociales del lenguaje y no en la literatura como un contenido primordial, lo cual implica que debe prestar mayor énfasis en temas como redacción, argumentación y el uso social, legal y académico de la

lengua. Señalaron que el desconocimiento de títulos literarios no refleja que el estudiante carece del dominio del contenido evaluado mediante este modelo de ítems, por lo cual no lo consideran como un aprendizaje básico y fundamental para el aprendizaje de la asignatura (indicadores 5, 6 y 7).

Otros problemas que encontraron estuvieron relacionados con la estrategia evaluativa empleada y las características de los ítems hijos. Mencionaron que la estrategia evaluativa utilizada en este modelo de ítems se aleja a la realidad de la forma en que se trabaja en el aula debido a que al abordar el tema se presentan fragmentos de obras al estudiante para que busque las características y estructura de cada género. El uso de títulos de obras no es una práctica común debido a que la enseñanza del contenido se enfoca a identificar las características de los personajes y la voz narrativa, entre otras. Comentaron que la identificación de títulos es insuficiente para brindar evidencias de que el estudiante conoce las características y estructura de los géneros literarios (indicadores 11, 12 y 13).

Por otro lado, al revisar las distintas secciones que conforman al modelo de ítems encontraron que existe incongruencia en varias de ellas. Mediante la revisión del indicador 16 encontraron que la estrategia evaluativa indica que se generen ítems hijos en los que se presenten cuatro géneros literarios y dos columnas (una para las obras y otra para las clasificaciones), pero las reglas y los ítems hijos indican otra dirección. Sugirieron revisar este aspecto del modelo y realizar los ajustes necesarios.

Finalmente, los expertos señalaron que los elementos que fueron seleccionados del currículum para elaborar el banco de elementos cadena no son adecuados para el nivel académico de los estudiantes debido a que se parte de la noción de que conocen por nombre todas las obras literarias y saben clasificarlas. Debido a que ese no es el énfasis del programa de estudios y por ende no refleja los aprendizajes esperados, los ítems hijos que se generen serán

de nivel de dificultad alto. Agregaron que durante la educación secundaria no se revisan obras literarias con la finalidad de identificar o clasificar en los distintos géneros (indicadores 19 y 25).

Además, mencionaron que los elementos cadena seleccionados no son géneros sino subgéneros de la narrativa, por lo cual consideraron que los ítems hijos generados serán inviables para la evaluación del contenido que menciona el modelo de ítems. La sugerencia que realizaron para mejorar el modelo de ítems fue que se reestructurara en cuanto a su definición, estrategia evaluativa y banco de información, de tal forma que evalúe lo que realmente se pretende o espera que el estudiante domine respecto al tema géneros literarios (indicador 27).

ESP15. Reglamento o instructivo

En este modelo de ítems, los ítems hijos generados evalúan si el estudiante cuenta con los elementos para interpretar y reflexionar acerca de las reglas y principios que rigen la vida cotidiana. Un ejemplo de este tipo de ítems se muestra en la Figura 4.35, en donde se plantean diferentes acciones a realizar en función del cumplimiento de reglas establecidas dentro de un reglamento hospitalario.

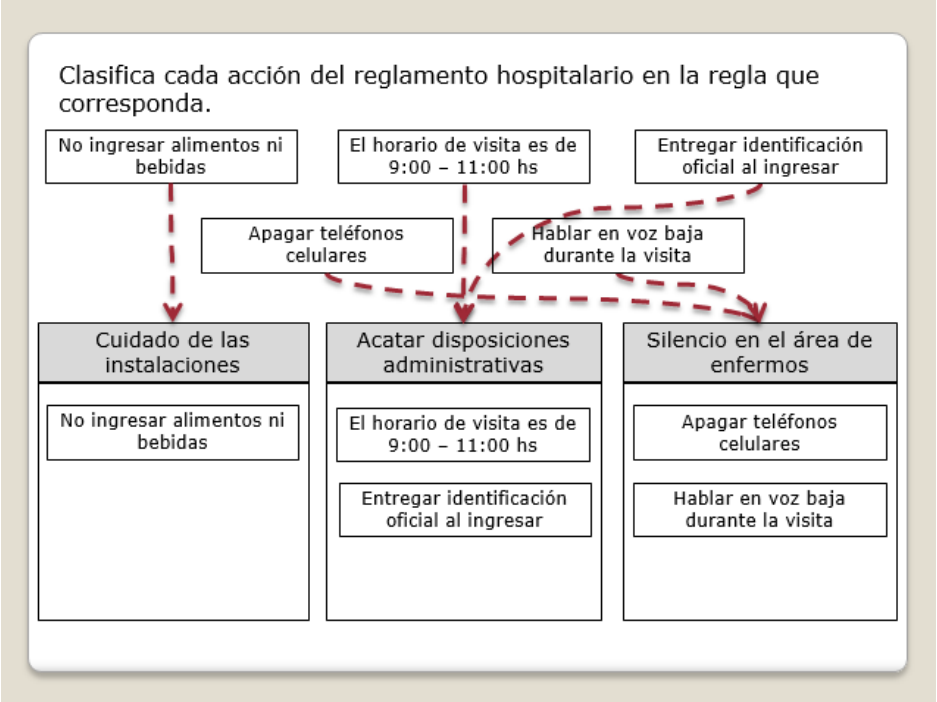


Figura 4.35. Ejemplo de ítem hijo del modelo ESP15: Reglamento o instructivo.

Para responder al ítem hijo, el estudiante debe identificar a qué regla pertenece cada una de las acciones mencionadas, seleccionarlas y moverlas hasta depositarlas dentro de la categoría a la que corresponden. La Figura 4.36 muestra las respuestas consensuadas por los expertos, a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

Figura 4.36. Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP15.

Dentro de los problemas que presenta este modelo de ítems, destacan aquellos relacionados con la pertinencia curricular y congruencia del contenido que evalúa (indicadores 4,

5, 6 y 7), ya que se trata de un tema que los expertos no consideran como un pilar para el dominio de la asignatura. Mencionaron que no lo consideran coherente con lo que se enseña en el aula, ya que la interpretación de reglas y su clasificación en categorías no es el énfasis del proceso de enseñanza-aprendizajes, sino la identificación de derechos y obligaciones, la estructura de un reglamento, así como el uso y aplicación de los modos verbales en la redacción de reglas para la conformación de un reglamente.

En cuanto a la estrategia evaluativa y las características de los ítems hijos (indicadores 12, 13, 14 y 25), los expertos señalaron que a diferencia de lo que se marca en la estrategia, el contenido se aborda en el aula desde un enfoque lingüístico, orientando las acciones hacia sus usos y aplicaciones sociales. También mencionaron que la base del reactivo no es lo suficientemente clara como para que el estudiante pueda responder a los ítems hijos, ya que no presenta fragmentos del reglamento que aborda, sino únicamente su título; sugirieron modificar la redacción y presentaron una propuesta específica para hacerlo. Por último, los expertos indicaron que el grado de dificultad de algunos de los ítems hijos que utiliza el Excoba para evaluar este contenido es mayor al de los ejercicios que se presentan al estudiante en los libros de texto.

ESP03. Gráficos, esquemas y diagramas en textos

En este modelo de ítems, se explora si el estudiante es capaz de buscar, seleccionar y registrar información a partir de textos y relacionarla con la que se presenta en diversos gráficos. En la Figura 4.37 se observa el ejemplo de un ítem hijo similar a los que se generan en el Excoba.

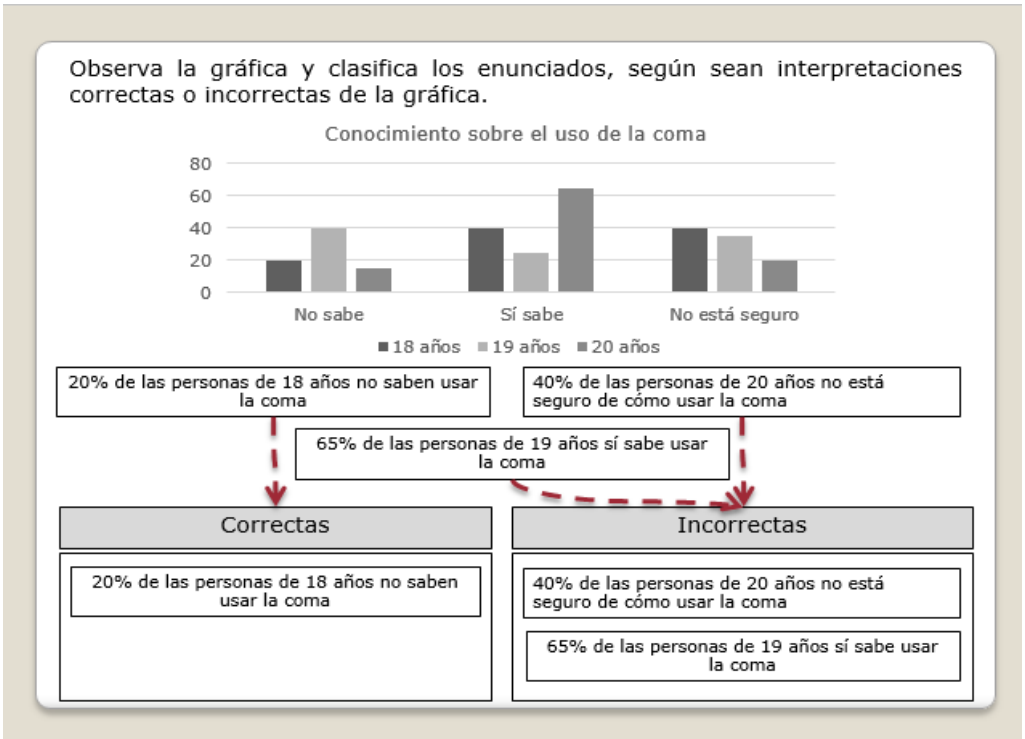


Figura 4.37. Ejemplo de ítem hijo del modelo ESP03: Gráficas, esquemas y diagramas en textos.

Lo que el estudiante debe hacer para responder al ítem hijo es observar la información contenida en la gráfica y valorar si cada uno de los enunciados que se disponen como opciones de respuesta, interpretan o no la información de la gráfica en forma correcta. Una vez que determine eso, deberá seleccionar cada uno de los recuadros con los enunciados, moverlos y depositarlos en la ubicación correspondiente a su respuesta.

Durante el proceso de validación, los expertos detectaron problemas con el modelo de ítems. La Figura 4.38 muestra las respuestas consensuadas por el panel de expertos a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

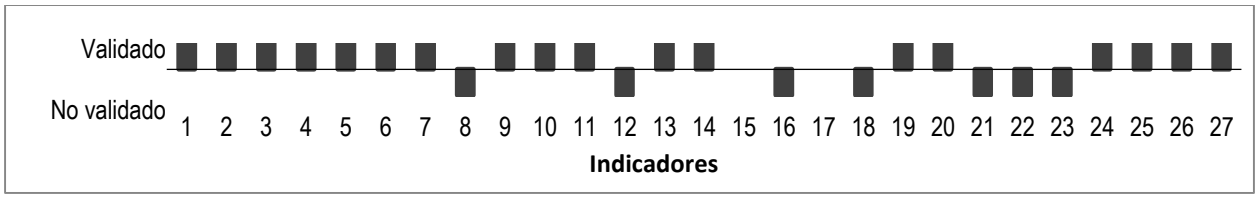


Figura 4.38. Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP03.

Uno de los aspectos que detectaron fue la inconsistencia entre la definición del contenido y su delimitación (indicador 8), ya que la primera menciona el uso de tres recursos visuales (gráficas, diagramas y esquemas), mientras que la segunda indica que se utilizará solamente una de las tres opciones. Por otro lado, también señalaron que la estrategia evaluativa se aleja de la forma en que se enseña en el aula (indicador 12), ya que generalmente se utilizan textos como acompañamiento de los gráficos, de tal manera que el estudiante los asocie y pueda hacer inferencias; sin embargo, en la estrategia evaluativa planteada en este modelo de ítems no se utilizan textos, sino contenidos de manera desarticulada en los mismos gráficos. Para que los ítems hijos de este modelo se acerquen a una evaluación auténtica, sugirieron que los ítems hijos tengan otra forma de operar y muestren un texto informativo junto con opciones de diferentes tipos de gráficos para que el estudiante elija el que mejor lo explique.

También mencionaron que no existe correspondencia entre la cantidad de elementos cadena que se indica en la estrategia evaluativa que se presentarán al estudiante, y las reglas para generar ítems. Sugirieron que se definiera la cantidad y se corrigiera el modelo de ítems (indicador 16). La calidad de los gráficos fue otro aspecto en el que detectaron problemas, ya que según indicaron los expertos, requieren revisión para mejorar su nitidez (indicador 18).

En los ítems hijos se encontraron tres tipos de problemas: la falta de representación respecto al contenido delimitado, errores ortográficos y tipográficos en dos de los gráficos, y uso de términos poco apropiados para el lenguaje de los estudiantes (indicadores 21, 22 y 23).

ESP01. Preguntas de acuerdo con propósitos específicos

En este modelo de ítems, se explora si el estudiante es capaz de identificar y discriminar los elementos esenciales para la búsqueda de información de un tema específico. En la Figura 4.39 se muestra un ejemplo de un ítem hijo similar a los que se generan en el Excoba.

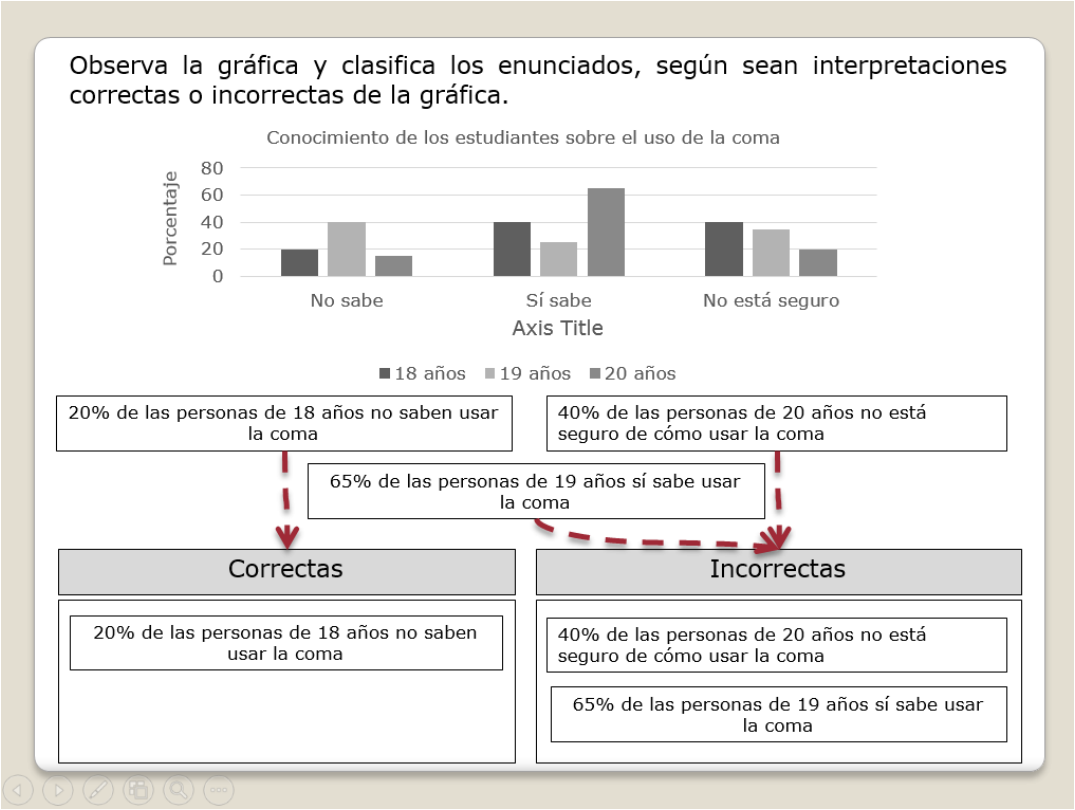


Figura 4.39. Ejemplo de ítem hijo del modelo MAT06: Sistemas de ecuaciones.

El modelo de ítems genera ítems hijos en donde se solicita al estudiante que clasifique enunciados según los considere como interpretaciones correctas o no de la información contenida en una gráfica. Para responder al ejemplo, el estudiante debe observar que los datos de la gráfica se encuentran agrupados de acuerdo con la edad y tipo de respuesta que emitieron algunos estudiantes sobre el conocimiento del uso de la coma. También debe notar que la

información está representada en porcentajes. Mediante la ubicación de la información en la gráfica y el manejo la información podrá determinar para cada enunciado si se trata de una interpretación correcta o incorrecta de la gráfica, con la finalidad de seleccionarlos, moverlos y clasificarlos en el lugar que corresponda.

En cuanto al proceso de validación, la Figura 4.40 muestra las respuestas consensuadas por los expertos a cada uno de los 27 indicadores que evaluaron mediante el formato de validación de ítems.

Figura 4.40. Opinión de los expertos en los 27 indicadores validados del modelo de ítems ESP01.

Dentro de los problemas que presenta este modelo de ítems en lo relacionado a la estrategia evaluativa (indicadores 11, 12 y 13), los expertos mencionaron que no es la más adecuada para el propósito que se persigue, debido a que el conocimiento que se pretende evaluar mediante estos ítems hijos implica las nociones del estudiante sobre la elaboración de preguntas para buscar información de temas de diversa naturaleza; sin embargo, los ítems hijos exploran contenidos muy puntuales y de otras asignaturas como Historia y Biología, que requieren del uso de la memoria y a su vez dependen de que se haya cubierto cabalmente el programa de estudios de dichas asignaturas.

En este sentido, los expertos consideraron que se corre el riesgo de que: a) el alumno utilice su memoria para emitir la respuesta y no aplique el conocimiento adquirido, y b) que de no haberse revisado completamente el currículum de todas las asignaturas, el estudiante no tenga nociones suficientes para responder. Para resolver esta situación sugirieron modificar la

estrategia evaluativa mediante la inclusión de preguntas como elementos cadena, en donde el estudiante haga una selección delas que considere más pertinentes para la obtención de información sobre un tema dado.

También mencionaron que no existe correspondencia entre las reglas para generar los ítems hijos y la estrategia evaluativa (indicador 16) debido a que detectaron una discrepancia entre ambas secciones del modelo de ítems, ante lo cual sugirieron una revisión y modificación según sea pertinente. En cuanto a los elementos que fueron seleccionados del currículum para conformar el banco de información (indicador 19), los expertos mencionaron que sí son pertinentes ya que las opciones de respuesta corresponden a cada tema que abordan, pero no permiten evaluar si el alumno es capaz de buscar información.

Resultados conjuntos

En este apartado se muestran los resultados derivados del análisis del área de Español en donde se contrastó la información derivada del trabajo con el panel de expertos, de tal manera que se tomaron los cuatro modelos de ítems que tuvieron por lo menos 20% de los indicadores de calidad no validados, y los ocho indicadores que no fueron validados en al menos 20% de los modelos de ítems de toda la asignatura. Se detectó la presencia de regularidades en el tipo de problemáticas que presentan. La Tabla 4.23 muestra esta información.

Tabla 4.23.
Modelos de ítems de Español y regularidades que presentaron

No.	Indicador	Modelos de ítems				Frecuencia relativa
		ESP01 Preguntas de acuerdo con propósitos específicos	ESP03 Gráficos, esquemas y diagramas en textos	ESP10 Géneros literarios	ESP15 Reglamento o instructivo	
12	Estrategia evaluativa semejante a como se enseña en aula	×	×	×	×	4 de 4
16	Reglas p/generar ítems hijos responden a estrategia evaluativa	×	×	×	----	3 de 4
4	Contenido coherente con lo que se enseña en aula	----	----	×	×	2 de 4
5	Dominio de contenido es básico para asignatura	----	----	×	×	
19	Banco de información corresponde al contenido seleccionado	×	----	×	----	
14	Base del reactivo clara y suficiente para emitir respuesta	----	----	----	×	1 de 4
18	Gráficos e imágenes apropiados	----	×	----	----	
27	Banco de información e Ítems hijos libres de otro tipo de errores	----	----	×	----	

Nota: Se muestra únicamente la información de los modelos de ítems que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Español. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

Se observa que existen coincidencias entre varios modelos de ítems, respecto a los tipos de problemas presentes en ellos. Por ejemplo, los cuatro modelos de ítems: ESP01 (Preguntas de acuerdo con propósitos específicos), ESP03 (Gráficos, esquemas y diagramas en textos), ESP10 (Géneros literarios), y ESP15 (Reglamento o instructivo), además de tener al menos 20% de los indicadores de calidad no validados, coincidieron en que tienen problemas con la correspondencia de la estrategia evaluativa para generar ítems hijos y la forma en que se enseña el contenido en el aula (indicador 12).

De igual manera, los modelos de ítems ESP01 (Preguntas de acuerdo con propósitos específicos), ESP03 (Gráficos, esquemas y diagramas en textos), ESP10 (Géneros literarios) no fueron validados en el indicador 16, ya que los expertos detectaron que en estos tres modelos las reglas para generar ítems hijos no responden cabalmente a la estrategia evaluativa.

Otra coincidencia que surgió de este análisis fue que los modelos de ítems ESP10 (Géneros literarios) y ESP15 (Reglamento o instructivo), coincidieron en que no fueron validados los indicadores 4 y 5, lo cual significa que en el caso de estos dos modelos, los expertos consideran que el contenido evaluado no es coherente con los contenidos que se enseñan en el aula y tampoco lo consideran básico para el dominio de la asignatura.

Por último, los modelos de ítems ESP01 (Preguntas de acuerdo con propósitos específicos) y ESP10 (Géneros literarios), presentaron problemas en cuanto a la correspondencia de los elementos cadena que conforman el banco de información, con el contenido que fue seleccionado del currículum (indicador 19). Lo anterior implica que en ambos modelos se producirán ítems hijos que no permiten evaluar el contenido que se señaló en los modelos.

En lo que respecta a los indicadores 14, 18 y 27, aunque no fueron validados por los expertos en más del 20% de los modelos de ítems del área de Español, en el caso de estos cuatro modelos no hubo coincidencias.

Resultados según las sugerencias del panel de expertos para el mejoramiento del área

Además de los problemas que fueron detectados por los expertos de manera puntual en cada modelo de ítems, hubo una serie de sugerencias que éstos realizaron para el mejoramiento del área de Español del Excoba.

Los expertos consideraron que en conjunto, 90% de los modelos de ítems de la asignatura de Español que revisaron representan las habilidades y conocimientos básicos que deben

dominar los estudiantes al término de la secundaria. Hubo dos modelos de ítems que consideraron que no son indispensables y pueden omitirse del examen. En su lugar, mencionaron que existen dos contenidos muy importantes que deberían incluirse: *Figuras retóricas* y *Redacción de la autobiografía*. Mencionaron que el primero es importante ya que su dominio es fundamental para la adquisición de nuevos aprendizajes, tal como la comprensión de los mensajes implícitos en la comunicación de diversas fuentes de información y textos (literarios, sociales, publicitarios y persuasivos, entre otros). El segundo contenido mencionado es considerado esencial debido a que los temas relacionados con el tema del ser permite la generación de espacios de expresión de sentimientos por ende del uso de la lengua (voz narrativa, recursos lingüísticos y uso de tiempos verbales, entre otros).

Al solicitar a los jueces la estimación de qué porcentaje de estudiantes que concluyen la secundaria consideran que responderían correctamente a los ítems hijos generados mediante los veinte modelos de ítems que conforman la asignatura de Español mencionaron que 70%, ya que en su experiencia la mayoría de los estudiantes no leen instrucciones correctamente y responden impulsivamente. Además, varios modelos de ítems no coinciden con la forma en que se enseña en el aula, ya sea en la naturaleza de los ejercicios o porque utilizan terminología que no corresponde a la que se utiliza en el aula al abordar los contenidos. Puntualizaron el caso del modelo de ítems ESP01 en el cual detectaron que para responder, el estudiante requiere nociones de los temas que abordan sus ítems hijos, los cuales no corresponden a la asignatura ni al contenido evaluado. También ejemplificaron con el caso del modelo ESP02, el cual utiliza incorrectamente el término *oración temática* en la base del reactivo, pudiendo causar confusiones en los estudiantes. Finalmente también mencionaron que el modelo ESP12 menciona en la base del reactivo la presencia de metáforas, pero realmente se trata de lenguaje figurado.

Finalmente, otro aspecto importante que destacaron y a su vez fue común a la mayoría de los modelos de ítems de tipo *categoría elementos*, es que no se estableció una misma regla para la cantidad de opciones de respuesta que se mostrarán al estudiante y a su vez clasificará dentro de una categoría. Consideraron que esto puede resultar en confusión por parte de los estudiantes, ya están acostumbrados a ubicar la mayor cantidad de respuestas en los espacios disponibles, pero en el Excoba hay ocasiones en las que se presentan ítems en los que quedan vacías algunas categorías de clasificación. Sugirieron establecer una misma regla para todos los modelos de tipo categoría elementos.

Resultados según la relación entre los problemas detectados y los contenidos evaluados

Una vez que se identificó el ámbito curricular al que pertenece cada modelo de ítems, se procedió a realizar un análisis para detectar si existe algún tipo de relación entre el contenido curricular que evalúan los modelos de ítems que presentaron mayor cantidad de indicadores no validados ($\geq 20\%$), su pertenencia a un ámbito dentro del currículum de la asignatura y el tipo de problema expresado en los indicadores que no fueron validados en 20% o más de los modelos de ítems. La Tabla 4.24 resume los hallazgos.

Tabla 4.24.
Modelos de ítems, ámbitos curriculares e indicadores no validados con mayor frecuencia en Español

Modelos de ítems	Contenido	Ámbito	No. de indicador							
			4	5	12	14	16	18	19	27
ESP01	Preguntas de acuerdo con propósitos específicos	Estudio	--	--	×	--	×	--	×	--
ESP03	Gráficos, esquemas y diagramas en textos		--	--	×	--	×	×	--	--
ESP10	Géneros literarios	Literatura	×	×	×	--	×	--	×	×
ESP15	Reglamento o instructivo	Participación ciudadana	×	×	×	×	--	--	--	--

Nota: Se muestra únicamente la información de los modelos de ítems que presentaron 20% o más indicadores no validados por los expertos, así como aquellos indicadores que no fueron validados en 20% o más modelos de ítems de Español. Los símbolos en las columnas representan lo siguiente: (×) = Indicador no validado, (---) = No aplica.

De los cuatro modelos de ítems que cumplieron con el criterio de 20% o más indicadores no validados por los expertos, se puede observar que dos corresponden al ámbito *Estudio* (ESP01 y ESP03), mientras que uno pertenece al ámbito *Literatura,* y el restante al ámbito *Participación ciudadana.*

El propósito de la asignatura es que los estudiantes adquieran destrezas dentro de las distintas prácticas sociales del lenguaje, de tal manera que logren una comprensión adecuada de su estructura y propiedades. Así su participación en las actividades académicas y de interacción social se podrá llevar a cabo de manera eficaz. El currículum y todos sus contenidos se encuentran organizados en tres grandes ámbitos: *Estudio*, *Literatura* y *Participación ciudadana*. Los dos modelos (ESP01 y ESP03), que más problemas presentaron en los indicadores con menor porcentaje de validación en el área de Español, también coincidieron en el ámbito curricular de pertenencia: *Estudio*.

En este ámbito, las prácticas sociales del lenguaje están orientadas hacia la reflexión respecto al contenido de textos, modos de expresión e intercambios comunicativos realizados a través de ellos. Se busca que los estudiantes adquieran un cabal dominio de la forma de expresión escrita, ya que se considera que es en donde se aprende a construir el lenguaje y se aplican las normas establecidas para ello. En este ámbito se procura que los estudiantes produzcan textos, preparen información y elaboren textos escritos expresando con claridad y coherencia sus ideas, utilizando vocabulario especializado y de carácter técnico, adquirido mediante la consulta de diversas fuentes de información.

Los modelos de ítems ESP01 y ESP03 coinciden en que ambos buscan evaluar si el estudiante cuenta con las competencias para realizar preguntas y obtener información sobre temas específicos, también exploran si cuenta con la habilidad para interpretar información contenida en formato gráfico. Los expertos indicaron que no consideran estos contenidos como

parte fundamental de la asignatura y tampoco encontraron similitud entre la estrategia evaluativa utilizada y la forma en que se enseña en el aula. Mencionaron que ambos modelos de ítems presentan una dirección o enfoque evaluativo que no solamente se aleja de la realidad del proceso de enseñanza-aprendizaje, sino que difiere sustancialmente de él.

Lo anterior sugiere que al menos en estos dos modelos de ítems sí existe relación entre el tipo de contenido al que corresponden en el currículum y la naturaleza de la problemática señalada por los expertos. La información proporcionada por los expertos indica la necesidad de realizar una revisión detallada de ambos modelos en cuanto al contenido que evalúan, su relevancia para ser incluidos en el examen y su coherencia con la forma en que se enseña en el aula. En el caso de los modelos de ítems ESP01 y ESP03 no hubo coincidencia en la organización curricular, pero sí en el tipo de problemas encontrados por los expertos, aspecto que fue analizado en el apartado de resultados conjuntos.

Etapa 4: Análisis de la totalidad de modelos de ítems como conjunto

En los resultados de la etapa 4 del MVCE se analizan los problemas encontrados en los 56 modelos de ítems incluidos en este estudio, con la finalidad de detectar regularidades, explorando la información en tres niveles: (1) su clasificación en el sistema GenerEx, (2) la ejecución que debe realizar el estudiante al responder, y (3) el tipo de conocimientos que exploran (declarativo, procedimental, esquemático, estratégico).

En primer lugar se muestra la distribución de los modelos de ítems según las categorías de análisis mencionadas. Además, se consideró esencial presentar el análisis de los 56 modelos de ítems según los indicadores que evaluaron los expertos, haciendo especial énfasis en aquellos que no fueron validados en al menos 20% de los modelos de ítems incluidos en este estudio.

Distribución de los modelos de ítems en el GenerEx

El Excoba utiliza el sistema informático llamado GenerEx para generar ítems hijos a partir de las reglas y restricciones que se encuentran descritas en los modelos de ítems. Estas reglas sirven para hacer combinaciones e intercambios de los elementos cadena e integrales que conforman el banco de información asociado al contenido curricular evaluado. Existen 22 tipos de modelos de ítems en el GenerEx, clasificados así debido a que requieren distinta programación para que se puedan producir los ítems hijos. Los modelos de ítems de las asignaturas de Matemáticas, Historia, Química y Español implican el uso de 14 de ellos (ver Figura 4.41). Si se desea revisar con mayor detalle la información, el lector puede remitirse a la Tabla 3.2 ubicada en el Capítulo 3. En ella se explica con detalle la forma en que opera cada uno de los tipos de modelos según el GenerEx.

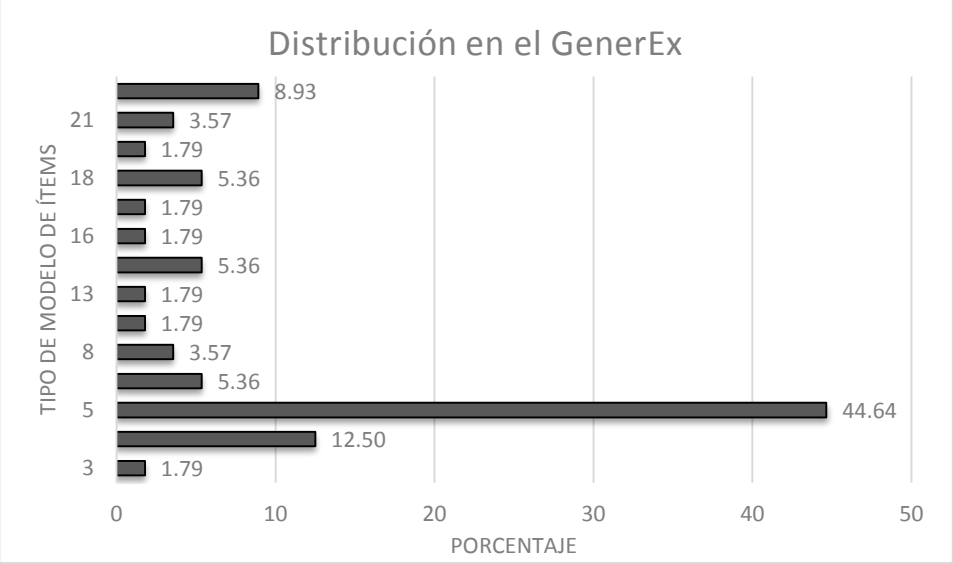


Figura 4.41. Distribución de los 56 modelos de ítems del Excoba por tipo de programación en el GenerEx. Se presentan los 14 tipos de modelos de ítems que se utilizaron para generar los ítems hijos de Matemáticas, Historia, Química y Español. Los números representan los nombres asignados a los modelos en el GenerEx: 3= Elemento imagen, 4= Selección de elementos, 5= Elemento categoría, 7= RN/fórmulas, 8= RN/ecuaciones, 10= RN/pendiente, 13= RN/Sucesiones, 15= RN/etiquetas, 16= Selección frases, 17= Orden elemento múltiple, 18= Frase imagen, 20= RN/gráficas, 21= RN/R. algebraica, y 22= R. algebraica.

En primer lugar, se puede observar que 44.64% (25) de la totalidad de los modelos de ítems de las asignaturas mencionadas son del tipo *Elemento categoría,* el cual se utiliza para generar ítems hijos en las cuatro asignaturas. Estos modelos están programados para utilizar la información alimentada en el banco de elementos cadena e integrales y mediante asociación, clasificarlos en distintas categorías, a las cuales se asocian elementos cadena e integrales que son seleccionados de manera aleatoria por el sistema con la finalidad de generar los ítems hijos.

De los modelos de ítems, 12.5% (7) son del tipo *Selección elementos*, empleado únicamente en las asignaturas de Química y Español. En este tipo de modelos el sistema elige al azar las categorías y elementos para presentarlos en la interfaz. El tercer tipo de modelos más frecuentes y utilizados únicamente en la asignatura de Matemáticas son los denominados *R. algebraica,* en donde 8.93% (5) requieren una programación apoyada por un editor de ecuaciones, de tal manera que conforme se escribe la respuesta, se transforma a formato de ecuación.

Hubo tres tipos de modelos cuya distribución en el GenerEx fue de 5.36% (3) en cada uno: 1) *RN/fórmulas* utilizado únicamente en la asignatura de Matemáticas, donde el sistema intercambia los elementos integrales de la base del reactivo mediante el uso de fórmulas y reglas para hacer cálculos y determinar sus valores; 2) *RN/etiquetas* también utilizado únicamente en la asignatura de Matemáticas, utiliza imágenes o figuras que contienen datos, los cuales son sustituidos por el sistema para llevar a cabo el proceso de aleatorización de los elementos, y *Frase imagen,* empleado en los modelos de ítems de las asignaturas de Química y Español, donde el sistema elige textos al azar en los cuales marca algunos segmentos para que el estudiante elija los que considere correctos, los seleccione, mueva y coloque en un espacio de la interfaz.

También se observa que 3.57% (2) de los modelos son de dos tipos: RN/ecuaciones y RN/R. algebraica. Ambos son utilizados únicamente en la asignatura de Matemáticas; en el primer tipo se elige al azar una ecuación del banco de elementos integrales. En el segundo tipo el sistema

implica una programación donde se deben realizar dos acciones principales: transformar las respuestas que escriba el examinado de un formato numérico a uno algebraico, y permitir seleccionar imágenes de la interfaz, de tal manera que la respuesta escrita se acompañe de la selección de un gráfico.

Por último, hay seis tipos de modelos de ítems en el GenerEx que incluyen a 1.79% (1) de los modelos de ítems, cada uno: 1) *Elemento imagen*, utilizado únicamente en la asignatura de Historia y que requiere la asociación de elementos con segmentos en una imagen, de tal manera que el estudiante pueda colocarlos dentro de ella; 2) *RN/Pendiente* que es empleado para un modelo de ítems de la asignatura de Matemáticas, en donde se utilizan gráficas de funciones para calcular la pendiente de cada recta que se presentará como imagen en la interfaz; 3) *RN/sucesiones*, también empleado únicamente en un modelo de ítems de Matemáticas, en donde el sistema selecciona un grupo de números dentro de una serie previamente alimentada al banco de elementos, con la finalidad de presentarlos al estudiante y que escriba el siguiente número de la serie; 4) *Selección frases*, utilizado en un modelo de ítems de Español, que implica una programación en la que un texto estará dividido en segmentos que serán marcados para que el examinado los seleccione y registre su respuesta; 5) *Orden elementos múltiple*, empleado únicamente en un modelo de ítems de la asignatura de Química, en donde el sistema selecciona un grupo de imágenes del banco de información, las cuales siguen un orden preestablecido según su grado de posesión de una propiedad química, y 6) *RN/*gráficas, que se utiliza únicamente en el área de Matemáticas e implica el uso de gráficas dinámicas, las cuales son modificadas cuando el sistema modifica los valores de sus variables.

Distribución de los modelos de ítems por tipo de ejecución

Como se ha mencionado en diversos apartados, los modelos de ítems del Excoba requieren que el estudiante realice cuatro tipos de ejecución para que emita y registre su respuesta: arrastre, selección, escritura y mixta. La Figura 4.42 muestra la distribución de los 56 modelos.

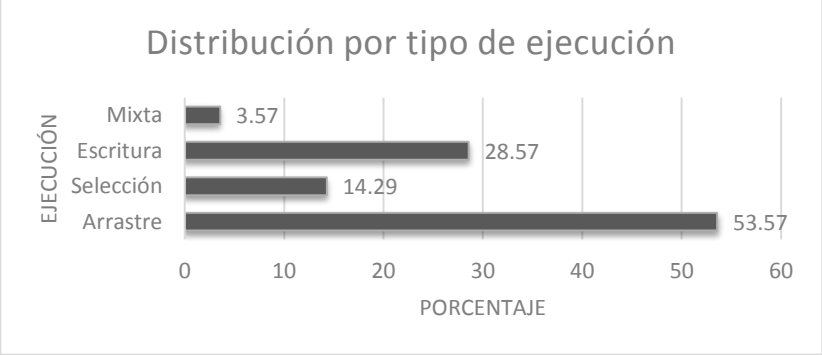


Figura 4.42. Distribución de los 56 modelos de ítems del Excoba por tipo de ejecución.

Más de la mitad (53.57% [30]) de los modelos de ítems requieren que el estudiante utilice el ratón para mover y colocar elementos en distintas ubicaciones de la interfaz; 28.57% (16) implican la escritura libre de la respuesta, donde el estudiante registra su respuesta mediante el uso del teclado de la computadora; 14.29% (8) requieren que el estudiante coloque el ratón en un lugar de la interfaz y seleccione segmentos de ella para señalar y registrar su respuesta, y 3.57% (2) modelos de ítems son de respuesta mixta, donde se deben realizar dos acciones para registrar la respuesta: escribir y mover elementos.

Distribución de los modelos de ítems por tipo de conocimiento

La clasificación de los tipos de conocimientos que explora el Excoba incluye cuatro niveles: declarativo, procedimental, esquemático y estratégico. La Figura 4.43 muestra la distribución de los 56 modelos de ítems incluidos en el análisis.

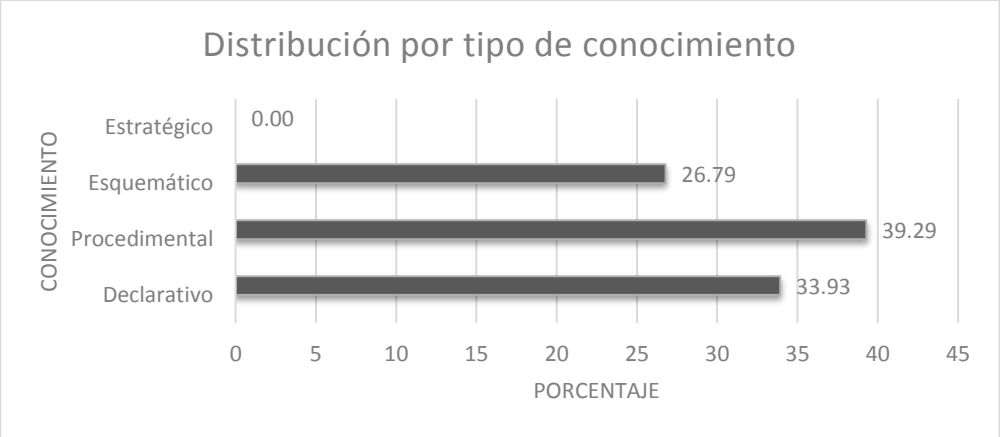


Figura 4.43. Distribución de los 56 modelos de ítems del Excoba por tipo de conocimiento.

Se observa que la distribución fue similar en cuanto a los conocimientos de tipo *declarativo*, *procedimental* y *esquemático*. En primer lugar, 39.29% (22) de los modelos exploran un nivel de conocimiento *declarativo* donde el estudiante debe emplear datos y hechos, hacer uso de vocabulario y terminología asociada al contenido, manejar conceptos básicos y definiciones para clasificar los elementos que se le presentan en los ítems hijos.

También se observa que 33.29% (19) de los modelos de ítems evalúan conocimientos de tipo *procedimental*, los cuales requieren que el estudiante recurra al uso de su conocimiento de los métodos y técnicas, así como las secuencias de pasos necesarios para llegar a la solución de un problema. El tercer grupo que se observa en la gráfica es el de los modelos de ítems que exploran conocimientos de tipo *esquemático*, donde fueron clasificados 26.79% (15). En ellos el estudiante debe tener la capacidad de sintetizar y organizar la información que conoce sobre el tema, de tal manera que la utilice para explicar un problema y llegar a su solución.

Los conocimientos de tipo *estratégico* no fueron evaluados por ningún modelo de ítems de las asignaturas de Matemáticas, Historia, Química y Español del Excoba. Esto puede ser debido a que se trata de un instrumento que evalúa competencias básicas y los conocimientos de tipo

estratégico son considerados de orden superior, ya que requieren que el estudiante reflexione sobre los distintos procesos necesarios para la solución de una tarea o problema, la planeación muy detallada de estrategias y la integración de los otros tres tipos de conocimientos.

Problemas más frecuentes en los 56 modelos de ítems

Un nivel de análisis que se consideró muy importante incluir fue el de los problemas que se presentaron con mayor frecuencia en los 56 modelos de ítems analizados en ese estudio. Mediante una revisión de las opiniones de los expertos de cada una de las cuatro áreas, se detectaron aquellos indicadores que no fueron validados en al menos 20% de los 56 modelos de ítems. La Figura 4.44 muestra de manera gráfica una síntesis de los resultados obtenidos.

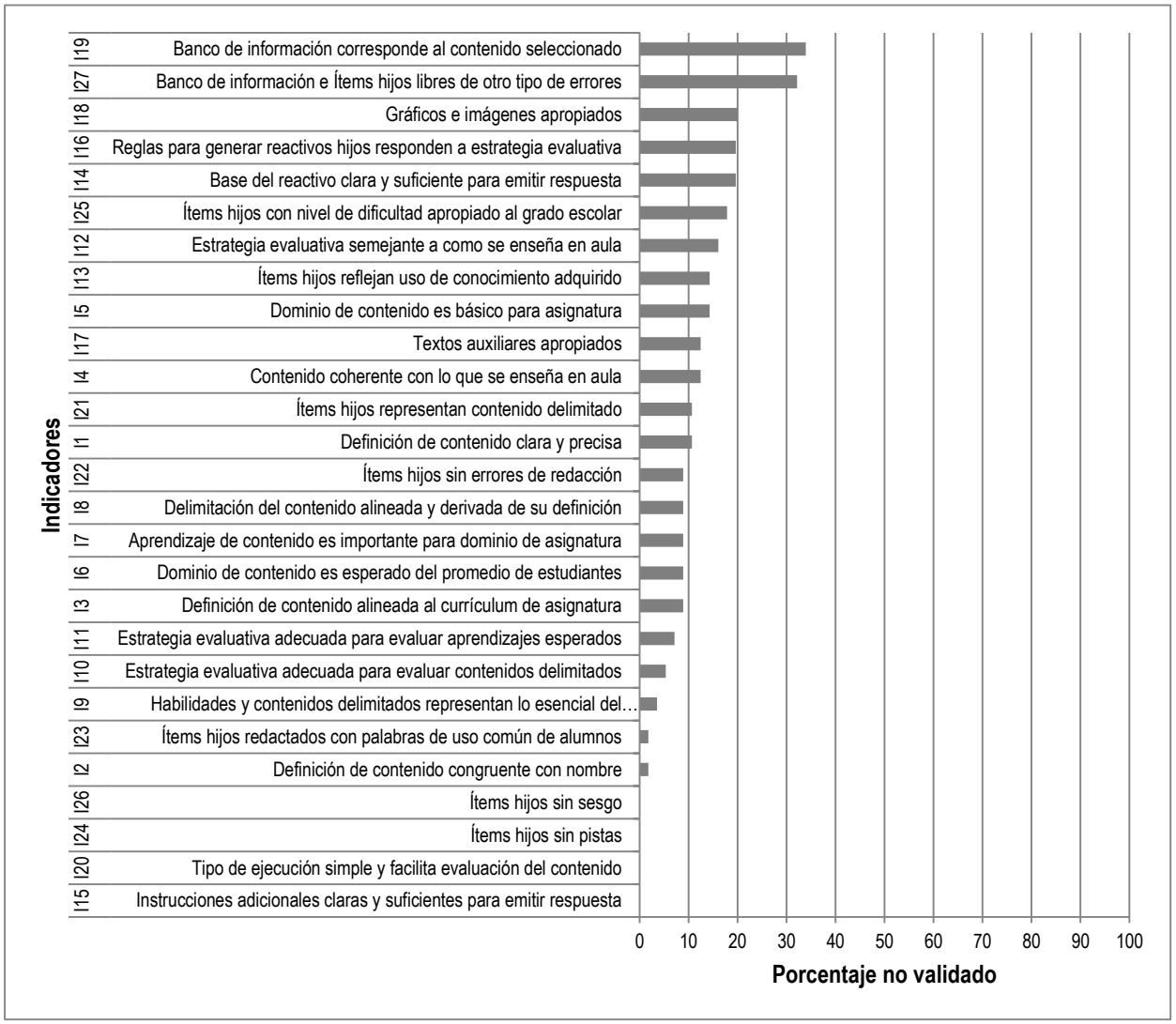


Figura 4.44. Porcentaje de modelos de ítems no validados por los expertos en cada indicador evaluado.

En primer lugar, se puede observar que el indicador 19 fue el que los expertos señalaron con mayor frecuencia, ya que estimaron que no se cumplía en 34% (19) de los modelos de ítems. Este indicador se relaciona con problemas en los ítems hijos y los elementos cadena que se utilizan para generarlos. Esto significa que en más de la tercera parte de los 56 modelos de ítems los expertos identificaron problemas específicos, tales como: la presencia de elementos dentro del banco de información poco relevantes o alejados de la intención evaluativa y/o de los

contenidos de las asignaturas analizadas; problemas de redacción que podrían generar confusión en los estudiantes o se alejan de la forma en que se plantean los problemas y ejercicios en el aula.

También se detectaron elementos marcados como respuestas correctas cuando no lo son, o cercanía conceptual entre ellos, lo cual dificulta la distinción entre correctos e incorrectos; falta de homogeneidad en los niveles de dificultad de los elementos cadena; elementos cadena o integrales que no corresponden al grado escolar del estudiante o a contenidos que marca el programa, y uso de términos poco familiares para los alumnos. Lo anterior sugiere una revisión detallada de los 19 modelos de ítems en los que se detectó este problema, para realizar las modificaciones necesarias.

El segundo indicador que sobresale por no haber sido validado por los expertos en 32% (18) de los modelos de ítems fue el 27. Este indicador aborda otro tipo de problemáticas no detectadas mediante el resto de los indicadores, pero que se consideraron importantes en el proceso de evaluación de los modelos de ítems. Entre otras cosas, los expertos detectaron que hay elementos dentro del banco de información de varios modelos que son inviables para la producción de ítems que evalúen el contenido mencionado en ellos, particularmente en la asignatura de Matemáticas.

También se señalaron inconsistencias en los niveles de dificultad entre familias de ítems pertenecientes a un mismo modelo, así como falta de claridad en el tipo de respuesta que se le solicita al estudiante. En el caso particular de Matemáticas, mencionaron que el estudiante podría optar por responder escribiendo fracciones o también decimales, pero se desconoce cuál es el tipo de respuesta esperada. Otro aspecto similar señalado es la ausencia de información respecto a la cantidad de decimales que se puede escribir en la respuesta.

En la asignatura de Español, los expertos mencionaron que se requiere una mayor revisión de la redacción en los textos auxiliares. También recomendaron la revisión exhaustiva o

reestructuración de algunos modelos (p. ej., ESP10) por considerar que tanto la definición como la estrategia evaluativa y el banco de elementos cadena que los integran son insuficientes para evaluar el domino del contenido que se marca.

El tercer indicador que no fue validado con frecuencia fue el 18, siendo señalado en 20% de los modelos de ítems en los que fue evaluado. Es importante mencionar que debido a que explora aspectos de calidad en el diseño y nivel de complejidad de los gráficos o imágenes utilizadas por el generador de reactivos, solamente fue necesario evaluarlo en 15 de los 56 modelos de ítems, ya que únicamente esa cantidad emplea gráficos. Los paneles de expertos señalaron problemas en tres modelos de ítems de las asignaturas de Historia, Química y Español, relacionados con el diseño, nitidez, uso de colores, nivel de complejidad y calidad general de los gráficos empleados para generar los ítems hijos

Otro indicador que no fue validado en 20% (11) de los 56 modelos de ítems fue el 16, que evalúa si las reglas establecidas para generar los reactivos hijos responden a la estrategia evaluativa. Los expertos encontraron que en algunos casos no existe correspondencia, ya que detectaron modelos de ítems en donde la descripción de la estrategia y la operación de la misma presentan contradicciones. Mencionaron que en el caso de los 11 modelos de ítems señalados, el ítem hijo final que se presenta al examinado no refleja cabalmente la estrategia evaluativa planteada en el modelo. Sugirieron que se revise este aspecto con la finalidad de homologar todas las secciones.

Finalmente, el indicador 14 también fue señalado por los expertos en 20% (11) de los modelos de ítems. En este caso, los problemas detectados fueron en relación con la claridad y suficiencia de la base del reactivo. Indicaron que los 11 modelos de ítems presentan los siguientes problemas: a) las instrucciones no contienen una explicación lo suficientemente clara para que el estudiante comprenda lo que debe realizar para emitir sus respuestas, b) la base del reactivo

presenta problemas de redacción y/o uso incorrecto de términos que pueden causar confusión en los examinados por no corresponder al lenguaje utilizado en el proceso de enseñanza - aprendizaje. En este sentido, los expertos sugirieron que se prestara atención especial a la redacción utilizada en la base de los reactivos y que se revisara con minuciosidad.

5. Discusión y conclusiones

El Excoba/GenerEx representa el primer proyecto mexicano para generar de manera automática exámenes destinados a los procesos de admisión a las instituciones educativas de México, en los niveles de educación media superior y superior. Por el gran impacto que tiene, es indispensable aportar evidencias de su validez. Los estudios de Ferreyra (2014) y Pérez (2013) han aportado evidencias, respectivamente, de la estructura interna de este instrumento y de los procesos cognitivos que se demandan a los estudiantes para responder las tareas evaluativas. Sin embargo, no se había abordado el tema de su validez de contenido. Por tal motivo, el objetivo de este trabajo fue proponer y probar una metodología para obtener evidencias de validez de contenido de los modelos de ítems del Excoba.

En los siguientes apartados se discuten los beneficios, alcances y limitaciones de esta metodología, así como algunas propuestas para estudios futuros que mejoren y fortalezcan los procesos para obtener evidencias de validez dentro del campo de la generación automática de ítems (GAI). Es importante recordar algunos aspectos relacionados con el Excoba, ya que la metodología propuesta se implementó utilizando como insumo un grupo de modelos de ítems que son parte de su estructura y son representativos de las distintas áreas del conocimiento que lo conforman. Por ello, se presenta una breve semblanza de las características y los principios teóricos que fundamentan el Excoba. Asimismo, se retoma la descripción de la metodología utilizada inicialmente y que se adaptó para crear la propuesta planteada aquí.

El Excoba es un examen normativo que se aplica, a gran escala, de manera computarizada. Sus resultados se emplean como criterios para seleccionar a estudiantes que ingresan a la educación media superior, por lo que es considerado como un instrumento de alto impacto. Se encuentra alineado al currículum de educación básica mexicana, ya que evalúa las competencias básicas que se espera que dominen los estudiantes al concluir este nivel de estudios. Debido a

que fue creado bajo los principios de la GAI, los ítems que lo conforman son generados mediante modelos de ítems, los cuales son documentos que contienen los lineamientos para generar automáticamente grandes cantidades de reactivos equivalentes en cuanto a conceptos.

Para elaborar el Excoba y sus modelos de ítems, los desarrolladores utilizaron como fundamento un modelo para la planeación y el desarrollo de instrumentos de evaluación, propuesto originalmente por Nitko en 1994, el cual fue adaptado por primera vez en México por Contreras en 2000, y posteriormente por el INEE en 2005. Dicho modelo fue retomado por los desarrolladores del Excoba, debido a que tiene dos características que lo hacen distinto a los instrumentos tradicionales: 1) su fundamento teórico es la GAI, en la que se utiliza tecnología computarizada y modelos de ítems para generar grandes cantidades de versiones de ítems de un mismo contenido curricular, lo cual lo convierte en un generador de exámenes, y 2) busca aproximarse lo más posible a la evaluación auténtica, mediante el uso de ítems de respuesta construida y semiconstruida.

A lo largo de este trabajo se revisó cómo la GAI es un campo emergente que permite la elaboración de instrumentos, desde dos aproximaciones: teoría fuerte y teoría débil. Cuando se utiliza teoría fuerte, los esfuerzos de los desarrolladores se centran en dos aspectos del instrumento: 1) establecer un modelo cognitivo como punto de partida para descubrir cuáles son los mecanismos cognitivos subyacentes al proceso de solución de los ítems generados, y 2) identificar cuáles son los componentes que determinan los niveles de dificultad de los ítems, con la finalidad de predecir y controlar sus propiedades psicométricas, y así generar ítems *isomorfos*, como les llamó Bejar (2002).

Contrario a lo propuesto por distintos investigadores (véase Bejar, 1993; Gierl y Lai, 2012; Gitomer y Bennett, 2002), si se emplea GAI de teoría débil, como es el caso del Excoba, no es necesario realizar un mapeo de los procesos cognitivos subyacentes al contenido evaluado o a las respuestas de los estudiantes examinados. Se elaboran modelos de ítems que contienen un

conjunto de reglas claras y precisas que dirigen el diseño de los reactivos. En esta aproximación el valor del modelo de ítems es alto, ya que contiene todas las variables necesarias para crear las distintas versiones de los ítems.

En la revisión de literatura en el campo de la GAI se encontró que la validación de instrumentos elaborados mediante sus principios, representa un reto especial para el campo de la psicometría y la evaluación educativa. Esto se debe a que se parte de la premisa de que un modelo de ítems garantiza isomorfismo en todos los ítems que sean generados a partir de él. Además, en instrumentos desarrollados mediante GAI, presentar evidencia empírica respecto a las propiedades psicométricas del conjunto de ítems que conforman un examen resulta insuficiente, porque se debe contar con métodos que aborden la obtención de evidencias de validez de los modelos de ítems, de sus componentes y de los ítems que se generen a partir de ellos.

Asimismo, se hizo evidente cómo durante los últimos 15 años diversos estudios (Bejar, 2003; Singley y Bennett, 2002) han documentado la importancia de establecer un procedimiento riguroso y sistemático para la elaboración de reactivos. También se observó que dichos estudios han propuesto el uso de modelos de ítems para elaborar ítems isomorfos, describiendo procedimientos para su generación automática. También se han realizado estudios como el de Gierl et al. (2008), que proponen taxonomías en la elaboración de distintos tipos de modelos de ítems.

Del GAI Excoba se documentaron dos estudios para la obtención de evidencias de validez cognitiva y de la estructura interna (Pérez, 2013; Ferreyra, 2014); sin embargo fue evidente que no se han realizado estudios que aborden la validación de los contenidos de los modelos de ítems. Por ello, se planteó la necesidad de proponer una metodología que permita contar con evidencias de validez de contenido del instrumento, en cuanto a su desarrollo y a sus modelos de ítems.

Para lograr lo anterior, se tomó como punto de partida la metodología propuesta por Lynn en 1986, que consta de dos fases y cinco etapas. Se llevó a cabo una adaptación a dicha metodología, cuyo producto fue un modelo que consta de dos fases y cuatro etapas, denominado *Modelo para la obtención de evidencias de validez de contenido del Excoba* (MVCE).

Las fases de dicho modelo contemplaron dos acciones importantes en la obtención de evidencias de validez de contenido: 1) la documentación del procedimiento que se siguió para el diseño y elaboración del Excoba, en la que se describió el modelo utilizado por los autores del instrumento, que para fines de esta investigación se denominó *Modelo de elaboración del Excoba* (MEE), y 2) la obtención de evidencias de validez de contenido de 56 modelos de ítems, de cuatro de las áreas que conforman al Excoba (Matemáticas, Historia, Química y Español), cuyo desarrollo implicó el trabajo con paneles de expertos en dichas disciplinas.

De acuerdo con la propuesta de Lynn, parte de las evidencias de validez de contenido se encuentran implícitas durante la etapa de desarrollo de un instrumento. Por eso, la primera fase del MVCE se destinó a la revisión de la literatura y los documentos existentes sobre el diseño del Excoba. Entre los resultados obtenidos se describió el modelo utilizado por los autores del instrumento, el cual se basa en una metodología de diseño robusta y que ha sido probada en otros escenarios evaluativos de orden nacional e internacional (Contreras, 2000; INEE, 2005; Nitko, 1994).

Los estándares publicados en el año 2014 por la APA-AERA y NCME indican que el marco conceptual de una prueba ayuda a precisar el tipo de evidencias que serán recolectadas con la finalidad de evaluar las interpretaciones que de ellas se hagan. Asimismo, señalan que el desarrollo de un instrumento debe incluir una descripción detallada de las habilidades, los conocimientos, procesos, las competencias o características que se pretende medir.

En este sentido, uno de los hallazgos más importantes de esta fase del MVCE fue la confirmación de que al utilizar paneles de especialistas de diferentes disciplinas, los autores del Excoba garantizaron una selección sistemática, colegiada y cuidadosa de los contenidos curriculares que conforman el examen. El trabajo realizado por los expertos que participaron en el diseño del Excoba implicó un cuidadoso análisis reticular del programa de estudios mexicano de la educación básica, de cada una de las asignaturas que conforman la estructura conceptual del instrumento. Por tanto, la elaboración de los modelos de ítems que conforman la primera versión del mismo y que generan las decenas de versiones de los ítems hijos, pasó por controles de calidad rigurosos.

Por otro lado, al analizar el Modelo de Elaboración del Excoba, se observó que no todas las fases se llevaron a cabo en secuencia. Algunos procesos se realizaron en forma paralela, como la creación del sistema informático GenerEx, su interfaz para administrar exámenes y la elaboración de los modelos de ítems por parte de los paneles de expertos. Las fases y etapas del modelo que utilizaron los autores del Excoba para su elaboración se presentan de manera secuenciada; sin embargo, se encontró que el modelo no es lineal, ya que para iniciar una etapa no es necesario que concluya por completo la anterior.

Por ejemplo, el modelo contempla en fases distintas la construcción del GenerEx y la construcción de los modelos de ítems; sin embargo, debido a que el sistema informático está diseñado para cumplir los requerimientos de los modelos de ítems y generar ítems hijos con las características visuales y funcionales que en ellos se exigen, fue necesario trabajar simultáneamente ambas fases. Ello implicó efectuar múltiples ajustes en los modelos de ítems y en la misma programación del sistema. Al analizar estas fases del modelo se pudo determinar cómo fue que los desarrolladores crearon los 22 tipos de modelos programáticos diferentes que se encuentran alojados en el GenerEx, y que permiten operar los 120 modelos de ítems que conforman la estructura conceptual del Excoba.

El proceso de documentar lo anterior fue de gran relevancia. Si bien el Excoba fue diseñado y elaborado con suficiente rigor metodológico, para cumplir las normas técnicas de calidad establecidas por organismos internacionales, no fue sino hasta que se llevó a cabo la presente investigación que se logró documentar –y con ello respaldar–, las acciones y los esfuerzos en el desarrollo de estos instrumentos.

La segunda fase de implementación del MVCE, propuesto en esta investigación, implicó el análisis de los contenidos y modelos de ítems de cuatro de las ocho áreas que conforman el Excoba: Matemáticas, Historia, Química y Español. En total se analizaron 56 modelos de ítems, los cuales representan 47% de los modelos del instrumento. Lo anterior se realizó de acuerdo con los siguientes criterios:

1) Análisis de cada área en forma independiente. Para ello se clasificaron los modelos de ítems según su organización en el programa de estudios, sus requerimientos programáticos en el GenerEx, el tipo de ejecución que realizan los estudiantes al responder, y el tipo de conocimientos que exploran los ítems hijos.
2) Análisis de cada área mediante el trabajo con paneles de expertos. Con esto se determinó la pertinencia y congruencia de los contenidos seleccionados del currículum, así como de la estructura de los modelos de ítems y las características de los ítems hijos que de ellos derivan.
3) Análisis agrupado de los 56 modelos de ítems. Con esto se identificaron regularidades o problemas en común entre los modelos.

A continuación se describen los principales hallazgos. Se abordan según la etapa de la fase 2 del MVCE a la que pertenecen.

5.1. Síntesis de los resultados de la fase 2 del MVCE, etapa 1: análisis de los modelos según las distintas categorías de clasificación

El análisis de los modelos de ítems, según el tipo de programación que se requiere para generar los ítems hijos, permitió identificar que los autores del Excoba enfrentaron retos en el diseño y la generación de los mismos. Los expertos que elaboraron los modelos de ítems y establecieron las reglas y restricciones para su generación presentaron ideas innovadoras y alternativas para su evaluación, buscando el mayor grado de proximidad a la forma auténtica en la que los estudiantes aprenden y son evaluados.

Debido a que la elaboración del Excoba bajo los principios de la GAI es una experiencia inédita, muchos de los retos enfrentados se relacionan con la puesta en marcha y operación de las estrategias evaluativas descritas en los distintos modelos de ítems. Esto implicó vigilar con detenimiento la elaboración del sistema GenerEx. Para ello, se siguió un proceso iterativo mediante el cual las distintas versiones se fueron mejorando significativamente, hasta llegar al sistema actual que permite generar los ítems hijos con las características previstas.

Se encontró que la mayor parte de los modelos de ítems incluidos en este estudio (44.64%) corresponden a un tipo denominado *elemento categoría*. Este tipo utiliza reglas de clasificación para presentar los ítems hijos.

Otro hallazgo importante fue que de las 14 clasificaciones que fueron utilizadas por el GenerEx para operar los 56 modelos de ítems analizados, 42.85% (6) fueron creados para atender las necesidades de un solo modelo. Tal es el caso del tipo llamado *RN/Pendiente*, que se utiliza exclusivamente para generar ítems de un solo modelo de Matemáticas, llamado *MAT19: Pendiente de una recta*. La creación de esta clasificación dentro del GenerEx fue necesaria, ya que dicho modelo establece en su estrategia evaluativa, que se presente un plano cartesiano en la interfaz, en la cual debe aparecer una recta que el sistema genera mediante el empleo de una

fórmula para calcular las coordenadas, y así mostrar los puntos por los cuales traza una pendiente. Ninguna de las 21 clasificaciones de modelos del GenerEx existentes podía atender las necesidades del modelo MAT19, por lo que fue necesario crear esta nueva clasificación.

Lo anterior respalda lo que Solano-Flores et al. señalaron en 1999, respecto al estricto cuidado que se debe tener al elaborar los modelos de ítems. Estos autores afirmaron que los modelos son buenas herramientas para desarrollar instrumentos de evaluación, siempre y cuando sean construidos minuciosamente y midan dominios específicos del aprendizaje. Ello sugiere que la elaboración de los modelos es una tarea compleja, en la que se debe tener cuidado en el nivel de detalle con que se precisan las instrucciones para generar ítems.

En cuanto al tipo de ejecución, es importante recordar que el Excoba es un examen que se aplica y administra mediante el uso de computadoras, y su operación está condicionada a las características de las herramientas informáticas y los accesorios que las acompañan. Por ello, tanto el diseño de los modelos de ítems, como del sistema GenerEx, contemplaron el uso del ratón (mouse) y el teclado para la emisión y registro de las respuestas de los examinados. Esto influyó directamente en el diseño que se siguió para construir el instrumento. En este sentido, y con la finalidad de aproximarse a una evaluación auténtica y alejarse de los ítems de opción múltiple, se encontró que los ítems del Excoba exigen del estudiante cuatro diferentes tipos de ejecución para emitir su respuesta.

En primer lugar, se encontró que hay modelos que generan ítems de *arrastre*, en los que el estudiante mueve elementos mediante el uso del ratón y los deposita en algún espacio de la interfaz. Existen los modelos de *selección*, los cuales generan ítems en los que se utiliza el ratón para seleccionar elementos que se despliegan en ventanas emergentes, cuando se coloca y sostiene el puntero del cursor en algún espacio previamente programado para mostrar dichas ventanas. El Excoba también cuenta con los modelos que generan ítems de *escritura*, en los

cuales se escribe la respuesta mediante el uso del teclado. Por último, un cuarto tipo de ejecución llamada *mixta,* en el que se combina la escritura con el movimiento y la colocación de elementos.

En cuanto a la distribución de los modelos de ítems, de acuerdo con los cuatro tipos mencionados, el hallazgo más relevante fue que más de la mitad (53.57%) requieren que el estudiante realice la acción de mover elementos y depositarlos en lugares específicos de la interfaz para registrar sus respuestas.

Otro hallazgo importante que se logró mediante el análisis de las características de los modelos de ítems, fue respecto al tipo de conocimientos que exploran los ítems hijos que de ellos derivan. De acuerdo con la taxonomía que propuso Ruiz-Primo (2007) existen cuatro tipos de conocimientos, según el nivel de demanda cognitiva que se requiera para ejecutar una tarea: declarativo, procedimental, esquemático y estratégico. El Excoba evalúa conocimientos de los tres primeros tipos, y excluye modelos de ítems que generen ítems hijos en los que la demanda cognitiva es de tipo estratégico.

Al analizar los 56 modelos de ítems se encontraron diferencias entre las cuatro áreas Matemáticas evalúa principalmente conocimientos de tipo procedimental, mientras que Historia únicamente aborda conocimientos esquemáticos. Química y Español se centran en evaluar contenidos declarativos.

Debido a que se trata de un instrumento que evalúa competencias básicas, y a que los conocimientos de tipo estratégico se consideran de orden superior porque implican la integración de los primeros tres (declarativo, procedimental y esquemático), los resultados indican que sí existe congruencia entre los fundamentos teóricos del Excoba y su estructura conceptual, al mantener el nivel de demanda cognitiva en un nivel principalmente procedimental. Sin embargo, no se observó equilibrio en la distribución de los tipos de conocimientos evaluados en estas cuatro áreas del examen.

5.2. Síntesis de los resultados del trabajo realizado con los paneles de expertos

La fase de obtención de evidencias de validez de contenido requirió del trabajo de paneles de expertos para cada una de las cuatro áreas incluidas en este estudio. De acuerdo con el modelo propuesto por Lynn en 1986, es fundamental la participación de expertos para determinar la validez de los contenidos que forman parte de la estructura de un examen. En el caso del desarrollo del Excoba y GenerEx, el trabajo con los paneles de expertos aportó elementos indispensables para determinar tres aspectos fundamentales: a) si los contenidos seleccionados del currículum de la educación básica en las áreas de Matemáticas, Historia, Química y Español de secundaria son pertinentes, representativos e indispensables para reflejar las competencias y los aprendizajes esperados en los estudiantes; b) si los modelos de ítems que conforman el Excoba son suficientes en cuanto a su estructura, estrategia evaluativa, reglas y elementos para generar ítems hijos equivalentes en contenido, y a la vez se aproximan a un proceso de evaluación auténtica, y c) si los modelos generan ítems equivalentes en cuanto a sus contenidos.

Para garantizar que el proceso de trabajo con paneles de expertos se llevara a cabo adecuadamente, se siguieron las recomendaciones de diversos autores citados a lo largo de este documento (Gable y Wolf, 1993; Lynn, 1986; McGartland et al., 2003; Sireci y Faulkner-Bond, 2014; Walz, Strickland y Lenz, 1991). Entre ellas, destacan aspectos como la importancia de seleccionar cuidadosamente a los participantes, asegurarse que posean una amplia experiencia y conocimiento en el tema que corresponda y en los contenidos evaluados; así como prestar atención al proceso de capacitación de los expertos, en cuanto al trabajo y las actividades que desarrollarán.

Uno de los aspectos primordiales que en la selección de los participantes fue que contaran con una amplia experiencia en el manejo no solo del programa de estudios y los contenidos de

la asignatura de su especialidad, sino en la enseñanza de las competencias que el currículum marca. Siguiendo los planteamientos de Sireci y Faulkner-Bond (2014), fue imprescindible llevar a cabo un proceso de capacitación para contextualizar a los expertos en diversas temáticas.

Durante los cursos de capacitación se abordaron temas sobre los procesos asociados a la elaboración de instrumentos de gran escala y al uso de un modelo específico para ello, la importancia que reviste mantener una estricta vigilancia de los estándares de calidad internacional establecidos para instrumentos de evaluación educativa, así como las implicaciones de la tarea que realizarían como evaluadores expertos. Con la implementación de estos controles de calidad y la sensibilización de los expertos en la naturaleza de la actividad a realizar, la evaluación de los modelos de ítems se llevó a cabo de manera minuciosa. Se dedicó tiempo exhaustivo a la revisión, discusión y el análisis de cada uno de los elementos cadena e integrales que conforman los modelos de ítems, y se pudo obtener información respecto a su correspondencia, congruencia, pertinencia y viabilidad.

Entre los hallazgos sobresalientes, los expertos estimaron que entre 75% y 100% de los modelos de ítems representan las habilidades y los conocimientos básicos que deben dominar los estudiantes al término de la secundaria. Por otra parte, el porcentaje de estudiantes que consideran que responderían correctamente los ítems hijos fue considerablemente menor, en un rango entre 50% y 70%.

También se logró identificar la presencia de patrones en cuanto al tipo de problemas que se presentaron en los modelos de ítems. Dichos problemas se relacionan con: la coherencia de la estrategia utilizada para evaluar algunos contenidos y la forma en la que se enseñan en el aula; las características de los ítems hijos y su capacidad para evaluar si los estudiantes saben utilizar el conocimiento adquirido en el aula; la falta de claridad en la base de algunos ítems hijos; la

calidad de los gráficos, y los niveles de dificultad de los elementos que conforman el banco de información de algunos modelos de ítems.

Si bien la mayoría de los contenidos seleccionados del currículum durante la elaboración del instrumento fueron ratificados como esenciales por los expertos que participaron en este estudio, también se pudieron identificar en cada área aquellos que no lo son y que, a su juicio, impedían la inclusión de algunos verdaderamente esenciales. Un ejemplo de ello se presentó en el área de Español, donde fueron identificados cinco contenidos que, en opinión de los expertos, son importantes mas no esenciales para el aprendizaje de la asignatura en conjunto: *Géneros literarios, Reglamento o instructivo, Documentos que establecen derechos y obligaciones, Llenado de formatos, y Recursos lingüísticos y visuales en los mensajes publicitarios*. En tales casos se solicitaron sugerencias de contenidos más representativos de lo que el estudiante debe aprender.

Finalmente, en cuanto a la equivalencia de los ítems hijos de un mismo modelo los resultados permitieron confirmar que –salvo casos muy específicos en los que se encontraron elementos del banco de información que no representan puntualmente los contenidos marcados en el programa de estudios– la gran mayoría de los modelos de ítems de las cuatro áreas analizadas sí son equivalentes estructural y conceptualmente. Lo anterior sugiere que el nivel de detalle y especificidad de los distintos componentes de los modelos de ítems debe ser alto. Esto es, entre más explícitas sean las reglas para generar ítems, mayor relación tendrán entre sí los ítems que deriven de un mismo modelo.

Es importante señalar que los resultados presentados en este trabajo pueden considerarse como parciales, ya que los datos sobre la validez de contenido de los modelos de ítems no son de todas las áreas que conforman el generador de exámenes Excoba. Sin embargo, el análisis de la estructura y los contenidos de los modelos de ítems que fueron presentados, y que son

esenciales para todo instrumento de evaluación, arrojaron información que apoya la noción de que se trata de un modelo metodológico que puede emplearse en la evaluación de instrumentos creados bajo los principios de la GAI de teoría débil.

5.3. Alcances y limitaciones de la metodología utilizada

Hasta el momento los procesos de obtención de evidencias de validez se han centrado principalmente en la revisión de la estructura interna de las dimensiones que conforman los constructos evaluados. Se pudo constatar que, a excepción de dos estudios para obtener evidencias de validez del Excoba (Pérez, 2013; Ferreyra, 2014), en la extensa literatura revisada no hay información que documente estudios de validez de contenido en instrumentos de este tipo. Lo anterior mostró cómo el Excoba –además de contar con un procedimiento de planeación, diseño y elaboración riguroso–, ha sido objeto de estudios respecto a la validez de su estructura interna y de los procesos cognitivos que subyacen a las respuestas emitidas por los estudiantes.

El presente trabajo de investigación representa una importante aportación, en términos de obtener los distintos tipos de evidencias de validez necesarios para considerar que los resultados derivados de la aplicación de un instrumento elaborado mediante los principios de la GAI de teoría débil son útiles y ayudan a la toma de decisiones. La información generada no solo es útil para los autores del instrumento, sino que permea en el trabajo de todos aquellos que incursionen en la elaboración de instrumentos de gran escala bajo los principios de la GAI. Esto se debe a que representa una primera propuesta metodológica para obtener evidencias de validez de contenido en exámenes no adaptativos, generados mediante los principios mencionados.

Arendasy (2005) ha señalado que el futuro de los instrumentos de evaluación de habilidades psicológicas se encuentra dentro del campo de la GAI, debido a factores que tienen que ver con controles de calidad. El Excoba procura cumplir criterios de calidad, tales como la

cuidadosa selección de los contenidos evaluados a través de la revisión exhaustiva del currículum de la educación básica, el análisis de los contenidos y modelos de ítems mediante la evaluación por paneles de expertos, y la creación de un sistema informático que permite la implementación de reglas y restricciones para el establecimiento de procedimientos para la generación de ítems hijos equivalentes.

Al implementar el MVCE se pudo constatar cómo el Excoba cumple con los procesos que propuso Luecht en 2012. En primer lugar, durante su planeación y diseño se contemplaron aspectos como la elaboración de mapas de constructos y la forma en la que los datos serían interpretados. Debido a su alineación con el currículum fue imprescindible definir a priori la importancia que revisten los contenidos que marca el programa de estudio, así como determinar y justificar los motivos por los que se consideraron representativos del currículum de las asignaturas contempladas en la estructura del instrumento. Por ello, se puede decir que el punto de partida para la elaboración del Excoba fue firme y evita una inversión económica sustancial en la elaboración y validación de sus modelos de ítems e ítems hijos a futuro.

En segundo lugar, el Excoba contiene el equivalente de lo que Luecht (2012) llamó *modelos de tareas* y *mapas de modelos*, ya que cuenta con 120 modelos de ítems diferentes, clasificados en 22 tipos según la programación requerida para su funcionamiento. Operan mediante el uso de reglas que fueron establecidas para generar decenas o cientos de ítems hijos. Además, el Excoba utiliza plantillas de ítems en los modelos, las cuales contienen los elementos cadena e integrales que al combinarse generan grandes cantidades de ítems hijos equivalentes en contenido, lo cual es congruente con el tercer proceso propuesto por Luecht.

Por otro lado, mediante la revisión y documentación de su diseño se pudo constatar que durante la elaboración del Excoba se realizó un estudio que permitió establecer un primer modelo de aproximación para obtener evidencias de validez de constructo, con el que se dieron a conocer las propiedades métricas de las distintas versiones de los exámenes que genera, así

como de sus familias de ítems e ítems hijos. Los resultados de dicho estudio también mostraron qué tan isomorfos son los ítems hijos de una misma familia de reactivos, tomando en cuenta su grado de dificultad y la competencia que los define (Ferreyra, 2014).

Finalmente, mediante la implementación del MVCE se determinó la medida en la que los modelos de ítems se aproximan a procesos de evaluación auténtica, evalúan los aprendizajes considerados indispensables para cimentar la adquisición de nuevos conocimientos, y son apropiados para el nivel y grado escolar de un estudiante que concluye su educación secundaria.

De acuerdo con Porter (2002), es muy importante conocer el grado de alineación que existe entre los contenidos evaluados en un examen y el estilo instruccional empleado por el maestro. En este sentido, el MVCE permitió que durante la fase de trabajo con los paneles de expertos se obtuviera información respecto a su nivel de percepción de la alineación entre la forma en la que se enseña en el aula y la estrategia evaluativa para generar los ítems hijos. Esta información permitió saber que desde la perspectiva de los expertos –aunque en casos específicos y no en el instrumento en general–, sí existe diferencia entre la alineación curricular y algunos de los modelos de ítems de las distintas asignaturas del Excoba incluidas en el estudio.

Si bien la metodología propuesta para obtener evidencias de validez de contenido del Excoba proporcionó información para mejorar el instrumento, también reflejó algunas limitaciones. Una de ellas es de naturaleza práctica, porque está relacionada con la calidad del trabajo efectuado por los paneles de expertos. La recomendación de diversos especialistas en la materia es que se consideren de 3 a 10 expertos; sin embargo cada panel de esta investigación estuvo conformado por 2 integrantes.

La situación anterior se debió a factores como la falta de acceso del investigador a escenarios educativos en los que se encuentran colectivos de docentes, la disponibilidad de tiempo de los expertos, así como el hecho de no haber remuneración económica por su

participación. A pesar de ello, la información proporcionada por los expertos que integraron los paneles fue muy enriquecedora, porque permitió detectar aspectos del instrumento que pasarían desapercibidos sin su intervención. Uno de esos aspectos son los contenidos que fueron excluidos del Excoba, y que son considerados imprescindibles en el currículum. De esta manera, se confirmó que una de las situaciones que más repercuten en los resultados obtenidos en un proceso de evaluación y análisis como el que se realizó, es la adecuada selección de los integrantes que conforman los paneles de expertos.

5.4. Sugerencias para nuevas líneas de investigación

La metodología propuesta en este trabajo representa una primera aproximación a la obtención de evidencias de validez de contenido de un generador automático de ítems. Por tanto, para fortalecerla es recomendable que se lleven a cabo estudios posteriores.

Un análisis muy importante que debe realizarse es la exploración de los tipos de conocimiento que evalúan los distintos modelos de ítems del Excoba. En los análisis realizados en este estudio se detectó que los ítems hijos que derivan de los modelos exploran tres tipos de conocimientos, según la taxonomía propuesta por Ruiz-Primo en 2007. Los conocimientos de tipo estratégico implican que el estudiante ponga en práctica habilidades de reflexión sobre los procesos implicados en la toma de decisiones, que utilice estrategias para evaluar alternativas y elegir la solución más adecuada a un problema; de tal manera que se observe la presencia combinada de habilidades relacionadas con los conocimientos de tipo declarativo, procedimental y esquemático, como cimientos del conocimiento de tipo estratégico.

Asimismo, será muy útil un análisis a profundidad que verifique si el Excoba contiene modelos de ítems que evalúan conocimientos de tipo estratégico, ya que puede generar una alternativa a la elaboración de exámenes que exploren niveles cognitivos de orden complejo.

Definir qué tipos de conocimientos serán evaluados en un instrumento es de gran importancia, ya que este tipo de análisis ayuda a guiar la toma de decisiones durante la etapa de diseño, en el sentido de que se puede definir a priori el nivel de demanda cognitiva que se desea en los ítems hijos. Por ello, también es importante realizar un estudio de validez cognitiva de la totalidad de los modelos de ítems del Excoba. Y aunque existe ya una investigación que analizó la validez del proceso y de la estructura cognitiva de los ítems, únicamente incluyó ítems que evalúan competencias del área denominada *habilidades cuantitativas,* quedando aún por analizar el resto de las áreas del instrumento.

Algo muy importante que no fue posible realizar en este estudio fue incluir la totalidad de las áreas contempladas por el Excoba. Sería de gran beneficio contar con un análisis de toda la estructura conceptual del instrumento, en el cual se incluyan los 64 modelos de ítems restantes que no fueron analizados mediante la metodología propuesta. Esto proporcionaría dos tipos de beneficios importantes: por un lado, permitiría contar con información valiosa para apoyar ajustes y mejorías al instrumento, dando pie a obtener evidencias de validez de los contenidos de todos sus modelos de ítems, y no solamente de áreas específicas de su estructura conceptual; por otro lado, el MVCE debe seguir siendo probado para modificarlo y hacerlo más eficiente. Implementarlo en el resto de los modelos de ítems del Excoba permitirá obtener información que ayude a este proceso.

Otra línea de investigación que debe abordarse es la congruencia entre el instrumento y el estilo instruccional de los docentes. Si bien el MVCE permitió conocer la percepción de los expertos respecto al grado de alineación entre la estrategia evaluativa empleada en los modelos de ítems y la forma en la que se enseñan los contenidos en el aula, es de gran relevancia realizar un estudio para explore en qué medida se encuentran alineados aspectos como la forma en la que los docentes enseñan el currículum mexicano, los materiales que utilizan durante el proceso de enseñanza-aprendizaje, los contenidos y aprendizajes esperados que marca el currículum, así

como los resultados de dicho proceso. Un contraste entre estos elementos y las características de los ítems hijos que se generan mediante los modelos de ítems del Excoba sería de gran relevancia para mejorar el instrumento.

Cabe recordar que los estudios de Pérez (2013) y Ferreyra (2014) fueron realizados para conocer los procesos cognitivos subyacentes a las respuestas de los examinados y la estructura interna de los ítems del Excoba. Ambos aportaron modelos metodológicos que permiten contar con procedimientos sistemáticos para obtener evidencias de validez de la estructura cognitiva de los ítems del Excoba, y de la estructura interna del instrumento desde distintas aproximaciones analíticas.

El MVCE propuesto en esta investigación incluye un nivel de análisis que explora los tipos de conocimientos y procesos cognitivos involucrados durante la solución de problemas. Asimismo, mediante la operación del modelo se obtuvo información relacionada con la percepción de los expertos sobre el grado de dificultad de los ítems hijos generados mediante los modelos analizados, así como las características estructurales que pudieran alterar las respuestas de los examinados. Esta información requiere analizarse de manera conjunta con los resultados obtenidos en los estudios de Pérez (2013) y Ferreyra (2014). Por ello, se propone un estudio de *metaevaluación* en el que confluyan las tres investigaciones de validez realizadas con los modelos e ítems del Excoba, para buscar coincidencias en los resultados.

Al llevar a cabo la presente investigación, se pudo observar que los distintos instrumentos elaborados mediante los principios de la GAI coinciden en que el uso de modelos de ítems permite generar grandes cantidades de ítems isomorfos; sin embargo, la revisión de literatura sobre las propuestas existentes para evaluar el aprendizaje en estudiantes, permitió evidenciar que la mayoría de los instrumentos aún utilizan formatos de reactivos de opción múltiple y en algunos casos de respuesta construida.

El presente trabajo abordó un campo de la evaluación novedoso y poco explorado: la validación de ítems y pruebas generadas con la GAI. Por este motivo, los procedimientos para obtener evidencias de validez son actualmente los mismos que se utilizan para los ítems y pruebas generadas de forma tradicional. Mediante este trabajo se propuso un modelo que se adscribe a dichos procedimientos de validación y que se aplicó al caso concreto de un generador de exámenes llamado Excoba.

Durante la implementación de la metodología se realizaron diferentes ajustes y se descubrieron elementos que representan una aportación importante a esta ciencia emergente, llamada así por Haladyna y Rodríguez (2013). Sin embargo, aún falta mucho por hacer, pues aunque los distintos esfuerzos de distintos investigadores por sistematizar los procedimientos implementados son alentadores, siguen siendo incipientes.

Referencias

American Educational Research Association, American Psychological Association y National Council on Measurement in Education [AERA, APA y NCME] (1985). *Standards for educational and Psychological Testing*. Washington D. C.: American Educational Research Association.

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

American Educational Research Association, American Psychological Association y National Council on Measurement in Education [AERA, APA y NCME] (2014). *Standards for educational and Psychological Testing*. Washington D. C.: American Educational Research Association.

Anastasi, A. (1988). *Psychological testing*. Sexta edición. Nueva York: Mac-millan.

Anastasi, A. (1992). Tests and assessment. What counselors should know about the use and interpretatios of psychological tests. *Journal of counseling and development*, 70(5), 610-615 (EJ446913)

Archbald, D. A. y Newmann, F. M. (1988). *Beyond standardized testing: Assessing authentic academic achievement in the secondary school*. Reston, Virginia: National Association of Secondary School Principals.

Arendasy, M. (2005). Automatic Generation of Rasch-Calibrated Items: Figural Matrices Test GEOM and Endless-Loops Test EC. *International Journal of Testing*, 5(3) 197–224.

Arendasy, M. y Sommer, M. (2012). Using automatic item generation to meet the increasing item demands of high-stakes educational and occupational assessment. *Learning and individual differences*, 22, 112-117.

Backhoff, E. E., Ibarra, R. M. A. y Rosas, M. M. (1996). Desarrollo y validación del sistema computarizado de exámenes (SICODEX). *Revista de la Educación Superior*. México: Asociación Nacional de Universidades e Instituciones de Educación Superior, XXV, I (97).

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

Backhoff, E. E., Larrazolo, R. N., Ramírez, J. L., Rosas, M. M. y Tirado, S. F. (2015). *Excoba: Examen de Competencias Básicas*. México: Instituto Nacional de Derechos de Autor.

Backhoff, E. E. y Tirado, S. F. (1992). Desarrollo del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). *Revista de la Educación Superior*, XXI, 3(83). México: Asociación Nacional de Universidades e Instituciones de Educación Superior.

Backhoff, E. E., Sánchez, A., Peón, M., Monroy, L. y Tanamachi, M. (2006). Diseño y Desarrollo de los Exámenes de la Calidad y el Logro Educativos *Revista Mexicana de Investigación Educativa*, abril-junio 2006, 11(29), 617-638.

Baker, E. L. (2010). What Probably Works in Alternative Assessment. *National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. 20 pp. (ED512658).

Barab, S., Thomas, M., Dodge, T., Carteaux, R., y Tuzun, H. (2005). Making Learning Fun: Quest Atlantis, a Game without Guns. *Educational Technology Research and Development*, 53(1), 86-107.

Barton. P. (1999). Too Much Testing of the Wrong Kind; Too Little of the Right kind in K-12 education. *A Policy information perspective*. Educational Testing Service, Princeton, NJ. Policy Information Center. (ED430052).

Bejar, I. I. (1993). A generative approach to psychological and educational measurement. En N. Frederikson, R. J. Mislevy e I. I. Bejar (Eds.). *Test theory for a new generation of tests* (pp. 323-359). Mahwah, NJ: Erlbaum.

Bejar, I. I. (1996). *Generative response modeling: Leveraging the computer as a test delivery medium.* (Reporte de investigación de ETS 96-13). Princeton, NJ: Educational Testing Service.

Bejar, I. I. (2002). Generative testing: from conception to implementation. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development,* (pp. 199-218). Mahwah, NJ: Lawrence Erlbaum Associates.

Bejar, I. I., Lawless, R. R., Morley, M. E., Wagner, M. E., Bennett, R. E., y Revuelta, J. (2003). A feasibility study of on-the-fly item generation in adaptive testing. *Journal of Technology, Learning, and Assessment*, 2(3), 1-29.

Bezruczko, N. (2014). Automatic item generation implemented for measuring artistic judgment aptitude. *Journal of applied measurement*, 15(1), 1-25.

Bormuth, J. R. (1970). *On a theory of achievement test items*. Chicago: University of Chicago Press.

Borsboom, D. (2008). Latent variable theory. *Measurement*, 6, 25–53.

Borsboom, D. y Mellenbergh, G. J. (2004). The concept of validity. *Psychological review*, 111(4), 1061-1071.

Boyd-Batstone, P. (2004). Focused anecdotal records assessment: A tool for standards-based, authentic assessment. *International Reading Association*, 58(3), 230–240. (EJ684387)

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

Bullens, D. (2002). *Authentic assessment: Change for the future*. Proyecto de investigación acción para obtener el grado de Master en Artes. Saint Xavier University. Chicago, Illinois: EE UU. (ED468067).

Centro Nacional de Evaluación para la Educación Superior (2013). *Guía del examen nacional de ingreso a la educación media superior EXANI-I*. Vigésima edición. México: Centro Nacional de Evaluación para la Educación Superior, A. C.

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

College Board (2011a). *Guía de estudio para presentar las Pruebas para el Ingreso al Nivel de Educación Media Superior PIENSE II*. The College Board Puerto Rico y América Latina.

College Board (2011b). *Guía de estudio para presentar la Prueba de Aptitud Académica*. The College Board Puerto Rico y América Latina.

College Board (2011c). *Prueba de Aptitud Académica (PAA).* The College Board Puerto Rico y América Latina. Recuperado de: http://oprla.collegeboard.com/ptorico/pr/program/paa.html

College Board (2012). *Guía de estudio para tomar las pruebas del Programa de Evaluación y Admisión Universitaria 2012-2013*. The College Board Puerto Rico y América Latina.

Committee on the Foundations of Assessment (2001). *Knowing What Students Know: The Science and Design of Educational Assessment*. James W. Pellegrino, Naomi Chudowsky, and Robert Glaser (Eds.). Board on Testing and Assessment, Center for Education, National Research Council. National academy press: Washington, D.C.

Consejo Asesor Externo del Centro Nacional de Evaluación para la Educación Superior [CENEVAL] (2000). *Estándares de calidad para instrumentos de evaluación educativa*. México: Centro Nacional de Evaluación para la Educación Superior, A.C. (Ceneval).

Contreras Niño, L. A. (2000). *Desarrollo y Pilotaje de un Examen de Español para la Educación Primaria de Baja California* (Tesis de maestría). Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California, México.

Costa, A. (Ed.), Kallick, B. (1992). Reassessing assessment. En A. Costa, J. Bellanca y R. Fogarty (Eds.). *If minds mater: a foreword to the future,* Vol. 2 (pp. 275-280). EE UU: IRI/Skylight.

Cronbach, L. J. (1971). Test validation. En R. L. Thorndike (Ed.), *Educational Measurement* (2a. ed., pp. 443-507). Washington, DC: American Council on Education-The Oryc Press.

Cronbach, L. J. y Meehl, P. E. (1955). Construct validity in psychological tests. *Psychologycal Bulletin*, 52(4), 281-302.

Cumming, J. J., y Maxwell, G. S. (1999). Contextualizing Authentic Assessment. *Assessment in Education: Principles, Policies, and Practices,* 6(2), 177–194.

Draaijer, S. y Hartog, R. J. M. (2007). Design patterns for digital item types in higher education. E-*Journal of instructional science and technology*, 10(1), 1-32.

Drasgow, F., Luecht, R. M. y Bennett, R. (2006). Technology and testing. En R. L. Brennan (Ed.), *Educational measurement* (4ta ed., pp. 471-516). Washington, DC: American Council on Education.

Embretson, S. E. (1999). Generating items during testing: psychometric issues and models. *Psychometrika*, 64(4), 407-433.

Embretson, S. E. (2002). Generating abstract reasoning items with cognitive theory. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 219-250). Mahwah, NJ: Lawrence Erlbaum Associates.

Enright, M. K., Morley, M., y Sheehan, K. M. (2002). Items by design: The impact of systematic feature variation of item statistical characteristics. *Applied measurement in education*, 15(1), 49-74.

Ferreyra, F. (2014). *Modelo para la validación empírica del EXHCOBA-R, producido por un generador automático de reactivo* (Tesis doctoral). Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California, México.

Finn, P. J. (1975). A question writing algorithm. *Journal of reading behavior*, 4, 341-367.

Gable, R. y Wolf, J. (1993). *Instrument development in the affective domain: meassuring attitudes and values in corporate and school settings*. Boston: Kluwer Academic.

Geerlings, H., Glass, C. A. W. y Van der Linden, W. J. (2011). Modeling rule-based item generation. *Psychometrika*, 76(2), 337-359.

Gierl, M. J. (2007). Making diagnostic inferences about cognitive attributes using the rule-space model and attribute hierarchy method. *Journal of educational measurement*, 44, 325-340.

Gierl, M. J. y Hollis, L. (2012). Using weak and strong theory to create item models for automatic item generation. Some practical guidelines with examples. En Mark J. Gierl y Thomas M. Haladyna (Eds.), *Automatic Item Generation: Theory and practice* (pp. 26-39). New York: Routledge.

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

Gierl, M. J., Lai, H. y Breithaupt, K. (2012, abril). *Methods for creating and evaluating the item model structure used in automatic item generation*. Ponencia presentada en la reunión anual del National Council on Measurement in Education en Vancouver, Canadá.

Gierl, M., Leighton, J., Changjiang, W., Jiawen, Z., Rebecca, G., y Tan, A. (2009). *Validating Cognitive Models of Task Performance in Algebra on the SAT*. Research Report 2009-3. College Board, Research Report (Vol. No. 2009–3). New York.

Gierl, M. J., Zhou, J., y Alves, C. (2008). Developing a taxonomy of item model types to promote assessment engineering. *The Journal of Technology, Learning, and Assessment*, 7(2).

Gitomer, D. H. y Bennett, R. E. (2002). *Unmasking Constructs Through New Technology, Measurement Theory, and Cognitive Science* (Memorandum de investigación, Febrero de 2002, RM-02-01). Educational Testing Service. Statistics and research division. Princeton, NJ.

Glas, C. A. W. y Van der Linden, W. J. (2003). Computarized adaptative Testing with item cloning. *Applied Psychological Measurement*, 27, 247-261.

González-Montesinos, M. (2004). *Defining and measuring academic standards for higher education: a formative study at the University of Sonora* (Tesis doctoral no publicada). University of Arizona, Tucson, AZ, Estados Unidos de América.

Goodrich, H. (2001). *Project Zero Research projects: rubrics and self-assessment project*. Recuperado de: http://www.pz.harvard.edu/rubrics.php

Gulikers, J. M., Bastiaens, T.J., y Kirshner, P.A. (2004). A Five-Dimensional Framework for Authentic Assessment. *Educational Technology Research and Development*, 52(3), 67–87. (EJ732658).

Guttman, L. (1959). Introduction to facet design and analysis. *Acta Psychologica*, 15, 130-138.

Haladyna, T. M. (1991). Generic questioning strategies for linking teaching and testing. *Educational technology: research and development*, 39, 73-81.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items.* Tercera edición. Estados Unidos: Routledge.

Haladyna, T. M. (2012). Automatic item generation: A historical perspective. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 13-25). Nueva York: Routledge.

Haladyna, T. M. y Rodríguez, M. (2013). *Developing and validating test items*. New York: Routledge.

Haladyna, T. M. y Shindoll, R. R. (1989). Item shells: A method for writing effective multiple-choice test items. *Evaluation and the Health Professions*, 12, 97-106.

Haynes, S., Richard, D. y Kubany, E. (1995) Content Validity in Psychological Assessment: a functional approach to concepts and methods. *Psychological Assessment* 7(3), 238-247.

Hernández, S. R, Fernández, C. C. y Baptista, L. P. (2006) *Metodología de la investigación*. Cuarta edición.

Hively, W. (1974). Introduction to domain-referenced testing. *Educational technology*, 14(6), 5-10.

Hively, W., Patterson, H. L. y Page, S. H. (1968). A "universe-defined" system of arithmetic achievement tests. *Journal of Educational Measurement*, 5, 275-290.

Holling, H., Berling, J. P. y Zeuch, N. (2009). Automatic item generation for probability word problems. *Studies in Educational Evaluation*, 35, 71-76.

Hombo, C. y Dresher, A. (2001). *A simulation study of the impact of automatic item generation under NAEP-like data conditions*. Ponencia presentada en Annual Meeting of the National Council on Measurement in Education en Seatle, WA, Estados Unidos.

Instituto Nacional para la Evaluación de la Educación (INEE). (2005). Plan general de evaluación del aprendizaje. Dirección de pruebas y medición. Dirección de proyectos internacionales y especiales. México: INEE.

Irvine, S. H. (2002). The foundations of item generation in mass testing. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 3-34). Mahwah, NJ: Lawrence Erlbaum Associates.

Irvine, S. H. y Kyllonen, P. C. (Eds.). (2002). *Item generation for test development*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Jensen, A. R. (1980). Bias in mental testing. Nueva York: The free press.

Kallic, B. (1992). Evaluation: a colaborative process. En A. Costa, J. Bellanca y R. Fogarty (Eds.). *If minds mater: a foreword to the future* Vol. 2 (pp. 313-319). EE UU: IRI/Skylight

Kerlinger, F. N. y Lee, H. B. (2002). *Investigación del comportamiento*. Cuarta edición. México: McGraw-Hill pp. 603-625.

Keyser, S. y Howell, S. (2008). The State of Authentic Assessment. Recuperado de base de datos ERIC. (ED503679) http://files.eric.ed.gov/fulltext/ED503679.pdf.

LaDuca, A., Staples, W. I., Templeton, B. y Holzman, G. B. (1986). Item modeling procedure for constructing content-equivalent multiple-choice questions. *Medical Education*, 20, 53-56.

Lai, H., Alves, C. y Gierl, M. J. (2009, junio). *Using automatic item generation to address item demands for CAT.* Ponencia presentada en el 2009 GMAC Conference on Computerized Adaptive Testing.

Lawshe, C. (1975). A quantitative approach to content validity. *Personnel Psychology*. 28, 536-575.

Linn, R. L. (2006). The Standards for Educational and Psychological Testing: guidance in test development. En S. M. Downing y T. M. Haladyna (Eds.), *Handbook of test development* (pp. 27-38). New Jersey: Lawrence Erlbaum Associates.

Liu, M., y Haertel, G. (2011). *Design Patterns: A Tool to Support Assessment Task Authoring* (Reporte técnico 11: Evaluación a gran escala). Menlo Park, CA: SRI International.

Lord, F. M. (1971). The self-scoring flexi level test. *Journal of Educational Measurement*, 31, 3–31.

Lowyck, J. y Poysa, J. (2001). Design of collaborative learning environments. Computers in Human Behavior, *17*(5-6), 507-516.

Luecht, R. M. (2006a, mayo). *Engineering the test: Principled item design to automated test assembly.* Ponencia presentada en la reunion annual de la Society for Industrial and Organizational Psychology en Dallas, Texas, Estados Unidos.

Luecht, R. M. (2006b, septiembre). *Assessment Engineering: an emerging discipline*. Ponencia presentada en el Centre for research in applied measurement and Evaluation, Universidad de Alberta, Edmonton, Alberta, Canadá.

Luecht, R. M. (2007a, abril). *Assessment engineering in language testing: From data models and templates to psychometrics*. Ponencia invitada en la reunion anual del National Council on Measurement in Education, Chicago, IL.

Luecht, R. M. (2008a, febrero). *Assessment engineering*. Ponencia presentada en el panel: Assessment Engineering: Moving from Theory to Practice en la reunion annual de la Association of Test Publishers, Dallas, TX, Estados Unidos.

Luecht, R. M. (2008b, octubre). *Assessment engineering in test design, development, assembly, and scoring*. Conferencia magistral presentada en la reunión anual de la East Coast Language Testing Organizations (ECOLT), Washington, D. C. Estados Unidos.

Luecht, R. M. (2009, junio). *Adaptive computer-based tasks under an assessment engineering paradigm.* Ponencia presentada en el Congreso de la GMAC CAT, Minneapolis, MN, Estados Unidos.

Luecht, R. M. (2010, abril). *Controlling difficulty and security for complex computerized performance exercises using assessment engineering*. Ponencia presentada en la Reunión anual del National Council on Measurement in Education, Denver, CO, Estados Unidos.

Luecht, R. M. (2011, marzo). *Assessment design and development, version 2.0: from art to engineering*. Conferencia de clausura presentada en la reunión anual de la Association of Test Publishers, Phoenix, Arizona, Estados Unidos.

Luecht, R. M. (2012). An introduction to assessment engineering for automated item generation. En Mark J. Gierl y Thomas M. Haladyna (Eds.), *Automatic Item Generation: Theory and practice* (pp. 59-76). New York: Routledge.

Luecht, R. M., Burke, M., y Devore, R. (2009, Abril). *Task modeling of complex computer-based performance exercises*. Ponencia presentada en la reunión anual del National Council on Measurement in Education, San Diego, CA.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research*, 35(6), 382-385.

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

Markle, S. M. y Tiemann, P. W. (1970). *Really understanding concepts*. Champaign, Ill: Stipes.

Markus, K. y Borsboom, D. (2013). Reflective measurement models, behavior domains, and common causes. *New Ideas in Psychology* 31, 54–64.

Martínez, M. A., Hernández, M. J. y Hernández, M. V. (2006). *Psicometría*. Madrid: Alianza.

McGartland, D., Berg-Weger, M., Tebb, S., Lee, S. y Rauch, S. (2003). Objectifying Content Validity: conducting a Content Validity Study in Social work research. *Social work research,* 27(2), 94-104.

Messick, S. (1981). Evidence and ethics in the evaluation of tests. *Educational researcher*, 10(9), 9-20.

Messick, S. (1993). Validity. En R. L. Linn (Ed.), *Educational Measurement*. Tercera edición (pp. 13-103). Nueva York: American Council on Education y Orix Press.

Miller, M.D., Linn, R.L., y Gronlund, N.E. (2008). *Measurement and Assessment in Teaching*. Décima edición. EE UU: Pearson.

Minsky, M. (1974). *A framework for representing knowledge* (Memo no. 306). MIT Artificial intelligence laboratory. Cambridge, Massachusetts.

Mislevy, R. J. y Rinconscente, M. M. (2006). Evidence-centered assessment design. En S. Downing y T. Haladyna (Eds.), *Handbook of test development* (pp. 61-90). Nueva Jersey, EE UU: Lawrence Erlbaum Associates.

Nitko, A. (1994, julio). *A Model for Developing Curriculum-Driven Criterion-Referenced and Norm-Referenced National Examinations for Certification and Selection of Students*. Ponencia presentada en la Conferencia Internacional sobre Evaluación y Medición Educativas, de la Asociación para el Estudio de la Evaluación Educativa en Sudáfrica (ASSESA), Pretoria, Sudáfrica.

Nunnally, J. y Bernstein, I. (1994). *Psychometric theory*. Tercera edición. Nueva York: McGraw-Hill.

Osburn, H. G. (1968). Item sampling for achievement testing. *Educational and psychological measurement*, 28, 95-104.

Patterson, G. (1993). Orderly change in a stable world: The antisocial trait as a chimera. *Journal of Consulting and Clinical Psychology*, 61, 911-919.

Pérez, J. C. (2013). *Análisis de validez del proceso y de la estructura cognitiva de una prueba de Habilidades Cuantitativas* (Tesis doctoral). Instituto de Investigación y Desarrollo Educativo de la Universidad Autónoma de Baja California, México.

Popham, W. J. (1990). *Modern Educational Measurement: A practitioner's Perspective*. Segunda edición. Indiana, EE UU: Prentice Hall.

Popham, W. J. (2000). *Modern Educational Measurement: Practical guidelines for educational leaders*. Tercera edición. Boston, EE UU: Allyn and Bacon.

Porter, A. C. (2002). Measuring the content of instruction: uses in research and practice. *Educational researcher,* 31(7) 3-14.

Quellmalz, E., Schank, P., Hinojosa, T., y Padilla, C. (1999). Performance assessment links in science. *Practical Assessment, Research & Evaluation*, 6(10).

Randal, K. (2004). Cleveland's Good for Business: Students learn real business skills in imaginary Texas town. *Feature story archives.* Archivo electrónico del 26 de abril - 3 de mayo, Recuperado de http://www.utexas.edu/features/archive/2004/cleveland.html

Revuelta, J. y Ponsoda, V. (1998). Un test adaptativo informatizado de análisis lógico basado en la generación automática de ítems. *Psicothema*, 10, 753-760.

Rinconscente, M. M., Mislevy, R. J., Hamel, L. y grupo PADI (2005). *An introduction to PADI task templates* (Reporte técnico 3). Menlo Park, CA: SRI International.

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 Psicología*, 40, 85–95.

Roid, G. H. y Finn, P. (1977). *Algorithms for developing test questions from sentences in instructional materials* (Reporte interno). San Diego, California: Navy personnel research and development center.

Roid, G. H. y Haladyna, T. M. (1978).The use of domains and item forms in the formative evaluation of instructional materials. *Educational and psychological measurement*, 38, 19-28.

Ruiz-Primo, M. A. (2007). Assessment in science and mathematics: lessons learned. En M. Hoepfl y M. R. Lindstrom (Eds.) *Assessment of technology education 56th yearbook, 2007.* Council on technology teacher education (pp. 203-232). McGraw Hill: EEUU.

Rule, A. (2006). The components of authentic learning. *Journal of Authentic Learning*, 3(1), 1–10.

Savery, J. R., y Duffy, T. M. (2001). *Problem based learning: an instructional model and its constructivist framework* (Reporte técnico 16-01). Bloomington, Indiana: Center for Research on Learning and Technology, Indiana University.

Secretaría de Educación Pública (SEP) (2006a). Educación básica. Secundaria. Ciencias. *Programas de estudio 2006*. México: SEP. ISBN 968-9076-07-8.

Secretaría de Educación Pública (SEP) (2006b). Educación básica. Secundaria. Español. *Programas de estudio 2006*. México: SEP. ISBN 968-9076-04-3.

Secretaría de Educación Pública (SEP) (2006c). Educación básica. Secundaria. Matemáticas. *Programas de estudio 2006*. México: SEP. ISBN 968-9076-02-7.

Secretaría de Educación Pública (SEP) (2008). Educación básica. Secundaria. Historia. *Programas de estudio 2006*. México: SEP. ISBN 978-970-829-003-6.

Shea, J. A, Poniatowski, P. A., Day, S. C., Langdon, L. O., LaDuca, A. y Norcini, J. J. (1992). An adaptation of item modeling for developing test-item Banks. *Teaching and learning in medicine,* 4, 19-24.

Simon, M. G. (1989, marzo). *Use of Item Shells to Construct Higher Level Test-Items in Reading Comprehension*. Ponencia presentada en la reunión anual de la American Educational Research Association, San Francisco, California, EE UU.

Singley, M. K., y Bennett, R. E. (2002). Item generation and beyond: Applications of schema theory to mathematics assessment. En S. H. Irvine y P. C. Kyllonen (Eds.), *Item generation for test development* (pp. 361-384). Mahwah, NJ: Lawrence Erlbaum Associates.

Sinharay, S. y Johnson, M. (2012). Statistical modeling of Automatic Item Generation. En M. J. Gierl y T. M. Haladyna (Eds.), *Automatic item generation: Theory and practice* (pp. 183-195). Nueva York: Routledge.

Sireci, S. y Faulkner-Bond, M. (2014) Validity evidence based on test content. *Psicothema*, 26(1), 100-107.

Smith, G. y McCarthy, D. (1995). Methodological considerations in the refinement of clinical assessment instruments. *Psychological Assessment*, 7, 300-308.

Snavely, L., y Wright, C. (2003). Research Portfolio Use in Undergraduate Honors Education: Assessment Tool and Model for Future Work. *Journal of Academic Librarianship*, 29(5), 298-303.

Solano-Flores, G. (1991). *Diseño lógico de exámenes*. México: Trillas.

Solano-Flores, G., Jovanovic, J. Shavelson, R. J. y Bachman, M. (1999). On the development and evaluation of a shell for generating science performance assessment. *International Journal of Science Education*, 21(3), 293-315.

Solano-Flores, G. y Shavelson, R. J. (1997). Development of performance assessments in science: conceptual, practical, and logistical issues. *Educational measurement: issues and practice*, 16(3), 13-22.

Solano-Flores, G., Shavelson, R. J. y Schneider, S. A. (2001). Expanding the notion of assessment shell: from task development tool to instrument for guiding the process of science assessment development. *Revista Electrónica de Investigación Educativa*, 3(1). Recuperado de http://redie.uabc.mx/vol3no1/contents-solano.html

Suen, H. (1990). *Principles of test theories*. Hillsdale, Nueva Jersey: Erlbaum.

Suen, H. y Ary, D. (1989). *Analyzing quantitative observation data*. Hillsdale, Nueva Jersey: Erlbaum.

Svinicki, M. (2004). Authentic Assessment: Testing in Reality. *New Directions for Teaching and Learning*, 100, 23-29.

Tirado, F. (2010, abril). *Reactivos estructurales constructivos*. Ponencia presentada en la Segunda Reunión del Consejo consultivo del EXHCOBA, Baja California, México.

Tirado, F., Backhoff, E. y Larrazolo, N. (2014, noviembre). *Innovación en la evaluación: aprovechando las herramientas digitales*. Ponencia presentada en el XI Foro de evaluación educativa, Puerto Vallarta, México.

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*, 2(3), 67-84.

Universidad Nacional Autónoma de México (UNAM) (2014). Secretaría de Desarrollo Institucional. Dirección General de Evaluación Educativa. *Examen de ingreso a la licenciatura de la UNAM, ¿cómo me preparo?* Recuperado de https://www.seil.dgee.unam.mx:8443/seil-web/pre-registro/info.jsp

Vendlinski, T., Underdahl, J. y Simpson, E. (2002, junio). *Authentic Assessment of Student Understanding in Near-Real Time!* Ponencia presentada en la National Education Computing Conference, San Antonio, Texas.

Walsh, W. (1995). *Tests and assessment*. Nueva York: Prentice-Hall.

Walz, C., Strickland, O. y Lenz, E. (1991). *Measurement in nursing research* (Segunda ed.). Filadelfia: F. A. Davis.

Wellington, P., Thomas, I., Powell, I. y Clarke, B. (2002). Authentic Assessment Applied to Engineering and Business Undergraduate Consulting Teams. *International Journal of Engineering Education*, 18(2), 168–179.

Wiggins, G. (1990). The case of authentic assessment. *Practical Assessment, research and evaluation*, 2(2). Recuperado de http://PAREonline.net/getvn.asp?v=2&n=2

Wilson, M. (2005). *Constructing measures: an item response modeling approach*. Mahwaj, NJ: Laurence Erlbaum Associates.

Zhou, J. (2009). *A Review of Assessment Engineering Principles with Select Applications to the Certified Public Accountant Examination* (Reporte técnico no. W0903). The American Institute of Certified Public Accountants.

Apéndices

Apéndice 1. Modelo de Elaboración del Examen de competencias Básicas (Excoba)

Examen de Competencias Básicas

excoba

Modelo de Elaboración del Excoba (MEE)

Fases	Etapas	Participantes	Procedimientos	Productos
1. Planeación general	1. Diseño del plan general de evaluación	• Consejo consultivo • Comité técnico • Personal técnico	1. Documentación 2. Seminarios 3. Trabajo colegiado 4. Reuniones periódicas de trabajo 5. Pruebas del funcionamiento del sistema	1. Manual Técnico para el Diseño del Excoba
	2. Diseño y desarrollo del sistema informático GenerEx	• Especialistas en bases de datos • Especialistas en diseño gráfico • Especialistas en sistemas de información • Personal técnico		2. Cuestionarios de contexto del estudiante (versión computarizada)
	3. Diseño y elaboración de cuestionarios de contexto	• Especialistas en diseño de cuestionarios de contexto • Especialistas en evaluación del aprendizaje • Personal técnico		3. Plataforma informática y sistema computarizado del examen 4. Plataforma informática y sistema computarizado de los cuestionario de contexto 5. Sistema para calificar exámenes y generar reportes estadísticos 6. Sistema de información general de usuarios del Excoba
2. Estructuración del Excoba	4. Diseño del Excoba	• Comité técnico • Comités académicos • Personal técnico	1. Análisis curricular 2. Preparación de materiales (retículas y formatos) 3. Capacitación 4. Trabajo colegiado 5. Documentación del proceso	7. Manual Técnico para el Diseño del Excoba 8. Tabla de Contenidos de cada área del examen con justificaciones

3. Construcción del Excoba	5. Elaboración de modelos de ítems con plantillas y generadores de ítems	• Comité técnico • Comités Elaboradores de modelos de ítems • Personal técnico	1. Preparación de materiales (formatos y Tablas de contenidos) 2. Capacitación 3. Trabajo colegiado 4. Revisión y corrección de los modelos de ítems 5. Documentación del proceso	9. Manual técnico para la Elaboración de modelos de ítems con plantillas y generadores de ítems 10. Modelos de ítems de cada área del examen 11. Versiones de modelos de ítems con sus plantillas y generadores de ítems
4. Construcción de la interfaz	6. Programación del GenerEx	• Comité técnico • Especialistas en bases de datos • Personal técnico	1. Especificación de requerimientos del GenerEx 2. Programación del GenerEx para diferentes tipos de reactivos 3. Pruebas y ajustes al GenerEx 4. Documentación del GenerEx	12. Sistema informático GenerEx 13. Manual Técnico del GenerEx
	7. Diseño gráfico de la interfaz	• Comité técnico • Especialistas en diseño gráfico • Personal técnico	1. Especificación de requerimientos de la interfaz 2. Diseño y elaboración de la interfaz 3. Documentación de la interfaz	14. Interfaz gráfica del examen. 15. Manual Técnico de la Interfaz.
	8. Montaje en red del examen.	• Comité técnico. • Especialistas en sistemas de información. • Personal técnico.	1. Alimentación de la información para la generación de ítems 2. Programación y montaje de los modelos de ítems en el sistema GenerEx 3. Pruebas de funcionamiento del examen	16. Sistema computarizado del Excoba 17. Manual Técnico del Sistema
5. Administración del Excoba	9. Piloteo	• Comité técnico. • Especialistas en sistemas de información. • Estudiantes. • Personal técnico.	1. Administración del examen a muestra de estudiantes. 2. Generación de bases de datos	18. Manual Técnico para la Administración del examen 19. Base de datos con resultados del piloteo. 20. Base de datos con resultados de los Cuestionarios de Contexto.
	10. Administración real del examen.	• Investigadores asociados especialistas en tecnología y procesos de evaluación.	1. Capacitación de personal técnico. 2. Instalación del examen.	21. Manual de capacitación para personal técnico. 22. Base de datos de resultados.

		• Personal de soporte técnico. • Personal de instituciones usuarias.	3. Administración del examen a estudiantes.	23. informes de resultados a instituciones usuarias.
6. Análisis e interpretación de resultados del Excoba	11. Análisis psicométricos para comité técnico.	• Comité Técnico. • Asesores en medición.	1. Análisis estadísticos de resultados. 2. Reuniones informativas.	24. Informe Técnico sobre el Comportamiento Psicométrico del examen.
	12. Análisis de resultados para el Consejo Consultivo.	• Comité Técnico. • Consejo consultivo.	1. Análisis estadísticos de resultados. 2. Reuniones informativas.	25. Informe Técnico sobre los resultados de la evaluación asociados con las variables de contexto.
7. Recopilación de evidencias de Validez del Excoba	13. Estudios de validación.	• Comité Técnico. • Asesores en validación. • Personal técnico.	1. Documentación. 2. Investigación.	26. Informes Técnicos de Estudios de Validación.

Apéndice 2. Formato de currículum vitae

Examen de Competencias Básicas

excoba

FORMATO DE CURRÍCULUM VITAE

INFORMACIÓN DEL ESPECIALISTA					
NOMBRE(s)		APELLIDOS:			
FECHA DE NACIMIENTO			TELÉFONO CELULAR		
CORREO ELECTRÓNICO			TELÉFONO OFICINA		
ÁREA DE ESPECIALIZACIÓN					
CIUDAD DE RESIDENCIA					
ÁREA DE ESPECIALIDAD (ASIGNATURA)					
OCUPACIÓN					
ESCUELA					
NIVEL EDUCATIVO	PRIMARIA ( ) SECUNDARIA ( ) BACHILLERATO ( ) UNIVERSIDAD ( )				
TIPO DE ESCUELA	PÚBLICA ( ) PRIVADA ( ) OTRA ( )				
CATEGORÍA LABORAL					
DEDICACIÓN	Tiempo completo	( )	Medio tiempo	( )	
	Otro (especificar)				
GRADO(S) ESCOLAR(ES) QUE IMPARTE					
ASIGNATURA(S) QUE IMPARTE					

ACTIVIDADES PROFESIONALES		
FECHAS	**PUESTO**	**INSTITUCIÓN**

Examen de Competencias Básicas

FORMATO DE CURRÍCULUM VITAE

FORMACIÓN ACADÉMICA		
NOMBRE DEL GRADO	**INSTITUCIÓN**	**FECHA DE TITULACIÓN**
LICENCIATURA EN:		
MAESTRÍA EN:		
DOCTORADO EN:		
OTROS ESTUDIOS		

CURSOS IMPARTIDOS		
NOMBRE	**DIRIGIDO A**	**LUGAR/INSTITUCIÓN**

CURSOS A LOS QUE HA ASISTIDO	
NOMBRE	**LUGAR/INSTITUCIÓN**

OTRAS ACTIVIDADES O INFORMACIÓN CURRICULAR RELEVANTE		

Apéndice 3. Protocolo para la evaluación de los modelos de ítems

Examen de Competencias Básicas

PROCESO DE VALIDACIÓN DE LOS CONTENIDOS DE LOS MODELOS DE ÍTEMS
DEL EXAMEN DE COMPETENCIAS BÁSICAS (Excoba)

Formato de evaluación

Clave del modelo de ítem evaluado: _____ Nombre del evaluador: _____

Respuesta al reactivo presentado: _____ Fecha: _____

Observaciones al ítem hijo resuelto:

_____

_____

Instrucciones: a continuación se presentan una serie de afirmaciones cuya finalidad es evaluar la información contenida en las diferentes secciones del modelo de ítem. Lea cada afirmación y responda marcando con una X la columna correspondiente. En caso de que su respuesta sea *EN DESACUERDO*, incluya comentarios que justifiquen su opinión.

PRIMERA SECCIÓN Considerando la información que se encuentra en la sección denominada: ***Datos de identificación del contenido a evaluar***, responda a las siguientes afirmaciones:	**De Acuerdo**	**En Desacuerdo**	**N/A**	**COMENTARIOS**
1. La definición de este contenido se presenta de manera clara y precisa.				

2. La definición del contenido es congruente con su nombre.				
3. La definición de este contenido está alineada al currículum de la asignatura.				
4. El contenido es coherente con lo que se enseña en el aula.				
5. El dominio de este contenido es un aprendizaje básico y fundamental para el aprendizaje de la asignatura.				
6. El dominio de este contenido es un aprendizaje que se espera del promedio de los estudiantes.				
7. El aprendizaje de este contenido es importante para el dominio de la asignatura.				

SEGUNDA SECCIÓN Considerando la información que se encuentra en la sección denominada: ***Características del contenido curricular evaluado***, responda a las siguientes afirmaciones:	**De Acuerdo**	**En Desacuerdo**	**N/A**	**COMENTARIOS**
8. La delimitación que se hizo del contenido está alineada y deriva de su definición.				
9. Las habilidades y contenidos delimitados representan lo esencial del contenido definido inicialmente.				

TERCERA SECCIÓN Considerando la información que se encuentra en la sección denominada: ***Generador de ítems***, responda a las siguientes afirmaciones	**De Acuerdo**	**En Desacuerdo**	**N/A**	**COMENTARIOS**
10. La estrategia evaluativa utilizada es adecuada para evaluar los contenidos delimitados.				
11. La estrategia evaluativa utilizada es adecuada para evaluar los aprendizajes esperados del alumno.				
12. La estrategia evaluativa utilizada se asemeja a la forma en la que el proceso de enseñanza-aprendizaje se lleva a cabo dentro del salón de clases.				
13. Los ítems hijos requieren que al responder, el estudiante demuestre que sabe utilizar el conocimiento adquirido dentro del aula.				
14. La base del reactivo es clara y suficiente para que el estudiante emita su respuesta.				
15. Las instrucciones adicionales a la base del reactivo son claras y suficientes para que el estudiante emita su respuesta.				
16. Las reglas establecidas para generar ítems hijos responden a la estrategia evaluativa.				
17. Los textos auxiliares utilizados por el generador de ítems son apropiados para evaluar los aprendizajes esperados (diseño, nivel de complejidad).				
18. Los gráficos o imágenes utilizadas por el generador de ítems son apropiados para evaluar los aprendizajes esperados (diseño, nitidez, edición, colores, nivel de complejidad).				

19. Los elementos seleccionados del currículum para elaborar el banco de información son pertinentes y permiten evaluar el contenido al que corresponden.				
20. Las acciones que el alumno debe realizar para responder los ítems hijos (arrastre, selección, escritura) son simples y facilitan la evaluación del contenido curricular.				
21. Los ítems hijos representan adecuadamente el contenido delimitado.				
22. Los ítems hijos no contienen errores gramaticales ni faltas de ortografía en su redacción.				
23. Los ítems hijos están redactados con palabras familiares y de uso común de los alumnos.				
24. Los ítems hijos no contienen pistas en su redacción o en los elementos auxiliares (tablas, gráficos), que ayuden a contestarlos correctamente aun sin tener el conocimiento o habilidad.				
25. Los ítems hijos tienen un nivel de dificultad adecuado y son apropiados para el grado escolar del contenido que se evalúa.				
26. El vocabulario utilizado en los ítems hijos, así como el diseño de los elementos gráficos, no sesga en contra de grupos de personas debido a edad, sexo, región, clase social o raza.				
27. El banco de elementos y/o ítems hijos están libres de otro tipo de errores no contemplados en los puntos anteriores.				

CUARTA SECCIÓN

28. Estime en qué porcentaje los 20 modelos de ítem revisados, representan las habilidades y conocimientos básicos que deben dominar los estudiantes al término de la primaria/secundaria.	

29. ¿Considera usted que hay otros contenidos que deberían incluirse en el examen? ¿Cuáles?	
30. ¿Qué porcentaje de estudiantes que concluyen la primaria/secundaria considera Ud. que responderían correctamente este tipo de ítems?	

Firma del evaluador

Apéndice 4. Formato de compromiso de confidencialidad

Examen de Competencias Básicas

excoba

COMPROMISO DE CONFIDENCIALIDAD

Mediante la presente me comprometo a cumplir con las normas de confidencialidad en el tratamiento de los materiales de los Exámenes de Competencias Básicas (Excoba).

He leído y me comprometo a cumplir la *Normatividad en el Trabajo de Validación de los Contenidos de los modelos de ítems* que aparece en el recuadro de abajo. Asimismo, me comprometo a seguir las indicaciones del personal del Excoba.

Al firmar el presente compromiso acepto que las personas que se encuentran asociadas a los distintos proyectos de investigación y desarrollo del Excoba, se reservan el derecho de reportar cualquier violación del mismo a los autores del Excoba, quienes ostentan el derecho de autor de dicho examen.

Normatividad en el Trabajo de Validación de los Contenidos de los modelos de ítems

1. No divulgar por ningún medio el contenido del Excoba, los materiales de trabajo, ni los contenidos de las discusiones de grupo.
2. Mantener todo el tiempo en la sala de juntas el material asignado.
3. No reproducir por ningún medio los materiales propios del Excoba, ni aquellos utilizados para su diseño.
4. Abstenerse de usar en la sala de juntas grabadoras, cámaras fotográficas, de video o cualquier otro aparato que se especifica en la norma número 3.
5. Abstenerse de abrir las argollas, de arrancar o retirar hojas, y/o cambiar el orden del material incluido en las carpetas de trabajo, en caso de que las hubiere.
6. Escribir con letra clara y legible, y procurar trabajar limpiamente evitando tachaduras.
7. En su caso, completar todos los formatos asignados.
8. Apegarse puntualmente a los tiempos para las actividades indicadas por el coordinador del grupo de evaluadores, incluyendo descansos y comidas.

Nombre Firma Fecha