



Universidad Autónoma de Baja California
Instituto de Investigación y Desarrollo Educativo

**“Desarrollo de un procedimiento para la evaluación de ítems
mediante el juicio de expertos”**

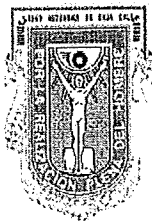
T E S I S

QUE PARA OBTENER EL GRADO DE
MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

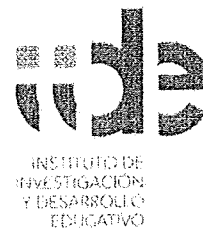
Ángela Aldana Torres

Ensenada B. C., septiembre de 2008.



Universidad Autónoma de Baja California
Instituto de Investigación y Desarrollo Educativo

Maestría en Ciencias Educativas



**“Desarrollo de un procedimiento para la evaluación de ítems mediante el
juicio de expertos”**

T E S I S

que para obtener el grado de

MAESTRA EN CIENCIAS EDUCATIVAS

Presenta

Ángela Aldana Torres

APROBADO POR:

M. C. Luis Ángel Contreras Niño
(Director de Tesis)

Dr. Eduardo Backhoff Escudero
Sinodal

Dr. Joaquín Caso Niebla
Sinodal

ÍNDICE

		Página
Capítulo 1.	Introducción	
	1.1 Contexto	11
	1.2 Estudios previos	13
	1.3 Planteamiento del problema	15
	1.4 Objetivos	17
	1.5 Justificación	18
	1.6 Supuestos del estudio	20
	1.7 Limitaciones	21
Capítulo 2.	Marco conceptual	
	2.1 Normatividad internacional para la elaboración, evaluación y uso de pruebas	23
	2.2 Modelos para validar ítems de pruebas de gran escala	26
	2.2.1 Validez cultural y de género	28
	2.2.2 Validez de contenido	36
Capítulo 3.	Método	
	3.1 Participantes	48
	3.2 Materiales e instrumentos	
	3.2.1 Manual del participante	48
	3.2.2 Formato de evaluación de reactivos	50
	3.2.3 Presentación en Power Point	56
	3.2.4 Cámara y videos	56
	3.2.5 Cuestionario para evaluar la operación del módulo	54
	3.3 Procedimientos	
	3.3.1 Capacitación	54
	3.3.2 Elaboración del formato de evaluación de reactivos	59
	3.3.3 Revisión de ítems	62
	3.3.4 Aplicación del cuestionario para evaluar la operación del módulo	60
	3.3.5 Análisis estadístico de reactivos	64
	3.3.6 Transcripción y análisis de videos	65
Capítulo 4.	Resultados	
	4.1 Videos de la capacitación	66
	4.2 Errores registrados por los jueces en el formato de evaluación de reactivos	69
	4.3 Evaluación de la sesión de capacitación y asesoría	74
	4.3.1 Manual técnico del participante	78
	4.3.2 Operación de la capacitación	79
	4.3.3 Operación de la evaluación de ítems	79
	4.3.4 Resultados de la evaluación de los ítems por el comité	80
	4.4 Ítem análisis	81
	4.4.1 Dificultad	82
	4.4.2 Discriminación	83
Capítulo 5.	Conclusiones y recomendaciones	85
	Bibliografía	94
	Anexos	99

ÍNDICE DE FIGURAS

Figura	Descripción	Página
1	Versión sintética del Formato de evaluación de reactivos.	53
2	Esquema para la capacitación para validar ítems mediante juicios.	55
3	Áreas a explorar del reactivo, según las aportaciones de Hambleton, Popham, Jornet y Suárez, Nitko y Solano-Flores.	61
4	Factores e indicadores para evaluar la capacitación y asesoría al comité	63
5	Errores identificados por la profesora de español en los ítems, según el formato de evaluación de reactivos.	70
6	Errores identificados por el profesor de matemáticas 1 en los ítems, según el formato de evaluación de reactivos.	70
7	Errores identificados por el profesor de matemáticas 2 en los ítems, según el formato de evaluación de reactivos.	70
8	Errores identificados por el profesor de matemáticas 3 en los ítems, según el formato de evaluación de reactivos.	70
9	Errores identificados por el psicómetra en los ítems, según el formato de evaluación de reactivos.	71
10	Errores identificados mediante el consenso en la revisión grupal de los ítems, según el formato de evaluación de reactivos.	71
11	Frecuencia de los errores identificados en los ítems, por tipo de error, según el formato de evaluación de reactivos.	80
12	Reactivos con error consensuado por área	73
13	Tipo de errores por área consensuados en cada ítem	74
14	Respuestas al cuestionario para evaluar la operación del módulo.	83

ÍNDICE DE TABLAS

Tabla	Descripción	Página
1	Recomendaciones para la elaboración de reactivos propuestas por Haladyna, Haladyna y Merino, 2002, Bothell, INEE (2002), Garden y Orpwood (1996), el INCE (Mullis et al., 2002) y Hambleton (1993).	26
2	Ventajas y desventajas de la revisión mediante jueces según Hambleton (1993).	33
3	Estándares y calidad técnica del contenido del test (Contreras, 2000).	45
4	Áreas a explorar en el formato de evaluación de reactivos (Hambleton, 1993; Popham, 1991; Suárez y Jornet, 1990a y 1990b; Solano-Flores y col., 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000).	51
5	Respuestas al cuestionario para evaluar la operación del módulo.	76
6	Respuestas de los jueces a las preguntas abiertas del cuestionario para evaluar la operación del módulo.	77
7	Comparación entre los resultados del ítem análisis de los 15 reactivos revisados y los 17 que no fueron revisados.	81

ANEXOS

Tabla	Descripción	Página
1	Reglas para la construcción de reactivos propuestas por Haladyna, Haladyna y Merino (2002), Bothell (2001), Garden y Orpwood (1996) y el INCE (Mullis et al., 2002), ETS (2003), Hambleton (1993) y el INEE (2002).	101
2	Manual técnico del participante.	108
3	Cuestionario para evaluar el módulo.	131
4	Presentación en Power Point empleada durante la capacitación.	135
5	Comentarios de los jueces en los reactivos evaluados según el protocolo evaluativo.	139
6	Cuestionarios para evaluar el módulo contestados por los participantes.	145
7	Formatos de evaluación de reactivos contestados por los participantes.	158
8	Notas sobre la videograbación.	166
9	Síntesis de la videograbación sobre la evaluación de los reactivos.	168
10	Resultados del ítem análisis mediante el software del IIDE-UABC.	179
11	Resultados del ítem análisis mediante el ITEMAN.	182

Resumen

El presente estudio es una extensión del modelo para elaborar exámenes criterios de gran escala para evaluar los aprendizajes, el cual se ha desarrollado en el Instituto de Investigación y Desarrollo Educativo de Universidad Autónoma de Baja California (IIDE-UABC) (Contreras, 2000). El objetivo general del proyecto fue diseñar, operar y evaluar un módulo para capacitar y asesorar a los miembros de un comité para realizar la validación, mediante operaciones de juicio, de ítems que se emplean en la evaluación criterial del aprendizaje a gran escala.

El método empleado en la evaluación de los reactivos es a priori a la aplicación de la prueba. Se basa en la revisión de jueces quienes previa capacitación exploraron un conjunto de ítems en los aspectos curricular, sociocultural y psicométrico. Cabe señalar que la capacitación de los jueces se realizó conforme a la propuesta de Solano-Flores (Solano-Flores y colaboradores, 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000) y del Instituto Nacional para la Evaluación de la Educación (INEE, 2005).

Los participantes en el estudio -integrantes del comité de jueces- fueron dos profesores reconocidos por su experiencia en la enseñanza de las matemáticas, dos profesores expertos en Cálculo integral, una profesora experta en español y una persona con conocimiento de psicometría.

Por su parte, se enfatiza el hecho de que se revisaron 15 ítems que forman parte de un examen de Cálculo integral II, elaborado por un cuerpo colegiado de la Facultad de Ingeniería (FI) de la Universidad Autónoma de Baja California, unidad Mexicali (Contreras, Encinas y De Las Fuentes, 2005).

Los procedimientos desarrollados para lograr el objetivo general del estudio consistieron en el diseño de cuatro materiales que apoyaron en la operación y evaluación del módulo:

- El manual técnico del participante, documento fundamental para el desarrollo de la capacitación, que incluye aspectos conceptuales de la validación de

reactivos, el procedimiento para la revisión de los ítems, conclusiones y un glosario para apoyar a los participantes.

- El formato de evaluación de reactivos, protocolo que contiene 9 afirmaciones que exploran al ítem en cuanto a su congruencia curricular, elementos socioculturales y congruencia psicométrica.
- Una presentación en Power Point de doce diapositivas que apoya en la capacitación.
- El cuestionario para evaluar la operación del módulo, que cuestiona mediante 32 preguntas a los participantes sobre los materiales utilizados y los resultados obtenidos tras la operación del módulo.

Además, se elaboraron procedimientos para llevar a cabo la capacitación, la revisión de reactivos en modalidad individual y grupal, así como la filmación de las sesiones. Los principales resultados obtenidos son:

- Una síntesis de la transcripción de las dudas de los participantes, los comentarios y del proceso de evaluación de reactivos.
- El registro de los errores identificados por los jueces en el formato de evaluación de reactivos.
- La evaluación de la sesión de capacitación y asesoría.

A esto se añade la presentación de los resultados del ítem análisis de la aplicación a gran escala de los 15 reactivos revisados en 2006, y una comparación preliminar con 17 reactivos de la misma prueba que no fueron evaluados por los jueces, ni modificados, pero que fueron aplicados a 344 alumnos de la FI.

En síntesis, la finalidad del estudio fue documentar, con base en la evidencia proporcionada por los procedimientos desarrollados, una experiencia de validación de reactivos mediante juicios. Lo anterior, para aportar evidencia de este método en el perfeccionamiento psicométrico, cultural y curricular de ítems que evalúan el dominio que tienen los examinados sobre el currículum.

DEDICATORIA

A mi familia por todo el cariño y apoyo que me han brindado en la realización de todos los proyectos que he emprendido.

AGRADECIMIENTOS

Al Consejo Nacional de Ciencia y Tecnología por su apoyo para mejorar mi formación en este posgrado.

Al Instituto de Investigación y Desarrollo Educativo de la UABC y a todo su personal por darme la oportunidad de continuar mi formación académica en su programa de maestría.

Al Instituto Nacional para la Evaluación de la Educación por auxiliarme tan gentilmente en el préstamo de material en DVD sobre capacitación de validez de ítems mediante juicios.

Al CECYTE BC por todo el apoyo que he recibido durante la realización de este proyecto.

Al Dr. Guillermo Solano Flores por la revisión de este trabajo, aportando observaciones, el enfoque documental de la tesis y sobre todo la metodología para la revisión de ítems.

Al M.C. Luis Ángel Contreras Niño por el enorme apoyo, paciencia, experiencia y guía académica que me hicieron crecer, aprender y ver con diferentes ojos la importancia de la justicia en la evaluación educativa.

A mis asesores de tesis: Dr. Eduardo Backhoff Escudero y M.C. Joaquín Caso Niebla por compartir ideas, sugerencias y propuestas que orientaron el rumbo de este proyecto de investigación, así como por compartir su valioso tiempo y experiencias.

Al personal de la Facultad de Ingeniería de la UABC y del CECYTE BC que participaron como jueces en el comité revisor de reactivos.

Al Ing. José Carlos Jiménez Payán por el enorme apoyo que me ha brindado en mi desarrollo académico y profesional.

A Carlos Zamora por la confianza que ha tenido en esta profesora y por su valioso apoyo en mi formación académica y laboral.

A Malibé Rosas, a mis compañeros y amigos de la Maestría en Ciencias Educativas, en especial a Brenda Boroel, y Antelmo Castro por su amistad y apoyo incondicional.

Capítulo 1

Introducción

1.1 Contexto

El Programa Nacional de Educación 2001-2006 señalaba dentro de sus objetivos particulares la necesidad de contar con un organismo que fortalezca el sistema de evaluación nacional, para lo cual apoya la creación en 2002 del Instituto Nacional para la Evaluación de la Educación, a quien se le confiere la responsabilidad de establecer criterios relativos al uso y la difusión de la evaluación en el país.

El mismo programa expresaba que existen áreas de oportunidad en lo que respecta a instrumentos y sistemas de evaluación; ya que los mecanismos utilizados hasta la fecha no han sido lo suficientemente claros ni sistemáticos en la rendición de cuentas del sistema educativo nacional, por lo que sus resultados sólo han contribuido a desalentar los esfuerzos que se han hecho en materia educativa (Secretaría de Educación Pública, 2001, p.121).

En congruencia con los lineamientos nacionales en materia de evaluación, en junio de 2003 se publica el Plan de Desarrollo Institucional (PDI) 2003-2006 de la UABC, el cual incluye una iniciativa designada Fortalecimiento de los procesos y mecanismos institucionales de seguimiento y evaluación del funcionamiento de los procesos académicos y administrativos, cuya finalidad es crear un sistema de indicadores y un esquema de evaluación de los diferentes procesos institucionales, promoviendo un uso permanente de la información generada que sirva para apoyar la toma de decisiones (Universidad Autónoma de Baja California, 2003).

La propuesta del PDI es evaluar de manera sistemática los logros alcanzados en el proceso de aprendizaje por los alumnos, para lo cual propone el diseño de un programa de evaluación del aprendizaje que proporcione la información necesaria, asegure la calidad académica de la UABC y rinda cuentas a la sociedad en general. Con el propósito de cumplir lo planteado al respecto en el PDI, en marzo de 2004 la UABC publicó una convocatoria enfocada a directivos, cuerpos colegiados y académicos de la Institución, a participar en el diseño de instrumentos de evaluación colegiada del aprendizaje

que contribuyan al mejoramiento continuo de la calidad del aprendizaje (Universidad Autónoma de Baja California, 2004).

Por su parte, dentro del sistema de educación medio superior estatal, se realizaron acciones encaminadas a mejorar los sistemas de evaluación del aprendizaje. El Colegio de Bachilleres del Estado de Baja California (COBACH) y el Colegio de Estudios Científicos y Tecnológicos del Estado de Baja California (CECYTE) solicitaron el apoyo del IIDE-UABC para capacitar y asesorar a sus docentes en la elaboración de exámenes de gran escala para evaluar el aprendizaje de sus alumnos. Tal colaboración inició en septiembre de 2003 y finalizó en diciembre de 2004 para el caso del CECYTE, pero continuó hasta mediados de 2005 con el COBACH.

1.2 Estudios previos

Como refiere el título de este trabajo, el módulo que se propone elaborar, operar y evaluar es un desarrollo complementario del trabajo realizado por Contreras y colaboradores (2000, 2004 y 2005), en el IIIDE-UABC, mismo que a su vez está:

(...) basado en la metodología propuesta por Anthony Nitko en 1994, complementado por la metodología [...] propuesta por James Popham (1991); por el esquema metodológico aportado por Ronald Berk y colaboradores para la construcción de tests criterioles (1984); así como en los planteamientos psicométricos propuestos por Kellaghan y Greaney para el caso de países en desarrollo (1992; 1995). (Contreras, 2000, p. 8)

Para elaborar un examen criterial con base en el modelo que se ha desarrollado en el IIIDE-UABC, es necesario organizar y capacitar a comités de especialistas que realicen actividades como coordinar la construcción del examen, diseñar el instrumento, elaborar especificaciones e ítems y aplicar los modelos de examen, entre otras.

Según dicho modelo, el comité coordinador del examen está integrado fundamentalmente por jefes de docencia, coordinadores académicos o directores de una institución educativa; mientras que los comités de diseño de las pruebas, de elaboración de especificaciones y de elaboración de ítems se conforman principalmente por profesores con experiencia y dominio de una materia determinada.

Experiencias de evaluación similares se han desarrollado en Estados Unidos, tales como las diseñadas por Popham (1991), Hambleton (1988), así como por la National Assessment of Educational Progress (NAEP), el Center for Research on Evaluation, Standards and Student Testing (CRESST) y el National Assessment Governing Board (NAGB) (Contreras, 2000, p. 22).

En España, el entonces Instituto Nacional de Calidad y Evaluación (INCE) del Ministerio de Educación y Ciencia de España (actual Instituto de Evaluación) elaboró exámenes para evaluar los aprendizajes a nivel primaria (citado en Contreras, 2000). Mientras que en Botswana, Nitko (1994) ha aplicado su modelo de desarrollo de exámenes criterioles centrados en el currículum en la

educación básica de ese país, así como también se ha adaptado para el caso de Jamaica.

En nuestro país, son escasos los aportes de la investigación educativa en este campo; Contreras (2000, pp. 23-25) refiere algunos de los escasos trabajos realizados por Shmelkes en 1994 y 1999; Ornelas, en 1994; Ruiz y Martínez, en 1996; Zorrilla en 1996 y Backhoff en 1996. Los proyectos señalados anteriormente tienen características diversas, pero destacan las observaciones relativas al interés de los investigadores por diseñar e implementar evaluación a gran escala, contribuyendo poco a poco con una cultura evaluativa en México. A pesar de tales esfuerzos, a excepción del caso de los exámenes desarrollados en Aguascalientes por Ruiz y Martínez (1996), es visible una falta de alineación de las pruebas al currículo de la educación básica. Otra excepción a esta situación es el caso de los exámenes nacionales para educación básica que desarrolla actualmente el INEE, denominados Exámenes de la Calidad y el Logro Educativos (EXCALE) (Backhoff, 2005).

En el ámbito regional, hay algunas experiencias de aplicación del modelo que desarrolla el IIDE-UABC en instituciones educativas: el Examen de Español para la Educación Primaria en Baja California (Contreras, 2000), los exámenes sumarios elaborados entre 2003 y 2005 por el COBACH y los exámenes de certificación del CECYTE BC (Contreras y Hurtado, 2004), así como los exámenes colegiados desarrollados en 2004 por la UABC (Contreras y colaboradores, 2005).

1.3 Planteamiento del problema

Los exámenes a gran escala evalúan los aprendizajes que logran cientos o miles de alumnos, lo que los define como instrumentos de alto impacto potencial que pueden tener en las vidas de las personas afectadas por su aplicación y por las decisiones que son tomadas por las autoridades gubernamentales y educativas con base en los resultados que arrojan.

Por ello, un instrumento de esa naturaleza debe ser válido psicométrica y culturalmente, para poder asegurar su objetividad y la justicia a los examinados, y además debe evitar la penalización y ofensa a la población examinada (Popham, 1991). Por lo tanto, el diseño y operación de un módulo para validar ítems mediante juicios de expertos, es necesario para garantizar el rigor conceptual y metodológico del instrumento de evaluación.

El modelo para diseñar exámenes criterios de gran escala para evaluar los aprendizajes del IIDE-UABC, considera procedimientos específicos para la producción y pilotaje de ítems y modelos de examen, a fin de explorar sus propiedades psicométricas. Ello resulta necesario para desarrollar una prueba, pero es insuficiente para construir un examen criterial de calidad; pues aún se requiere diseñar un módulo dedicado especialmente a la validación de los ítems, como lo establece el modelo original de Nitko (1994).

Así, continuar con el desarrollo y validación de ese modelo para evaluar el aprendizaje a gran escala es importante porque permitirá superar una de sus grandes limitaciones; es decir, el escaso énfasis en la obtención de evidencias de validez de ítems mediante juicios de expertos.

Además, un estudio con estas características es pertinente debido a las necesidades de formación y asesoría en el campo evaluativo que diversas instituciones le plantean continuamente al IIDE-UABC. Así, resulta importante continuar el desarrollo de estos trabajos de investigación y transferencia tecnológica, por el impacto académico del trabajo y los recursos que aportan al Instituto los convenios de colaboración implicados.

Capítulo 1 Introducción

Finalmente, el efecto de los resultados del estudio que se pretende realizar podrá extenderse a otras instituciones educativas, ya que el modelo ha sido aplicado en otros niveles educativos (básico, medio superior y superior), de ahí la trascendencia de contar con una sólida metodología para desarrollar instrumentos evaluativos de calidad técnica con los cuales se pueda dar cuenta cabalmente del dominio que logran los estudiantes sobre el currículum.

1.4 Objetivos

General:

- Diseñar, operar y evaluar un módulo para capacitar y asesorar a los miembros de comités que validan mediante jueceo ítems empleados en la evaluación criterial del aprendizaje a gran escala.

Específicos:

- Establecer áreas a explorar del reactivo.
- Establecer criterios para la selección de comités evaluadores.
- Diseñar y probar procedimientos, instrumentos y materiales para capacitar y asesorar a los comités evaluadores de ítems.

1.5 Justificación

El modelo que desarrolla el IIDE-UABC, adaptado a partir de la metodología propuesta por Nitko (1994), necesita incorporar formalmente un procedimiento para la validación de los reactivos, porque hasta el momento esta tarea ha recaído en las mismas personas que elaboran los ítems, lo cual resulta inconveniente pues se convierten en juez y parte, y por ello es insuficiente a la luz de los estándares psicométricos establecidos para esta clase de instrumentos, a fin de propiciar la obtención de evidencias sobre su calidad técnica.

Considerando la importancia que tiene la validez de una prueba de gran escala y en particular la de tipo cultural, como lo demuestran las investigaciones de Solano y colaboradores (2000, 2001), en este trabajo se continúa el desarrollo del modelo citado en un aspecto de gran relevancia; es decir, el diseño, operación y evaluación de un módulo para capacitar y asesorar a los miembros de comités cuya única finalidad será validar ítems mediante juicios.

En síntesis, el desarrollo y evaluación de un módulo para validar mediante jueces los ítems empleados en la evaluación a gran escala, es importante por las siguientes razones:

- Se mejora el modelo del IIDE-UABC, pues a partir del procedimiento general propuesto para la elaboración de exámenes criteriosales a gran escala, se incorpora una alternativa formal para validar ítems mediante el jueceo de comités.
- Se mejora la calidad técnica de los ítems que así lo requieran, al ser detectada mediante juicios de expertos la posible existencia de su funcionamiento diferencial y al ser enviados esos ítems a sus elaboradores para su corrección.
- Se asegura la congruencia y coherencia de los ítems respecto al currículum, con la ayuda de profesores en servicio. Adicionalmente, la calidad psicométrica de los reactivos será evaluada de acuerdo a las convenciones internacionales, con el apoyo de un psicómetra. Además,

Capítulo 1 Introducción

se considerará la influencia de los elementos socioculturales siguiendo la línea de Solano y colaboradores, con la intención de ser justos en la evaluación, dando a todos los examinados la oportunidad de resolver una prueba sin que su sexo, religión o contexto cultural lo limiten.

1.6 Supuestos

- La capacitación y asesoría que reciben los miembros de los comités de evaluación que juzgan los ítems empleados en la evaluación criterial a gran escala de los aprendizajes, afectará de manera significativa la calidad de sus juicios sobre los ítems y en consecuencia la calidad de estos últimos.
- Los materiales y procedimientos empleados para capacitar y asesorar a los miembros de los comités evaluadores de los ítems empleados en la evaluación criterial de los aprendizajes a gran escala, facilitarán de manera significativa su tarea al emitir sus juicios sobre dichos reactivos.

1.7 Limitaciones

Aunque sería deseable, el presente estudio no se propone explorar el estatus de diferentes modos de integración y funcionamiento de comités de expertos para la validación de ítems. A cambio, se trabajó con la integración de un solo comité. Lo anterior debido a limitaciones de tiempo y recursos para la realización de este proyecto de investigación. Sin embargo, se sugiere que se desarrolle este objetivo en un estudio posterior; de ese modo se podrán comparar los resultados obtenidos con dos o tres configuraciones de comités o de formas de operación. Así, se podría tener evidencia sobre la integración y funcionamiento más adecuado para la revisión de ítems mediante juicios; es decir, del comité que contenga a los integrantes idóneos y que presenten mejores resultados.

Por último, el objetivo específico que enuncia el establecer criterios de decisión para determinar la validez de juicios sobre los ítems de una prueba, se trabajó de modo parcial, ya que el único criterio considerado para determinar la validez de un juicio emitido por los jueces fue el consenso, el cual como se verá en el capítulo de resultados se dio en prácticamente todos los casos.

Capítulo 2

Marco conceptual

2.1 Normatividad internacional para la elaboración, evaluación y uso de pruebas

Dado el alto impacto que pueden ocasionar las pruebas de gran escala en las vidas de los examinados, profesores y autoridades educativas, se ha considerado necesario normar su desarrollo. Entre los esfuerzos en ese sentido, destaca la tarea que han emprendido la *American Psychological Association* (APA), la *American Educational Research Association* (AERA) y el *Nacional Council on Measurement in Education* (NCME) las cuales integraron un comité conjunto para normar las prácticas relativas a la elaboración y uso de instrumentos de evaluación psicológica. El comité conjunto publicó en 1954, los primeros estándares internacionales para tests educativos y psicológicos, mismos que mencionan los criterios técnicos que norman la construcción, evaluación, documentación y usos de pruebas (APA, 1999). Dentro de los criterios más importantes, se encuentran los de validez, confiabilidad y errores de medición; desarrollo y revisión del instrumento de evaluación; administración, calificación y generación de informes de resultados; uso justo y adecuado de la evaluación, así como las responsabilidades de los usuarios del instrumento.

De particular importancia para este trabajo son los estándares correspondientes a la obtención de evidencias que permitan hacer inferencias válidas a partir de los resultados en el examen, sobre el dominio de los contenidos curriculares por parte de los examinados; particularmente, las evidencias relacionadas con el contenido del examen (validez de contenido) y el estándar relativo al desarrollo y revisión del instrumento de evaluación por parte de comités de expertos que juzgan los ítems (validez cultural, estudio de sesgo, etc.).

Respecto a las reglas para construir pruebas para evaluar el aprendizaje mediante reactivos de opción múltiple, existe en el contexto internacional una serie de recomendaciones. En este apartado se revisarán las aportaciones hechas al respecto en Estados Unidos por Bothell (2001), Haladyna, Haladyna y Merino (2002); en España por el INCE (Mullis *et al.*, 2002) respecto a la prueba *Trends in International Mathematics and Science Study* (TIMSS); en Estados

Unidos por el *Educational Testing Service* (ETS, 2003) sobre la prueba *National Assessment of Educational Progress* (NAEP) y en México por el Instituto Nacional para la Evaluación de la Educación (INEE, 2002).

Haladyna y Dowling proponen 43 reglas para facilitar la construcción de ítems de opción múltiple, las cuales están basadas en investigaciones realizadas, en libros de texto sobre la elaboración de exámenes, así como en estudios sobre la preparación de pruebas de opción múltiple (ver tabla No. 1 y anexo No. 1). Sus reglas se centran en el contenido del reactivo, el formato, el estilo, el cuerpo del reactivo y en las opciones de respuesta. Para el objetivo que aquí compete, se retoman 30 de sus reglas (Haladyna, Haladyna y Merino, 2002).

A su vez, Bothell (2001) propone 14 reglas para escribir ítems de opción múltiple, las cuales versan sobre el contenido del reactivo, el estilo, la redacción de la base del reactivo y de las opciones de respuesta, la respuesta correcta, el evitar pistas en los distractores y proporcionar al examinado instrucciones claras (ver tabla No. 1 y anexo No. 2).

De acuerdo con Garden y Orpwood (1996) y el INCE (Mullis *et al.*, 2002), en el TIMSS se consideraron los siguientes factores para decidir los dominios de contenido, los temas y objetivos a evaluar:

- Los contenidos estaban incluidos en los currículos de un número significativo de países participantes (70% de ellos).
- La alineación de los dominios de contenido con las categorías de resultados de TIMSS 1995 y TIMSS 1999.
- La importancia de los contenidos para desarrollos futuros en la educación matemática y científica.
- La adaptación de los contenidos a las poblaciones de examinados.
- Su adecuación para ser evaluados en un estudio internacional de gran escala.
- Su contribución al equilibrio global y a la cobertura de los dominios de contenido y cognitivos.

De conformidad con lo anterior, en la tabla No. 1 se puntualizan las recomendaciones que hacen dichos investigadores para la elaboración de reactivos a gran escala (ver también anexo No. 3), las cuales se enfocan en el

contenido del reactivo, la redacción, la respuesta correcta, el uso del lenguaje, y añaden los estándares psicométricos y el sesgo cultural.

Otro organismo que hace recomendaciones para la elaboración de reactivos es el ETS; esto con la finalidad de garantizar la justicia en la evaluación (ver tabla No. 1 y anexo No.4). El ETS elaboró una serie de lineamientos a seguir durante el proceso de construcción del examen que se agrupan en seis temáticas: a) tratar a la gente con respeto; b) minimizar los efectos del conocimiento o habilidad sobre un constructo irrelevante; c) evitar el uso de materiales que sean innecesarios, controversiales, inflamatorios, ofensivos o molestos para los examinados; d) emplear la terminología apropiada; e) evitar estereotipos y f) representar la diversidad en la caracterización de las personas.

En México, el INEE (2002) diseñó una normatividad para la construcción de reactivos de las pruebas nacionales, la cual se centra en la congruencia del ítem con su especificación, el contenido de los reactivos, el uso del lenguaje, la redacción, los materiales, los formatos, las características de los reactivos de opción múltiple, y en recomendaciones para las pruebas de matemáticas y de español (ver tabla No. 1 y anexo No. 5).

Capítulo 2. Marco conceptual

Tabla No. 1 Recomendaciones para la elaboración de reactivos propuestas por Haladyna, Haladyna y Merino, 2002, Bothell, INEE (2002), Garden y Orpwood (1996), el INCE (Mullis et al., 2002) y Hambleton (1993).

Categorías	Haladyna, Haladyna y Merino (2002)	Bothell (2001)	Garden y Orpwood (1996) y el INCE (Mullis et al., 2002)	ETS (2003)	Hambleton (1993)	INEE (2002)
El contenido del reactivo	X	X	X	X		X
El Formato	X					X
El Estilo	X	X				X
La redacción de la base del reactivo	X	X	X			X
Redactar las opciones de respuesta	X	X	X			X
Respuesta correcta	X	X	X		X	
Pistas en los distractores	X	X				X
Instrucciones	X	X	X			X
Uso del lenguaje			X	X	X	X
Formato				X	X	X
Estándares psicométricos			X			
Sesgo cultural			X	X	X	
Representación estereotipada				X	X	

2.2 Modelos para validar ítems de pruebas de gran escala

En este trabajo se propone explorar la validez de los reactivos en por lo menos tres áreas: congruencia curricular, elementos socioculturales y congruencia con las convenciones psicométricas, con la finalidad de detectar previo a su aplicación un posible funcionamiento diferencial de los ítems que ponga en ventaja a unos examinados sobre otros que tienen la misma habilidad general que mide la prueba. A continuación se describe lo que diferentes autores sostienen como los principales problemas a detectar en los ítems en cada una de estas áreas:

- **Congruencia curricular:** Un examen que no está alineado al currículum no examina de manera justa a la población puesto que no se centra en los contenidos esenciales que debe dominar el examinado y que supuestamente le fueron enseñados (Nitko, 1994). En consecuencia, se debe probar que cada ítem evalúa algún aspecto importante del currículum.
- **Elementos socioculturales:** Existe un funcionamiento diferencial del ítem si las condiciones que permiten a los estudiantes realizar un logro no son las mismas para todos, ni todos los estudiantes tienen el mismo acceso a contenido instruccional adecuado. Intrínsecamente, las consecuencias de las pruebas pueden repercutir de manera diferente según el género, etnia, raza, idioma y clase social del grupo (Sexton y Solano-Flores, 1999).
- **Apego a las convenciones psicométricas:** Existen estándares que deben atender los elaboradores de los exámenes para que al redactar los ítems estos sean justos en su aplicación, de tal manera que todos los examinados tengan la misma oportunidad para demostrar lo que saben, así como de ser informados de la naturaleza de la prueba y de la manera en que se emplearán los resultados (Joint Committee on Testing Practices, 2005).

En general, para determinar si existe un funcionamiento diferencial del ítem (FDI) pueden seguirse dos métodos: empíricos como el delta-plot, chi-cuadrada o Modelos IRT (Suárez y Jornet, 1990a) o de jueceo (Cole y Moss, 1993; Title, 1990).

Como ya se señaló, el modelo que se quiere complementar ya contempla ambos métodos, tanto el jueceo como la prueba empírica; sin embargo, es necesario un análisis más profundo de los ítems que sea realizado formalmente por un comité externo integrado por especialistas que revisen, para cada uno de los ítems, las tres áreas antes mencionadas; y mediante el jueceo por consenso emitan su dictamen sobre los reactivos (Title, 1990).

La integración del citado comité evaluador puede variar considerablemente, pero en general se trata de comités pequeños y conformados por especialistas en contenido. Sin embargo, Solano-Flores y colaboradores (2005) establecieron comités para evaluar la traducción mexicana de la prueba TIMSS-95, en los que participaron profesores en servicio de las diferentes áreas evaluadas, especialistas en el currículum de la educación básica, un lingüista, un traductor, un psicólogo y un psicómetra.

La obtención de evidencias de validez de ítems de pruebas de gran escala ha sido estudiada por varios autores, quienes además de proponer métodos empíricos para obtener datos que fundamenten juicios y acciones para su mejora, optan a favor de métodos de jueceo para realizarla. En este apartado se revisarán brevemente los aportes en este campo de los psicómetras norteamericanos Popham (1991), Michener y Kesselman (1986), Title (1990), Childs (1990), Zimmerman, Sudweeks, Shelley y Wood (1990), Hambleton (1993), Sireci (1995a), Sireci y Geisenger (1995b), Solano-Flores (2000, 2001 y 2002); Sexton y Solano-Flores (1999 y 2002); Nelson-Barber y Solano-Flores (2000), Nitko (1994), así como de los españoles Suárez y Jornet (1990a y 1990b).

2.2.1 Validez cultural y de género

Popham (1991) recomienda la revisión de reactivos mediante jueces, los cuales constituirán un panel de revisión de sesgo por miembros de minorías y por revisoras femeninas, para asegurar que se considerará a la vez el sesgo de género. El número de integrantes con los que trabajó Popham varió de 8 hasta 35 miembros.

El método que utilizó Popham consistió en dar a los panelistas orientación acerca del tipo de sesgo que se busca; para después hacer preguntas en relación al sesgo no deseado, las cuales deben contestar con un sí o no. De esta manera, si el panel es cuidadosamente seleccionado, orientado y asistido durante el proceso de revisión, se identificarán los ítems con sesgo para ser eliminados de la prueba o corregidos.

Por su parte, Title (1990) define al sesgo como un prejuicio o como una inclinación o dirección en particular. Decir que un examen presenta sesgo significa que prejuzga o es injusto para grupos o individuos que tienen características diferentes al grueso de los examinados. En los Estados Unidos estos grupos incluyen minorías étnicas, mujeres, individuos cuyo primer idioma no es el inglés y personas que presentan alguna discapacidad.

Las acusaciones de sesgo se basan en el examen de los ítems de las pruebas, en busca de diferencias grupales en ejecuciones promedio y el uso de los exámenes. El sesgo en pruebas ha sido estudiado en exámenes empleados para seleccionar la admisión de estudiantes en educación media superior; también para ubicar a los estudiantes en clases de educación especial, en la certificación de competencias mínimas o estándares de logro en educación secundaria, evaluación de programas educativos y asesorías de carrera (Title, 1990).

El sesgo en pruebas, desde una perspectiva orientada hacia el constructo, ha sido examinado en estudios de pruebas empleadas en investigaciones culturales y en los primeros acercamientos para desarrollar exámenes culturales. Las primeras investigaciones de medición de la inteligencia realizadas durante la primera década de 1900, identificaron problemas en la evaluación de niños pertenecientes a diferentes grupos; por ejemplo, aquellos cuya lengua materna no era el inglés. Con el desarrollo de exámenes mentales grupales y las primeras pruebas a gran escala, como el Army Alpha en la Primera Guerra Mundial, las mediciones ampliaron su uso y la atención del público.

Title (1990) señala que en la segunda década de 1900, la cultura a la que pertenecía la población examinada era considerada al realizar interpretaciones de los puntajes de pruebas mentales.

En los años 30 del siglo pasado, hubo estudios sobre el efecto del idioma y la cultura en las puntuaciones de las pruebas. Durante la década de los 40 y a principios de los 50, se efectuaron de nuevo estudios sobre las diferencias raciales en la medición de la inteligencia. Havighurst, Davis y colaboradores

(citados por Title, 1990), trataron de desarrollar tests mentales que fueran justos culturalmente.

En ausencia de un criterio externo, Title (1990) considera que el estudio de sesgo en pruebas es un área activa de investigación. Para los exámenes más usados en educación como los de logro, desarrollo de habilidades, o mediciones de intereses básicos, no hay estándares externos para fundamentar los juicios de sesgo o justicia. Los métodos estadísticos, fuente principal de investigación, asumen que los ítems de toda una prueba no presentan sesgo. Sin embargo, la definición estadística no es la única que ilustra el estudio del sesgo. Los críticos han explorado el contenido de las pruebas desde diversas perspectivas, haciendo juicios sobre cómo se retrata a las minorías y a las mujeres en estereotipos o como son ampliamente representadas en sus roles ocupacionales, educacionales, familiares o recreativos. Otra forma de detectar el sesgo se basa en el análisis comparativo de ítems de las pruebas estandarizadas con el currículum, los materiales didácticos y la enseñanza.

Para Title (1990), la planeación del examen, su diseño, la revisión de los reactivos y el desarrollo de normas y escalas son etapas en el proceso de desarrollo de una prueba, en el que se hacen juicios que afectan las percepciones del sesgo.

Los procedimientos usados para juzgar el sesgo en el examen incluyen formas de revisión y lineamientos enfocados en estereotipos identificados de mujeres, minorías y discapacitados. Ahí se les pide a los jueces que identifiquen los ítems que puedan ser más o menos familiares para ciertos grupos.

El análisis puede usar categorías como las siguientes: el papel que representan los adultos, los niños, los géneros masculino y femenino; si su papel es principal o secundario, los tipos de medio ambiente, el comportamiento exhibido y los tipos de consecuencias del comportamiento.

A pesar de que en la actualidad los jueces son usados de manera más consistente en el desarrollo de las pruebas, hay poca investigación que se cuestione aspectos como: ¿quién es un juez "experto"?, ¿qué tanto acuerdo debe existir entre los jueces? o ¿el acuerdo debe ser importante? Por ello, Title

(1990) considera que las aproximaciones estadísticas para determinar el sesgo no deben sustituir las revisiones de los jueces. Sin embargo, no se espera que los procedimientos estadísticos y de jueceo identifiquen como sesgados a los mismos ítems. Ambas aproximaciones son necesarias, a pesar de que ninguna sea el criterio final para aceptar o rechazar reactivos en una medición particular.

De manera similar, Childs (1990) considera que debido a que los resultados de los exámenes a menudo afectan el futuro educativo de los examinados, los exámenes deben proveer igualdad de oportunidades para que los estudiantes demuestren sus habilidades y conocimiento. Por ello, es indispensable estudiar el tema de género y justicia dentro de los exámenes, el cual trata acerca de las diferencias entre las oportunidades para hombres y mujeres. Según el autor, un examen está sesgado si un hombre y una mujer que tienen el mismo nivel de habilidad obtienen puntajes diferentes. Las condiciones bajo las cuales una prueba es administrada, el vocabulario empleado en los ítems e incluso la actitud del estudiante hacia el examen afectan los resultados de una prueba.

Mientras que el sesgo es una característica del examen por sí mismo, la justicia se refiere a la manera en que se emplean los resultados de una prueba. Aunque un examen esté libre de sesgo, puede ser usado para dar ventaja o desventaja a los miembros de un género. Por ello, los desarrolladores de pruebas tratan asegurarse de que las preguntas de sus exámenes no están sesgadas y que sus usos no tratarán de manera injusta a los miembros de un género. Sin embargo, ningún examen es perfecto. Una revisión cuidadosa de cada pregunta de un examen puede alertar sobre un posible sesgo. Al respecto, se debe revisar si los reactivos contienen:

- material o referencias que puedan ser ofensivos a los miembros de un género.
- referencias a objetos e ideas que sean más familiares a los hombres o a las mujeres.
- representación inadecuada de hombres y mujeres como actores en los reactivos o representación estereotipada de roles.

Si las preguntas involucran objetos e ideas que son más familiares o menos ofensivas para los miembros de un género, entonces el examen será más fácil para los individuos de ese género. Los estándares para el logro en ese examen podrían ser injustos para los individuos del género a quienes les es menos familiar o se sientan más ofendidos por los objetos e ideas discutidas, porque les será más difícil demostrar su dominio sobre el material (Childs, 1990).

Hambleton (1993) también recomienda dos métodos para detectar un posible funcionamiento diferencial del ítem (FDI): los empíricos y la revisión mediante jueces. A pesar de que existen desacuerdos entre ambos métodos, el investigador considera que deben emplearse de manera complementaria para liberar al examen de reactivos que ofendan o penalicen a miembros de una población.

Los métodos empíricos que recomienda son los modelos IRT y el método Mantel-Haenszel.

Por su parte, la revisión mediante jueces la considera uno de los métodos más efectivos para detectar un posible funcionamiento diferencial del ítem. Consiste en juzgar cada ítem subjetivamente en presencia o ausencia de cualquier característica que pueda llevarnos a la exhibición del FDI. Dichos métodos son también llamados de jueceo, y requieren que un número de jueces realicen tareas relativas a la revisión de una serie de ítems que conforman una prueba.

Los jueces pueden actuar de manera independiente. Sin embargo, es más frecuente que los jueces se reúnan en grupos para tener discusiones más productivas. Además, para asuntos de estereotipos y de representaciones justas, la atención de los jueces debe centrarse en asegurar que los examinados que pertenezcan a los grupos mayoritarios y minoritarios tengan la misma familiaridad y experiencia con el contenido del ítem (Title, 1990). En este sentido, las aproximaciones de los jueces para detectar el FDI juegan el papel de establecer la validez de constructo, contenido y de cara o aparente.

Hambleton (1993) agrupa los aspectos del ítem que deben ser revisados en dos grandes categorías: 1) Representación estereotipada o inadecuada y 2) Sesgo étnico, cultural, religioso y de clase.

En ellas, se plantean una serie de interrogantes que exploran si el ítem presenta algún posible error que afecte a la población examinada (ver tabla No. 1 y anexo No. 1).

Las principales ventajas y desventajas del método de revisión mediante jueces se presentan en la tabla No. 2.

Tabla No. 2 Ventajas y desventajas de la revisión mediante jueces según Hambleton (1993).

Ventajas	Desventajas
▪ Con frecuencia es más barato que el método estadístico el cual requiere la recolección de datos. ▪ Su uso apropiado le da al desarrollador del examen validez de cara que es valiosa y poderosa en el área social, ética y política en que se desenvuelve. ▪ Puede realizarse antes de la aplicación de la prueba; así, los ítems identificados como pobres o defectuosos pueden eliminarse. ▪ Si los jueces son expertos en contenido, puede asegurar la validez de contenido de la prueba.	▪ No siempre coincide con la detección mediante análisis estadístico. ▪ El gasto que implica reunir a los jueces. ▪ El tiempo y el gasto involucrado en la capacitación de los jueces. ▪ La susceptibilidad de los jueces a fatigarse, aburrirse y otras condiciones que podrían interferir con la validez y confiabilidad de sus juicios.

Solano-Flores y colaboradores (2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000) han estudiado la obtención de evidencias de validez desde una perspectiva cultural. Sostienen que esta forma de aproximarse al estudio de la validez de pruebas debe ser incorporada formalmente en el desarrollo de las prácticas de evaluación. El objetivo de su trabajo (Sexton y Solano-Flores, 1999) es contribuir a la obtención de pruebas equitativas, a través de una perspectiva sociocultural para tratar la diversidad cultural en las evaluaciones.

Dichos autores realizaron investigación con la finalidad de detectar qué tan efectivamente los profesores evaluaban la apropiación cultural de los ítems

de ciencias y matemáticas. Es decir, si la experiencia de enseñar en un contexto sociocultural determinado les permitía identificar maneras efectivas de mejorar los reactivos de acuerdo a la noción de validez cultural, y cómo la perspectiva cultural de los profesores y su comprensión de grupos culturales diversos se relaciona con la evaluación de matemáticas y ciencias, así como con las prácticas docentes que se llevan a cabo.

En esencia, la metodología propuesta por Sexton y Solano-Flores consiste en reunir a un grupo de profesores cuya función es examinar los ítems desde las perspectivas de estructura y complejidad; relevancia cultural y sensibilidad; y contenido.

A continuación se definen brevemente cada una de esas perspectivas:

- **Estructura y complejidad.** Qué tan apropiados son los ítems en cuanto a su estructura, vocabulario y grado de dificultad.
- **Relevancia cultural y sensibilidad.** Qué tan relevantes son para la cultura de los estudiantes los problemas y las situaciones presentadas.
- **Contenido.** En qué grado los conceptos manejados son parte de las experiencias primarias de los estudiantes.

Los resultados que obtuvieron los investigadores (Sexton y Solano-Flores, 1999) fueron los siguientes:

- Los antecedentes culturales de los profesores y sus experiencias con los estudiantes pertenecientes a una minoría cultural parecen influenciar los asuntos y el interés al examinar la apropiación cultural de ciencia y matemáticas.
- El nivel de análisis de cómo y por qué las características específicas de un ítem pueden actuar en contra de una minoría de estudiantes parece ser superficial para todos los profesores, sin importar sus antecedentes culturales.

Para Solano-Flores y Sexton (1999), hay dos maneras de validar ítems: la primera mediante prueba empírica, basada en observaciones y datos

estadísticos, y la otra que se realiza mediante juicios, los cuales pueden ser formales o informales.

Los juicios formales pueden ser aparentes, acertados, falsos positivos, falsos negativos, sobre el contenido, sobre el texto e imagen y sobre el contexto y adecuación cultural. Los investigadores antes citados centran su atención en los tres últimos tipos:

En lo que respecta al contenido, estudia el currículum formal; es decir, la validez cognoscitiva y el currículum implementado. Además, cuestiona si la dificultad es acorde al grado escolar del examinado. Estudiar el currículum formal implica evaluar el contenido en el currículum y en los libros de texto. Mientras que estudiar la validez cognoscitiva se refiere a las habilidades y conocimientos requeridos para resolver un reactivo. Por su parte, el currículum implementado hace referencia a si se evalúa lo que enseña la escuela y el profesor.

Los aspectos que se evalúan en cuanto a texto e imagen son tres: formato, demanda lingüística y validez cultural. Para que un formato sea válido debe estar libre de errores ortográficos, de redacción, de formato e impresión. Para satisfacer el criterio de demanda lingüística debe tener redacción simple e incluir sólo los términos técnicos necesarios. En cuanto al aspecto de validez cultural, uno de los más importantes para los investigadores, se enfoca en el uso del idioma y la epistemología. El uso del idioma especifica que deben utilizarse palabras comunes y claras que signifiquen lo mismo entre todos los grupos escolares. Mientras que la epistemología se refiere a que el ítem se interprete igualmente entre todos los individuos de todos los grupos sociales.

Cuando estudia el contexto y adecuación cultural lo hace con el objetivo de indagar sobre el contexto social del examinado; es decir, si contiene situaciones cotidianas de todos los grupos sociales; también examina el contexto del reactivo, que debe estar libre de pistas. Un tercer punto dentro del contexto es el tratamiento; o sea que trate con respeto a todos los grupos sociales; y por último, la equidad: que esté libre de estereotipos étnicos, de género, regionales y socioeconómicos.

2.2.2 Validez de contenido

Para validar el contenido de reactivos empleados en pruebas para evaluar el aprendizaje, Popham (1991) propone también la realización del análisis de ítems mediante métodos empíricos y métodos de juicios. Los primeros son a posteriori; se aplican con mayor frecuencia en pruebas normativas y consisten en ensayar el ítem con examinados para reunir y analizar los datos sobre su ejecución y buscar posibles fallas.

Por su parte, los métodos de juicio son a priori; se emplean con mayor frecuencia en pruebas criterioles y están orientados a la búsqueda de sesgo que ofenda con base en características personales, retratando de forma estereotipada a miembros de un grupo social o penalizando a un grupo de examinados haciéndolos ejecutar un ítem menos bien que otro grupo.

Para Popham, los juicios humanos pueden ser realizados por el elaborador del ítem, jueces independientes o bien los propios examinados.

En caso de que el propio elaborador del ítem se convierta en juez, debe dejar “dormir” el ítem por algún tiempo, digamos una o dos semanas antes de volverlo a revisar formalmente.

Un juez independiente debe ser un experto en contenido, sin interés particular en los ítems o miembro de un grupo potencialmente afectado. Tanto el elaborador del ítem como el juez independiente se darán a la tarea de revisar que el ítem sea congruente con la especificación; que el reactivo se adhiera a las reglas para redactar ítems (sin determinantes específicos, ni otras fallas); detectar errores de contenido, incluso “finos”; e identificar la ofensa en el ítem mediante la revisión de la prueba en su conjunto, para alertar así a revisores y escritores.

Un examinado también puede ser juez, pero solo en situaciones en que no se penalice la ejecución defectuosa. Su tarea como juez es aportar información sobre la claridad de las instrucciones, el tiempo para completar el ítem, ítems confusos, ítems con más de una respuesta correcta (o sin ella) y sobre aquellas palabras que no se entendieron en el reactivo.

Angoff (1989) realizó un estudio para probar la hipótesis que Traynor elaboró en 1985 cuando sostuvo que los ítems del *Test of English as a Foreign Language* (TOEFL) contenían referencia a los estadounidenses, sus lugares y costumbres, así como que daba preferencia a los examinados que han pasado algún tiempo viviendo en los Estados Unidos de Norteamérica. En este estudio se trabajó con dos muestras de examinados. El primer grupo estuvo conformado por aquellos que habían vivido en Estados Unidos más de un año. Mientras que el segundo lo integraron residentes de menos de un mes en aquel país.

Según Ericsson, Molloy, Anderson, Urquhart y Hale (citados por Angoff, 1989) la familiaridad del examinado con el contenido y el idioma tienen un efecto en la ejecución, por lo que Angoff se preguntó si el contexto de un ítem (no su contenido) se orienta a favor de aquellos individuos que están relativamente mejor informados del contexto del ítem.

En el estudio se compararon las puntuaciones medias de cada sección del TOEFL y luego se hizo un estudio de funcionamiento diferencial de ítems (FDI).

Cuando se puntuaron los ítems se les pidió a los examinados que juzgaran cada ítem y expresaran sí: 1) se refería a algún aspecto de la vida estadounidense; 2) no se refería o solo se describía; y 3) era ambiguo al respecto. Después, los ítems fueron clasificados del 1 al 5, donde 1 significó "fuertemente estadounidense" y 5 "no estadounidense". Como las respuestas obtenidas no fueron lo suficientemente explícitas y había áreas de ambigüedad, se requirió una instrucción más detallada para hacer los juicios.

Por su parte, el FDI se hizo mediante el procedimiento Mantel-Haenszel (MH). Se compararon los puntajes dados por los jueces con el fin de determinar si era más clara su referencia o los valores MH. Los resultados de ambos procedimientos demostraron que no había evidencia de sesgo contextual en el TOEFL (Angoff, 1989).

De igual importancia para la validez de contenido es el estudio realizado por Michener y Kesselman (1986) para determinar las funciones técnicas del trabajo mecánico de una compañía privada, como resultado de una

especificación de competencia técnica (ECT). La ECT identificó el conocimiento y las habilidades significativas organizadas en el dominio de contenidos.

Un procedimiento de validación de contenido determinó la relevancia del programa de entrenamiento mecánico y comprobó la validez del contenido analizado mediante la revisión de los planes de clase y otros materiales instruccionales para producir un inventario preliminar de los elementos de entrenamiento para cada curso.

Los expertos en la materia valoraron cada elemento en el inventario de contenido de acuerdo a su importancia para la tarea.

A partir de un caso similar, Zimmerman, Sudweeks, Shelley y Wood (1990) recomiendan que antes de administrar una prueba se obtenga una realimentación sobre su efectividad mediante la consulta a personal ajeno a su elaboración, como pueden ser colegas que opinen acerca de la dificultad de los ítems, editores que evalúen la claridad de la prueba, especialistas en elaboración de exámenes, quienes pueden determinar algunos factores en el formato de la prueba que pueden hacer variar el puntaje de los estudiantes; así como estudiantes que indiquen si el dominio evaluativo de la prueba fue estudiado en el salón de clase.

Además, sugieren crear una tabla que presente el número de ítems que evalúan cada área de contenido y cada tarea, para determinar si todas las áreas fueron evaluadas correctamente. Así como aplicar una pequeña muestra de ítems a los estudiantes a examinar y revisar los ítems mientras el tiempo lo permita.

En el contexto de la evaluación criterial de gran escala en España, Suárez y Jornet (1990a y 1990b) coinciden en que el análisis de sesgo puede realizarse desde dos perspectivas complementarias: métodos empíricos y revisión lógica.

Comentan que los métodos empíricos se basan en procesamiento estadísticos como el método Delta-Plot, Chi-cuadrada y modelos IRT, con la finalidad de mejorar el test con relación a variables externas al mismo y en relación a características de subpoblaciones.

Una segunda propuesta para analizar el sesgo recomendada por los investigadores es la revisión del examen por un grupo de expertos, de preferencia que sean las mismas personas que realicen el análisis y la especificación de contenido. En este sentido, el grupo de expertos revisará la congruencia ítem-objetivo, la calidad técnica del reactivo y la posible existencia de sesgo.

La congruencia ítem-objetivo es la representatividad global de un ítem respecto al objetivo del que se deriva. La representatividad está prácticamente asegurada en el caso de objetivos muy operativizados y reglas precisas de generación, pero es más problemática si se parte de unos objetivos poco definidos.

La calidad técnica se revisa a fin de asegurar que se cumplen las especificaciones técnicas sobre temas de medida.

El sesgo es considerado a fin de evitar que la formulación de los ítems pueda beneficiar o perjudicar a determinados subgrupos de la población a que pertenecen, en relación con otros. Además, se deberá afrontar la detección de jueces que emitan valoraciones extremas y se les pedirá que analicen las posibles consecuencias de la aplicación del examen sobre los estudiantes.

A pesar de que existe el acuerdo en que los exámenes deben demostrar una representación adecuada del dominio del contenido a evaluar, hay pocos procedimientos disponibles para examinar si un examen cumple sus fines. Debido a esto, Sireci y Geisenger (1993) realizaron un estudio en donde emplearon dos métodos para validar contenidos y compararon la información obtenida. Cada método fue realizado por un grupo de 15 expertos en la materia quienes evaluaron dos exámenes: el Auditing Section from the Uniform Certified Public Account Examination y un examen nacional estandarizado de logro sobre estudios sociales.

Los resultados obtenidos mediante correlación y el análisis cluster, sugieren que la estructura del contenido de un examen puede ser evaluada adecuadamente cuando se utiliza el análisis de la similaridad de la información del ítem proporcionada por los expertos en la materia y que dicho procedimiento

debe ser empleado para complementar el análisis de la relevancia del ítem, más que para reemplazarlo (Sireci y Geisinger, 1995b).

Los investigadores mencionan en su reporte dos métodos previos para validar reactivos, los empíricos y los subjetivos. Los empíricos analizan la respuesta del examinado para descubrir su estructura. Por su parte, los subjetivos analizan los datos que proveen los expertos en la materia, quienes estiman la relevancia de los ítems con relación a los objetivos de contenido señalados en las especificaciones del examen.

Los métodos empíricos emplean análisis factorial o escalas multidimensionales para realizar la correlación inter-ítem matricial derivada de la respuesta de los examinados a los ítems. Los factores o dimensiones resultantes son comparados con la estructura del dominio del contenido que persigue el examen, expresado en sus especificaciones.

Según Sireci y Geisinger (1995), ambos métodos son criticados porque el grado de relevancia del ítem y su correspondencia con el dominio del contenido son independientes de la ejecución del examinado en el ítem. De este modo, consideran que hay variables que son irrelevantes para evaluar la representatividad del contenido, tales como la dificultad, el nivel de habilidades, la variabilidad de la población examinada o la motivación, pero sí afectan el factor o las soluciones dimensionales. Aunque esos análisis son relevantes para obtener evidencias sobre la validez de constructo o de criterio, no son el centro de las evaluaciones sobre el contenido de una prueba.

Por su parte, los métodos subjetivos determinan si los ítems de un examen representan las áreas de contenido que pretende medir. Los expertos en contenido revisan los ítems y hacen juicios sobre lo apropiado de cada ítem, cotejándolo o comparándolo con su especificación. En esta línea, Hambleton (citado por Sireci y Geisinger, 1995b) desarrolló una escala de tres puntos, en donde se trata de identificar si el ítem mide cada área de contenido de la prueba. En particular, provee percepciones sobre la correspondencia del reactivo con su especificación. Sugiere que se midan las apreciaciones de los expertos mediante

escalas tipo Likert, se obtenga su media y se proporcione información descriptiva sobre las percepciones sobre el ítem en relación a su contenido.

De igual manera, Aiken (citado por Sireci y Geisinger, 1995b) obtiene un índice derivado de las puntuaciones de los expertos en la materia para evaluar la relevancia de un ítem dentro de un dominio de contenido particular. Ese índice cuenta el número de categorías usadas para puntuar cada reactivo y el número de jueces que respondieron a cada categoría. Se obtiene mediante la fórmula:

$$V = \frac{\sum_{i=1}^{c-1} in_i}{N(c-1)}$$

donde

c= número de categorías de relevancia en la escala de puntuación.

i= peso dado a cada categoría.

ni= número de jueces que revisan el ítem en la categoría (ith).

N= total de jueces.

A la categoría más baja se le da el valor de 0, a la siguiente 1 y así sucesivamente hasta llegar a la más alta, a la cual se le asigna el valor de c-1. Además, Aiken proporciona una fórmula para evaluar la significancia de la validez del índice cuando se trabaja con muchos jueces.

Hay dos limitaciones de este método: como los expertos en la materia están versados en las áreas de contenido, no pueden proveer una evaluación independiente de la estructura del contenido del test. Además, los expertos no evalúan todo el contenido del examen como una estructura, sino que cada ítem se revisa por separado; por lo tanto, no hay información sobre el contenido global de la estructura de la prueba.

Por su parte, Sireci y Geisinger trabajaron con una escala multidimensional (método MDS) con el propósito de descubrir las dimensiones obtenidas del análisis de los expertos en contenido, sobre la similaridad de todos los ítems que conforman una prueba. Así, los jueces revisan la similaridad de todos los posibles pares de reactivos. El método trata de evitar que el juez se

influncie por su conocimiento previo de la estructura del contenido definida por los diseñadores de la prueba.

Las limitaciones encontradas en este método fueron que el proceso no ha sido evaluado, sólo participaron tres jueces en un examen para un salón de clase local, y las dimensiones fueron interpretadas subjetivamente sin análisis complementario que apoyara sus interpretaciones.

No obstante, los resultados de este estudio demuestran que la estructura del contenido de una prueba puede ser evaluada adecuadamente, analizando la similitud de los datos provenientes de los expertos en la materia.

Las configuraciones de los ítems de un examen, que resultan de los análisis multidimensionales, proveen una ilustración libre de sesgo de la estructura del contenido de un examen. Cuando estas configuraciones sean comparadas con las especificaciones de la prueba, se mostrará la homogeneidad de las áreas del contenido (Sireci y Geisinger, 1995b). Según los investigadores, este procedimiento fue útil para identificar aquellas áreas de contenido que fueron representadas de manera pobre por un conjunto de ítems, así como aquellos ítems que no están elaborados conforme a sus especificaciones. Sin embargo, estos resultados no proveen información en relación a la estructura del contenido.

Para asegurar que el dominio de contenido sea representado adecuadamente, los desarrolladores de pruebas deben demostrar: que las especificaciones del contenido definen el dominio adecuadamente; que las especificaciones de la prueba representan las especificaciones del dominio adecuadamente; y, que los ítems de la prueba son representativos de las especificaciones de la prueba.

Para desarrollar especificaciones de exámenes, los desarrolladores de las pruebas deben realizar prácticas analíticas para conocer el conocimiento y las habilidades que se requieren para ejercer una profesión. En las pruebas educativas, las especificaciones son construidas con base en revisiones del currículum nacional, libros de texto y mediante el consenso de expertos en la materia a evaluar.

A pesar de que en la práctica los análisis, revisiones curriculares y los consensos de expertos ayudan a asegurar la adecuada definición del dominio, no aseguran por sí mismos la representación del dominio en el examen.

De acuerdo con los requerimientos de evidencia en relación al contenido que señalan la AERA, la APA y los estándares del NCME (citado por Sireci y Geisinger, 1995b), la primera tarea para los desarrolladores de ítems consiste en especificar adecuadamente el universo de contenido que una prueba intenta representar, dado el propósito de la misma. Otra tarea importante es determinar el grado en que el formato y las propiedades de respuesta de una muestra de ítems son representativos del universo.

Respecto al procedimiento que siguieron Sireci y Geisinger (1995b) para trabajar en el examen de ciencias sociales, a cada revisor se le dio un cuadernillo que contenía todos los pares posibles de 40 exámenes. Así cada cuadernillo tenía 780 pares. Cada ítem se evaluaba en una escala tipo Likert de 10 puntos, etiquetado el 1 como “fuertemente similar” y el 10 como “fuertemente diferente”. Además, los revisores:

- leyeron cada ítem e hicieron juicios en relación a la similaridad de pares de ítems con el contenido que estaba midiendo.
- Encerraron su puntuación en la escala de 10 puntos.
- No se les dio un criterio para establecer sus puntuaciones. Se usó esa ambigüedad para evitar sesgo a favor de las especificaciones de la prueba.
- Después de terminar lo anterior, contestaron un pequeño cuestionario que les preguntaba sobre el criterio que usaron para juzgar la similaridad entre los pares de ítems.
- Posteriormente, evaluaron la relevancia de cada ítem a la luz de cada área de contenido de las especificaciones de la prueba. Estas puntuaciones se hicieron mediante una escala de 10 puntos, donde 1 significó “nada relevante” y 10 “sumamente relevante”.
- Se le dio a cada experto en contenido una hoja con la descripción de cada área de contenido de la especificación de la prueba.

De los datos proporcionados por los jueces se obtuvieron:

- Una matriz triangular de la similaridad de las puntuaciones.
- Una lista de los criterios usados en la elaboración de las similaridades de las puntuaciones de los ítems.

- Una matriz rectangular de la relevancia de las puntuaciones de los ítems.

Continuando con el desarrollo de modelos psicométricos, Nitko (1994), plantea un modelo para crear exámenes nacionales alineados al currículum. El modelo consiste de nueve etapas y la séptima de ellas se refiere al análisis primario de resultados de un examen, cuyos procedimientos incluyen especificar estándares de calidad para el contenido del ítem, para la calidad técnica del ítem y para la calidad integral de los puntajes del examen.

Para Nitko, lo importante es garantizar la calidad de la relación ítem-representa-currículum, la cual solo es posible si el conjunto de los ítems que llamamos examen, representa al universo de contenido que llamamos currículum; entonces podemos tener seguridad de que la ejecución del examinado ante el test es una evidencia que nos permite inferir válidamente cómo sería su ejecución en el resto del dominio que llamamos currículum. Desde luego, la clave para garantizarlo es el juicio humano, el cual está presente a lo largo del proceso (Contreras, 2000).

Al revisar la congruencia ítem-especificación, los ítems elaborados son sometidos a una cuidadosa revisión de contenido, psicométrica y lógica contra las especificaciones correspondientes; además se consideran los estándares, medidas y criterios correspondientes a la calidad del contenido de los ítems del test y a la calidad técnica de cada ítem del test.

Los propósitos de este análisis son crear y mantener un programa de control de calidad y relevancia del examen así como definir las cualidades que el examen debe exhibir antes de que se utilice oficialmente para monitorear los aprendizajes.

Los estándares de calidad técnica que propone Nitko en su modelo se agrupan en calidad del contenido del test, calidad técnica de cada ítem y calidad de las calificaciones de la prueba.

Para efectos de este trabajo solo se considerarán las áreas de calidad del contenido de los ítems y la calidad técnica de cada ítem, las cuales son explicadas en la tabla No. 3 (Contreras, 2000; p. 31).

Capítulo 2. Marco conceptual

Para L'Allier, Miles y Welsh (1995) la validez de contenido es el grado en el que los ítems de una prueba evalúan el contenido específico de manera consistente con la naturaleza del contenido. Es particularmente importante en los exámenes de logro y habilidades. En estas situaciones es crítico que el instrumento de evaluación examine que el contenido sea consistente conforme a los objetivos que pretende evaluar.

Tabla No. 3 Estándares y calidad técnica del contenido del test (Contreras, 2000).

Área de control	Estándar	Medida	Criterio que debe satisfacer
Calidad del contenido de los ítems del test	1. Corrección del contenido 2. Corrección de la respuesta correcta 3. Relevancia e importancia de la tarea a ejecutar 4. Congruencia del ítem del test con el contenido 5. Correspondencia del ítem del test con su especificación	1. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 2. Juicio de expertos en contenido (sí-no) 3. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 4. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos) 5. Clasificación de cada ítem por expertos en contenido (de 0 a 4 puntos)	1. Promedio de clasificación de 3.5 por ítem 2. Todos confirman que la respuesta correcta lo es, o es la mejor opción 3. Promedio de clasificación de 3.5 por ítem 4. Promedio de clasificación de 3.5 por ítem 5. Promedio de clasificación de 3.5 por ítem
Calidad técnica de cada ítem del test	1. Escritura de ítems sin defecto 2. Vocabulario apropiado 3. Dificultad apropiada 4. Discriminación apropiada 5. Evitación de estereotipos étnicos y de género (ofensa) 6. Evitación de sesgo	1. Revisión del ítem por un profesional de la escritura de ítems 2. Todas las palabras en el ítem deben estar incluidas en los materiales usados en la instrucción (juicio del Comité Diseñador: sí-no) 3. Valor "p" del ítem, obtenido de la muestra de ensayo 4. Índice de discriminación 5. Juicios del comité Diseñador 6. Juicios del Comité Diseñador	1. Cada ítem debe estar exento de cualquier defecto de escritura 2. Cada ítem contiene, exclusivamente, palabras incluidas en los materiales empleados en clase 3. $.05 < p < .95$ 4. $rbis > .2$ 5. Ningún ítem juzgado contiene estereotipos 6. Ningún ítem juzgado pone en ventaja a algún grupo

La validez de contenido será baja si los reactivos del examen: 1) indebidamente enfatizan algunos temas, conceptos o procedimientos, o 2) contienen afirmaciones, condiciones o contextos no relacionados con lo que ha sido enseñado, o 3) cubre temas, conceptos o procedimientos que no han sido

enseñados en la clase. Por lo tanto, la validez de contenido no evalúa que tan bien se examina un contenido, sino si evalúa el contenido específico.

La validez de contenido se realiza generalmente mediante la relación de la lista de criterios (desarrollada en los objetivos durante el diseño del proceso instruccional) con una descripción de la manera en que cada criterio será medido. A diferencia de otras formas de validez, la validez de contenido no se expresa como un coeficiente de correlación numérico. En su lugar, se expresa por una comparación objetiva de los reactivos del examen con el contenido del currículum como un porcentaje. Por ejemplo, el examen cubrió el 75% del contenido estudiado.

El proceso completo de diseño, usado por NETg (L'Allier, Miles y Welsh, 1995), para obtener evidencia de validez de contenido, empieza por un análisis de las tareas y continúa con el desarrollo de los objetivos. Después, la validez de contenido se establece mediante la relación de los reactivos de la prueba con sus objetivos, tarea efectuada por el equipo a cargo del control de calidad durante varias revisiones.

Una vez descritos los principales aspectos conceptuales que respaldan este documento, en el siguiente capítulo se presenta la parte metódica en donde se exponen los detalles de los participantes, los materiales e instrumentos empleados así como los procedimientos desarrollados. Cabe señalar que en el siguiente apartado, se consideran las recomendaciones propuestas por la normatividad internacional para la elaboración, evaluación y uso de pruebas, así como los modelos para validar ítems de pruebas de gran escala respecto a aspectos culturales, de género y de contenido.

Capítulo 3

Método

Se evaluó la complejidad cognitiva y cultural de 15 de 32 reactivos del Examen Colegiado de Matemáticas II (Cálculo Integral), elaborado entre 2005 y 2006 por un cuerpo colegiado de la Facultad de Ingeniería Mexicali de la UABC. La revisión se realizó en mayo de 2006 en dicha unidad académica, antes de que la prueba se aplicara a 344 alumnos que cursaban la materia en el tronco común de ingeniería.

3.1 Participantes

Los participantes en el estudio conformaron un comité evaluador que estuvo integrado por seis participantes: dos profesores reconocidos por su experiencia en la enseñanza de las matemáticas, dos profesores expertos en cálculo integral, una persona experta en español y una persona con conocimiento de psicometría.

La capacitación se realizó como parte de las acciones previstas en un convenio de colaboración para la elaboración de exámenes colegiados, celebrado por el Instituto de Investigación y Desarrollo Educativo con la Facultad de Ingeniería de la UABC (Contreras *et al*, 2005). Por lo tanto, cuatro de los participantes fueron profesores de la misma facultad, mientras que la experta en español labora en el nivel medio superior y fue invitada debido a la experiencia que tiene como docente en esta área tanto en el nivel medio superior como en el superior.

3.2 Materiales e instrumentos

Para llevar a cabo la capacitación y asesoría al comité para la validación de ítems mediante jueceo, se elaboraron y utilizaron los siguientes materiales e instrumentos: manual del participante, formato de evaluación, presentación en Power Point, cámara y videos y cuestionario para evaluar la operación del módulo, mismos que se describen a continuación.

3.2.1 Manual del participante

Es el material fundamental empleado en la capacitación de los jueces. Se diseñó de conformidad con las normas técnicas establecidas por el Consejo de

Normalización y Certificación de Competencia laboral (CONOCER) (Excelencia Laboral, 2004) para el diseño y estructuración de programas de capacitación y fue utilizado como texto base para apoyar la capacitación de los miembros del comité evaluador de los reactivos, misma que fue impartida de manera presencial, con una duración de 6 horas efectivas.

Cabe señalar que para la elaboración del manual técnico, se documentó una experiencia de capacitación de jueces para la validación de ítems mediante juicios, que impartió Solano-Flores en el Instituto Nacional para la Evaluación de la Educación (INEE, 2005), la cual fue videograbada en formato DVD por dicha institución y amablemente facilitada para la realización del presente estudio. Los reactivos que se revisaron en tal capacitación formaron parte de la muestra de ítems de los exámenes EXCALE que fueron aplicados en mayo de 2005 a una muestra de los estudiantes de la educación primaria y secundaria del país, correspondientes a la materia de español.

El manual del participante está integrado por un apartado de antecedentes, una introducción al módulo, un primer capítulo denominado aspectos conceptuales, un segundo capítulo titulado evaluación de reactivos, las conclusiones generales, un glosario y las referencias bibliográficas (ver anexo No. 2). Enseguida se describen brevemente dichos apartados:

Los **antecedentes** ofrecen al participante un breve panorama de la evaluación criterial del aprendizaje a gran escala, de la metodología que se ha desarrollado para la elaboración de estos exámenes en el IIDE-UABC, así como experiencias similares de evaluación que se han desarrollado en el ámbito internacional, nacional y regional.

Por su parte, la **introducción al módulo** presenta al participante el objetivo de la capacitación y describe los temas a desarrollarse en cada apartado del manual.

El primer tema, denominado **aspectos conceptuales**, contiene algunas nociones conceptuales necesarias para entender la validación de ítems mediante juicios. Primero se abordan los principales tipos de validez

relacionados con la tarea de evaluación de reactivos. Posteriormente se exploran los métodos empíricos y de jueces para evaluar los ítems, los cuales permiten obtener evidencias sobre su validez. El segundo tema, **evaluación de reactivos**, presenta el formato de evaluación a utilizar; define algunos conceptos relativos al formato, explica la normatividad para su uso y los procedimientos para llevar a cabo la evaluación de los reactivos.

Finalmente, se presentan las **conclusiones generales** que engloban lo esencial de la capacitación; le sigue un **glosario** que define algunos términos que se podrían prestar a confusión entre los participantes y por último las **referencias bibliográficas** que se consultaron para elaborar el manual.

3.2.2 Formato de evaluación de reactivos

Consta de 9 afirmaciones que se integraron con base en las cuatro características más relevantes en las que coinciden los autores sobre validez de ítems: contenido, texto e imagen, congruencia con las convenciones psicométricas y contexto y adecuación cultural. Como se aprecia en la tabla No. 4, de cada área se desprenden una serie de indicadores que exploran ciertos aspectos de validez en los ítems.

El formato de evaluación de reactivos sintetiza en 9 afirmaciones los aspectos recomendados por los autores en la tabla No. 4 (Hambleton, 1993; Popham, 1991; Suárez y Jornet, 1990a y 1990b; Solano-Flores y col., 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000).

Capítulo 3. Método

Tabla No. 4 Áreas a explorar en el formato para la evaluación de reactivos (Hambleton, 1993; Popham, 1991; Suárez y Jornet, 1990a y 1990b; Solano-Flores y col., 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000).

Área a explorar	Indicador
Contenido	- El nivel de dificultad es apropiado al nivel del examinado. - Evalúa el dominio de conocimientos o procedimientos considerados en el currículum formal, los libros de texto y que son enseñados por los docentes (currículum implementado). - El conjunto de ítems representa apropiadamente al currículum. - El reactivo se resuelve utilizando determinadas habilidades o conocimientos.
Texto e imagen	- Está libre de errores ortográficos, de formato, de diseño y de impresión. - Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar. - La redacción es simple e incluye sólo los términos técnicos necesarios. - Utiliza un vocabulario que puede ser interpretado igualmente por examinados de todos los grupos sociales.
Congruencia con las convenciones psicométricas	- La redacción del ítem es clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es. - Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.
Contexto y adecuación cultural	- Contiene situaciones comunes a todos los grupos sociales. - Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros. - El reactivo trata con respeto a los miembros de todos los grupos sociales. - Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.

En cuanto al **contenido** del reactivo las afirmaciones evalúan el currículum formal, el currículum implementado y la validez cognoscitiva. En el **currículum formal**, valoran el contenido del currículum y de los libros de texto, así como si la dificultad es acorde con el grado escolar del examinado. Sobre el **currículum implementado** evalúan si el reactivo se centra en el conocimiento que se enseña en la escuela y que es enseñado por el profesor dentro del salón de clase. Por último, sobre el contenido se examina la **validez cognoscitiva**, es decir, si el reactivo sólo se resuelve mediante habilidades o conocimientos que se evalúan.

Por su parte, los juicios sobre el **texto e imagen** se enfocan en el formato, la demanda lingüística y la validez cultural. En el **formato** se busca que el reactivo esté libre de errores ortográficos, de redacción, de formato o de

impresión. Además, se indaga si las imágenes o gráficas que se presentan en el reactivo se enfocan sólo en el contenido a evaluar. A su vez, la **demand lingüística** se refiere a que el reactivo tenga redacción simple e incluya solo los términos técnicos necesarios. Mientras que los juicios sobre **validez cultural** revisan el uso del idioma y la epistemología: en el uso del idioma se pretende asegurar que el ítem contenga palabras comunes y claras que signifiquen lo mismo en todos los grupos sociales, mientras que en la epistemología se verifica que el reactivo pueda interpretarse igualmente por todos los individuos de todos los grupos sociales.

Respecto a las **convenciones psicométricas** se examina si la redacción del reactivo es clara, esencial y sin errores conceptuales; si la respuesta correcta realmente lo es. Si está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.

Finalmente, los juicios sobre el **contexto y adecuación cultural** indagan sobre el contexto social del examinado, el contexto del reactivo, el tratamiento y la equidad. En el **contexto social del examinado** se verifica que el reactivo contenga situaciones cotidianas en todos los grupos sociales. En el **contexto del reactivo** se busca que el reactivo esté libre de pistas que puedan beneficiar o perjudicar al examinado. La parte del **tratamiento** se refiere a que el reactivo trate con respeto a los miembros de todos los grupos sociales. Finalmente, la **equidad** indaga si el reactivo está libre de estereotipos de género, étnicos, regionales y socioeconómicos, que ofendan a los examinados.

En la figura No. 1 se presenta una versión comprimida del formato empleado para registrar los juicios de los integrantes del comité evaluador de ítems y en el anexo No. 12 se presenta el formato que se empleó durante la sesión de evaluación de reactivos.

Capítulo 3. Método

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	
2. Utiliza un vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículum, los libros de texto y que son enseñados por los docentes	
9. El conjunto de ítems representa apropiadamente al currículum.	

Figura No. 1 Versión sintética del Formato de evaluación de reactivos

3.2.3 Presentación en Power Point

Contiene el material de apoyo para la introducción del curso, mismo que resultó necesario para capacitar a los miembros del comité a fin de determinar la validez de ítems. Esta presentación consta de doce diapositivas que ilustran aspectos críticos de los temas que se desarrollan en el manual del participante (ver anexo No. 3).

3.2.4 Cámara y videos

Los videos contienen la información que documenta las sesiones de capacitación, la asesoría prestada a los miembros de los comités, así como el proceso y resultados de la evaluación de los ítems por los jueces, a fin de contar con datos para contrastar los supuestos del estudio, así como para tener evidencias que documentaran las decisiones que adoptaron los comités sobre los ítems.

3.2.5 Cuestionario para evaluar la operación del módulo.

Consta de 32 preguntas, de las cuales 26 son de tipo Likert con 4 opciones de respuesta y el resto son preguntas abiertas (ver anexo 3). El cuestionario indaga sobre los materiales empleados (manual técnico del participante, formato de evaluación de reactivos y la presentación en Power Point), así como los resultados de la sesión de capacitación, de la asesoría y de la validación de reactivos del examen de cálculo integral.

3.3 Procedimientos

3.3.1 Capacitación

La capacitación tuvo por objetivo preparar a los miembros del comité para validar, mediante jueceo, los ítems que forman parte del examen criterial de cálculo integral; prueba para aplicarse a gran escala, que fue diseñada por el colegio de profesores de matemáticas de la Facultad de Ingeniería Mexicali de la UABC.

La planeación didáctica de la sesión de capacitación se diseñó con base en la propuesta temática que Solano-Flores utilizó en el INEE en 2005, para lo cual se estructuró un esquema con los temas a desarrollar, así como de las actividades a realizar por los participantes (ver figura No. 2).

La capacitación constó de cuatro etapas: introducción, aspectos conceptuales, evaluación de reactivos y evaluación de la capacitación. Se invirtieron seis horas a estas actividades, las cuales se filmaron con la finalidad de documentar el procedimiento.

En la etapa denominada **aspectos conceptuales**, se revisaron algunas nociones teóricas necesarias para entender el proceso para determinar la validez de ítems mediante juicios. Primero se abordaron los tipos de validez; posteriormente se exploraron los métodos empíricos y de jueceo para determinarla, los cuales permiten obtener evidencias de validez de ítems.

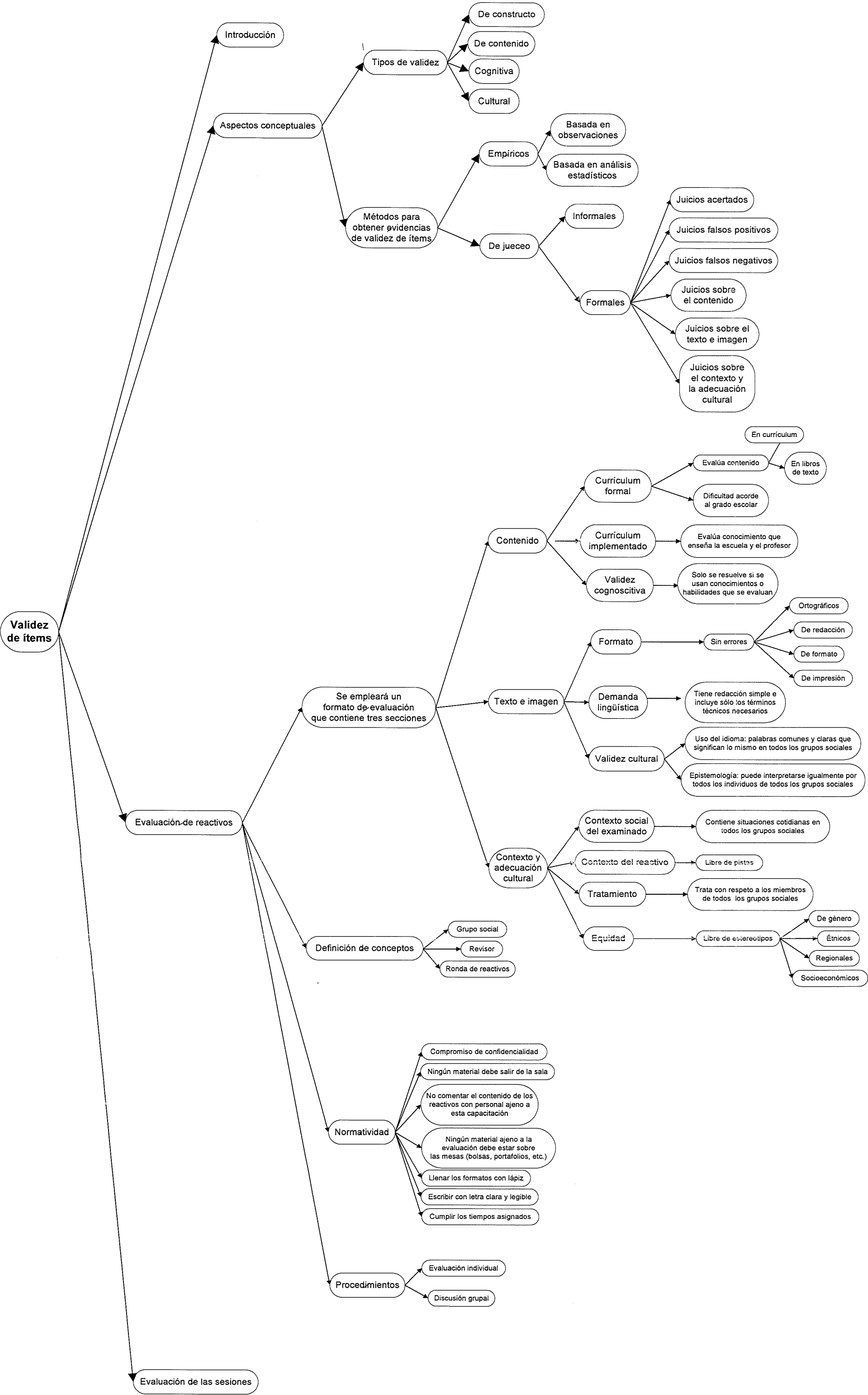


Figura No. 2 Esquema para la capacitación de los miembros del comité para validar ítems mediante juicios

Después, en la etapa de **evaluación de reactivos**, se examinó el formato a utilizar para evaluar los ítems; se definieron algunos conceptos relativos al formato, se explicó la normatividad a seguir y los procedimientos para llevar a cabo la evaluación de los reactivos.

Finalmente, se procedió a la **evaluación de la sesión**. Para ello se valoraron de manera individual tanto el proceso de capacitación como la asesoría, con fines de realimentación.

3.3.2 Elaboración del formato de evaluación de reactivos

El desarrollo del formato de evaluación, involucró una síntesis de las propuestas que diferentes autores plantean sobre validación de ítems mediante juicios, mismas que se describieron en el capítulo 2 y se plasman en la tabla No. 4. Así, se consideraron las aportaciones de Hambleton (1993), Popham (1991), Suárez y Jornet (1990a y 1990b) y Solano-Flores y colaboradores (2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000).

A continuación se presentan las propuestas que se retomaron para elaborar el formato de evaluación de reactivos:

- La identificación de estereotipos, la presencia de sesgo étnico, cultural o de clase, la ofensa, la penalización, el contenido y la revisión del formato (Hambleton, 1993).
- La ofensa y la penalización, el contenido, la congruencia curricular, respuesta correcta, uso del vocabulario y las propiedades psicométricas (Popham, 1991).
- La congruencia ítem-objetivo, la calidad técnica y el sesgo (Suárez y Jornet, 1990a y 1990b).
- La redacción, la dificultad apropiada, la evitación de estereotipos y la ofensa (Nitko, 1994).
- El contenido; el texto e imagen y el contexto y la adecuación cultural (Solano-Flores y colaboradores, 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000).

Si bien las cuatro áreas a explorar que considera el formato de evaluación de reactivos son propuestas por Solano-Flores y colaboradores (Solano-Flores y colaboradores, 2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000), los enunciados que lo integran están asociados con esa clasificación y han sido complementados con otros aspectos que estos autores no consideran pero que son relevantes.

Una síntesis de las sugerencias mencionadas con anterioridad se encuentra en la figura No. 3, instrumento auxiliar para la elaboración del protocolo. Ahí se pueden identificar de manera gráfica las propuestas de los autores, se observan las semejanzas y diferencias. De esos conceptos surgen las ideas clave que fueron integradas en los 9 enunciados del formato de evaluación de reactivos.



Figura No. 3. Áreas a explorar del reactivo, según las aportaciones de Hambleton, Popham, Jornet y Suárez, Nitko y Solano-Flores.

3.3.3 Revisión de ítems

Se comentó con los participantes que para la evaluación de reactivos se trabajaría con un formato simple, que sería fácil de llenar, requeriría poco entrenamiento por parte del juez, permitiría integrar juicios e identificaría a los reactivos de mejor calidad para incluirlos en los exámenes, así como mejorar los ítems con fallas.

Para llenar el formato, se informó a los jueces que su labor como revisores consistiría en inspeccionar el reactivo y marcar con una cruz en el cuadro correspondiente, si estaban o no de acuerdo con cada una de las afirmaciones. Si marcaban con un “no” una aseveración, tendrían que justificar ese “no” y decir cuál es el error que identificaron. Además, se expuso la importancia de que la cruz no invada otros cuadros y que las posibles anotaciones que se hagan sean con letra legible, de lo contrario habría dificultad para entender sus opiniones.

Continuando con la evaluación, se añadió que son tres las secciones que integran el formato de evaluación: 1) contenido; 2) texto e imagen; y 3) contexto y adecuación cultural.

Para iniciar el trabajo de evaluación de reactivos, se presentó la **normatividad** que regiría la revisión. En especial se consideró el compromiso de confidencialidad, por lo que se explicó que ningún material debería salir la sala de juntas donde se llevó a cabo la sesión y que no se debía comentar el contenido de los reactivos con nadie ajeno al equipo de revisores.

De igual forma, se añadió que para que la capacitación cumpliera con sus objetivos, habría otros compromisos que deberían asumirse, tales como dejar las bolsas, portafolios, cuadernos o cualquier otro objeto que no fueran los reactivos, formatos de evaluación y lápices en una mesa o en otro lugar para que no interfieran el trabajo de revisión. También se señaló que todos los formatos debían ser llenados con lápiz y con letra clara para facilitar la revisión.

Un último punto esencial para realizar el trabajo de la manera más eficiente posible, se argumentó que era cumplir con puntualidad los tiempos asignados para cada tarea, ya sea para trabajar o para descansar. Para que la

normatividad anterior se cumpliera, se cuidaron dos condiciones: primero, se les dieron lápices, borradores y sacapuntas a los revisores, posteriormente se entregaron reactivos preasignados para iniciar la labor de evaluación.

El procedimiento que desarrollaron como evaluadores constó de dos etapas: La primera consistió en realizar una **evaluación individual**, en la que cada revisor recibió sus reactivos, los contestó como si fuera un examinado y llenó el formato de evaluación descrito anteriormente. Lo anterior tuvo como objetivo identificar reactivos que presentaran algún problema. Para realizar la evaluación individual, los revisores:

- Leyeron cuidadosamente y contestaron el reactivo.
- Respondieron (con lápiz) sí o no a cada pregunta de la evaluación, con base en su experiencia profesional.
- Justificaron cada respuesta “no” en el espacio asignado para ello.
- Agregaron los comentarios que consideraron necesarios en el cuerpo del reactivo. Ninguno de los jueces utilizó el reverso para escribir algún comentario.

Una vez realizado el procedimiento anterior, se llevó la segunda etapa, que fue una **revisión grupal** en la que se discutió cada reactivo identificado como problemático. Ahí, se les pidió que propusieran los reactivos que ameritaban una discusión. Tales ítems se proyectaron en pantalla y se discutieron en grupo.

Para coordinar esta discusión se consideró importante tomar en cuenta tres aspectos:

- Evitar que la discusión se desviara,
- Optimizar el tiempo de discusión y
- Llegar a un consenso sobre las características del reactivo que deberían modificarse.

3.3.4 Aplicación del cuestionario para evaluar el módulo

El cuestionario se empleó para valorar las actividades realizadas durante la operación del módulo, así como los materiales empleados (ver anexo No. 3). El esquema metodológico para elaborarlo consistió en desarrollar el cuestionario

Capítulo 3. Método

a partir una tabla con factores, indicadores y subindicadores a valorar sobre la capacitación, asesoría, evaluación de reactivos y materiales empleados, tal como se muestra a continuación en la figura No. 4:

Factores	Indicadores	Subindicador	Preguntas
Manual técnico del participante	Contenido	Conceptos y procedimientos	¿Presentó conceptos y procedimientos que resultaron claros? ¿Los conceptos y procedimientos incluidos en el manual facilitaron la tarea de evaluación de ítems mediante jueceo? ¿En general los conceptos y procedimientos descritos resultaron complejos?
		Extensión	¿El contenido del manual resultó demasiado extenso para el tiempo en que se revisó?
	Protocolo	Utilidad para Identificar errores en los ítems	¿Los enunciados son claros? ¿Permiten identificar errores en los ítems?
	Ejemplos	Significatividad	¿Ayudan a entender los conceptos? ¿Son claros y significativos?
	Presentación de Power Point	Conceptos y procedimientos descritos	¿Las diapositivas apoyan la comprensión de conceptos y procedimientos descritos? ¿La presentación es completa? ¿El tipo de letra es legible?
	Estructura	Introducción	¿Facilitó la comprensión del contenido del manual? ¿Facilitó la comprensión de la actividad evaluativa a realizar?
		Metodología	¿Los procedimientos que describe son claros y significativos? ¿Se explican con la profundidad suficiente los procedimientos a realizar? ¿Algún procedimiento resultó confuso o complejo?
		Ejercicios evaluativos	¿Los ejercicios presentados apoyaron en la comprensión de conceptos y procedimientos? ¿Los ejercicios presentados son suficientes?
Operación de la capacitación	Instructor	Dominio del tema	¿Demostró el instructor dominio del tema? ¿Cuando se le preguntó, resolvió de manera clara la duda en cuestión? ¿El vocabulario que utilizó fue innecesariamente complejo?
		Fluidez	¿Su discurso fue fluido?
		Claridad	¿Fue clara su exposición?

Capítulo 3. Método

Factores	Indicadores	Subindicador	Preguntas
		Seguridad	¿Mostraba seguridad al desarrollar su actividad como instructor?
		Coherencia	¿Hubo coherencia en sus fundamentaciones?
		Didáctica	¿Propició el instructor la participación de todos?
	Inducción	Objetivos de la capacitación	¿Se hizo una presentación de los miembros del comité y del instructor? ¿Se presentaron los objetivos de la capacitación? ¿Quedó una idea clara sobre la intención de la capacitación?
	Desarrollo	Actividades realizadas	¿Las actividades realizadas permitieron que usted lograra un aprendizaje significativo? ¿Durante la capacitación se resolvieron sus dudas? ¿La normatividad sobre la confidencialidad quedó explicada claramente?
	Cierre	Conclusiones	¿Se presentaron conclusiones sobre los puntos más importantes al finalizar la sesión? ¿Se lograron los objetivos propuestos al inicio de la instrucción?
Operación de la evaluación de ítems	Tiempo destinado a la capacitación	Suficiencia	¿El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems? ¿A qué parte le dedicaría más tiempo?
	Sesión de evaluación de reactivos	Conducción de la sesión de evaluación	¿La sesión se condujo de manera que siempre se logró el consenso? ¿Se respetaron los tiempos asignados para realizar las actividades? ¿Se comprendió lo que se realizaría en la sesión de evaluación?
	Integración del comité	Miembros que participaron	¿Considera adecuada la constitución del comité? ¿Todos los miembros participaron?
Resultados de la evaluación de los ítems por el comité	Productos obtenidos	Logro de objetivos	¿Se obtuvieron los productos esperados? ¿El trabajo de los jueces ayudó a mejorar la calidad de los reactivos?
	Importancia del trabajo	Resultados obtenidos	¿Considera que se mejoraron los ítems? ¿Considera que el proceso de capacitación jueces es necesario para evaluar los reactivos efectivamente? ¿Considera que vale la pena la revisión de los ítems por parte de jueces? ¿Los resultados obtenidos lo convencen?

Factores	Indicadores	Subindicador	Preguntas
			de la importancia de la validación de ítems? ¿La utilización del protocolo produjo los resultados esperados?

Figura No. 4 Factores e indicadores para evaluar la capacitación y asesoría al comité

Como puede observarse en la figura No. 4, para diseñar el cuestionario se consideraron cuatro factores evaluativos: el manual técnico del participante, la operación de la capacitación, la operación de la evaluación de ítems y los resultados de la evaluación de los ítems por el comité. Enseguida se comentan brevemente:

- Para evaluar el **manual técnico del participante** se consideraron indicadores que valoran la claridad y complejidad de los conceptos y procedimientos revisados, así como su extensión; la utilidad del protocolo para identificar errores en los ítems, la significatividad de los ejemplos presentados, los conceptos y procedimientos descritos en la presentación de Power Point y la estructura propia del manual (introducción, metodología y ejercicios evaluativos).
- Sobre la **operación de la capacitación** se evaluó el dominio del tema, la fluidez, claridad, seguridad, coherencia y didáctica de la instructora; además, se valoró si en la inducción se presentaron los objetivos de la capacitación, las actividades realizadas durante el desarrollo de la sesión, las conclusiones presentadas en el cierre y se preguntó si el tiempo destinado a la capacitación fue el suficiente para lograr sus objetivos.
- Por su parte, en la **operación de la evaluación de los ítems** se indagó sobre la conducción de la sesión de evaluación y la participación en ella de los miembros en la actividad.
- Finalmente, sobre los **resultados de la evaluación de ítems** por los miembros del comité se preguntó a los participantes si se lograron los objetivos y su apreciación sobre la importancia del trabajo realizado.

3.3.5 Análisis estadístico de reactivos

Con la finalidad de obtener alguna evidencia empírica sobre las diferencias, en cuanto a las propiedades psicométricas de los ítems, entre el conjunto de 15 reactivos que fueron revisados mediante jueceo y los otros 17 que no recibieron este tratamiento, tras la aplicación a gran escala del examen se realizó un ítem análisis con dos programas: ITEMAN (Assessment Systems Corporation, 1993) con el que se determinó la dificultad y el poder discriminativo de todos los reactivos. Además, se trabajó con el software para analizar reactivos que se desarrolló en el IIIDE-UABC.

3.3.6 Transcripción y análisis de videos

El proceso de capacitación, revisión de reactivos y evaluación del módulo se filmaron para tener evidencia que apoyara la identificación de los resultados del uso de los formatos de evaluación de reactivos por los jueces, así como las respuestas al cuestionario para evaluar la operación del módulo. Para este propósito, los videos fueron proyectados un mínimo de tres veces y de ahí se rescató información sobre las dudas de los participantes, sus aportaciones y el proceso de revisión de ítems. Sólo esos datos se transcribieron ya que son los que aportan la información más relevante de lo ocurrido en el módulo.

Capítulo 4

Resultados

La aplicación de los procedimientos descritos en el capítulo anterior, así como la administración de los materiales diseñados, produjo los resultados que se presentan a continuación.

4.1 Videos de capacitación, formato de evaluación de reactivos y cuestionario de evaluación del módulo.

La capacitación se realizó en mayo de 2006 en la Facultad de Ingeniería de la UABC en la ciudad de Mexicali, B. C. A la capacitación asistieron 5 docentes de la Facultad de Ingeniería y una docente del CECYTE BC. Uno de los docentes de la UABC, abandonó la sala a las 12:30 p.m. y ya no regresó; aunado a esto, otro docente de la misma unidad académica no continuó después de la hora de comida. Posteriormente, un docente que no asistió al inicio de la capacitación se integró cuando se estaba realizando la sesión de evaluación de reactivos. Cabe señalar que, aunque participó en la discusión, este juez no contestó el formato de evaluación de reactivos, ni el cuestionario de evaluación del módulo.

Que algunos participantes dejen el módulo de capacitación implica algunas dificultades para el desarrollo de la sesión de revisión de reactivos, ya que al disminuirse el número de participantes se tienen menos juicios, se presentan más interrupciones en el proceso y se extiende la duración de las sesiones. Por lo tanto, se recomienda a las personas que en el futuro conduzcan talleres para capacitar a los revisores de ítems, formalizar por escrito el compromiso de permanencia de los participantes, así como apagar celulares para evitar distracciones durante la operación del módulo. También se puede realizar la actividad fuera de la ciudad de los participantes, así se llevaría a cabo la sesión de manera integral y sin interrupciones.

El desarrollo de la capacitación tuvo una duración de dos horas (ver anexo No. 2). Después, se procedió con la sesión de evaluación de reactivos, durante la cual los jueces revisaron una muestra de 15 reactivos de un examen que consta de 32 ítems. Finalmente, los participantes contestaron el cuestionario de evaluación del módulo y de esta manera concluyó la operación del módulo por parte del comité.

Es importante mencionar que buena parte de los resultados descritos en esta sección forman parte de una síntesis elaborada con base en la transcripción de los videos que se tomaron (ver Anexo No. 8), mismos que muestran los comentarios, participaciones, dudas y sugerencias más relevantes del comité evaluador de reactivos.

A continuación, se ilustra el procedimiento desarrollado para analizar los resultados obtenidos mediante la aplicación de los tres instrumentos descritos en el método. Si bien con los 15 ítems se realizaron los mismos pasos, para ilustrar el análisis efectuado en cada uno de ellos en este apartado solo se presentará, para el caso del reactivo No. 1, la triangulación de las evidencias proporcionadas por la narración de los videos, el formato de evaluación de reactivos y el cuestionario de evaluación del módulo.

En primer lugar, de acuerdo con la videograbación, la profesora de español sugirió mejorar la redacción y los demás jueces estuvieron de acuerdo en agregar el término “ecuaciones” o “expresiones” a la base del reactivo, porque estaba incompleta. En este sentido, las respuestas que dieron los jueces en el formato de evaluación de reactivos corresponden con lo narrado en el video, ya que 4 jueces señalaron un error en la redacción e hicieron propuestas para mejorar la base del reactivo.

En segundo lugar, los videos presentan una discusión que giró en torno al contenido a evaluar en el reactivo No. 1. Los jueces argumentaron que el contenido esencial era la integración y no la aritmética, por lo que acordaron cambiar el exponente fraccionario del distractor “A” por un número entero para evitar que los alumnos se detengan en esa parte del reactivo y demuestren su dominio sobre la derivación. De igual forma, las respuestas de los jueces 2 y 3 al formato de evaluación de reactivos, respaldan que se identificó el error, ya que el contenido a evaluar en el primer distractor no corresponde al nivel cognitivo a que apela el ítem.

Cabe señalar que ambas problemáticas son congruentes con lo expresado en un tercer instrumento: el cuestionario para evaluar la operación del módulo. Ahí, los jueces comentaron que los ítems se mejoraron en su redacción, en el formato y en el contenido. Además, expresaron que la sesión de capacitación los ayudó a

Capítulo 4. Resultados

comprender la importancia de la revisión de ítems para mejorar la calidad técnica de los reactivos, evitar el sesgo y la discriminación, asegurar que el examen sea justo, así como contar con un comité revisor integrado por diversos especialistas.

4.2 Errores registrados por los jueces en el formato de evaluación de reactivos

Como ya se comentó, en la sesión de evaluación de reactivos los jueces revisaron 15 reactivos del Examen Colegiado de Matemáticas II que consta de 32 ítems. Posterior a la revisión individual de cada uno de ellos, los jueces registraron los errores identificados en cada ítem en el formato de evaluación de reactivos. Como puede observarse en los registros de los jueces (ver figuras No. 5 a la No. 10 y el anexo No. 5), en los ítems 5, 6, 9 y 10 no se encontraron errores; no obstante, el resto de los ítems sí presentan observaciones del tipo 1 al 9, conforme al formato de evaluación de reactivos.

Cabe señalar que, durante la revisión individual, los jueces comentaron otros errores que no encajaban en alguno de los 9 tipos considerados en el formato, por lo que la instructora sugirió que los anotaran en el espacio del error 9, el cual no sería utilizado porque se examinaron solamente 15 ítems de la prueba. Para fines de esta presentación de resultados, se agregó una columna a los tipos de error, denominada 10, en la cual se registró la falta de tiempo para responder el reactivo o la propuesta para eliminar un reactivo.

Una vez concluida la revisión individual los jueces procedieron a la fase de revisión grupal, en la cual se decidió por consenso cuáles errores se registrarían una vez que fueron suficientemente discutidos y justificados.

En las figuras No. 5 a la No. 9 se aprecian los errores que identificó cada juez en la fase de revisión individual y que registró en el formato de evaluación de reactivos. Cuando no se identificó error alguno, la figura indica "sin error". Por otro lado, es conveniente señalar que en la figura No. 10 también se representó de manera gráfica el consenso al que llegaron los jueces respecto a cada ítem evaluado; ahí se observan los errores identificados durante la sesión de revisión grupal y que se acordaron con base en la opinión y argumentación de los participantes.

A su vez, en la figura No. 11 se aprecia la frecuencia y el porcentaje de ítems con error por tipo de error identificado.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4	Sin error									
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Figura No. 5 Errores identificados por la profesora de español en los ítems, según el formato de evaluación de reactivos.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Figura No. 6 Errores identificados por el profesor de matemáticas 1 en los ítems, según el formato de evaluación de reactivos.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Figura No. 7 Errores identificados por el profesor de matemáticas 2 en los ítems, según el formato de evaluación de reactivos.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8										
9										
10										
11										
12										
13										
14										
15										

Figura No. 8 Errores identificados por el profesor de matemáticas 3 en los ítems, según el formato de evaluación de reactivos.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4	Sin error									
5										
6										
7										
8	Sin error									
9	Sin error									
10	Sin error									
11										
12	Sin error									
13										
14	Sin error									
15										

Figura No. 9 Errores identificados por el psicómetra en los ítems, según el formato de evaluación de reactivos.

Ítem	Tipo de error									
	1	2	3	4	5	6	7	8	9	10
1										
2										
3										
4										
5										
6										
7										
8	Sin error									
9										
10										
11										
12										
13	Sin error									
14	Sin error									
15	Sin error									

Figura No. 10 Errores identificados mediante el consenso en la revisión grupal de los ítems, según el formato de evaluación de reactivos.

Tipo de error	N	%
1	12	80.00
2	9	60.00
3	1	6.66
4	2	13.33
5	0	0.00
6	0	0.00
7	1	6.66
8	3	20.00
9	0	0.00
10	9	60.00

10→ Otro tipo de error. En este caso fue la falta de tiempo para responder el reactivo o la propuesta de eliminar el ítem.

Figura No. 11 Número y porcentajes de ítems con error identificado, por tipo de error, según el formato de evaluación de reactivos.

Como puede observarse en la figuras No. 10 y 11, hubo 3 tipos de error que no fueron identificados por los jueces, tal es el caso de las afirmaciones No. 5, 6 y 9.

Por otra parte, en el anexo No. 5 se presentan los comentarios que los jueces escribieron en el formato de evaluación de reactivos. En él pueden observarse algunas semejanzas en los errores identificados en cada reactivo y en las propuestas para mejorar los mismos. Tal es el caso del reactivo no. 1, en donde 4 de los 5 jueces detectaron errores de texto e imagen, por lo que en el consenso determinaron que se agregaría un término a la base del reactivo para hacer más clara la redacción. En el mismo anexo, también se aprecian diferencias en cuanto a los registros. Si revisamos los comentarios del ítem 10, por ejemplo, un juez identificó un error de texto e imagen, mientras otro encontró fallas en el contenido y los demás revisores no hallaron fallas.

Considerando a los 15 ítems, se agruparon los errores por área con la finalidad de visualizar cuáles son aspectos a mejorar de los reactivos. Como puede observarse en la figura No. 10, según el consenso de los jueces se presentaron errores en el contenido, el texto e imagen así como en la congruencia con las convenciones psicométricas. Para el caso de **contenido**, fueron 12 los ítems que presentaron alguna falla relacionada con el nivel cognitivo de los reactivos y en un caso se expresó que ese contenido no era abordado por todos los profesores en las clases.

En **texto e imagen**, fueron 13 los ítems identificados con error debido a que faltaban signos de puntuación, se hicieron propuestas para mejorar la redacción, hubo fallas en el formato y diseño, así como en las gráficas.

En **congruencia con las convenciones psicométricas** se encontraron 3 errores en un distractor, la redacción poco clara de un ítem y en un reactivo sin respuesta correcta. Una aportación que dio el trabajo grupal al formato de evaluación de reactivos fue el sustituir el último enunciado en donde se indagaba si el conjunto de ítems representa al currículo por una afirmación en el que los examinadores de reactivos pusieran algún otro error que no señalara el formato.

Debido a la naturaleza del contenido de la prueba, no se identificaron errores en el **contexto y adecuación cultural**, ya que el lenguaje empleado en la base del

reactivo sólo hace referencia a las funciones y expresiones propias del cálculo integral. Sin embargo, los jueces comentaron durante la capacitación que es importante conocer esta área para así evitar el sesgo y una posible penalización+ que ponga en ventaja o desventaja a algunos examinados respecto a otros en futuros ítems que elaboren.

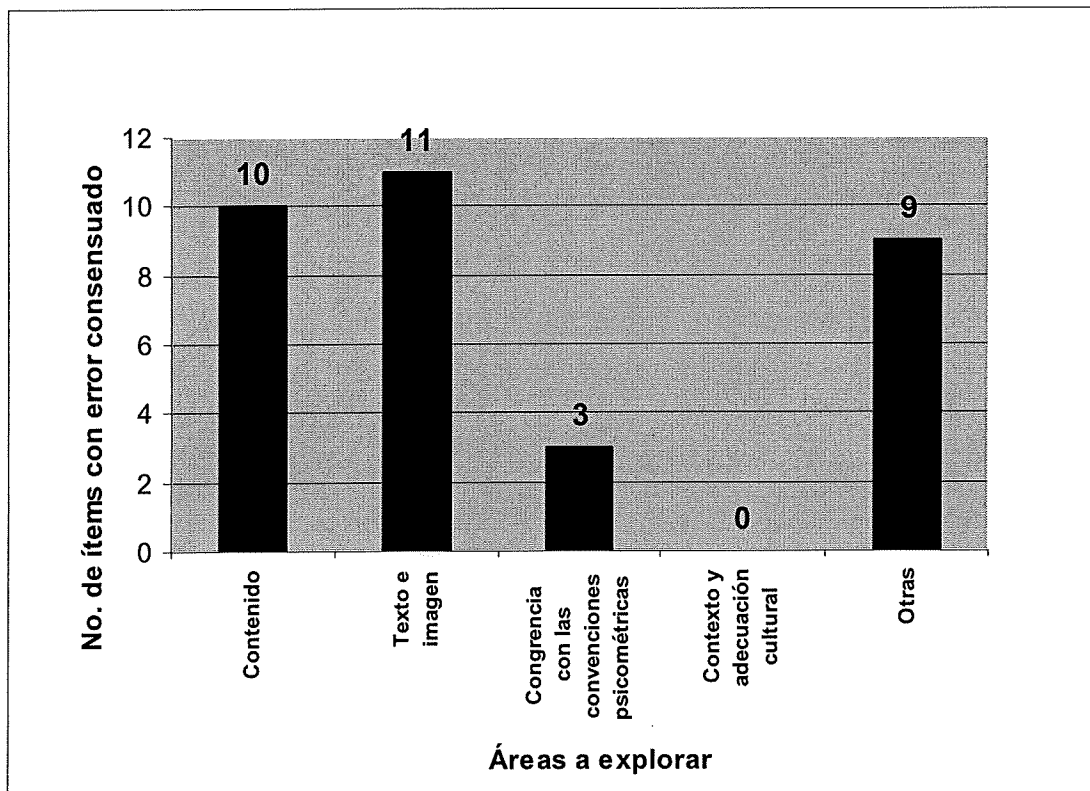


Fig 12. Reactivos con error consensuado por área

Además, la figura 12 presenta los cuatro posibles errores por área que puede tener un ítem. En el reactivo 11 por ejemplo, se observa que hay fallas en texto e imagen, congruencia con las convenciones psicométricas y contexto y adecuación cultural. Asimismo, los reactivos 5 y 15 presentan otros tipos de errores que no estaban contemplados en el formato de evaluación de reactivos y que fueron especificados por los jueces en el formato de evaluación de ítems.

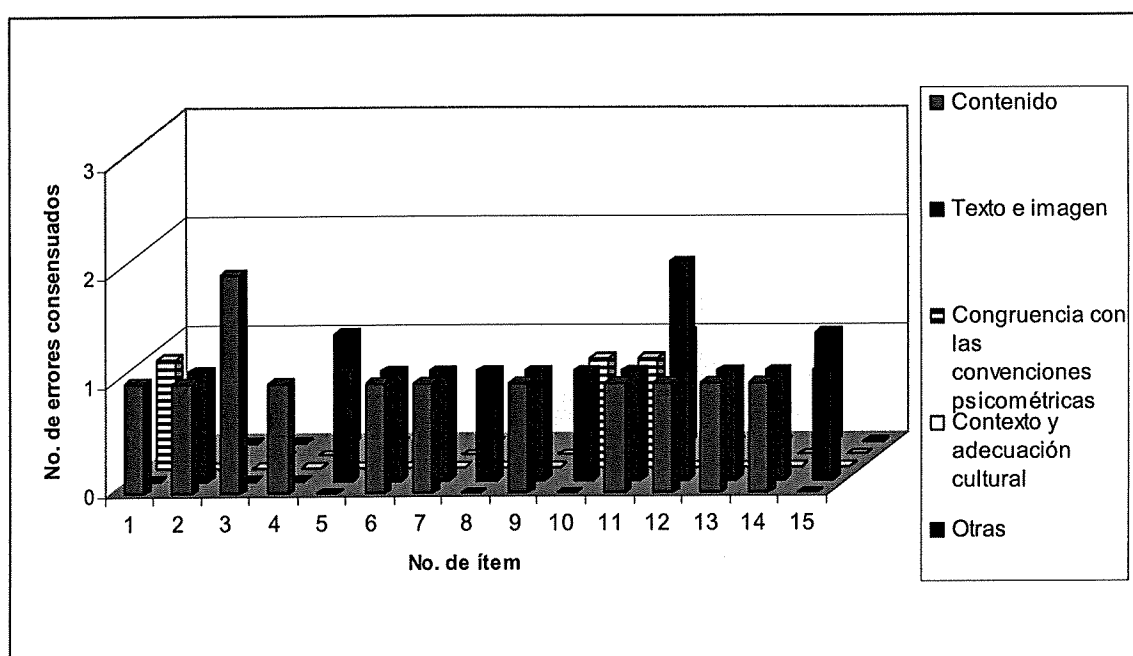


Fig 13 Tipo de errores por área consensuados en cada ítem

4.3 Evaluación de la sesión de capacitación y asesoría

Ahora bien, las respuestas emitidas por los participantes en el cuestionario para evaluar la sesión de capacitación y asesoría, se clasificaron de acuerdo a su frecuencia (ver la tabla No. 12 y el anexo No. 5). De acuerdo con la información aportada por los participantes, resulta claro que la capacitación, asesoría y evaluación fueron adecuadas, ya que en las 28 preguntas expresadas con un carácter positivo, prevaleció el **siempre** en un 96.42% de los casos, el **casi siempre** lo expresaron un 35.71% de los participantes, el **algunas veces** un 3.57% de ellos y el **nunca** no se utilizó en estos cuestionamientos. Por su parte, en las 4 indagaciones redactadas con un carácter negativo, el **nunca** fue señalado por el 75% de los participantes, el **algunas veces** por el 50% de ellos, el casi siempre no se registró y el siempre en un 25%.

De conformidad con el procedimiento correspondiente descrito en el método, se obtuvieron los resultados sobre las opiniones de los participantes referidas a los indicadores y subindicadores del cuestionario para evaluar la operación del módulo.

En la tabla No. 5 se presentan las observaciones al manual técnico del participante, a la operación de la capacitación, a la operación de la evaluación de los ítems y a los resultados de la evaluación de los mismos. Por su parte, las respuestas a las preguntas abiertas correspondientes a los reactivos 27 al 32 del cuestionario, también corroboran que la capacitación fue adecuada; es decir, se lograron los objetivos planteados para la capacitación y además ésta apoyó en la evaluación de reactivos, tal como puede observarse en la tabla No. 6.

Tabla No. 5 Respuestas al cuestionario para evaluar la operación del módulo.

No.	Preguntas	Nunca	Algunas veces	Casi Siempre	Siempre
1	El instructor presentó conceptos y procedimientos que resultaron claros				100%
2	Los conceptos y procedimientos incluidos en el manual facilitaron la tarea de evaluación de ítems mediante jueceo.				100%
3	En general, los conceptos y procedimientos descritos resultaron complejos.	75%	25%		
4	El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.	75%	25%		
5	Los enunciados del protocolo permiten identificar errores en los ítems.			33%	67%
6	Los ejemplos seleccionados fueron claros.			33%	67%
7	Los ejemplos proporcionados fueron significativos.				100%
8	Las diapositivas en la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.				100%
9	La introducción del manual del participante facilitó la comprensión de su contenido.			25%	75%
10	En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.				100%
11	Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.	100%			
12	Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.			25%	75%
13	Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.				100%
14	El instructor demostró dominio del tema.				100%
15	Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.				100%
16	El instructor utilizó vocabulario que fue innecesariamente complejo.				100%
17	La exposición del instructor fue clara.				100%
18	Las actividades realizadas permitieron que lograra un aprendizaje significativo.				100%
19	Durante la capacitación se resolvieron sus dudas.			25%	100%
20	La normatividad sobre la confidencialidad quedó explicada claramente.				100%
21	El instructor tuvo coherencia en sus fundamentaciones.				100%
22	Quedó una idea clara sobre la intención de la capacitación.				100%
23	Se lograron los objetivos propuestos al inicio de la instrucción.			67%	33%
24	El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.		25%	75%	
25	La sesión de evaluación se condujo de manera que siempre se logró el consenso.			25%	75%
26	Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).			25%	75%
27	¿Se comprendió lo que se realizaría en la sesión de evaluación?				100%
28	Después del trabajo desarrollado, ¿considera que se mejoraron los ítems?				100%
29	¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?				100%
30	¿Considera que valió la pena la revisión de los ítems por los jueces? ¿Por qué?				100%
31	¿Los resultados obtenidos lo convencer de la importancia de la validación de ítems?				100%
32	¿La utilización del protocolo produjo los resultados esperados?			25%	75%

Capítulo 4. Resultados

Tabla No. 6 Respuestas de los jueces a las preguntas abiertas del cuestionario para evaluar la operación del módulo.

No.	Preguntas	Respuestas
27	¿Se comprendió lo que se realizaría en la sesión de evaluación?	a. Sí. b. Sí, la instructora fue lo suficientemente clara y precisa para plantear los objetivos de la sesión. c. Sí. d. Sí, había que realizar una evaluación individual de los reactivos y luego otra grupal.
28	Después del trabajo desarrollado, ¿considera que se mejoraron los ítems?	a. Así es, la mejora fue tanto en su redacción como en sugerencias de presentación, por ejemplo en gráficas, y también en contenido. b. Sí, se notaron muchos detalles que antes no nos fueron tan obvios. c. Sí. d. Sí, mejoraron en cuando a su redacción, dificultad y nivel cognitivo.
29	¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?	a. Sí, en el proceso. b. Sí, claro. c. Sí. d. Sí, porque a pesar de que ya habían realizado una revisión informal, no fue tan completa como la de esta sesión. La capacitación ayudó a que los profesores comprendieran la importancia de la revisión de ítems, así como a identificar posibles errores en éstos.
30	¿Considera que valió la pena la revisión de los ítems por los jueces? ¿Por qué?	a. Sí. b. Porque todo esto enriquece nuestro trabajo; se pudieron detectar algunos reactivos que no son claros e incluso errores tanto de redacción como de concepto. c. Valió la pena en cuanto a todo lo que se alcanzó a revisar. d. Sí, porque la integración del comité ayudó a que distintos especialistas colaboraran, discutieran y llegaran a un consenso sobre los reactivos a revisar.
31	¿Los resultados obtenidos lo convencen de la importancia de la validación de ítems?	a. Totalmente. El taller sube el nivel de conciencia sobre la instrucción y mejora los reactivos. b. Sí. c. Sí convencen nada más faltó tiempo de revisar ítems. d. Sí, la validación de ítems mediante jueces es un procedimiento importante y necesario para mejorar la calidad técnica de los reactivos, evitar el sesgo, la discriminación y asegurar que el examen evalúe de manera justa a los examinados.
32	¿La utilización del protocolo produjo los resultados esperados?	a. El protocolo debe llevar un punto que considere el tiempo de ejecución del reactivo. b. Considerablemente, pues nos da la pauta para dicha revisión. c. Sí. d. Sí, los jueces nos sentimos satisfechos con el trabajo realizado ya que se identificaron errores en los ítems y se hicieron propuestas para mejorarlos.

4.3.1 Manual técnico del participante

- **Contenido.** Respecto al contenido del manual técnico del participante, los jueces expresaron en el cuestionario para evaluar la operación del módulo que los conceptos y procedimientos incluidos en el manual facilitaron la tarea de evaluación de ítems mediante operaciones de juicio. Sin embargo, comentaron que algunas veces los conceptos y procedimientos resultaron complejos.
- **Extensión.** De igual modo, añadieron que algunas veces el contenido del manual resultó demasiado extenso para el tiempo en que se revisó.
- **Protocolo.** Para tres jueces, los enunciados del protocolo *siempre* permitieron identificar errores en los ítems; mientras que para uno de ellos, su respuesta fue *casi siempre*.
- **Ejemplos.** Para tres jueces, los ejemplos seleccionados siempre fueron claros; mientras que para uno de ellos, su respuesta fue *casi siempre*. No obstante, todos declararon que los ejemplos proporcionados fueron significativos.
- **Presentación de Power Point.** Al respecto, los jueces expresaron que las diapositivas de la presentación en Power Point *siempre* los apoyaron la comprensión de los conceptos y procedimientos descritos.
- **Estructura.** Sobre la *introducción* del manual, dos jueces consideraron que ese apartado *siempre* facilitó la comprensión de su contenido; a diferencia de uno de ellos quien expresó que *casi siempre* recibió ese apoyo de la introducción.
- Respecto a la *metodología* que presenta el manual del participante, los cuatro jueces dijeron que los procedimientos a realizar se explicaron siempre con la profundidad *suficiente*. Finalmente, sobre los *ejercicios* evaluativos, los participantes opinaron que siempre fueron *suficientes* para comprender la tarea a realizar; sin embargo, solo tres jueces los valoraron como un apoyo permanente en la comprensión de conceptos y procedimientos; a diferencia de un juez que opinó que *casi siempre* recibió ese apoyo.

- **Sugerencias.** Los participantes sugirieron que el manual se entregara antes de la capacitación para que tuvieran más tiempo de revisarlo y de comentar las dudas cuando se trabajara con la presentación de Power Point.

4.3.2 Operación de la capacitación

- **Instructor.** Según los participantes, la instructora siempre demostró dominio del tema. Respecto a la resolución de dudas de manera clara, dos jueces expresaron que la instructora siempre lo hizo; no obstante, uno de los jueces afirmó que el instructor no resolvió de manera clara una duda en cuestión. Por último, dos jueces consideraron que la instructora no utilizó un vocabulario innecesariamente complejo; sin embargo un juez afirmó que sí lo hizo. Respecto a la claridad, los jueces comentaron que la exposición siempre fue clara. A su vez, añadieron que la instructora siempre tuvo coherencia en sus fundamentaciones.
- **Inducción.** Los participantes afirmaron que la idea sobre la intención de la capacitación quedó clara.
- **Desarrollo.** Los jueces opinaron que las actividades realizadas permitieron que se lograra un aprendizaje significativo. Sobre las dudas que surgieron en el desarrollo, dos jueces expresaron que siempre se resolvieron; mientras que uno comentó que casi siempre.
- **Cierre.** En esta parte, los jueces opinaron que se lograron los objetivos de la capacitación propuestos al inicio de la instrucción.
- **Tiempo destinado a la capacitación.** Dos jueces estimaron que casi siempre el tiempo fue suficiente para realizar adecuadamente la evaluación de los ítems; no obstante, el tercer juez opinó que fue suficiente solo algunas veces.

4.3.3 Operación de la evaluación de ítems

Todos los jueces afirmaron haber **comprendido** lo que se realizaría en la sesión de evaluación de reactivos; como lo expresa el juez 2, la instructora fue lo

suficientemente clara y precisa para plantear los objetivos de la sesión. Respecto a la **conducción** de la sesión de evaluación, dos jueces opinaron que siempre se logró el consenso, mientras que el tercer juez comentó que casi siempre se llegó al acuerdo.

Sobre los **tiempos asignados** para realizar las actividades de la sesión de evaluación, dos jueces dijeron que casi siempre se respetó el tiempo asignado y un tercer juez expresó que algunas veces se respetó este tiempo.

4.3.4 Resultados de la evaluación de los ítems por el comité

- **Productos obtenidos.** Una vez terminada la revisión individual y grupal de los 15 reactivos, los jueces consideraron que se mejoraron los ítems. El juez 2 expresó que la mejora fue tanto en su redacción como en sugerencias de presentación de gráficas y de contenido. Por su parte, el juez 3 añadió que con la revisión de reactivos se percataron de muchos detalles que antes no les habían sido tan obvios.
- **Importancia del trabajo.** Los participantes expresaron que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente; que valió la pena la revisión de los ítems por los jueces. Los participantes expresaron que las **actividades realizadas** enriquecen el nivel del examen, ya que se pudieron detectar algunos reactivos que no son claros e incluso errores tanto de redacción como de concepto. Además, comentaron que sí valió la pena en cuanto a todo lo que se alcanzó a revisar.
- Para los jueces, los **resultados obtenidos** los convencieron de la importancia de la validación de ítems; expresaron que el taller subía el nivel de conciencia sobre la instrucción y mejora los reactivos; sin embargo comentaron que faltó más tiempo para revisar todos los ítems de la prueba.
- Finalmente, los jueces expresaron que la utilización del **protocolo** produjo los resultados esperados, porque les da la pauta para la revisión de los ítems. Sugirieron que se le agregue otro enunciado en donde se considere el tiempo de ejecución del reactivo.

4.4 Ítem análisis

Como evidencia adicional, en este apartado se presenta el ítem análisis de los resultados del pilotaje del examen colegiado de Matemáticas II, mismo que fue aplicado en octubre de 2006 a 344 alumnos del segundo semestre de la Facultad de Ingeniería de Mexicali de la UABC (ver tabla No. 7).

Tabla No. 7 Comparación entre los resultados del ítem análisis de los 15 reactivos revisados y modificados y los 17 que no fueron revisados

Revisados (k=15) No revisados (k=17)	SOFTWARE ELABORADO EN EL IIDE-UABC				ITEMAN			
	Reactivos revisados		Reactivos no revisados		Reactivos revisados		Reactivos no revisados	
	P	D	P	D	p	rbis	p	rbis
1	.63	.53	.38	.39	.63	.49	.38	.33
2	.46	.41	.42	.44	.46	.46	.42	.38
3	.27	.44	.35	.16	.26	.42	.35	.15
4	.55	.60	.47	.42	.55	.57	.47	.39
5	.42	.61	.45	.30	.42	.58	.45	.26
6	.29	.56	.62	.30	.29	.55	.62	.27
7	.25	.31	.18	.23	.25	.29	.18	.19
8	.22	.40	.55	.32	.22	.38	.55	.26
9	.43	.35	.46	.47	.43	.35	.46	.45
10	.76	.32	.42	.35	.76	.28	.42	.32
11	.73	.31	.43	.35	.73	.31	.43	.28
12	.61	.44	.26	.33	.61	.41	.26	.35
13	.57	.30	.42	.51	.57	.29	.42	.52
14	.36	.44	.26	.25	.36	.43	.26	.21
15	.39	.40	.24	.29	.39	.40	.24	.27
16			.37	.37			.37	.33
17			.23	.20			.23	.22
Media	.46	.43	.38	.33	.46	.37	.38	.29

P= Índice de dificultad

D= Índice de discriminación

r_{bis}= Coeficiente de discriminación

No satisface el criterio

Las respuestas capturadas por el cuerpo colegiado de la Facultad de Ingeniería de la UABC de Mexicali presentan información sobre los 15 reactivos que se revisaron y modificaron para mejorarse, con base en las observaciones realizadas durante la sesión de evaluación de ítems en mayo de 2006, así como de

los otros 17 reactivos de la prueba, los cuales no fueron revisados mediante jueces ni se modificaron.

Con esta información se procedió a realizar el análisis de los ítems mediante dos programas: un *software* desarrollado en el IIDE-UABC y el ITEMAN versión 3.50 (Assessment Systems Corporation, 1993). El *software* del IIDE-UABC calcula el índice de discriminación (**D**) considerando la diferencia de la dificultad (**p**) entre el grupo alto (27%) y el grupo bajo (27%). Por su parte, el ITEMAN realiza una correlación punto biserial (r_{bis}) que determina el coeficiente de discriminación estableciendo una relación entre la respuesta de cada ítem y las puntuaciones obtenidas por todos los examinados en la prueba (ver anexos No. 10 y No. 11). Debido a esta característica, se le considera un método más duro al incluir a toda la población examinada.

Como puede observarse en la tabla No. 6, existe una ligera diferencia entre los valores que arroja el *software* del IIDE-UABC y el ITEMAN, debido a la población que considera cada programa para calcular el coeficiente de discriminación. Para efectos de este análisis, se consideran solamente los datos proporcionados por el ITEMAN, debido a que es una herramienta más precisa.

El panorama general de la tabla No. 6 presenta a dos conjuntos de reactivos con buenas propiedades psicométricas, a excepción de los ítems No. 18 y No. 22 del grupo de 17 reactivos, los cuales carecen de discriminación apropiada porque no satisfacen el criterio de calidad técnica de ítems (Nitko, 1994) ya que el valor del coeficiente de discriminación **D** es menor que .20.

4.4.1 Dificultad

Examinando la dificultad, como señala Nitko (1994), para el caso de una prueba criterial se considera apropiada si en cada ítem **p** está entre .05 y .95, requisito que cumplen los 32 reactivos. En la tabla No. 6 se observa que existen diferencias en las medias de dificultad de los dos grupos de ítems; así el conjunto de 15 reactivos tiene una dificultad media de 0.46, mientras el otro grupo tiene una media de 0.38, lo cual es una diferencia de 8 centésimas. De ahí que los 15 reactivos modificados son contestados correctamente por un mayor número de examinados, aunque el hecho de que el grupo de los 17 reactivos sea más difícil no

es un factor determinante para decir que alguno de los conjuntos discrimina mejor. Más bien, estos resultados pueden interpretarse como producto del progresivo incremento de dificultad que caracteriza a los contenidos de matemáticas, mismos que generalmente son requisitos unos con respecto a otros.

4.4.2 Discriminación

El estándar propuesto por autores como Nitko señala que para que un ítem discrimine de manera apropiada su valor D debe ser mayor o igual a .20.

Como puede observarse en la tabla No. 6, en el grupo de los 15 reactivos revisados los que tienen un coeficiente de discriminación (r_{bis}) de 0.40 y superior son los número 1, 2, 3, 4, 5, 6, 12, 14 y 15; esto es, el 60% de los ítems evaluados tienen una discriminación muy buena.

Por otra parte, dentro del mismo grupo, los reactivos 14 y 15 tienen una discriminación de .36 y .39, que representan el 13.33% del conjunto, lo cual permite afirmar que son razonablemente buenos, pero que pueden mejorarse. Al igual que los reactivos 7, 10 y 13 cuyo valor r_{bis} es .29, .28 y .29, respectivamente, lo cual permite ponerles una bandera de alerta, porque a pesar de ubicarse dentro del estándar, su valor D indica que necesitan mejorarse.

En contraste, en el grupo de los 17 reactivos, los ítems 24 y 28 tiene muy buena discriminación (.45 y .52), representando el 11.76% del conjunto. Además, el 35.29% de los ítems, específicamente el 16, 17, 19, 25, 27 y 31, tienen una discriminación que está entre .33 y .38, por lo que son considerados razonablemente buenos pero pueden mejorarse.

Por el contrario, los reactivos 18 y 22 no cumplen con el criterio de calidad técnica debido a que la discriminación es menor que .20 (ver tabla No. 7). Debido a esto, éstos dos ítem requieren mejorarse mediante revisión; o bien, eliminarse de la prueba.

La media del coeficiente de discriminación de los reactivos modificados (.37) indica que los ítems son razonablemente buenos, mientras que el valor medio de los 17 reactivos (.29) manifiesta que los ítems son marginales y necesitan mejorarse; lo anterior con base en la ejecución de los 344 estudiantes en un ítem y el resultado en el examen.

Como se ha hecho notar, los resultados del ítem análisis presentan diferencias en los atributos de dificultad y discriminación de los dos subconjuntos de reactivos. En términos de dificultad, el conjunto de los 15 reactivos revisados fue contestado mejor por los examinados y es ligeramente más fácil que el grupo de 17 reactivos. De la misma manera, con base en la ejecución de los 344 estudiantes en un ítem y en el examen, los reactivos revisados discriminan mejor entre los alumnos que dominan y los que no dominan los contenidos del examen de Cálculo integral II; esto lo indica el valor medio del coeficiente de discriminación biserial puntual, que permite clasificarlos como ítems muy buenos.

Por último, aunque no fue un objetivo específico del estudio, se realizó una primera evaluación de naturaleza cualitativa sobre los méritos del módulo con base en las características psicométricas de los reactivos. Como puede apreciarse en la figura No. 6, el análisis de los ítems arrojó diferencias pequeñas en los valores p y R_{bis} en los reactivos revisados y los no revisados. En ese sentido, se identificaron tres reactivos de los 17 que no fueron evaluados por el comité, mismos que no satisfacen el criterio de calidad técnica de ítems ya que el valor del coeficiente de discriminación D fue menor que .20. El hecho de que las diferencias en los valores p y R_{bis} sean pequeñas, puede estar relacionado con la bondad de la metodología empleada para desarrollar toda la prueba de cálculo integral. No obstante, como ya se comentó, la mejora en la calidad de los primeros 15 reactivos puede ser atribuida a las operaciones de juicio realizadas, según lo expresaron los miembros del comité durante la sesión y en las respuestas que dieron en el cuestionario de evaluación del módulo. De la misma manera, aunque la evidencia que se presenta en el análisis de reactivos no es proporcionada por las opiniones de los jueces, sino que fue aportada por los 344 examinados que respondieron la prueba, confirma que la operación del módulo logró resultados positivos para mejorar la calidad de los reactivos.

Capítulo 5. Conclusiones y recomendaciones

En este apartado, se retoman la parte metódica empleada para desarrollar el módulo, los resultados obtenidos en la operación y evaluación de dicho módulo y la relación de ambos aspectos con las nociones conceptuales propuestas; todo ello a la luz de los objetivos que orientaron el estudio. Además, a partir de las experiencias obtenidas al operar el módulo se formulan algunas sugerencias para mejorarlo.

Para lograr lo anterior, a continuación se presentan los puntos más relevantes del estudio desarrollado:

En el capítulo 1, denominado *Introducción*, se propuso como objetivo general el diseñar, operar y evaluar un módulo para capacitar y asesorar a los miembros de un comité, para validar mediante jueceo ítems empleados en la evaluación criterial del aprendizaje a gran escala. Como se recordará, para lograr ese propósito se fijaron los siguientes objetivos específicos:

- Establecer áreas a explorar del reactivo.
- Establecer criterios para la selección de comités evaluadores.
- Diseñar y probar procedimientos, instrumentos y materiales para capacitar y asesorar a los comités evaluadores de ítems.

a) En cuanto al objetivo específico de establecer las áreas a explorar del reactivo, su logro se documenta en los capítulos 2 al 4, los cuales exponen con detalle el desarrollo de las actividades enfocadas en la selección de las áreas a valorar en los ítems.

Así, en el capítulo 2, se exponen las sugerencias internacionales para la elaboración, evaluación y uso de pruebas, así como los modelos para validar ítems de exámenes a gran escala, mismas que fueron consideradas al establecer las áreas a explorar del reactivo. Como resultado, en ese apartado se determinó que serían tres los aspectos a evaluar en los ítems: congruencia curricular (Nitko, 1994), elementos socioculturales (Sexton y Solano-Flores, 1999) y estándares psicométricos (Joint Committee on Testing Practices, 2005).

En este punto, cabe señalar que se diseñó un protocolo para detectar errores en los ítems correspondientes a tales áreas. Por lo que, una vez establecidas en el

capítulo 2 las áreas a explorar de los reactivos, el capítulo 3 describe el procedimiento desarrollado para el diseño del formato de evaluación de reactivos. Ahí, se menciona que dicho protocolo recupera las principales aportaciones teóricas hechas por los psicómetras que recomiendan la validación de ítems mediante juicios (ver 3.2.2 y 3.3.2).

Seguidamente, en el capítulo 4 se muestran los resultados obtenidos al operar el formato; es decir, los errores que identificaron los jueces en los 15 ítems revisados y los comentarios que, en su caso, expresaron al respecto (ver figuras No. 5 a la No. 10 y anexos No. 5 y No. 6). Además, en ese capítulo se narran las observaciones que los jueces hicieron sobre la claridad y la efectividad del formato empleado para evaluar los ítems. De manera específica, tres de cuatro jueces declararon que el formato es claro y que permite identificar errores en los reactivos.

b) Por su parte, el segundo objetivo específico del estudio, señala el establecimiento de criterios para la selección de comités evaluadores. Para lograr esta meta, dentro de la revisión de los modelos conceptuales que se presentó en el capítulo 2, para validar ítems de pruebas a gran escala, se comentan las experiencias que algunos psicómetras han tenido al integrar comités para validar ítems mediante juicios (Popham (1991), Michener y Kesselman (1986), Title (1990), Childs (1990), Zimmerman, Sudweeks, Shelley y Wood (1990), Hambleton (1993), Sireci (1995a), Sireci y Geisinger (1995b), Solano-Flores (2000, 2001 y 2002; Sexton y Solano-Flores, 1999 y 2002; Nelson-Barber y Solano-Flores, 2000), Nitko (1994), así como de los españoles Suárez y Jornet (1990a y 1990b).

Antes de operar el módulo, se consideró la posibilidad de variar la integración del comité evaluador de reactivos, con el fin de tener evidencias de cada arreglo y estimar cuál conformación producía mejores resultados. Sin embargo, debido a limitantes de recursos para este proyecto y en búsqueda de una propuesta de integración de un comité que considerara al personal necesario para que pueda ser retomado en otros estudios, solo se trabajó con una modalidad, en la cual participaron un profesor de español, dos profesores especialistas en Matemáticas II, dos profesores que imparten la asignatura y son reconocidos dentro de la Facultad de Ingeniería de la UABC y un psicómetra. Para cumplir con lo recomendado para

detectar sesgo de género se consideró la incorporación al comité tanto de hombres como mujeres.

Como resultado de la operación del comité de jueces, se logró la revisión de una muestra de 15 ítems del examen de Cálculo Integral II, a pesar de que dos jueces abandonaron la sesión y uno más se unió al equipo después de impartida la capacitación. La interacción entre los jueces y el proceso de toma de decisiones que adoptaron se ilustran en el capítulo 4 (ver 4.1). Ahí se narran las dudas y comentarios surgidos durante la capacitación, la revisión de ítems y la evaluación de la sesión.

Para concluir esta sección, podemos decir que a pesar de que no se identificó algún sesgo de género en los ítems, el comité estuvo de acuerdo en una integración diferenciada y expresaron que esta experiencia los sensibilizó en ese aspecto.

c) Por otra parte, el tercer objetivo específico de diseñar y probar procedimientos para capacitar y asesorar a los comités evaluadores se cumplió, tal como se describe en la parte metódica (ver 3.3.3). Las temáticas desarrolladas en la capacitación fueron: una introducción para contextualizar a los participantes; la revisión de aspectos conceptuales relativos a la validación de reactivos de pruebas para evaluar el aprendizaje de gran escala; la explicación de la validación de reactivos; y por último, el cierre de la sesión que consistió en evaluar las actividades realizadas.

Una representación gráfica de la capacitación puede observarse en la figura No. 2, en donde se visualizan los cuatro grandes momentos anteriormente descritos, así como las nociones conceptuales revisadas en cada etapa. De hecho, se considera que tal figura puede entregarse a los participantes antes de iniciar la operación del módulo, para que los jueces tengan una guía conceptual y procedimental de apoyo.

Seguidamente, el capítulo 3 presenta los materiales e instrumentos empleados, así como el procedimiento para desarrollar el material didáctico que apoyó el proceso de enseñanza aprendizaje durante la sesión de capacitación y revisión de ítems.

Los materiales elaborados fueron: el manual técnico del participante (ver anexo No. 2), el formato de evaluación de reactivos (figura No. 1), la presentación en Power Point (ver anexo No. 4) y el cuestionario para evaluar la operación del módulo (ver anexo No. 3). Por su parte, se emplearon como instrumentos una cámara y videos para filmar la operación del módulo, teniendo así un registro de lo acontecido durante la sesión.

En primer lugar, las actividades del comité se explicaron en el procedimiento de la capacitación. Ahí se estableció que los jueces debían recibir la capacitación para posteriormente realizar una revisión individual y otra grupal. En la revisión individual, su actividad consistió en contestar los ítems como si fueran los examinados. Después, apoyándose en el formato de evaluación de reactivos, evaluaron al ítem en los aspectos sociocultural, congruencia curricular y apego a las convenciones psicométricas. Luego, efectuaron la revisión grupal en donde se discutió reactivo por reactivo, retomando los errores identificados de manera individual y su justificación con el propósito de lograr un consenso para cada reactivo (ver figuras No. 5 a la No. 10). De esta manera fue posible elaborar una propuesta de las modificaciones necesarias que también consideraron la sugerencia de eliminar un ítem de la prueba (ver anexo No. 7).

La última actividad que realizaron los jueces fue valorar las actividades realizadas, los materiales empleados y los resultados obtenidos. El instrumento que emplearon para tal efecto fue el cuestionario para evaluar la operación del módulo (ver anexo No. 3), diseñado con base en las características del módulo, siguiendo el esquema metodológico de la figura No. 4. Como se mencionó en el capítulo 4 (ver sección 4.3), las respuestas de los jueces al cuestionario de evaluación indican que la capacitación, los materiales empleados y la sesión de revisión de ítems fueron adecuados, sin embargo hay algunas áreas de oportunidad que pueden mejorarse, tal como se aprecia en la tabla No. 5, específicamente el manual del participante y la extensión de la sesión de revisión de reactivos.

En cuanto al diseño del manual del participante para apoyar la capacitación, asesoría y evaluación de ítems, el procedimiento desarrollado para su elaboración

se describe en el capítulo 3 (ver sección 3.3.3) y los comentarios que los jueces emitieron sobre este material se presentan en la tabla No. 5.

Seguidamente, el capítulo 4 presenta las opiniones de los jueces sobre el manual técnico del participante respecto a su contenido, extensión y estructura. Es en esta sección donde los jueces sugieren que el manual técnico del participante se entregue antes de la capacitación para que así puedan revisarlo con más tiempo y llegar contextualizados a la actividad de evaluación.

Ahora bien, otro material que fue importante durante la capacitación fue la presentación en Power Point (ver anexo No. 4), la cual constó de 13 diapositivas que contienen los puntos más relevantes de las temáticas que se manejaron en la capacitación. En el capítulo 4, se presentan los comentarios que los jueces hicieron sobre este material en el cuestionario para evaluar la operación de módulo; coincidieron en que las diapositivas ayudaron en la comprensión de los conceptos y procedimientos descritos por la instructora.

De la misma manera, se observó y tomó registro mediante videograbación de las sesiones de capacitación y evaluación de reactivos. En el capítulo 3 se describen la cámara y los videos como instrumentos para recabar información, mientras que en el capítulo 4 se presenta una síntesis de los comentarios hechos por los jueces y de la discusión grupal de los reactivos (ver sección 4.1 y anexo No. 8).

Con base en esos resultados, se establecieron procedimientos analíticos para codificar y analizar la información. Como ya se señaló, en vez de realizar una transcripción completa de los videos, estos fueron vistos en tres ocasiones y en cada una de ellas se tomó nota de las principales observaciones, dudas y del proceso para llegar al consenso en la revisión grupal (ver Anexos No. 8 y No. 9). Para el caso del cuestionario de evaluación del módulo, se retomaron los factores, indicadores y subindicadores para presentar primero de manera gráfica y luego en forma narrativa, los resultados obtenidos.

Haciendo un balance general, podemos decir que el objetivo general propuesto en el presente trabajo se cumplió satisfactoriamente: El módulo fue diseñado de

conformidad con la normatividad internacional desarrollada para la evaluación de ítems mediante juicios y su elaboración estuvo apegada a los lineamientos técnicos propuestos por los principales modelos conceptuales y metodológicos desarrollados por especialistas que se han destacado en este campo; Además, se desarrollaron los materiales y procedimientos necesarios a fin de que el módulo fuera operado de manera apropiada. Finalmente, se elaboraron y aplicaron los instrumentos y procedimientos necesarios para obtener datos que permitieron evaluar los puntos fuertes y debilidades del módulo para capacitar y asesorar a los miembros de un comité, para validar mediante jueceo ítems empleados en la evaluación criterial del aprendizaje a gran escala.

Recomendaciones

Con base en el análisis de los resultados obtenidos y en la reflexión sobre el conjunto de experiencias obtenidas a lo largo del desarrollo del trabajo, se proponen las siguientes recomendaciones:

- **Operación del módulo.** Respecto a la operación del módulo, se sugiere que se asigne más tiempo. Para ello se estima un mínimo de dos sesiones con 8 horas diarias de actividad. También se recomienda que durante la revisión individual se trabaje con un máximo de 5 reactivos al mismo tiempo; esto con el fin de que los evaluadores puedan registrar los comentarios de inmediato en el formato de evaluación, y así no se olvide información y se ahorre tiempo.
- **Materiales e instrumentos.** Los participantes sugirieron que el manual técnico del participante se entregue a los miembros del comité por lo menos una semana antes de que se opere el módulo, a fin de que tengan tiempo de leerlo detenidamente, revisar el glosario, hacer observaciones y plantear las dudas correspondientes durante la exposición de la presentación. Además, propusieron que el mismo manual incluya el formato de evaluación de reactivos con la explicación de cada enunciado y que se incluyan más ejemplos para facilitar la comprensión e identificar rápidamente un error similar. Ahí mismo, se propone también que se agregue una columna al formato de evaluación que diga “el ítem está bien”, considerando que por omisión no hay error.
- **Otras investigaciones.** Se proponen cuatro recomendaciones para futuros trabajos de investigación relacionados:
 - Revisar el total de reactivos de una prueba, hacer las correcciones sugeridas por los jueces y pilotear en dos grupos el modelo corregido y otro sin modificaciones. Ello con el objetivo de tener más reactivos al momento de comparar los resultados del examen, los resultados del ítem análisis y observar otras posibles diferencias entre los exámenes.
 - Diseñar y probar diferentes arreglos para la integración y funcionamiento de comités evaluadores de ítems. En el presente

estudio se probó un solo arreglo; sin embargo, dos o tres configuraciones de comités diferentes podrían proporcionar información adicional sobre la integración ideal, en términos de eficacia y eficiencia, de un comité para validar ítems mediante juicios. Se recomienda que los docentes que participen en los arreglos tengan amplia experiencia y dominio del contenido de los reactivos.

- Profundizar en el establecimiento de criterios de decisión para determinar la validez de juicios sobre los ítems de una prueba, emitidos por los expertos. Como se mencionó, en este trabajo de investigación se aceptó como criterio de validez la decisión tomada en consenso; no obstante, el considerar otros criterios podría aportar información valiosa para la revisión de ítems mediante juicios.

Bibliografía

Bibliografía

- Angoff, W. (1989). Context bias in the test of English as a Foreign Language. (ERIC No. ED393941).
- APA (1999). *Standards for educational and psychological testing*. Prepared by a joint committee of the American Psychological Association, American Educational Research Association, National Council on Measurement in Education. Washington, D.C.: American Psychological Association pp. 1-169.
- Assessment Systems Corporation (1993). Item and Test Analysis Program: ITEMAN. Versión 3.50. Microcat Testing System.
- Backhoff, E. (2005). Mesa Pública de Análisis sobre Evaluación Educativa. *Exámenes de la Calidad y el Logro Educativo (Excale), La Nueva Generación de Pruebas del INEE*. Consultado el 16 de agosto de 2005 en: <http://multimedia.ilce.edu.mx/inee/pdf/mesaexcale.pdf>
- Bothell, T. (2001). 14 rules for writing multiple-choice questions. Birmingham Young University.
- Contreras, L.A. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California*. Tesis de maestría no publicada. Universidad Autónoma de Baja California. Ensenada, B. C. México.
- Contreras, L.A. y Hurtado, T. (2004). Colaboración interinstitucional para elaborar pruebas criterios de gran escala: el caso de los exámenes sumarios del Colegio de Bachilleres de Baja California. *Memoria del VI Foro de Evaluación Educativa*. México. Centro Nacional de Evaluación para la Educación Superior.
- Contreras, L.A., Encinas, J.A. y De Las Fuentes, B.M. (2005). *Evaluación Colegiada del Aprendizaje en la Universidad Autónoma de Baja California: el caso del examen de Matemáticas I de la Facultad de Ingeniería Mexicali*. Memoria del VIII Congreso Nacional de Investigación Educativa. COMIE.
- Cole, N. y Moss, P. (1993). *Bias in test use*. En Linn, Educational measurement (3ª ed.). Oryx Press: Nueva York. Pp. 201-217.
- Childs, R. (1990). Gender Bias and Fairness (ERIC No. ED328610).
- Educational Testing Service (2003). Fairness review guidelines. Consultado el 14 de abril de 2006 en: http://www.ets.org/Media/About_ETTS/pdf/overview.pdf

Bibliografía

- Excelencia Laboral, S. C. (2004). *Normas técnicas de competencia laboral. Consejo de normalización y certificación de competencia laboral*. Consultado el 12 de mayo de 2005 en: <http://www.excela.com.mx/imagenes/Pdf/CRCH0542.01.pdf>
- Haladyna, T.; Haladyna, R.; Merino, C. (2002). Preparación de preguntas de opciones múltiples para medir el aprendizaje de los estudiantes. OEI-Revista Iberoamericana de Educación. Consultado el 12 de marzo de 2006 en el sitio web: <http://www.campus-oei.org/revista/deloslectores/267Haladyna.PDF>
- Hambleton, R. (1993). Advances in the Detection of Differentially Functioning Test Items. *Laboratory of Psychometric and Evaluative Research Report No. 237*. University of Massachusetts, School of Education: Amherst, MA. Pp. 25-33.
- Mullis, I. et. al. (2002). Marcos teóricos y especificaciones de evaluación de TIMSS 2003. Autor: Madrid. Consultado el 1 de abril de 2006 en: <http://www.ince.mec.es/pub/marcosteoricostimss2003.pdf>
- Instituto Nacional para la Evaluación de la Educación (2002). *Estándares de calidad para instrumentos de evaluación educativa*. México: INEE.
- Instituto Nacional para la Evaluación de la Educación (2005). *Validación de reactivos para la construcción de los exámenes EXCALE que fueron aplicados en mayo de 2005 a una muestra de egresados de la educación primaria y secundaria del país, para evaluar los conocimientos del Español [DVD]*.
- Joint Committee on Testing Practices (2005). *Code of Fair Testing Practices in Education*. American Psychological Association. Consultado el 12 de mayo de 2005 en: <http://www.apa.org/science/fairtestcode.html>
- L'Allier, J., Miles, K. y Welsh, S. (1995). *The NETg Mastery Test strategy*. Naperville, IL: NETg.
- Michener, S.; Kesselman, G. (1986). Content-related validation techniques applied to training evaluation. (ERIC No. ED282097).
- Nelson-Barber, S. y Solano-Flores, G. (2000). *Assessing the cultural validity of Science and Mathematics assessments*. Poster Summary, National Science Foundation REC Research PI Meeting, Arlington, Virginia, January 11-12.
- Nitko, A. J. (1994). *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).

Bibliografía

- Popham, J. (1991). *Modern Educational Measurement: a practitioner's perspective* (2a. ed.). Allyn and Bacon: Needham Heights, MA.
- Title, C. K. (1990). *Test bias*. En Walberg, H. J. y Haertel, G. D. *The International Encyclopedia of Educational Evaluation*. Great Britain: Pergamon Press LTD.
- Ruiz, G. y Martínez, F. (1996). Evaluación de niveles de aprendizaje en dos grados de la educación primaria y en cuatro contextos socioeconómicos del estado de Aguascalientes. Ponencia presentada en II Foro Nacional de Evaluación Educativa. Memorias. Centro Nacional de Evaluación para la Educación Superior
- Secretaría de Educación Pública (2001). *Programa Nacional de Educación 2001-2006*. Primera edición, septiembre de 2001. México pp.121-184.
- Sexton, U. y Solano-Flores, G. (1999). *A Comparative Study of Teachers' Cultural Perspectives Across Different Cultures: Preliminary Results*. Poster presented at the Annual Meetin of the American Educational Research Association. Seattle, WA. April 10-14.
- Sexton, U. y Solano-Flores, G. (2002). *Cultural validity in assessment: A cross-cultural study on the interpretation of Mathematics and Science test items*. Paper presented at the 2002 meeting of the American Educational Research Association. New Orleans, LA, April 1-5.
- Sireci, S. (1995a). The Central Role of Content Representation in Test Validity. (ERIC No. ED387508).
- Sireci, S. y Geisinger, K. (1995b). Using Subject-Matter Experts to Assess Content Representation: An MDS Analysis. *Applied Psychological Measurement*. Volumen 19 (3), pp. 241-55. (ERIC No. EJ526849).
- Solano-Flores, G. (2002). *Cultural validity: The need for a socio-cultural perspective in Educational Measurement*. Paper presented at the Annual Meeting of the American Educational Research Association. New Orleans, LA, April 1-5, 2002.
- Solano-Flores, G. y Nelson-Barber, S. (2000). *Cultural Validity of Assessments and Assessment Development Procedures*. Paper presented at the 2000 American Educational Research Association Meeting. New Orleans, LA, April 24-28.
- Solano-Flores, G. y Nelson-Barber, S. (2001). *On the Cultural Validity of Science Assesments*. *Journal of Research in Science Teaching*, 38(5), 553-573.
- Solano-Flores, G. Contreras-Niño, L.A y Backhoff, E. (2005). *The Mexican Translation of TIMSS-1995: Lessons on Test Translation From a Post-Mortem Study*. Paper Presented the Annual Meeting of the National Council on Measurement in Education, Montreal, Canada, April 12-14, 2005.

Bibliografía

- Suárez, J. y Jornet, J. et al. (1990a). *Estudio sobre las características métricas de los distractores y del sesgo respecto a algunas variables diferenciales en el apartado de completar frases de la subprueba de vocabulario escrito de la BPL-FCI*. En *Revista de Investigación Educativa*. Vol. 9, no. 16, pp. 611-618.
- Suárez, J. y Jornet, J. et al. (1990b). Algunas notas de reflexión metodológica acerca del estudio de distractores y el sesgo de ítems en tests educativos y psicológicos. En *Revista de Investigación Educativa*, Vol. 8, no. 16, pp. 551-559.
- Title, C. K. (1990). Test bias. En Walberg, H. J. y Haertel, G. D. *The International Encyclopedia of Educational Evaluation*. Great Britain: Pergamon Press LTD.
- Universidad Autónoma de Baja California (2003). *Plan de Desarrollo Institucional 2003-2006* / Universidad Autónoma de Baja California. México: UABC.
- Universidad Autónoma de Baja California (2004). *Convocatoria para el Diseño de Instrumentos de Evaluación Colegiada del Aprendizaje* / Universidad Autónoma de Baja California. México: UABC.
- Zimmerman, B., Sudweeks, R., Shelley, M. y Wood, B. (1990). How to prepare better tests: guidelines for University Faculty. Brigham Young University Testing Services and The Department for Instructional Science.

Anexos

Anexo No. 1

**Reglas propuestas por psicómetras
destacados para la construcción de
reactivos**

Reglas propuestas por Haladyna y Dowling para la construcción de reactivos.

Aspecto:	Reglas:
El contenido del reactivo	• Cada pregunta debe reflejar un contenido específico y un sólo comportamiento mental específico, tal como se requiere en las especificaciones del examen. • Basar cada pregunta en algún punto importante del contenido a aprender; evitar contenido trivial. • Usar material novedoso para medir aprendizaje de un nivel cognitivo superior. Parafrasear el lenguaje usado en el libro de texto o durante la instrucción, para evitar la simple memorización del material. • Relacionar el contenido que evalúa cada pregunta del examen. • Evitar un contenido demasiado específico o demasiado general. • Evitar preguntas basadas en opiniones. • Evitar preguntas capciosas. • Mantener un vocabulario sencillo y apropiado para cada grupo de examinados.
El Formato	• Usar tres versiones de reactivos: con formato de pregunta, para completar y la que requiere la mejor respuesta. • Colocar la pregunta verticalmente en lugar de horizontalmente.
El Estilo	• Editar y corregir las preguntas. • Usar gramática, puntuación, mayúsculas y ortografía correctas. • Emplear solo la cantidad de palabras necesarias en cada pregunta.
La redacción de la base del reactivo	• Emplear instrucciones claras. • Incluir la idea central en la base del reactivo en vez de incluirla en las opciones. • Evitar palabras innecesarias. • Afirmar en lugar de negar y evitar palabras como no y excepto.
Redactar las opciones de respuesta	• Emplear tantas opciones de respuesta viables como pueda, pero la investigación sugiere que tres opciones son suficientes. • Asegurarse que sólo una de las opciones sea la respuesta correcta. • Variar el lugar de la opción correcta según el número de opciones. • Ubicar las opciones en un orden lógico, cronológico o numérico. • Mantener las opciones independientes de la base del reactivo, evitar traslapes. • Mantener las opciones homogéneas en términos de contenido y de estructura gramatical. • Mantener homogeneidad en la extensión de las opciones. • Nunca se debe usar como última opción: "ninguna de las opciones anteriores". • Evitar: todas las opciones anteriores. • Escribir opciones que afirman; evitar negativos. • Evitar pistas que indican la opción correcta, como. . . ▪ Palabras inclusivas como "siempre", "nunca", "completamente" y "absolutamente". ▪ Asociaciones lingüísticas, opciones con palabras idénticas o que se parecen a palabras en la base del reactivo. ▪ Inconsistencias gramaticales que dan una pista sobre la opción correcta. ▪ Una opción de respuesta notablemente correcta. ▪ Opciones en pares o en tríos que identifican claramente la opción correcta. ▪ Opciones absurdas o ridículas. • Asegurar que todos los distractores sean opciones verosímiles. • Usar como distractores los errores típicos que cometen los estudiantes durante el aprendizaje.

Reglas propuestas por Bothell (2001) para escribir ítems de opción múltiple.

- Usar distractores plausibles que incluyan opciones de respuesta incorrectas. Las respuestas deben ser acordes al contenido y las opciones de respuesta deben proporcionar distractores realistas.
- Usar formatos de pregunta. Los expertos recomiendan que los ítems sean preguntas en vez de enunciados incompletos.
- Enfatizar el nivel más alto de pensamiento. Emplear preguntas en que los estudiantes apelen a la aplicación de su memoria, con el fin de que recuerden principios, reglas o hechos en el contexto de su vida diaria. Para lograrlo, se debe retomar el concepto a evaluar en una situación cotidiana para el estudiante, con el fin de que transfiera su dominio del concepto.
- Mantener las opciones de respuesta en una extensión similar. Evitar que la respuesta correcta sea más corta o más extensa que las otras opciones.
- Equilibrar el lugar donde se coloca la respuesta correcta. Las respuestas correctas se ubican por lo regular en la segunda o tercera opción.
- Ser gramaticalmente correcto. Usar vocabulario simple, preciso y libre de ambigüedades.
- Evitar pistas que guíen hacia la respuesta correcta. Evitar que una pregunta de la prueba sea la respuesta de otro reactivo. Se sugiere que la prueba sea revisada por alguien que pueda encontrar errores, pistas, o problemas de puntuación antes de que se administre la prueba a los estudiantes. Asimismo, deben evitarse los extremos como nunca, siempre y solamente. Por último, deben evitarse también las palabras sin sentido y los enunciados incongruentes.
- Evitar preguntas de negación. 31 de 35 expertos en elaboración de exámenes recomiendan evitar preguntas de negación. Los estudiantes podrían encontrar la respuesta incorrecta sin conocer la correcta.
- Emplear solo una respuesta correcta o estar seguro de que la mejor opción está claramente planteada. El emplear varias respuestas correctas disminuirá la discriminación entre los estudiantes.
- Dar instrucciones claras.
- Emplear un problema sencillo, bien definido y que incluya la idea principal en la pregunta. Los examinados deben saber cuál es el problema sin haber leído las opciones de respuesta.
- Evitar la opción "todas las anteriores. Los estudiantes deben reconocer dos opciones correctas que lleven a la respuesta correcta.
- Evitar la opción "ninguna de las anteriores".
- No usar preguntas de opción múltiple cuando otros tipos de reactivos son más apropiados.

Factores considerados por Garden y Orpwood (1996) y el INCE (Mullis *et al.*, 2002), en el TIMSS.

- Los ítems presentan cuatro o cinco opciones de respuesta, de las cuales sólo una es correcta.
- No se emplea la opción de respuesta “no sé” o “ninguna de las anteriores”.
- Las instrucciones dadas a los estudiantes no sugieren que deban adivinar cuando no saben la respuesta.
- Evalúan cualquiera de los rendimientos en los dominios cognitivos.
- Es importante que las características lingüísticas de las preguntas sean adecuadas al desarrollo de los estudiantes. Por tanto, las preguntas están redactadas de manera clara y concisa.
- Las opciones de respuesta están redactadas de manera sencilla para disminuir la carga lectora de la pregunta.
- Las opciones incorrectas están redactadas para que sean plausibles, pero no para engañar.
- Para los estudiantes que no conozcan este formato de preguntas, las instrucciones que aparecen al inicio de la prueba incluyen unos ítems de elección múltiple de muestra para ejemplificar cómo seleccionar y especificar una respuesta.
- Un ítem puede ser eliminado por recomendación de por lo menos el 30% de los países.
- Debe tener valor p mayor de .20 (para ítems de 5 opciones de respuesta) o de .25 (para 4 opciones de respuesta).
- Debe tener correlación biserial positiva para las respuestas correctas y negativa para todos los distractores.
- Los ítems fueron revisados para evitar sesgo cultural o de otro tipo.

Lineamientos del ETS (2003) para la construcción de la prueba NAEP.

Aspecto:	Reglas:
Tratar a la gente con respeto	• En toda imagen y uso de lenguaje debe mostrarse respeto a todos los grupos sociales. • En los roles sociales no deben representarse ejemplos de trabajos que ocupen un lugar bajo en el estatus social o ser representados a los trabajadores como tontos, criminales o carentes de ética. • Además, no debe bromearse sobre los problemas de un grupo social. • El etnocentrismo debe evitarse en todo momento. Ningún grupo debe mostrarse como superior o inferior a otro.
Minimizar los efectos del conocimiento o habilidad sobre un constructo irrelevante	• El uso de mapas, gráficas y otros estímulos visuales son aceptables solamente si la habilidad para leer y entenderlos es parte del constructo, de lo contrario el conocimiento o habilidad para responder el ítem no estará distribuida equitativamente. • Evitar el uso palabras difíciles, figuras del lenguaje, idiomáticos o estructuras sintácticas que dificulten el lenguaje, a menos que el propósito del ítem sea medir la habilidad lingüística del examinado. • Evitar el uso de palabras o temas que sean asociadas con una clase social (elitismo). • Evitar el uso de tecnicismos. • Evitar el uso innecesario de temas bélicos. • Evitar el uso de palabras, frases y conceptos que sean conocidos por habitantes de una región (regionalismos). • Evitar el uso de conocimiento religioso, a menos que la prueba lo requiera. • Evitar el uso de habilidades espaciales, a menos que sea el objetivo del ítem. • Evitar el uso de preguntas relacionadas con la cultura norteamericana, a menos que sea el propósito del ítem.
Uso de materiales innecesarios u ofensivos para los examinados	• Evitar el uso de vocabulario e imágenes relativos a temas controversiales, a menos que sean necesarios para la evaluación. En este caso se encuentran: aborto, abuso de gente o animales, eutanasia, experimentación dolorosa en seres humanos o animales, genocidio, comportamiento sexual, cacería como deporte, violación, satanismo, tortura y cacería de brujas. • Tener extrema precaución en temas como accidentes, enfermedades, desastres naturales, muerte, evolución, religión, esclavitud, suicidio o comportamiento auto-destructivo, violencia y sufrimiento. • Evitar el uso de materiales que describan personas de otros países de manera ofensiva. • Evitar el uso de temas problemáticos.
Emplear la terminología apropiada	• No emplear etiquetas innecesarias para nombrar a las personas. • Usar la terminología apropiada para personas con capacidades diferentes. Al respecto, evitar el uso de términos negativos o que refuercen juicios negativos. • Nombrar a las personas por su nacionalidad y no por su continente de origen. • Usar la terminología apropiada de acuerdo con la orientación sexual de las personas. Evitar el uso del término obsoleto "homosexual". • Usar la terminología apropiada para personas de la tercera edad.
Estereotipos	• Evitar el uso de estereotipos en el lenguaje referidos a la edad, género, capacidades, nacionalidad, raza, religión, status socioeconómico u orientación sexual.
Diversidad social	• Equilibrar la participación de género, grupos étnicos y raciales.

Normatividad del INEE para la construcción de reactivos de las pruebas nacionales.

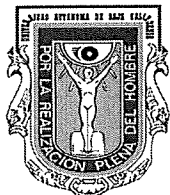
Aspecto:	Reglas:
Congruencia del ítem con su especificación	• Tanto la base de la pregunta como los distractores deben redactarse de conformidad estricta con su correspondiente especificación. • Ningún ítem debe ser más o menos difícil de lo que señale su especificación. • En los casos en que resulte necesario cambiar la dificultad del ítem, porqué así lo indique el análisis de la prueba empírica, la modificación se efectuará incrementando la dificultad de las opciones de respuesta y nunca formulando la base de la pregunta de manera más ambigua, mediante el uso de un vocabulario más sofisticado o creando situaciones más complejas. • Debe evitarse el uso de palabras y frases literales que aparecen en los libros de texto. • Se utilizarán ejemplos que aparecen en los textos, guías y otros materiales didácticos elaborados para apoyar la educación primaria, como base para construir ítems.
Contenido de los reactivos	• Reflejar el currículo nacional. • Centrarse en los contenidos curriculares que por su nivel de importancia y frecuente implementación dentro del aula son los más significativos en el nivel escolar al que se dirige la prueba. • Evaluar lo que es esencial y consecuente con un determinado aprendizaje, evitando tener que recurrir a conocimientos adquiridos fuera del salón de clases. • Considerar los contenidos, el vocabulario y los procesos de aprendizaje previos, pero sólo para lograr el andamiaje preciso con el contenido a evaluar y partir de lo que el alumno ya conoce. • Evitar requerir del estudiante el uso de una estrategia en particular para resolver los problemas que se le planteen. • Durante el desarrollo del ítem, debe evitarse entrar en detalles del conocimiento que resulten demasiado específicos, así como emplear información basada en opiniones. • Cada ítem debe propiciar en el alumno un sentimiento de éxito, estimular el mejoramiento y reducir las presiones y la tensión emocional; es decir, debe constituir un estímulo y un reto intelectual para los niños.
Uso del lenguaje	• Ser sencillo y acorde al nivel escolar al que se dirige la prueba. • Tener calidad ortográfica y de redacción. • Estar libre de expresiones idiomáticas locales que dificulten su comprensión. • Evitar todo tipo de sesgo (cultural, social, de género, etcétera); es decir, vocabulario o cualquier tipo de representación que ofenda a un grupo de estudiantes en particular, y/o que facilite o dificulte contestar correctamente. • Contemplar el uso de nombres propios representativos de la diversidad cultural de México. Se preferirán nombres cortos que sean fáciles de leer.
Redacción	• Tener oraciones breves y sintaxis sencilla. • Especificar claramente lo que se espera como respuesta y ayudar al estudiante a enfocar su atención en la tarea a realizar. • Considerar el uso de viñetas para el caso específico de instrucciones complejas. • Contemplar un tiempo mínimo de lectura, lo cual exige, por lo tanto, precisión, sencillez y claridad de ideas. • Utilizar una combinación semántica y sintáctica simple. • Ser siempre afirmativa. • Usar verbos en imperativo al inicio de la instrucción o instrucciones. • Establecer consignas directas y concisas. • La base de cada ítem será construida en forma de pregunta. • Cada ítem debe ser redactado de manera independiente de los demás reactivos. • Formular el reactivo en forma positiva evitando la palabra NO en la base de la pregunta. • Evitar el SI condicional en la base de la pregunta.

Aspecto:	Reglas:
	• Evitar mezclar el modo condicional y el modo interrogativo en una misma oración • El ítem debe tener unidad; es decir, incluye todos los elementos de información necesarios para responderlo.
Materiales	• Ser concisos. • Ser conceptualmente correctos. • Ser claros y simples, con el mínimo de distractores o información irrelevante. • Cumplir su función de apoyo y nunca “engañar” a los estudiantes para que se confundan o seleccionen una respuesta incorrecta. • Ser realistas y cercanos a la cotidianidad de los niños y/o adolescentes, particularmente en el caso de representaciones gráficas como los dibujos o las fotografías. • Ser atractivos y acordes al nivel escolar a evaluar.
Formatos	• Se evitará usar margen justificado. • El tipo de letra será de un tamaño de 12 puntos
Preguntas de opción múltiple	• En promedio, el reactivo debe requerirle al estudiante un minuto para su respuesta. Ocasionalmente podrá utilizar 2 ó 3 minutos. • Cada reactivo deberá contener una base (pregunta, afirmación, consigna, etcétera) y cuatro posibles respuestas (tres plausibles, pero sólo una considerada correcta). • La base del reactivo deberá presentar la información suficientemente clara a fin de que el estudiante sepa qué hacer antes de leer las opciones de respuesta. • En la base del reactivo se evitará incluir información irrelevante. • Los distractores deberán reflejar errores comunes o malos entendidos, concepciones simples o ingenuas, u otro tipo de error, de tal forma que las respuestas correctas demuestren lo que los alumnos realmente saben o pueden hacer. • Los distractores no deberán tener la intención de confundir al estudiante, por lo que no deberán ser variantes cercanas a la respuesta correcta. • Todos los distractores deberán tener congruencia gramatical con la pregunta. • No se deberán utilizar alternativas como: “todas las anteriores”, “ninguna de las anteriores”, “las dos de arriba”... • El diseño de las respuestas no deberá inducir a que los estudiantes descarten o eliminen alguna alternativa de respuesta o a que identifiquen la respuesta correcta debido a características superficiales y triviales de las otras alternativas, o debido a la complejidad sintáctica y conceptual de las mismas. • Los reactivos no deberán permitir que las respuestas correctas se obtengan mediante procedimientos incorrectos. • Preferentemente, las alternativas de respuesta deberán tener la misma longitud y formato, así como ser sintáctica y semánticamente paralelas. En caso de que esto no sea posible, podrá haber dos pares de alternativas con las mismas características. (Ejemplo: dos respuestas cortas y dos largas). • Las letras A, B, C y D deberán ser utilizadas para las alternativas de respuesta y no como encabezados o viñetas en el reactivo. • En el diseño de la prueba, las alternativas de respuesta deberán ordenarse considerando su longitud, cantidad o cronología. • En las especificaciones de reactivos las opciones deberán indicarse con balazos, no con letras. • En las especificaciones de reactivos, la primera opción deberá ser siempre la respuesta correcta. Deberá además estar indicada con un asterisco. • Debe cuidarse que ningún ítem proporcione determinantes específicos; es decir, pistas que indiquen la respuesta correcta o que invaliden ciertas opciones incorrectas que se ofrecen a los estudiantes, y que son independientes del dominio que tienen del contenido temático que se está explorando. A continuación se presentan los determinantes específicos más comunes: ▪ Se proporcionan claves contextuales que permiten identificar que una opción no puede ser correcta o, en su caso, que se trata de la respuesta correcta.

Aspecto:	Reglas:
	▪ Se proporcionan claves gramaticales que indican que una opción es la correcta o que no lo es. Por ejemplo, la base informativa de un ítem demanda que la respuesta correcta sea un sustantivo con número singular; sin embargo, salvo la opción correcta, las demás opciones presentan sustantivos cuyo número es plural. ▪ La longitud de una opción es considerablemente más larga o más corta que las demás. ▪ Se emplean determinantes como "nunca", "siempre" o "algunas veces". ▪ Se presentan asociaciones evidentes, opciones absurdas o indicadores semánticos muy generales o específicos.
Pruebas de matemáticas	• El uso del lenguaje deberá corresponder siempre a un grado menor al que se dirige el examen. • En la base de la pregunta, y para el caso particular del uso de números de más de tres dígitos enteros (que no refieran años calendario), se deberán utilizar espacios para agrupar los dígitos. Por ejemplo: 26 000. • Se podrán utilizar abreviaturas de medidas estándar; sin embargo, cuando pueda haber una posible confusión, será necesario evitarlas. Por ejemplo: pulgada (para sustituir la abreviatura "pul"). • Las variables deberán usar tipografía itálica y mayúscula. • La variable X no deberá usarse en los reactivos donde el símbolo x se utilice para indicar una multiplicación. • Toda fracción quedará representada mediante una línea horizontal para separar al numerador del denominador. Por ejemplo, $\frac{1}{2}$ • Los reactivos no deberán evaluar definiciones de vocabulario. • Las características de las cifras y símbolos dependerán del eje y del grado escolar, tomándose las correspondientes de los libros de texto • En la solución de problemas escritos o en contexto, la temática deberá corresponder a situaciones de la vida real, procurando que las medidas no sean irreales.
Pruebas de español	• Los reactivos deberán usar consignas por separado: una para la acción de lectura y otra para la actividad a desarrollar a partir del texto. • Los textos de lectura tendrán las siguientes características: ▪ Deberán ser tomados de los libros del alumno editados por la SEP, así como de libros, revistas o periódicos de reconocido prestigio. Siempre será necesario referir la fuente bibliográfica. ▪ Se considerará en todos los casos el uso de textos completos (no fragmentos o adaptaciones). ▪ Los textos informativos (artículos de divulgación, editoriales, artículos de opinión, reportajes, etcétera) deberán contener información fidedigna y de interés para los estudiantes del nivel educativo a evaluar. ▪ Los textos literarios deberán ser estéticos y cumplir la función recreativa que marca el plan y programa de estudio.

Anexo No. 2

Manual técnico del participante



Universidad Autónoma de Baja California
Instituto de Investigación y Desarrollo Educativo



Módulo 4

Curso-taller para la validación de ítems mediante juicios

Manual del participante

Mexicali, B. C., a 12 de mayo de 2006

Objetivo general:

Al finalizar presente módulo del curso-taller, los participantes validarán mediante jueceo ítems empleados en la evaluación criterial del aprendizaje a gran escala.

ÍNDICE

	Página
Antecedentes	4
Introducción al módulo	6
1 Aspectos conceptuales	7
1.1 Tipos de evidencias de validez relacionadas con la validación de ítems	7
1.1.1 Validez de constructo	7
1.1.2 Validez de contenido	8
1.1.3 Validez cultural	8
1.1.4 Validez cognitiva	9
1.2 Métodos para obtener evidencias de validez de ítems	9
1.2.1 Métodos empíricos	10
1.2.2 Métodos de jueceo	10
2. Evaluación de reactivos	11
2.1 Formato de evaluación	12
2.1.1 Contenido	12
2.1.2 Texto e imagen	13
2.1.3 Contexto y adecuación cultural	13
2.2 Definición de conceptos	14
2.2.1 Grupo social	14
2.2.2 Revisor	14
2.2.3 Ronda de reactivos	15
2.3 Normatividad	15
2.3.1 Lineamientos normativos	15
2.4 Procedimientos	16
2.4.1 Evaluación individual	16
2.4.2 Evaluación grupal	17
3. Conclusiones generales	17
4. Glosario	20
5. Referencias bibliográficas	24

Antecedentes

El presente módulo es un desarrollo complementario del trabajo realizado por Contreras y colaboradores (2000, 2004 y 2005), en el Instituto de Investigación y Desarrollo Educativo de la UABC, mismo que a su vez está:

(...) basado en la metodología propuesta por Anthony Nitko en 1994, complementado por la metodología [...] propuesta por James Popham (1991); por el esquema metodológico aportado por Ronald Berk y colaboradores para la construcción de tests criterioles (1984); así como en los planteamientos psicométricos propuestos por Kellaghan y Greaney para el caso de países en desarrollo (1992; 1995). (Contreras, *op.cit*, p. 8)

Para elaborar un examen criterial con base en el modelo que se ha estado desarrollando, es necesario organizar y capacitar a comités que realicen actividades como coordinar la construcción del examen, diseñar el instrumento, elaborar especificaciones de reactivos, elaborar los ítems y aplicar los modelos de examen, entre otras.

Según dicho modelo, el comité coordinador del examen puede estar integrado por jefes de área de docencia, coordinadores académicos o directores de planeación académica de una institución educativa; mientras que los comités de diseño de las pruebas, elaboración de especificaciones y elaboración de ítems se conforman básicamente por profesores con experiencia docente y dominio de una materia determinada.

Esta experiencia de evaluación parte de las que se han desarrollado principalmente en Estados Unidos, tales como las diseñadas por Popham (1991), Hambleton (1988), así como por la National Assessment of Educational Progress (NAEP), el Center for Research on Evaluation, Standards and Student Testing (CRESST) y el National Assessment Governing Board (NAGB) (Contreras, 2000, p. 22).

En España, también se han elaborado exámenes para evaluar los aprendizajes a nivel primaria; por ejemplo, por el INCE del Ministerio de Educación y Ciencia de España (Mullis *et al.*, 1995). Mientras que en Botswana, Anthony Nitko

(1994) ha aplicado su modelo de desarrollo de exámenes criteriosales centrados en el currículum de la educación básica, así como también se ha adaptado para el caso de Jamaica (Nitko, op.cit.).

En nuestro país, son escasos los aportes de la investigación educativa en este campo; Contreras (2000, pp. 23-25) refiere algunos de los escasos trabajos realizados, a partir de la década de los noventa del siglo pasado, por Schmelkes en 1994 y 1999; Ornelas, en 1994; Ruiz y Martínez, en 1996; Zorrilla en 1996 y Backhoff en 1996. Los proyectos señalados anteriormente tienen características diversas, pero destacan las observaciones relativas al interés de los investigadores por diseñar e implementar evaluación a gran escala, contribuyendo poco a poco a desarrollar una cultura evaluativa en México. A pesar de tales esfuerzos, a excepción del caso de los exámenes desarrollados en Aguascalientes, es visible una falta de alineación de las pruebas al currículum de la educación básica. Otra excepción a esta situación es el caso de los exámenes nacionales para educación básica que desarrolló entre 2004 y 2005 el Instituto Nacional para la Evaluación de la Educación (INEE), denominados Excale (Backhoff, 2005).

En el ámbito regional, hay algunas experiencias de aplicación del modelo que desarrolla el IIIDE-UABC en instituciones educativas: el Examen de Español para la Educación Primaria en Baja California (Contreras: 2000), los exámenes sumarios elaborados entre 2003 y 2005 por el COBACH y los exámenes de certificación del CECYTE BC (Contreras y Hurtado, 2004), así como los exámenes colegiados desarrollados entre 2004 y 2005 por la UABC (Contreras y colaboradores, 2005).

Introducción al módulo

Esta capacitación se realiza con el objetivo de preparar a un comité para validar mediante jueceo, ítems empleados en la evaluación criterial del aprendizaje a gran escala.

En la primera parte, denominada **aspectos conceptuales**, se revisarán algunas nociones conceptuales necesarias para entender la validación de ítems mediante juicios. Primero se abordarán los principales tipos de validez relacionados con la tarea de evaluación de reactivos. Posteriormente se explorarán los métodos empíricos y de jueceo, los cuales permiten obtener evidencias de validez de ítems. Después, en la etapa de **evaluación de reactivos**, se examinará el formato a utilizar, se definirán algunos conceptos relativos al formato, se explicará la normatividad para su uso y los procedimientos para llevar a cabo la evaluación de los reactivos. Finalmente, en la sección de **evaluación de sesiones**, se valorará de manera individual todo el proceso de capacitación y asesoría, con fines de realimentación y mejora del módulo.

Unidad 1. Aspectos conceptuales

Objetivo particular: El participante comprenderá aspectos conceptuales básicos para la realización de una validación de reactivos mediante jueceo.

Introducción a la Unidad

Dentro de los aspectos conceptuales a revisar se encuentran los tipos de validez de constructo, de contenido, cognoscitiva y cultural. Además, se abordarán los métodos para obtener evidencias de validez de ítems: empíricos y de jueceo.

Tema 1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Objetivos específicos: El participante examinará la validez de constructo, de contenido, la cultural y cognitiva con la finalidad de caracterizarlas así como de comprender en qué aspectos se centrará la validación de ítems mediante juicios.

Subtema 1.1.1 Validez de constructo

Consiste en obtener evidencia de que una prueba mide las características psicológicas de interés. La habilidad para razonar, la visualización espacial y la comprensión de la lectura son constructos, tal como también lo son las características de la personalidad, la inteligencia y la sociabilidad, entre otros (Popham, 1991).

Las estrategias empleadas para validar constructos buscan evidencia empírica que demuestre la existencia del constructo estudiado y que indique que una prueba está midiendo efectivamente un constructo determinado.

Para efectos de esta evaluación de reactivos, se evaluará a nivel ítem y no a nivel prueba, por lo que las revisiones de los ítems estarán enfocadas en la validez de contenido y cultural, específicamente.

Subtema 1.1.2 Validez de contenido

Se valida el contenido mismo de la prueba, es decir, si los reactivos evalúan la parte del currículum formal y la parte del currículum implementado que correspondan, por lo que se preguntará a profesores expertos en la materia si éste es un conocimiento que deben de dominar los alumnos y si ellos lo enseñan.

La validez de contenido es el grado en el que los ítems de una prueba evalúan el contenido específico de manera consistente con la naturaleza del contenido. Es particularmente importante en los exámenes de logro y habilidades. En estas situaciones es crítico que el instrumento de evaluación examine de manera consistente conforme a los objetivos que pretende evaluar (L'Allier y Welsh, 1995).

La validez de contenido será baja si los reactivos del examen: 1) indebidamente enfatizan algunos conceptos o procedimientos, o 2) contienen afirmaciones, condiciones o contextos no relacionados con lo que ha sido enseñado, o 3) cubre conceptos o procedimientos que no han sido enseñados en la clase. Por lo tanto, la validez de contenido no evalúa que tan bien se examina un contenido, sino si evalúa el contenido específico.

Subtema 1.1.3 Validez cultural

Después tenemos el aspecto de validez cultural. En la terminología cotidiana se menciona con frecuencia el sesgo cultural y el de género, entre otros. En esta evaluación de reactivos examinaremos que la falta de validez desde el punto de vista cultural puede muy bien producir un sesgo. Entonces, estaremos manejando los términos como opuestos uno del otro: si hay sesgo entonces no hay validez, o si hay un sesgo demasiado grande entonces eso puede afectar a la validez.

Lo que sucede con la validez cultural es que idealmente nos gustaría que siempre participaran miembros de todos los grupos culturales a los que se aplicará

la prueba, en el proceso de desarrollo de la misma; es decir, grupos raciales, socioeconómicos, de distintas regiones del país, etc. Esto es así porque cuando participan todos esos grupos en el proceso de evaluación, hay una mayor garantía de que los reactivos no van a estar sesgados a favor o en contra de algún grupo de la población que será examinada.

La evidencia más reciente demuestra que la integración diferenciada de grupos desarrolladores de pruebas es necesaria pero no es suficiente. Así, se le puede dar un reactivo a una persona que sea representante de un grupo o que tenga experiencia trabajando con él y se le puede solicitar que señale lo que a su juicio está mal en el ítem o que puede afectar a esos examinados; En dicho caso, muy probablemente ese revisor identificará cosas que no son realmente un problema o dejará sin identificar aspectos que son problemáticos.

Subtema 1.1.4 Validez de cognitiva

Es un procedimiento que consiste en entrevistar a los estudiantes, pedirles que piensen en voz alta cuando están resolviendo los ítems de una prueba. Mediante la información que proporcionan los examinados, se tiene evidencia de los procesos cognitivos que están utilizando y si éstos tienen que ver o no con el conocimiento que se pretende evaluar.

En ocasiones, los estudiantes llegan a la solución correcta sin haber mencionado para nada lo que tiene que ver con el reactivo, por lo que el ítem no está produciendo el pensamiento y la actividad mental que se supone debe suscitar en el estudiante.

Tema 1.2 Métodos para obtener evidencias de validez de ítems

Objetivos específicos: El participante conocerá los dos métodos generales que se emplean para obtener evidencias de validez de ítems, ubicando a los métodos de jueceo como los primordiales en la realización de las actividades que se desarrollarán en este módulo.

Subtema 1.2.1 Métodos empíricos

Los métodos empíricos se realizan después de aplicada la prueba; pueden basarse en observaciones y en análisis estadísticos. Para el cálculo del sesgo de los ítems mediante pruebas empíricas, existen varias propuestas (Suárez y Jornet, 1990).

- a) Método delta-plot (Angoff, 1972 y 1982) basado en la escalación absoluta de Thurstone (1975), que supone que las dos distribuciones de los índices de dificultad para dos subgrupos distintos, se ordenarán de forma independiente para cada uno de ellos.
- b) Técnicas de chi cuadrado que tiene en cuenta solos los aciertos del ítem, lo cual trajo varias críticas y Nunscrester (1977) y Camilli (1979) adoptan un chi cuadrado que incorpora al cálculo tanto las respuestas correctas como las incorrectas.
- c) Modelos de la teoría de la Respuesta al Ítem: Comparación de curvas características del ítem, igualdad de parámetros del test y comparación ajuste datos - modelo.

Subtema 1.2.2 Métodos de jueceo

Los métodos de jueceo se utilizan antes de pilotear los reactivos y pueden ser formales e informales.

Al efectuar una validación mediante el uso del juicio informal, se entrega al revisor un reactivo y se le pide que lo revise basado en su experiencia; es decir, que exprese cómo ve al reactivo, qué aciertos le encuentra al ítem, qué errores; incluso se le pregunta si identifica algún tipo de sesgo o cree que dicho reactivo afectaría a los examinados y cómo lo haría.

Por otra parte, en el proceso de validación de ítems mediante juicios formales, además de la participación de miembros de distintas comunidades culturales, se requiere del trabajo de especialistas, quienes también se darán a la tarea de revisar los reactivos y hacer análisis más sofisticados.

Los juicios formales sobre un ítem pueden ser acertados, falsos positivos o falsos negativos. Estos juicios se hacen sobre el contenido, sobre el texto e imagen y sobre el contexto y la adecuación cultural del reactivo. Los juicios **acertados** son aquellos en que se encontró un error existente en el ítem, mientras que en los juicios **falsos** hay un error por parte del revisor, quien identificó un problema pero los datos empíricos no lo respaldan (**falsos positivos**) o que no identificó un error pero la evidencia empírica obtenida tras el procesamiento estadístico demuestra que sí lo hay (**falsos negativos**).

Unidad 2. Evaluación de reactivos

Objetivo particular: Con base en el protocolo evaluativo diseñado, el participante evaluará el contenido, texto e imagen, así como el contexto y adecuación cultural de reactivos empleados en la evaluación del aprendizaje a gran escala.

Introducción a la Unidad

En esta unidad se revisará el formato para la evaluación de los reactivos que se empleará en las sesiones de evaluación, con la finalidad de que los revisores conozcan las categorías que lo integran, así como las subcategorías correspondientes en que deberán centrarse al realizar la revisión de los reactivos. Además, se examinarán los conceptos de revisor, grupo social y ronda de reactivos, necesarios para entender los procedimientos a desarrollar. También, se presentará la normatividad que regirá la sesión de evaluación y se explicarán los dos procedimientos que serán operados para la validación de reactivos.

Tema 2.1 Formato de evaluación

Objetivos específicos:

El participante comprenderá las categorías y subcategorías del formato de evaluación con el propósito de emplearlo eficazmente en la evaluación de ítems.

Para la evaluación de reactivos se trabajará con un formato simple que presenta una taxonomía de errores potenciales en el ítem, mismo que resulta fácil de llenar, requiere poco entrenamiento por parte del juez, permite integrar juicios e identificar a los reactivos de mejor calidad para incluirlos en los exámenes, así como detectar problemas específicos a fin de mejorar los ítems existentes.

Para contestar el formato, los revisores tendrán que inspeccionar el reactivo y después marcar, en su caso, la categoría y subcategoría donde existe un error en el ítem. Si el evaluador marca un error, debe describir brevemente en qué consiste

el error que identifica. Es importante que las posibles anotaciones que se hagan sean con letra legible, de lo contrario habrá dificultad para entender sus opiniones.

Son tres las secciones que integran el formato de evaluación de ítems: 1) contenido; 2) texto e imagen; y 3) contexto y adecuación cultural.

Subtema 2.1.1 Contenido

La **sección de contenido** está integrada por las afirmaciones 1 a la 4. Los juicios que se emiten sobre el contenido del ítem evalúan su correspondencia con el currículum formal, el currículum implementado y la validez cognoscitiva.

- En cuanto a la correspondencia con el currículum formal, valoran el contenido del ítem a la luz de las fuentes que definen currículum, tales como el programa y los libros de texto; también juzgan si la dificultad del contenido es acorde con el grado escolar del examinado.
- Los juicios sobre el currículum implementado evalúan si el reactivo se centra en el conocimiento que se enseña en la escuela y que se trabaja con el profesor dentro del salón de clase.
- Por último, los juicios sobre la validez cognoscitiva examinan si el reactivo únicamente se resuelve mediante el uso de las habilidades o conocimientos que se evalúan.

Subtema 2.1.2 Texto e imagen

La **sección de texto e imagen** del reactivo corresponde a las afirmaciones 5 a la 8. Los juicios sobre el texto e imagen se enfocan en el formato del ítem, la demanda lingüística que presenta al examinado y su validez cultural.

- En cuanto a su formato, se busca que el reactivo esté libre de errores ortográficos, de redacción, de formato gráfico o de impresión.
- La demanda lingüística se refiere a que el reactivo tenga una redacción simple e incluya solo los términos técnicos necesarios.
- Los juicios sobre validez cultural revisan el uso del idioma y la epistemología. Con el uso del idioma se pretende asegurar que el ítem contenga palabras comunes y claras que signifiquen lo mismo en

todos los grupos sociales, mientras que en la epistemología se verifica que el reactivo pueda interpretarse igualmente por todos los individuos de todos los grupos sociales.

Subtema 2.1.3 Contexto y adecuación cultural

Por su parte, la **sección de contexto y adecuación cultural** se explora en las afirmaciones 9 a la 12. Los juicios sobre el contexto y adecuación cultural indagan sobre el contexto social del examinado, el contexto del reactivo, el tratamiento y la equidad.

- En el contexto social del examinado verifican que el reactivo contenga situaciones cotidianas en todos los grupos sociales.
- En el contexto del reactivo se busca que el reactivo esté libre de pistas que puedan beneficiar o perjudicar al examinado.
- La parte del tratamiento se refiere a que el reactivo trate con respeto a los miembros de todos los grupos sociales. Finalmente, la equidad indaga si el reactivo está libre de estereotipos de género, étnicos, regionales y socioeconómicos, que ofendan a los examinados.

Por último, la afirmación número 13 se refiere a una categoría abierta que tiene como finalidad ofrecer un espacio para comentar algún otro tipo de error que los enunciados 1 al 12 no hayan identificado.

Tema 2.2 Definición de conceptos

Objetivos específicos:

El participante comprenderá el significado de grupo social, revisor de reactivos y ronda social.

El contexto de la revisión de ítems requiere precisar el significado de los conceptos de grupo social, revisor y ronda de reactivos, antes de iniciar la sesión de evaluación de ítems.

Subtema 2.2.1 Grupo social

Grupo social es una manera genérica de referirse a géneros, grupos étnicos o raciales, individuos de distintas regiones e individuos de distintos niveles socioeconómicos. Así, una persona no solo es indígena o no indígena, también es mujer u hombre, y pertenece a un grupo étnico o racial; y puede pertenecer además a un grupo socioeconómico determinado.

Subtema 2.2.2 Revisor

Un revisor es aquel que examina un reactivo, llena el formato de evaluación correspondiente y participa en la discusión grupal.

Subtema 2.2.3 Ronda de reactivos

Se define como el conjunto de reactivos que, previa revisión individual de ellos, se distribuye entre el grupo de revisores para que los examinen y evalúen simultáneamente.

Tema 2.3 Normatividad

Objetivos específicos:

El participante comprenderá los lineamientos normativos que regirán la revisión de los reactivos antes, durante y después de la sesión de evaluación.

Subtema 2.3.1 Lineamientos normativos

Para iniciar el trabajo de evaluación de reactivos, se debe conocer la **normatividad** que regirá la revisión. Al respecto, los principales lineamientos serán:

- Primero. Se deberá llenar y firmar un compromiso de confidencialidad, por parte de los revisores.
- Segundo. Debe quedar bien claro que ningún material de examen debe dejar la sala de reuniones donde se está trabajando.
- Tercero. No se deberá comentar el contenido de los reactivos con alguien ajeno al equipo de revisores.

- Cuarto. Es necesario dejar las bolsas, portafolios, cuadernos o cualquier otro objeto en un lugar diferente al que se está trabajando con los reactivos, formatos de evaluación y lápices, de manera que no interfieran con el trabajo de revisión.
- Quinto. Todos los formatos deben ser llenados con lápiz, no con pluma. Así se tiene la facilidad de borrar y escribir de nuevo sin hacer tachaduras.
- Sexto. La letra que se utilice al escribir debe ser lo más clara y legible posible, para facilitar la posterior revisión de los comentarios y sugerencias; de lo contrario, no se entenderá su recomendación y no habrá valido la pena el trabajo.
- Séptimo. Cumplir con puntualidad los tiempos asignados para cada tarea, ya sea para trabajar o para descansar, a fin de realizar el trabajo de la manera más eficiente posible.

Estas normas deben existir porque el trabajo que se está realizando es muy delicado debido a que los reactivos forman parte de una prueba que se aplicará a cientos o miles de examinados en un futuro. En consecuencia, debe protegerse la integridad y validez del instrumento.

Para que la normatividad anterior sea cumplida, se cuidarán dos condiciones:

- Se darán lápices, borradores y sacapuntas a los revisores;
- Posteriormente se entregarán reactivos preasignados para iniciar la labor de evaluación.

Tema 2.4 Procedimientos

Objetivos específicos: El participante comprenderá los procedimientos para revisar los reactivos, de manera individual y grupal.

El procedimiento que desarrollarán como evaluadores consta de dos etapas: una evaluación individual y otra evaluación grupal.

Subtema 2.4.1 Evaluación individual

En esta evaluación cada revisor recibirá sus reactivos. A continuación los contestará como si fuera un examinado y posteriormente llenará el formato de evaluación descrito previamente, con el propósito de identificar los problemas que presenta cada reactivo.

Para realizar la evaluación individual, los revisores deben:

1. Leer cuidadosamente y contestar el reactivo, como si fueran ellos los examinados.
2. Responder a lo que se solicita en cada categoría y subcategoría del formato de evaluación, con base en su experiencia profesional.
3. Justificar cada error identificado en el espacio asignado. Para ello debe escribirse con letra legible y hacer comentarios claros, libres de vaguedades.
4. Agregar los comentarios que se consideren necesarios en el cuerpo del reactivo. Sólo en casos extremos deberá usarse el reverso del formato para escribir.

Subtema 2.4.2 Evaluación grupal

Una vez realizado el procedimiento anterior, habrá una **revisión grupal** en la que se discutirán los reactivos que se identificaron como problemáticos. Ahí, se les pedirá que propongan los reactivos que requieren discusión. Tales ítems se proyectarán en pantalla y se discutirán en grupo.

En esta discusión es importante tomar en cuenta tres aspectos:

- que la discusión no se desvíe,
- optimizar el tiempo de discusión y
- que se llegue a un consenso sobre las características del reactivo que deben modificarse.

3. Conclusiones generales

Para validar –mediante jueceo- reactivos que se emplean en la evaluación criterial del aprendizaje a gran escala, es necesario que el comité evaluador conozca algunos aspectos conceptuales para su mejor comprensión de la validación de ítems. Estas nociones conceptuales son:

- los tipos de evidencia de validez relacionadas con la validación de ítems. Aquí se explora la validez de constructo, contenido, cultural y cognitiva.
- los métodos para obtener evidencias de validez de ítems, que pueden ser empíricos (a posteriori) o de jueceo (a priori).

Posteriormente, los participantes en la validación de ítems deben centrarse en la evaluación de reactivos, para lo cual es indispensable que dominen el formato de evaluación con el que trabajarán, que en este caso, está conformado por tres secciones: contenido, texto e imagen y contexto y adecuación cultural.

- **Contenido:** Explora si el contenido evaluado en el reactivo es representativo del currículum formal y del currículum implementado en el salón de clase.
- **Texto e imagen:** Indaga sobre el formato de la prueba, la demanda del ítem y la validez cultural.
- **Contexto y adecuación cultural:** Analizan el contexto social del examinado, el contexto del reactivo, el tratamiento y la equidad.

De igual forma, el concepto de grupo social, revisor y ronda de reactivos son presentados al examinado con el fin de que se familiarice con los términos a manejar durante la capacitación.

A su vez, es necesario que conozca los lineamientos normativos que regirán la sesión de evaluación de reactivos, a fin de que se proteja el material a validar y se actúe con ética profesional.

La evaluación de los reactivos se realizará en dos momentos: uno individual, en que el participante leerá los reactivos, los contestará como si fuera el examinado y anotará en el formato de evaluación sus juicios correspondientes; y otro momento

grupal, en donde se discutirán los reactivos que hayan sido identificados como problemáticos durante la evaluación individual.

Finalmente, el trabajo de validación mediante jueceo de ítems empleados en la evaluación criterial del aprendizaje a gran escala realizado en esta sesión, proporciona evidencia empírica que apela a la validez de contenido y cultural, con el fin de detectar errores en los ítems y mejorar así los instrumentos de evaluación antes de que sean contestados con los examinados.

4. Glosario

Adecuación cultural	Se refiere a que el reactivo esté acorde a la cultura del examinado. Para lo cual debe considerar su contexto social, estar libre de pistas que faciliten o dificulten su ejecución, manejar el lenguaje del examinado y estar libre de estereotipos sociales.
Constructo	Es una característica psicológica de interés. La habilidad para razonar, la visualización espacial y la comprensión de la lectura son constructos, tal como también lo son las características de la personalidad, la inteligencia y la sociabilidad, entre otros (Popham, 1991).
Contenido	De lo que trata una comunicación sobre algo; es decir, la sustancia, asunto, mensaje. También todo lo incluido en una colección (por ejemplo un programa, un curso, etc.).
Contexto social	Contexto social en que se desenvuelven los examinados. Se busca garantizar que el reactivo contenga situaciones comunes a todos los grupos sociales.
Currículum	Curso integrado de estudios académicos
Epistemología	Teoría filosófica del conocimiento. Teoría que explica cómo se conoce la realidad.
Equidad	La equidad indaga si el reactivo está libre de estereotipos de género, étnicos, regionales y socioeconómicos, que ofendan a los examinados
Evaluación de gran escala	Evaluaciones para los propósitos y audiencias fuera del aula. Normalmente se estandarizan las evaluaciones a gran escala en alguna medida y se aplican al mismo tiempo a grandes números de estudiantes. Los propósitos para la evaluación de gran escala incluyen asumir responsabilidad, evaluación de programas, certificación de competencias y selección de estudiantes para programas especiales. La evaluación de gran escala puede guiar las decisiones instruccionales en el aula, pero normalmente ése no es su propósito primario.
Evaluación referida a un criterio	Comparar la actuación de un estudiante con una descripción de la ejecución deseada. Por ejemplo, se esperaría que un estudiante tenga correctos tres de los cuatro ítems que miden un rasgo particular. Este método se contrasta a menudo con la evaluación referida a una norma.
Examen criterial	Instrumento para evaluar el aprendizaje cuyos resultados permiten comparar la ejecución de un examinado contra un criterio de ejecución preestablecido, usualmente el nivel de dominio del currículum.
Ítem o reactivo	Dominio que tiene un examinado de un contenido específico. En los ítems de opción múltiple, comúnmente incluye instrucciones para responder, una base, respuesta correcta y distractores, así como información textual, gráfica o tabular de referencia.
Jueceo	Técnica empleada antes del pilotaje de un examen que consiste en la evaluación reactivos para detectar posibles errores mediante el trabajo de un panel de jueces.

Sesgo	El sesgo es definido como prejuicio o como una inclinación o dirección en particular. Decir que un examen presenta sesgo significa que prejuzga o es injusto para grupos o individuos que tienen características diferentes al grueso de los examinados (Title, 1990).
Texto e imagen	Es el formato del reactivo. Considera la ortografía, la redacción, la forma, el diseño e y la impresión.
Tratamiento	La parte del tratamiento se refiere a que el reactivo trate con respeto a los miembros de todos los grupos sociales
Validez	Grado en que una prueba mide lo que dice. Por ejemplo, una evaluación válida de la solución de un problema matemático, mediría la habilidad de los estudiantes para resolver un problema, no la habilidad para leer y comprender el problema. Técnicamente, se refiere a la obtención de evidencias que soportan las inferencias basadas en los puntajes obtenidos en el test. Es el indicador más importante de la calidad de un examen.
Validez de contenido	La validez de contenido es el grado en el que los ítems de una prueba evalúan el contenido específico de manera consistente con la naturaleza del contenido. Es particularmente importante en los exámenes de logro y habilidades. En estas situaciones es crítico que el instrumento de evaluación examine de manera consistente conforme a los objetivos que pretende evaluar (L'Allier y Welsh, 1995).
Validez de constructo	Consiste en obtener evidencia empírica que demuestre la existencia del constructo estudiado y que indique que una prueba está midiendo efectivamente un constructo determinado.
Validez cultural	Es la efectividad con que los reactivos de una prueba consideran las influencias socioculturales; es decir, los valores, creencias, experiencias, patrones de comunicación, estilos de enseñanza-aprendizaje y epistemología inherentes, que moldean el pensamiento de los estudiantes y las formas en que entienden y responden a los ítems (Nelson-Barber y Solano-Flores, 2000).
Validez cognitiva	Es un procedimiento que consiste en entrevistar a los estudiantes, pedirles que piensen en voz alta cuando están resolviendo los ítems de una prueba. Mediante la información que proporcionan los examinados, se tiene evidencia de los procesos cognitivos que están utilizando y si éstos tienen que ver o no con el conocimiento que se pretende evaluar.

5. Referencias bibliográficas

- Backhoff, E. (2005). Mesa Pública de Análisis sobre Evaluación Educativa. *Exámenes de la Calidad y el Logro Educativo (Excale), La Nueva Generación de Pruebas del INEE*. Consultado el 16 de agosto de 2005 en: <http://multimedia.ilce.edu.mx/inee/pdf/mesaexcale.pdf>
- Contreras, L. A. (2000). *Desarrollo y pilotaje de un examen de español para la educación primaria en Baja California*. Tesis de maestría no publicada. Universidad Autónoma de Baja California. Ensenada, B. C. México.
- Contreras, L. A. y Hurtado, T. (2004). Colaboración interinstitucional para elaborar pruebas criterios de gran escala: el caso de los exámenes sumarios del Colegio de Bachilleres de Baja California. *Memoria del VI Foro de Evaluación Educativa*. México. Centro Nacional de Evaluación para la Educación Superior.
- Contreras, L.A., Encinas, J.A. y De Las Fuentes, B.M. (2005). *Evaluación Colegiada del Aprendizaje en la Universidad Autónoma de Baja California: el caso del examen de Matemáticas I de la Facultad de Ingeniería Mexicali*. Ponencia aceptada para su presentación en el VIII Congreso Nacional de Investigación Educativa, organizado por el Consejo Mexicano de Investigación Educativa, a celebrarse del 30 octubre al 02 de noviembre, en Hermosillo, Sonora.
- Hambleton, R.K. (1988). Criterion-referenced measurement. En Keeves, J.P. (Ed.) *Educational Research, Methodology and Measurement: an International Handbook*. (pp. 277-282). Washington: Pergamon Press.
- Nelson-Barber, S. y Solano-Flores, G. (2000). *Assessing the cultural validity of Science and Mathematics assessments*. Poster Summary, National Science Foundation REC Research PI Meeting, Arlington, Virginia, January 11-12.
- Nitko, A. J. (1994). *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).
- Popham, J. (1991). *Modern Educational Measurement: a practitioner's perspective* (2a ed.). Allyn and Bacon: Needham Heights, MA.
- Suárez, J. M., Jornet, J. M. et al. (1990). *Estudio sobre las características métricas de los distractores y del sesgo respecto a algunas variables diferenciales en el apartado de completar frases de la subprueba de vocabulario escrito de la BPL-FCI*. En Revista de Investigación Educativa. Vol. 9, no. 16, pp. 611-618.
- Suárez, J. M. y Jornet, J. M. et al. (1990). *Algunas notas de reflexión metodológica acerca del estudio de distractores y el sesgo de ítems en tests educativos y psicológicos*. En Revista de Investigación educativa, Vol. 8, no. 16, pp. 551-559.

Anexo No. 3

Cuestionario para evaluar el módulo



Cuestionario para evaluar la operación del módulo

El presente cuestionario se aplica con la finalidad de obtener información acerca de la calidad del módulo de capacitación en el que participó. La opinión que usted nos dé ayudará a mejorarlo. Favor de contestarlo sincera y objetivamente.

Instrucciones: Elija la opción que corresponda a su opinión sobre diversos aspectos del módulo.

1. El instructor presentó conceptos y procedimientos que resultaron claros.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
2. Los conceptos y procedimientos incluidos en el manual del participante facilitaron la tarea de evaluación de ítems mediante jueceo.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
3. En general, los conceptos y procedimientos descritos resultaron complejos.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
4. El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
5. Los enunciados del protocolo permiten identificar errores en los ítems.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
6. Los ejemplos proporcionados fueron claros.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
7. Los ejemplos proporcionados fueron significativos.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
8. Las diapositivas de la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
9. La introducción del manual de participante facilitó la comprensión de su contenido.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
10. En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
11. Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.
a) Siempre b) Casi siempre c) Algunas veces d) Nunca
12. Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

13. Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

14. El instructor demostró dominio del tema.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

15. Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

16. El instructor utilizó vocabulario que fue innecesariamente complejo.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

17. La exposición del instructor fue clara.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

18. Las actividades realizadas permitieron que lograra un aprendizaje significativo.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

19. Durante la capacitación se resolvieron sus dudas.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

20. La normatividad sobre la confidencialidad quedó explicada claramente.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

21. El instructor tuvo coherencia en sus fundamentaciones.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

22. Quedó una idea clara sobre la intención de la capacitación.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

23. Se lograron los objetivos propuestos al inicio de la instrucción.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

24. El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

25. La sesión de evaluación se condujo de manera que siempre se logro el consenso.

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

26. Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).

- a) Siempre b) Casi siempre c) Algunas veces d) Nunca

Preguntas abiertas:

27. ¿Se comprendió lo que se realizaría en la sesión de evaluación?

28. ¿Considera que se mejoraron los ítems?
29. ¿Considera que el proceso de capacitación jueces es necesario para evaluar los reactivos efectivamente?
30. ¿Considera que vale la pena la revisión de los ítems por parte de jueces? ¿Por qué?
31. ¿Los resultados obtenidos lo convencen de la importancia de la validación de ítems?
32. ¿La utilización del protocolo produjo los resultados esperados?

Anexo No. 4

Presentación en Power Point empleada durante la capacitación



Curso-taller para la validación de ítems mediante juicios

Lic. Ángela Aldana Torres
IIDE-UABC

Mexicali, B. C., a 12 de mayo de 2006.

Objetivo general

Al finalizar presente módulo del curso-taller, los participantes **validarán mediante jueceo** ítems empleados en la evaluación criterial del aprendizaje a gran escala.

Antecedentes

■ El presente módulo es un desarrollo complementario del trabajo realizado por Contreras y colaboradores (2000, 2004 y 2005), en el Instituto de Investigación y Desarrollo Educativo de la UABC.

■ Para elaborar un **examen criterial** con base en el modelo que se ha estado desarrollando, es necesario organizar y capacitar a comités que realicen actividades como coordinar la construcción del examen, diseñar el instrumento, elaborar especificaciones de reactivos, elaborar los ítems y aplicar los modelos de examen, entre otras.

■ Experiencias de evaluación similar se han desarrollado en el contexto internacional: Botswana (Nitko: 1994), España (Ministerio de Educación y Ciencia de España: 1995), Estados Unidos National Assessment of Educational Progress (NAEP), Jamaica (Ministry of Education and Culture: 1995) y México (Backhoff: 1996).

■ En el ámbito regional, hay experiencias de aplicación del modelo desarrollado por el IIDE en diversas instituciones educativas (COBACH, CECYTE BC, UABC).

Introducción al módulo

1. Aspectos conceptuales. Se revisarán algunas nociones conceptuales necesarias para entender la validación de ítems mediante juicios.

2. Evaluación de reactivos. Se examinará el formato a utilizar, se definirán algunos conceptos relativos al formato, se explicará la normalidad para su uso y los procedimientos para llevar a cabo la evaluación de los reactivos.

3. Evaluación de sesiones. Se valorará de manera individual todo el proceso de capacitación y asesoría, con fines de realimentación y mejora del módulo.

Unidad 1. Aspectos conceptuales

■ Dentro de los aspectos conceptuales a revisar se encuentran los tipos de **validez de constructo**, de **contenido**, **cognoscitiva** y **cultural**.

■ Además, se abordarán los **métodos para obtener evidencias de validez de ítems**: empíricos y de jueceo.

1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Validez de constructo

■ Obtener evidencia de que una prueba mide las características psicológicas de interés.

■ Demostrar la existencia del constructo estudiado y que indique que una prueba está midiendo efectivamente un constructo determinado.

■ Aquí se evaluará a nivel ítem.



1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Validez de contenido

- La validez de contenido será baja si los reactivos del examen:
 - 1) *indebidamente enfatizan* algunos conceptos o procedimientos,
 - 2) contienen afirmaciones, condiciones o contextos *no relacionados* con lo que ha sido enseñado, o
 - 3) cubre conceptos o procedimientos que *no han sido enseñados* en la clase.



1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Validez de contenido

- Es el grado en el que los ítems de una prueba evalúan el contenido específico de manera consistente con la naturaleza del contenido.
- Se valida si los reactivos evalúan la parte del currículum formal y la parte del currículum implementado que correspondan.
- Se pregunta a profesores expertos en la materia si ése es un conocimiento que deben de dominar los alumnos y si ellos lo enseñan.

1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Validez cognitiva

- Consiste en entrevistar a los estudiantes, pedirles que piensen en voz alta cuando están resolviendo los ítems de una prueba.
- Mediante la información que proporcionan los examinados, se tiene evidencia de los procesos cognitivos que están utilizando y si éstos tienen que ver o no con el conocimiento que se pretende evaluar.



1.1 Tipos de evidencias de validez relacionadas con la validación de ítems

Validez cultural

- Terminología cotidiana: sesgo cultural y el de género.
- Idealmente nos gustaría que *siempre* participaran miembros de todos los grupos culturales a los que se aplicará la prueba, en el proceso de desarrollo de la misma.
- La evidencia más reciente demuestra que la integración diferenciada de grupos desarrolladores de pruebas es necesaria pero no es suficiente.



1.2 Métodos para obtener evidencias de validez de ítems

Métodos empíricos

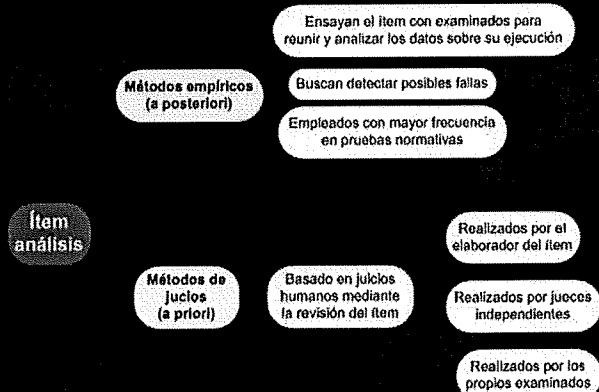
Para el cálculo del sesgo de los ítems mediante pruebas empíricas, existen varias propuestas (Suárez y Jornet, 1990):

- 1) Método delta-plot (Angoff, 1972 y 1982)
- 2) Técnicas de chi cuadrado que tiene en cuenta solos los aciertos del ítem.
- 3) Modelos de la teoría de la Respuesta al ítem:

Métodos de jueceo

- Se entrega al revisor un reactivo y se le pide que lo revise basado en su experiencia
- Cómo ve al reactivo, qué aciertos le encuentra al ítem, qué errores...
- Se requiere del trabajo de especialistas, quienes también se darán a la tarea de revisar los reactivos y hacer análisis más sofisticados.

1.2 Métodos para obtener evidencias de validez de ítems



1.2 Métodos para obtener evidencias de validez de ítems

Métodos de jueceo

- Pueden ser acertados, falsos positivos o falsos negativos.
- Se hacen sobre el contenido, sobre el texto e imagen y sobre el contexto y la adecuación cultural del reactivo.
 - Juicios acertados: aquellos en que se encontró un error existente en el ítem.
 - Juicios falsos: hay un error por parte del revisor, quien identificó un problema pero los datos empíricos no lo respaldan (falsos positivos) o que no identificó un error pero la evidencia empírica obtenida tras el procesamiento estadístico demuestra que sí lo hay (falsos negativos).

Anexo No. 5
Comentarios de los jueces en los
reactivos evaluados según el
protocolo evaluativo

Anexo No. 5 Comentarios de los jueces en los reactivos evaluados según el protocolo evaluativo empleado

Ítem	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Consenso
1	d. En la frase interrogativa, después de la palabra <u>siguientes</u> agregar un término, ya sea opciones, ecuaciones, para mayor especificación.	d. Siguientes... Confusión en A y B. La diferencia es álgebra.	d. Error en redacción. h. No evalúa en esencia, modificar los distractores.	Ninguno.	d. Agregar "opciones de respuesta" después de "siguientes".	b. Modificar el distractor A, ya que no corresponde al nivel cognitivo. Modificar el exponente de X (A). d. Agregar expresiones, ecuaciones.
2	a. Luego de $f(x) 1/\sqrt{x}$, el término A: la a debe ir con minúscula. d. El reactivo se escribiría así: Si <u>a</u> $f(x) 1/\sqrt{x}$ se integra la función $f(x)$ la solución general es:	a. ..Al <u>no</u> h. ¿Se está evaluando la regla de integración o el manejo de exponentes de álgebra?	d. Modificar la redacción.	i. Larga.	a. Después de la coma (,) va una "a" minúscula.	a. - " , al". (Agregar una coma antes de "al"). - Debe quedar: "Si se integra la función _____, la solución general es..."
3	d. La palabra dada se aprecia aislada.	a. $f(2)=4$ h. Se está evaluando aritmética.	h. Este tipo de ejercicio no es común analizarlo en el aula.	i. Larga en tiempo.	a. Después de -1 poner una coma (,). Quedaría $f'(x)=x^2-1$	b. - Cambiar $f(2)=4$, de tal manera que la respuesta sea un número entero. - Manejar solamente números enteros. h. No se hace en clase.
4	Ninguno.	a. Exponentes racionales enteros.	i. Mucho tiempo.	i. Larga en tiempo.	Ninguno.	b. Cambiar los exponentes a números enteros. i. Lleva mucho tiempo contestarlo, es de ejecución.
5	a. El reactivo, después de la palabra integral va coma.	a. Integral, la expresión.	d. Falta una coma (,).	Ninguno.	a. Poner una coma (,) después de "integral". Quedaría: "De acuerdo a las propiedades de la integral, (...)".	a. Debe quedar una coma (,) después de la palabra integral. "integral, "

Ítem	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Consenso
6	a. En el reactivo, luego de la palabra integral, va coma.	a. Integral, la expresión.	g. El distractor A me parece muy obvio! (cambiar el formato) por el de $\int()$ / $\int()$	Ninguno.	a. Agregar una coma (,) después de "integral". Quedaría: "De acuerdo a las propiedades de la integral, (...)".	a. Debe quedar: "La expresión integral es equivalente a: " o "Señala cuál de las siguientes expresiones es equivalente a tal". b. Cambiar el formato a sobre b (a / b)
7	a. Es más adecuado presentar el enunciado y abajo la gráfica, poner punto y aparte, y continuar con el enunciado que se refiera a las opciones. Se requiere mayor tamaño para las gráficas. En la palabra cual lleva acento la a. Que lse, en vez de que se te. d. La siguiente gráfica muestra la derivada de $f(x)$. A continuación, elige cuál de las opciones que se presentan es la gráfica correspondiente a la antiderivada de $df(x)/dx$	c. Gráficas con error en eje y.	d. No lo considero claro, creo que puede confundir al alumno (redacción).	i. Muy largo.	a. - Poner la gráfica debajo de la base del reactivo. - Las gráficas de las opciones de respuesta están muy chicas. d. Reducir la redacción, hacerla más concreta.	a. Quitar: "a continuación". Debe quedar: "La siguiente gráfica muestra la derivada de $f(x)$. Elige cuál de las opciones corresponde a $f(x)$. b. Las gráficas no son claras, les falta el eje de la "y" en la base del reactivo y distractores.
8	a. Bajar el reactivo, para que no esté tan junto a la presentación de arriba.	a. Solo $\mu=?$ h. Cambio por más sencillo.	Ninguno.	i. Muy largo en tiempo.	Ninguno.	a. El reactivo está muy pegado al encabezado.
9	a. La palabra Cual: a lleva acento.	a. Indique ¿cuál?	h. Modificar el grado de dificultad. i. Requiere más	Ninguno.	Ninguno.	a. Poner acento (') a cuál. h. El nivel de complejidad apela a

Ítem	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Consenso
			tiempo!			símbolos. No hay ejercicios parecidos en la clase, se recomienda cambiar por ejercicios más sencillos. i. El reactivo no se contesta en minuto y medio.
10	a. En el reactivo la palabra ésta se acentúa en la e.	h. 3+5+7+...	Ninguno.	Ninguno.	Ninguno.	a. - Poner acento a la e (ésta). - Debe quedar: "El desarrollo de la siguiente expresión _____ es....". g. En todos los distractores: cambiar las comas por signos de suma (+).
11	a. Al cerrar el signo interrogativo no se pone punto.	h. Considero que el reactivo debe de reemplazarse totalmente ya que el procedimiento evaluado escasamente se revisa en clase.	d. Tiene un error mínimo de redacción y en el punto después de la ? (pregunta). h. Considero que este tema no es abordado en clase de esta manera.	i. Largo en tiempo.	a. Quitar el punto después del signo de interrogación.	a. - Quitar el punto después del signo de interrogación. - Poner: "a la integral". d. No tiene respuesta correcta. h. El procedimiento escasamente se revisa en clase. i. En consenso recomiendan sacarlo, ya que no lo pudieron contestar. Se propone un reactivo en la representación de la gráfica.
12	Ninguno.	a. Dadas las integrales.... cambiar 5	d. Falta resaltar el 5 de la tercera integral y modificar el texto.	a. Está difícil ver claramente el factor 5 y responder en consecuencia.	Ninguno.	a. Debe quedar: "Dadas las integrales..." en la base del reactivo. Se agrega lo que está en negrita. $- \int_0^7 f(x) dx$ b. - La respuesta C cambiarla a 14.

Ítem	Juez 1	Juez 2	Juez 3	Juez 4	Juez 5	Consenso
13	a. La gráfica más amplia y acomodarla abajo. Lo mismo que en el ítem 7. Separar más la opción C de la 1) d. Lo mismo que en el ítem 7.	a. Las respuestas C y D muy juntas. c. Gráfica en detalle de títulos superiores e inferiores.	Ninguno.	Ninguno.	a. - Poner la gráfica debajo de la base del reactivo. - La opción de respuesta C está muy cerca del distractor D: separarlas. d. Reducir la redacción.	a. - La respuesta C y D están muy juntas, hay que separarlas. - Mejorar la redacción. - Bajar la pregunta. - Quitar a la gráfica la parte de arriba y abajo de F1, F2, F3. - Mover la leyenda de los ejes de las "y".
14	Ninguno.	h. Considero que debe quedarse hasta el planteo de la integral.	Ninguno.	i. Muy largo en tiempo.	Ninguno.	a. Debe quedar: "El área de la región delimitada...es planteada por...". Se quita el término "sobre". - Poner la "X" en itálica. i. - Se necesita mucho tiempo para contestar el reactivo. - Faltan reactivos de integrales definidas.
15	a. Muy junto la presentación de la pregunta.	a. Acotada se muestra a continuación 15→18 c. Gráfica no clara, ejes no claros, títulos y subtítulos	Ninguno.	b. Grado de dificultad elevado para el tiempo de discusión.	a. Poner la gráfica debajo de la base del reactivo. d. Modificar la redacción. Quedaría: "La siguiente gráfica muestra el primer cuadrante del eje y, de $y=x+1$ y $y=4$. ¿Cuál integral definida representa su área?".	a. - Después de "continuación" va punto y aparte. En el otro renglón empezará la pregunta. - Separar el encabezado del reactivo. - Hacer modificaciones a la gráfica, igual que las señaladas anteriormente.

- Está libre de errores ortográficos, de formato, de diseño y de impresión.
- Utiliza el vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.

- c. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.
- d. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.
- e. Está libre de estereotipos culturales, raciales de género, u otros que ofendan a la población examinada.
- f. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja a unos examinados sobre otros.
- g. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.
- h. Evalúa el dominio de conocimientos o procedimientos considerados en el currículum, los libros de texto y que son enseñados por los docentes.
- i. El conjunto de ítems representa al currículum

Anexo No. 6

Cuestionarios para evaluar el módulo contestados por los participantes

A



capturado

Cuestionario para evaluar la operación del módulo



El propósito del presente cuestionario es obtener información acerca de la calidad del módulo de capacitación en el que participó. La opinión que usted nos aporte ayudará a mejorarlo. Favor de contestarlo sincera y objetivamente.

Instrucciones: Elija la opción que mejor represente su opinión sobre diversos aspectos del módulo.

1. El instructor presentó conceptos y procedimientos que resultaron claros.
☒ Siempre () Casi siempre () Algunas veces () Nunca
2. Los conceptos y procedimientos incluidos en el manual del participante facilitaron la tarea de evaluación de ítems mediante jueceo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
3. En general, los conceptos y procedimientos descritos resultaron complejos.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca
4. El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca
5. Los enunciados del protocolo permiten identificar errores en los ítems.
☒ Siempre () Casi siempre () Algunas veces () Nunca
6. Los ejemplos proporcionados fueron claros.
☒ Siempre () Casi siempre () Algunas veces () Nunca
7. Los ejemplos proporcionados fueron significativos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
8. Las diapositivas de la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
9. La introducción del manual de participante facilitó la comprensión de su contenido.
☒ Siempre () Casi siempre () Algunas veces () Nunca
10. En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
11. Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca

Anexos

12. Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.

☒ Siempre () Casi siempre () Algunas veces () Nunca

13. Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.

☒ Siempre () Casi siempre () Algunas veces () Nunca

14. El instructor demostró dominio del tema.

☒ Siempre () Casi siempre () Algunas veces () Nunca

15. Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.

☒ Siempre () Casi siempre () Algunas veces () Nunca

16. El instructor utilizó vocabulario que fue innecesariamente complejo.

() Siempre () Casi siempre () Algunas veces ☒ Nunca

17. La exposición del instructor fue clara.

☒ Siempre () Casi siempre () Algunas veces () Nunca

18. Las actividades realizadas permitieron que lograra un aprendizaje significativo.

☒ Siempre () Casi siempre () Algunas veces () Nunca

19. Durante la capacitación se resolvieron sus dudas.

☒ Siempre () Casi siempre () Algunas veces () Nunca

20. La normatividad sobre la confidencialidad quedó explicada claramente.

☒ Siempre () Casi siempre () Algunas veces () Nunca

21. El instructor tuvo coherencia en sus fundamentaciones.

☒ Siempre () Casi siempre () Algunas veces () Nunca

22. Quedó una idea clara sobre la intención de la capacitación.

☒ Siempre () Casi siempre () Algunas veces () Nunca

23. Se lograron los objetivos propuestos al inicio de la instrucción.

☒ Siempre () Casi siempre () Algunas veces () Nunca

24. El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.

() Siempre ☒ Casi siempre () Algunas veces () Nunca

25. La sesión de evaluación se condujo de manera que siempre se logro el consenso.

☒ Siempre () Casi siempre () Algunas veces () Nunca

Anexos

26. Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).

☒ Siempre () Casi siempre () Algunas veces () Nunca

Preguntas abiertas:

27. ¿Se comprendió lo que se realizaría en la sesión de evaluación?

Sí, la instructora fue lo suficientemente clara y precisa para plantear los objetivos de la sesión.

28. Después del trabajo desarrollado, ¿considera que se mejoraron los ítems?

Así es. La negrita fue tanto en su redacción como en sugerencias de presentación gráfica como a fin de contenido.

29. ¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?

Sí. Sin el proceso, el juicio seguiría siendo empírico.

30. ¿Considera que valió la pena la revisión de los ítems por parte de jueces? ¿Por qué?

Sí

31. ¿Los resultados obtenidos lo convencer de la importancia de la validación de ítems?

Totalmente. El taller me sube al nivel de conciencia sobre la construcción y mejora del reactivo.

32. ¿La utilización del protocolo produjo los resultados esperados?

El protocolo me lleva un punto que considero el tiempo de ejecución del reactivo.



Cuestionario para evaluar la operación del módulo



El propósito del presente cuestionario es obtener información acerca de la calidad del módulo de capacitación en el que participó. La opinión que usted nos aporte ayudará a mejorarlo. Favor de contestarlo sincera y objetivamente.

Instrucciones: Elija la opción que mejor represente su opinión sobre diversos aspectos del módulo.

1. El instructor presentó conceptos y procedimientos que resultaron claros.
☒ Siempre () Casi siempre () Algunas veces () Nunca
2. Los conceptos y procedimientos incluidos en el manual del participante facilitaron la tarea de evaluación de ítems mediante jueceo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
3. En general, los conceptos y procedimientos descritos resultaron complejos.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca
4. El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca
5. Los enunciados del protocolo permiten identificar errores en los ítems.
☒ Siempre () Casi siempre () Algunas veces () Nunca
6. Los ejemplos proporcionados fueron claros.
☒ Siempre () Casi siempre () Algunas veces () Nunca
7. Los ejemplos proporcionados fueron significativos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
8. Las diapositivas de la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
9. La introducción del manual de participante facilitó la comprensión de su contenido.
 () Siempre ☒ Casi siempre () Algunas veces () Nunca
10. En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
11. Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.
 () Siempre () Casi siempre () Algunas veces ☒ Nunca

12. Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
13. Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
14. El instructor demostró dominio del tema.
☒ Siempre () Casi siempre () Algunas veces () Nunca
15. Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.
☒ Siempre () Casi siempre () Algunas veces () Nunca
16. El instructor utilizó vocabulario que fue innecesariamente complejo.
() Siempre () Casi siempre () Algunas veces ☒ Nunca
17. La exposición del instructor fue clara.
☒ Siempre () Casi siempre () Algunas veces () Nunca
18. Las actividades realizadas permitieron que lograra un aprendizaje significativo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
19. Durante la capacitación se resolvieron sus dudas.
☒ Siempre () Casi siempre () Algunas veces () Nunca
20. La normatividad sobre la confidencialidad quedó explicada claramente.
☒ Siempre () Casi siempre () Algunas veces () Nunca
21. El instructor tuvo coherencia en sus fundamentaciones.
☒ Siempre () Casi siempre () Algunas veces () Nunca
22. Quedó una idea clara sobre la intención de la capacitación.
☒ Siempre () Casi siempre () Algunas veces () Nunca
23. Se lograron los objetivos propuestos al inicio de la instrucción.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
24. El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
25. La sesión de evaluación se condujo de manera que siempre se logro el consenso.
() Siempre ☒ Casi siempre () Algunas veces () Nunca

26. Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).

☒ Siempre () Casi siempre () Algunas veces () Nunca

Preguntas abiertas:

27. ¿Se comprendió lo que se realizaría en la sesión de evaluación?

Si,

28. Después de trabajo desarrollado, ¿considera que se mejoraron los ítems?

Si, se notaron muchos detalles q' antes no nos fueron tan obvios

29. ¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?

Si, claro

30. ¿Considera que valió la pena la revisión de los ítems por parte de jueces? ¿Por qué?

Porque esto enriquece nuestro trabajo, se pudieron detectar algunos reactivos q' no son claros e incluso errores tanto de redacción como de concepto.

31. ¿Los resultados obtenidos lo convencer de la importancia de la validación de ítems?

Si

32. ¿La utilización del protocolo produjo los resultados esperados?

Considerablemente, pues no da la pauta por dicha revisión.

Muchas gracias



Cuestionario para evaluar la operación del módulo



C

El propósito del presente cuestionario es obtener información acerca de la calidad del módulo de capacitación en el que participó. La opinión que usted nos aporte ayudará a mejorarlo. Favor de contestarlo sincera y objetivamente.

Instrucciones: Elija la opción que mejor represente su opinión sobre diversos aspectos del módulo.

1. El instructor presentó conceptos y procedimientos que resultaron claros.
☒ Siempre () Casi siempre () Algunas veces () Nunca
2. Los conceptos y procedimientos incluidos en el manual del participante facilitaron la tarea de evaluación de ítems mediante jueceo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
3. En general, los conceptos y procedimientos descritos resultaron complejos.
☐ Siempre () Casi siempre ☒ Algunas veces () Nunca
4. El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.
☒ Siempre () Casi siempre () Algunas veces () Nunca
5. Los enunciados del protocolo permiten identificar errores en los ítems.
☐ Siempre ☒ Casi siempre () Algunas veces () Nunca
6. Los ejemplos proporcionados fueron claros.
☐ Siempre ☒ Casi siempre () Algunas veces () Nunca
7. Los ejemplos proporcionados fueron significativos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
8. Las diapositivas de la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
9. La introducción del manual de participante facilitó la comprensión de su contenido.
☒ Siempre () Casi siempre () Algunas veces () Nunca
10. En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
11. Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.
☐ Siempre () Casi siempre () Algunas veces ☒ Nunca

12. Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
13. Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
14. El instructor demostró dominio del tema.
☒ Siempre () Casi siempre () Algunas veces () Nunca
15. Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.
☒ Siempre () Casi siempre () Algunas veces () Nunca
16. El instructor utilizó vocabulario que fue innecesariamente complejo.
() Siempre () Casi siempre () Algunas veces ☒ Nunca
17. La exposición del instructor fue clara.
☒ Siempre () Casi siempre () Algunas veces () Nunca
18. Las actividades realizadas permitieron que lograra un aprendizaje significativo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
19. Durante la capacitación se resolvieron sus dudas.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
20. La normatividad sobre la confidencialidad quedó explicada claramente.
☒ Siempre () Casi siempre () Algunas veces () Nunca
21. El instructor tuvo coherencia en sus fundamentaciones.
☒ Siempre () Casi siempre () Algunas veces () Nunca
22. Quedó una idea clara sobre la intención de la capacitación.
☒ Siempre () Casi siempre () Algunas veces () Nunca
23. Se lograron los objetivos propuestos al inicio de la instrucción.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
24. El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.
() Siempre () Casi siempre ☒ Algunas veces () Nunca
25. La sesión de evaluación se condujo de manera que siempre se logro el consenso.
☒ Siempre () Casi siempre () Algunas veces () Nunca

26. Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).

() Siempre (☒) Casi siempre () Algunas veces () Nunca

Preguntas abiertas:

27. ¿Se comprendió lo que se realizaría en la sesión de evaluación?

Si.

28. Después de trabajo desarrollado, ¿considera que se mejoraron los ítems?

Si.

29. ¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?

Si.

30. ¿Considera que valió la pena la revisión de los ítems por parte de jueces? ¿Por qué?

Valió la pena en cuanto a todo lo que se alcanzó a revisar

31. ¿Los resultados obtenidos lo convencer de la importancia de la validación de ítems?

Si convencer. Nada más falta tiempo para revisar ítems.

32. ¿La utilización del protocolo produjo los resultados esperados?

Si



Cuestionario para evaluar la operación del módulo



El propósito del presente cuestionario es obtener información acerca de la calidad del módulo de capacitación en el que participó. La opinión que usted nos aporte ayudará a mejorarlo. Favor de contestarlo sincera y objetivamente.

Instrucciones: Elija la opción que mejor represente su opinión sobre diversos aspectos del módulo.

1. El instructor presentó conceptos y procedimientos que resultaron claros.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
2. Los conceptos y procedimientos incluidos en el manual del participante facilitaron la tarea de evaluación de ítems mediante jueceo.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
3. En general, los conceptos y procedimientos descritos resultaron complejos.
☐ Siempre ☐ Casi siempre ☐ Algunas veces ☒ Nunca
4. El contenido del manual del participante resultó demasiado extenso para el tiempo en que se revisó.
☐ Siempre ☐ Casi siempre ☐ Algunas veces ☒ Nunca
5. Los enunciados del protocolo permiten identificar errores en los ítems.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
6. Los ejemplos proporcionados fueron claros.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
7. Los ejemplos proporcionados fueron significativos.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
8. Las diapositivas de la presentación en Power Point apoyan la comprensión de conceptos y procedimientos descritos.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
9. La introducción del manual de participante facilitó la comprensión de su contenido.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
10. En el manual del participante se explican con la profundidad suficiente los procedimientos a realizar.
☒ Siempre ☐ Casi siempre ☐ Algunas veces ☐ Nunca
11. Algún procedimiento -explicado en el manual del participante- resultó confuso o complejo.
☐ Siempre ☐ Casi siempre ☐ Algunas veces ☒ Nunca

12. Los ejercicios presentados en el manual apoyaron en la comprensión de conceptos y procedimientos.
☒ Siempre () Casi siempre () Algunas veces () Nunca
13. Los ejercicios presentados fueron suficientes para comprender la tarea a realizar.
☒ Siempre () Casi siempre () Algunas veces () Nunca
14. El instructor demostró dominio del tema.
☒ Siempre () Casi siempre () Algunas veces () Nunca
15. Cuando se le preguntó, el instructor resolvió de manera clara la duda en cuestión.
☒ Siempre () Casi siempre () Algunas veces () Nunca
16. El instructor utilizó vocabulario que fue innecesariamente complejo.
() Siempre () Casi siempre () Algunas veces ☒ Nunca
17. La exposición del instructor fue clara.
() Siempre () Casi siempre () Algunas veces () Nunca
18. Las actividades realizadas permitieron que lograra un aprendizaje significativo.
☒ Siempre () Casi siempre () Algunas veces () Nunca
19. Durante la capacitación se resolvieron sus dudas.
☒ Siempre () Casi siempre () Algunas veces () Nunca
20. La normatividad sobre la confidencialidad quedó explicada claramente.
☒ Siempre () Casi siempre () Algunas veces () Nunca
21. El instructor tuvo coherencia en sus fundamentaciones.
☒ Siempre () Casi siempre () Algunas veces () Nunca
22. Quedó una idea clara sobre la intención de la capacitación.
☒ Siempre () Casi siempre () Algunas veces () Nunca
23. Se lograron los objetivos propuestos al inicio de la instrucción.
☒ Siempre () Casi siempre () Algunas veces () Nunca
24. El tiempo destinado a la capacitación fue el suficiente para realizar adecuadamente la evaluación de los ítems.
() Siempre ☒ Casi siempre () Algunas veces () Nunca
25. La sesión de evaluación se condujo de manera que siempre se logró el consenso.
☒ Siempre () Casi siempre () Algunas veces () Nunca

26. Se respetaron los tiempos asignados para realizar las actividades (sesión de evaluación).

(☒) Siempre () Casi siempre () Algunas veces () Nunca

Preguntas abiertas:

27. ¿Se comprendió lo que se realizaría en la sesión de evaluación?

Sí, había que realizar una evaluación individual de los reactivos y luego otra grupal.

28. Después de trabajo desarrollado, ¿considera que se mejoraron los ítems?

Sí, mejoraron en cuanto a su redacción, dificultad y nivel cognitivo.

29. ¿Considera que el proceso de capacitación de los jueces resultó necesario para evaluar los reactivos efectivamente?

Sí, porque a pesar de que ya habían realizado una revisión informal, no fue tan completa como la de esta sesión. La capacitación los ayudó a comprender la importancia de la revisión de ítems así como a identificar posibles errores en éstos.

30. ¿Considera que valió la pena la revisión de los ítems por parte de jueces? ¿Por qué?

Sí, porque la integración del comité ayudó a que distintos especialistas colaboraran, discutieran y llegaran a un consenso sobre los 15 reactivos a revisar.

31. ¿Los resultados obtenidos lo convencerán de la importancia de la validación de ítems?

Sí. La validación de ítems mediante jueces es un procedimiento importante y necesario para mejorar la calidad técnica de los reactivos, evitar el sesgo y la discriminación para asegurar que el examen evalúe de manera justa a los examinados.

32. ¿La utilización del protocolo produjo los resultados esperados?

Sí, los jueces nos sentimos satisfechos con el trabajo realizado ya que se identificaron errores en los ítems y se hicieron propuestas para mejorarlos.

Anexo No. 7

Formatos de evaluación de reactivos contestados por los participantes

A manera de ejemplo, se presenta un formato de evaluación de reactivos contestado por cada miembro del comité.

Nombre del evaluador:	Alvaro Encinas
No. de ítem:	1

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	✓
2. Utiliza un vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	✓
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	✓
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	siguientes... ✓ con fusión en A y B. Lodi Avonada es de la...
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	✓
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	✓
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	✓
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículum, los libros de texto y que son enseñados por los docentes	✓
9. El conjunto de ítems representa apropiadamente al currículum.	

Alvaro

2

Anexos

Nombre del evaluador: Humberto Bastidas
No. de ítem: 1

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	✓
2. Utiliza un vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	✓
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	✓
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	✓
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	✓
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	✓
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	✓
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículo, los libros de texto y que son enseñados por los docentes	✓
9. El conjunto de ítems representa apropiadamente al currículo.	

Anexos

Nombre del evaluador: Laura Carreño Zavala
 No. de ítem: 1

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	✓
2. Utiliza un vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	✓
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	✓
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	En la frase interrogativo, después de la palabra siguientes agregar un término, ya sea opciones, ecuaciones, para mayor especificación.
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	✓
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	✓
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículum, los libros de texto y que son enseñados por los docentes	
9. El conjunto de ítems representa apropiadamente al currículum.	

Anexos

Nombre del evaluador: <i>Ruth Elba Rivera C</i>
No. de ítem: <i>1</i>

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	/
2. Utiliza un vocabulario y un nivel de dificultad que resulten apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	✓
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	—
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	Error en redacción ✓
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	✓
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	✓
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	✓
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículo, los libros de texto y que son enseñados por los docentes	No evalúa en ciencias ✓ mod. distraído o bien
9. El conjunto de ítems representa apropiadamente al currículo.	

Anexos

Nombre del evaluador:

No. de ítem: 12

En su opinión, el ítem:	Comentarios
1. Está libre de errores ortográficos, de formato, de diseño y de impresión.	- Dadas <u>las integrales</u> --- base del reactivo (agregar) $\int 2f(x) dx$
2. Utiliza un vocabulario y un nivel de dificultad que resultan apropiados al nivel del examinado y puede ser interpretado igualmente por todos los grupos sociales.	- la resp. <u>C</u> cambiar a <u>14</u> - Los distractores no están bien a) 6 b) 3 c) 14 d) 7
3. Presenta únicamente imágenes o gráficas que apelan al contenido o habilidad a evaluar.	- Dar + espacios en blanco + las integrales están dif. de leer
4. La redacción del ítem debe ser clara, esencial y sin errores conceptuales; la respuesta correcta realmente lo es.	
5. Está libre de estereotipos culturales, raciales, de género, u otros que ofendan a la población examinada.	
6. Está libre de pistas y sesgos lingüísticos, socioeconómicos y otros que pongan en ventaja o desventaja a unos examinados sobre otros.	
7. Está exento de errores psicométricos, como opciones de longitud extrema, claves contextuales, asociaciones evidentes o absurdas, determinantes como "nunca" o "siempre", etc.	
8. Evalúa el dominio de conocimientos o procedimientos considerados en el currículo, los libros de texto y que son enseñados por los docentes	
9. El conjunto de ítems representa apropiadamente al currículo.	

Anexo No. 8

Notas sobre la videograbación

Fragmento de notas sobre la revisión de ítems.

Ángela: ¿los revisamos?

Voz de hombre: así es por ejemplo, entonces por ejemplo el 1er reactivo es en cuanto a la redacción dice: ¿Cuál de las siguientes es la derivada de la función?

Entonces ahí le falta ¿cuál de las siguientes "expresiones"? ¿cual de las siguientes "funciones"? ¿Cuál de las siguientes "opciones"?

Ángela: o ecuaciones...también podría ser?

Voz de hombre: también es valido, así es.

Voz de mujer: creo que puede ser ¿Cuál de las siguientes "operaciones"?

Voz de hombre: ¿ese comentario va en el punto cuatro o en el punto uno?.

Ángela: en el cuatro.

Vos de hombre: en el cuatro muy bien, luego otra; en donde va...el ocho es evaluar el dominio no? La pregunta, digo la respuesta correcta es la C verdad? Creo, no la marque, bueno no importa.

Voz de mujer: no, es la B.

Voz de hombre: perdón la b, no la marque, y la diferencia entre la A y la A, ahh si es, entre la A y la B, si se fijan es el coeficiente $\frac{2}{5}$ y el otro $\frac{5}{2}$, para diferenciar entre una respuesta y otra, la complicación vaya es de tipo algebraico, o mejor dicho de aritmético, no esta dirigida a la esencia de este reactivo, que es conocer la derivada o las derivadas así es. Entonces si el muchacho trae problemas de aritmética ahí y responde mal, entonces pues enmascaran el mal o buen manejo de la regla de antiderivación, con una dificultad de tipo aritmético Entre la A y la B, o sea si ahí el reactivo A, si el correcto dijimos que es A ¿verdad?

Voz de mujer: no, es B.

Voz de hombre: entonces el A habría que modificarlo para que quede bien claro que lo que está evaluando es la regla de antiderivación, no un asunto de aritmética.

Voz de mujer: **muy de acuerdo con lo que dices.**

Ángela: ¿Modificar la opción A?

Voz de hombre: Así es

Voz de mujer: sí, porque sí puede confundir al estudiante. Si te fijas sí son bastante parecidas y ahí el error sería más bien de aritmética y no de lo que es la integración, que es lo que queremos evaluar, yo siento que el problema debería, el hecho de poner el exponente fraccionario ya sabemos que les ocasiona muchos problemas a los estudiantes, entonces si de veras queremos saber si domina lo que es la integración no es necesario poner un exponente fraccionario.

Voz de hombre: estoy de acuerdo contigo. Entonces debe de ir en la propuesta, la propuesta sería: modificar el exponente de "x" a un número entero.

Voz de mujer: digo si no se quiere modificar, entonces sería cambiar.

Voz de hombre: estoy de acuerdo con la maestra, con eso se borra todas esas complicaciones por ejemplo: que en lugar de tener " x " a la $3/2$, por decir algo que sea " x " a la tres simplemente y ya nos vamos al foco, al blanco.

Voz de mujer: muy bien.

Ángela: ¿alguna otra observación maestra?

Voz de mujer: pues lo mismo ya había puesto aquí.

Ángela: muy bien, entonces pasamos, modificamos el distractor "A" ya que no corresponde al nivel cognitivo y el exponente de x , igual del distractor "A", agregar después de "siguientes", "expresiones" o "ecuaciones"...muy bien...algún otro comentario de alguno? Ok, entonces pasamos al siguiente reactivo (2) con observación que es...cuál maestro?

Voz de hombre: al dos, en el dos después dice: "si f de x igual a uno sobre raíz de x coma", ahí es a minúscula.

Voz de mujer: ¿pero quien iba a revisar la redacción, la lingüística?

Voz de mujer: a mi también me "volaba" algo, pero no me aterrice, yo lo que pensé que eran diminuto.

Voz de mujer: y ahí estaría bien, mejor se, yo creo que se podría dar el orden verdad? Si el reactivo se escribiera así: "si A a la primera se integra la función F de X la solución general es", porque así como que me suena: "si, y esto, al integrar la función" como que aislado.

Voz de hombre: a ver a ver, ¿cómo propones?

Voz de mujer: si digo, si y ya le pongo y ya saben la ecuación ¿no? coma, luego dice "al integrar la función" para mi sería mejor para el reactivo se escribiría así: "si a "tal tal" la fórmula ¿no?" se integra la función, para mi sería mejor "se integra".

Voz de hombre: ok a ver: "si F de X es igual a uno de X ".

Voz de mujer: o sea voltear el orden ¿no?, "si la integral de la función, o si se integra".

Voz de hombre: "si se integra, si se integra a ver, si se integra la función F de X igual a 1 sobre raíz de X " "si se integra" "la respuesta es" "o la respuesta sería" ¿que sería ahí? ¿es o sería? ..."la solución es"...la solución general es". La solución general tiene que ver con la letra C que parece llave, si suena muy bien.

Voz de mujer: "si se integra la función, F de X igual a (no audible) la solución general", tienes que poner otra vez F de X .

Voz de hombre: pues claro, está muy bien.

Ángela: alguna otra observación de la 2.

Voz de hombre: a ver estamos en la 2. Ahh sí, otra vez sería lo mismo que decía Ruth hace rato, de nueva cuenta la complejidad de "uno sobre raíz de X " va a pasar lo mismo con el manejo de los exponentes.

Voz de mujer: aunque te digo, ahí ya no se me hace tan grave porque es una sola expresión y no es tanto pues, esta bien incluirlos, pero de nuevo si es cierto, la complejidad, no nomás es la integral si no lo del manejo de exponentes.

Voz de hombre: y para diferenciar entre un distractor y otro, una respuesta y otra, otra vez esta, tiene el estudiante utilizar el arsenal de álgebra y aritmética, no el de cálculo.

Voz de mujer: si, porque si se puede prestar a confusión verdad?...por ejemplo entre la B y la D pudiera tener (no audible).

Ángela: entonces proponen cambiar las opciones de respuesta? O cual sería ahí la propuesta?

Voz de hombre: es que es una posibilidad, la otra es desde el mero principio, cambiar, modificar la función o sea la f de x , porque es un problema haz de cuenta muy semejante, al que comentamos hace rato del número uno.

Juez 5: disculpen, ahí cual es la discusión?

Ruth: ahh, no lo pusimos al tanto.

Ángela: si, mire lo que hicieron sus compañeros es revisar, revisaron los 15 primeros reactivos del examen, los contestaron como si fueran los examinados e hicieron anotaciones en el formato que vimos para evaluar el reactivo. Entonces ahorita estamos proponiendo qué reactivos vamos a revisar, qué problema identificaron en cada uno de los reactivos y por consenso cuál va a ser la propuesta para cambiarlo. Vamos ahorita a revisar estamos en el 2do, apenas.

Álvaro: o sea ya los revisamos en, como entre las....antes de salir con un cierto criterio por ahí, y estamos de alguna manera en un ejercicio parecido al que hicimos hace unos días allá abajo del modelo, estamos en el número dos.

Ángela: apenas en el número dos, entonces comentaban del 2do. reactivo? Cambiar la función?

Álvaro: una propuesta podría ser cambiar expresión "F de X" ajá, que es la función, porque lo estábamos comentando René es que los exponentes, cuando metemos ciertos exponentes en expresiones van a meter ruido pues y en lugar de estar evaluando la antiderivación vamos a estar evaluando el manejo de exponentes pues. Ya sabemos que por sí mismo ese tema es "pantanos" para muchos pues.

Ruth: pudiera darse el caso que el muchacho sepa integrar pero que no tenga la habilidad que es muy común con los alumnos de trabajar la aritmética o sea los exponentes y entonces este ahí nos va a quedar la duda podría hacerse pues, darse el caso de que nos va a quedar la duda si es porque no sabe integrar? o porque le falló a la hora de....

Álvaro: si buscar por ejemplo exponentes enteros pues.

Ruth: si como que manejando exponentes enteros te va a quedar más claro, si el muchacho contesta mal es porque no maneja la ley de antiderivación. Como sea es nuestra opinión falta la de los demás ¿no?

Álvaro: sí, va a quedar asentada obviamente, eso no quiere decir que se vaya a cambiar.

Ángela: claro. Entonces la propuesta el cambiar la función, muy bien, alguna otra observación del segundo reactivo.

Ruth: no, ya vimos también de la redacción ¿verdad?

Ángela: si, cambió en la redacción, muy bien entonces pasamos pues al tercer reactivo, ¿identificaron algún error, algún problema?

Anexo No. 9

Síntesis de la videograbación sobre la evaluación de los reactivos

Revisión previa de la prueba

Una vez terminada la exposición de la presentación en Power Point por parte de la instructora, uno de los docentes de la UABC expresó que dos semanas antes se reunieron los elaboradores de la prueba y revisaron con su criterio la misma prueba considerada para esta sesión. Por lo que el docente comentó que eso podría afectar la discusión grupal debido a que ya tuvieron una discusión previa. Ante esta inquietud, la instructora comentó que dos semanas es un tiempo bueno para quienes están participando como jueces en el módulo porque ya descasaron esos reactivos y podrían verlos con nuevos ojos.

En este sentido, el docente señaló que en esta revisión habría aspectos relevantes que la lingüista e instructora seguramente les observarían. No obstante, afirmó que los comentarios serían sobre detalles más pequeños que ellos no atendieron. Finalmente, la instructora le respondió al docente que se registraría ese comentario por si esa discusión previa aminoraba la discusión grupal.

Dudas

a) Sobre el procedimiento

Sobre el procedimiento y los tiempos asignados para las tareas, un docente insistió en que no se revisara todo el examen; sugería una exploración de menos reactivos con el objetivo de aprender la metodología y desarrollar la habilidad de analizar los reactivos. La instructora propuso al comité revisar solamente 15 reactivos de 32 que integraban el examen, considerando el limitado tiempo de la sesión. De conformidad con lo anterior, otro juez indagó sobre el tiempo que se tardaría en la discusión grupal, para lo cual se estimó un máximo de dos horas, debido al tiempo preasignado para la duración del módulo de validez de ítems.

Otra duda que surgió sobre el procedimiento fue que si llenarían el formato de evaluación una vez contestado el reactivo o hasta que acabaran los quince. Aquí se les explicó a los jueces que primero contestarían los 15 reactivos y después llenarían el formato de evaluación, recordándoles que tenían que justificar sus observaciones.

Una tercera duda en el procedimiento sobre la revisión grupal fue si alguno de los jueces estaría tomando nota sobre las decisiones que se tomarían por consenso. La instructora comentó que ella tomaría nota del consenso, la cual se complementaría con la videograbación de la actividad, ya que ese registro les ayudaría a mejorar sus reactivos.

b) Términos empleados

Durante la exposición de la presentación en Power Point, la única duda que surgió de uno de los participantes fue el término jueces. El docente dijo que esa palabra no existía, por lo que la instructora se dio a la tarea de explicar las operaciones de juicio que realizan los jueces, cómo se maneja el término en la literatura de los investigadores tanto en España como el término "judgement" en inglés.

El mismo docente preguntó por el término "gema", se le explicó que se emplea para designar a los reactivos que brillan por el número y/o gravedad de los errores que contienen. Para lo cual el profesor complementó el término añadiendo que es necesario pulir esas gemas para que queden como diamantes.

Comentarios de los participantes

a) Crítica a la actividad

Dentro de las resistencias encontradas para desarrollar esta revisión, se ubica el comentario de un participante, al expresar que las X y las Y son sólo X o Y; con lo cual estaba manifestando su desacuerdo por considerar aspectos culturales o lingüísticos en la revisión de la prueba de Matemáticas II. Como respuesta, la instructora señaló la importancia de la participación de la profesora experta en español debido a que un examen de matemáticas no está exento de algún posible problema de redacción. Se hizo énfasis en lo "posible" argumentando que la instructora y psicómetra en su momento, no participaba en el panel de expertos con el prejuicio de que los ítems tenían error, sino más bien en el ánimo de ayudarlos a identificar errores, en caso de que existieran en esa prueba.

b) Rol como participantes

Antes de iniciar la evaluación individual, en la cual los participantes contestaron la prueba como si fueran los examinados, la lingüista manifestó su

preocupación por esa actividad, pues dijo no saber de matemáticas. La instructora le indicó que su labor se enfocaría solamente en aspectos lingüísticos y de formato, mientras los demás compañeros estarían contestando los ítems. Un docente de la UABC argumentó sobre la necesidad de la participación de la profesora de español debido a que los docentes que elaboraron los reactivos no eran expertos en redacción.

Durante la revisión individual, en el proceso de contestar el ítem como si fueran los examinados, los docentes tardaron en promedio más de dos minutos en contestar cada reactivo. De ahí que en el tiempo estimado para contestar el examen, no lograron terminar, por lo que la instructora tomó la decisión de que contestaran solamente 15 reactivos y que sobre ellos se trabajaría en la revisión grupal.

Tras la experiencia de contestar el examen como si fueran los alumnos, los jueces reflexionaron en voz alta sobre la actividad. Uno de ellos comentó que siempre supo que no contestaría más de 20 reactivos en ese tiempo. En este sentido, concluyeron que algunos de los reactivos necesitan cinco minutos para contestarse y no el minuto y medio que estaba contemplado al diseñarlos. Del mismo modo, los docentes de matemáticas relacionaron los reactivos que ponen en los exámenes parciales en el salón de clases con los ítems contenidos en esta prueba. Por tanto, un juez concluyó que la tendencia del examen colegiado estaba siendo la de una prueba de ejecución y no de opción múltiple.

Enseguida, los jueces anotaron las observaciones de cada reactivo en el formato de evaluación. Posteriormente se inició la revisión grupal. En esta parte del procedimiento, la instructora se convirtió en la moderadora de la discusión, agregándose al comité en el papel de psicómetra.

Reactivo 1

El primer comentario que hizo la profesora de español y en el que todos estuvieron de acuerdo fue de redacción, los jueces propusieron agregar el término “ecuaciones” o “expresiones” a la base del reactivo, porque estaba incompleta. En segundo lugar, la discusión giró en torno a lo que estaban evaluando. Los jueces argumentaron que el contenido a evaluar era la integración y no aritmética, por lo

que acordaron cambiar el exponente fraccionario del distractor A, por un número entero para evitar que los alumnos se detengan en esa parte del reactivo y demuestre su dominio sobre la derivación.

Reactivo 2

Hubo dos discusiones sobre este reactivo, en el primero participó la profesora de español, quien propuso modificar la redacción de la base del reactivo para hacerla más clara al alumno. A esto se añadió una discusión sobre el nivel cognitivo necesario para responder el ítem. Los docentes de matemáticas aclararon que el contenido a evaluar en ese reactivo es antiderivación y no el manejo de exponentes. Se consideró que si un alumno no contestaba correctamente el reactivo, quedaría la duda si el problema con el examinado era aritmética o integración. Para solucionar tal problema, los jueces acordaron cambiar el exponente de la función por un número entero.

Cabe señalar que durante esta discusión se integró el quinto juez que también había sido invitado a colaborar en el módulo, pero que debido a que presentó un examen de oposición en la misma Facultad, no había estado en la capacitación, ni en la revisión individual. Sin embargo, conocía los reactivos porque participó en su elaboración.

Reactivo 3

En un caso similar al de los reactivos anteriores, en este ítem los jueces señalaron que de nuevo se estaban manejando exponentes fraccionarios, a lo que una profesora comentó que no es malo manejar algunos, pero no en todos los reactivos. Otro juez le respondió que en los libros de texto sí se incluyen problemas con números fraccionarios. No obstante, un tercer juez argumentó que estaba bien manejarlos en los exámenes tradicionales de ejecución, porque al desarrollar el procedimiento los docentes podrían identificar la parte en la que se equivocaron los examinados y valorar el resultado de la ejecución. Mientras que, en estos exámenes de opción múltiple, no podrían saber en qué parte fallaron los examinados; además, argumentó que para que el alumno demuestre si domina lo mínimo esencial de la integración no son necesarios los exponentes fraccionarios. Por eso, los jueces acordaron manejar únicamente exponentes enteros.

De manera adicional, los jueces discutieron sobre el manejo de ese tema dentro del salón de clases. Concluyeron que el contenido si bien es parte del programa, no es visto por todos los docentes.

Reactivo 4

Una vez más, la discusión retoma a los exponentes fraccionarios, por lo que los jueces recomiendan manejar solo exponentes enteros, para evitar que los estudiantes “naveguen en aguas pantanosas” y “truenen”.

Reactivo 5

En este reactivo la lingüista formuló una sugerencia de redacción, en la que se agregara una coma después del término “integral”, a lo que los profesores de matemáticas y la psicómetra estuvieron de acuerdo.

Reactivo 6

De la misma manera que en el ítem anterior, los jueces sugirieron agregar la coma, y observaron que la redacción de la base del reactivo era la misma que el ítem anterior. Comentaron si era “prudente”, “válido” o “bien visto” emplear una redacción idéntica, a lo cual la psicómetra preguntó si se estaba evaluando el mismo contenido y los profesores de matemáticas dijeron que sí. Para resolver esta inquietud, los jueces se dieron a la tarea de proponer otro enunciado para la base del reactivo.

Una observación más fue el formato en que se presentaba el cociente del distractor A. Los jueces en consenso sugirieron que una línea horizontal separara al numerador del denominador.

Reactivo 7

Los jueces identificaron dos errores: el formato de las gráficas que presenta y la redacción de la base del reactivo. En primer lugar, una gráfica estaba junto a la base del reactivo; por lo que dos jueces propusieron ubicarla debajo del enunciado. También comentaron que si los números que contenía la gráfica eran necesarios para los examinados, debían tener un tamaño más grande para ser legibles. Ante estos dos comentarios un tercer juez argumentó que el tamaño y posición de las gráficas se debía a la cantidad de hojas que tendría el examen, así como al número de copias necesarias para su aplicación.

Continuando con el formato, un juez señaló que a todas las gráficas les faltaba el eje de las "Y", y que los examinados se podrían confundir porque estaban acostumbrados a ver los ejes coordinados completos. Como resultado de estas observaciones, los jueces propusieron que se hicieran los cambios correspondientes.

El segundo error, identificado por la profesora de español, fue de redacción. Después de discutir entre los jueces y convencer a uno que no estaba de acuerdo, en consenso propusieron reducir el enunciado, quitar algunas palabras y hacerlo más claro para el estudiante.

Reactivo 8

El error identificado por los jueces consiste en que la base del reactivo está muy pegado al encabezado de la página, sugieren que se le de un poco más de espacio entre los textos.

Reactivo 9

Durante la revisión del reactivo 9 los jueces comentaron un error de ortografía, por lo que fue necesario poner el acento a la pregunta "¿cuál?". A esto se añadió un segundo comentario, que el ítem necesitaba más de un minuto y medio para contestarse. Además, que la sumatoria que exige al alumno es complicada, por lo que sugirieron una operación más sencilla ya que la del ítem requería que el examinado esté familiarizado con simbologías exponenciales. Los profesores de matemáticas comentaron que los ejercicios que se ven en clase no tienen relación con este ítem, por lo que sugirieron que se cambie por un reactivo más sencillo que apele al mismo contenido.

Reactivo 10

En este reactivo la profesora de español comentó que faltaba un acento en la base del reactivo. Los profesores de matemáticas retomaron la sugerencia y expresaron que no todos entendían lo mismo como estaba la redacción de la base del reactivo. Por lo que en consenso sugirieron que se modificara y quedara "El desarrollo de la siguiente expresión _____ es..."

Un segundo error identificado en el ítem fue que en los distractores, en vez de emplear comas, debían llevar signos de suma (+).

Reactivo 11

Al iniciar la discusión de este reactivo, uno de los jueces comentó que no pudo contestarlo; otro añadió que no estaba la respuesta correcta y que cuando se hizo la revisión informal ya habían acordado cambiar el reactivo por una representación gráfica. Sin embargo, los profesores de matemáticas se percataron de que el cambio se hizo pero por un reactivo más difícil. Consecuentemente, los jueces propusieron que se eliminara el reactivo ya que el contenido a evaluar escasamente se revisaba en el salón de clase y que además no contenía la respuesta correcta.

Enseguida, la profesora de español señaló que era necesario quitar un punto que aparecía después de un signo de interrogación. Por último, un profesor de matemáticas propuso que se completara “a la integral definida” ya que la base del reactivo no contenía esa frase.

Reactivo 12

En este ítem, la profesora de español propuso que se agregara a la base del reactivo “las integrales”, debido a que solo decía “dadas” y después se presentaban las integrales. Hubo una discusión por parte de los profesores de matemáticas, quienes objetaban que en los libros de texto se acostumbraba escribirlo así, aunque estaban de acuerdo en que parecía una frase incompleta y en consenso decidieron aceptar la sugerencia.

Después, los profesores de matemáticas comentaron sobre los distractores. Observaron que los límites de las integrales manejaban solamente números 5s y 7s; esto les daba pistas a los examinados, y les ayudaba a responder más fácilmente. Por ello, propusieron cambiar las opciones de respuesta, colocando en los distractores diferentes números.

Reactivo 13

La profesora de español y un profesor de matemáticas sugirieron dejar un espacio mayor entre las opciones de respuesta C y D, porque las integrales lo requieren para facilitar la lectura por parte del examinado. Los jueces estuvieron de acuerdo en hacer el cambio. Posteriormente la profesora de español comentó que era necesario poner un punto y aparte en la base del reactivo antes de presentar la

gráfica. A esta observación un profesor de matemáticas dijo que se había hecho así por cuestiones de espacio; que los programadores les habían dicho que sería difícil separar las frases de la gráfica.

A continuación, uno de los jueces comentó que las gráficas, al ser elaboradas en una calculadora graficadora, presentaban en la parte superior de la imagen los títulos y subtítulos, lo cual podría distraer a los examinados. Así, sugirieron cortar F1, F2 y F3.

De la misma manera, señalaron que la leyenda de los ejes de las "Y" estaba desfasada, por lo que propusieron que se moviera para no distraer al examinado tratando de ubicar dicho eje.

Reactivo 14

La primera observación hecha en este reactivo es el tiempo que requiere para contestarlo. Dos jueces estuvieron de acuerdo en que el ítem exigía varias operaciones para ser contestado. No obstante, un tercer juez opinó que no le parecía mucho trabajo porque demandaba solo tres procedimientos al examinado.

En segundo lugar, los jueces comentaron sobre la redacción, sugirieron que se mejorara y se redujera la base del reactivo, ya que tenía palabras y signos que no eran necesarios. Durante la discusión que se desarrolló en este reactivo, uno de los jueces –profesor de matemáticas- reflexionó sobre la finalidad de esta revisión; comentó que además de ubicar el contenido y el nivel cognitivo que demandaba, era importante identificar las fallas para poderlas corregir. De manera adicional, hasta este momento de la revisión, los jueces identificaron la falta de reactivos en la prueba para evaluar el dominio de las integrales definidas.

Reactivo 15

En este ítem, los jueces discutieron sobre la redacción de la base del reactivo. Acordaron que la pregunta iniciaría en el segundo renglón, después de la frase "a continuación". También señalaron que el encabezado de la prueba y el reactivo estaban muy juntos, dijeron que era necesario separarlos.

Por último, dialogaron respecto a la gráfica y acordaron respetar las decisiones acordadas en el reactivo 13.

Sesgo cultural y de género

Durante la discusión, los jueces comentaron que era importante la cuestión de género para la revisión de ítems. Sin embargo expresaron que para el examen que los ocupaba, no aplicaba ese tipo de validez. No obstante, agregaron que podría considerarse la validez cultural en exámenes que se aplican en la misma unidad académica pero que evalúan otro tipo de contenido.

Además, los comentarios registrados con la videograbación demuestran que el entrenamiento permitió que los jueces percibieron de manera diferente el sesgo cultural y de género, a pesar de que no hayan detectado esa problemática en la muestra de 15 reactivos evaluados.

Sugerencias y recomendaciones

Formato de evaluación de reactivos.

Uno de los participantes preguntó en qué parte del formato pondría una observación sobre un posible tiempo excesivo para contestar el reactivo. La instructora indagó si era cuestión de dificultad y el participante lo negó; comentó que podría ser un reactivo no muy difícil pero sí laborioso para contestarlo. Para atender esta necesidad, la instructora les pidió que en el espacio destinado a responder la afirmación número 9 del formato (el cual no sería utilizado en esta ocasión porque no se contaba con todos los ítems del examen) escribieran algún otro error que no se mencionara en algún otro enunciado del formato.

Palabras finales

Al finalizar la sesión, los jueces intercambiaron comentarios sobre la experiencia. Uno de ellos expresó que esta actividad le dio herramientas para ver los exámenes de diferente manera; también que es muy importante cuidar la redacción, las imágenes a emplear en los reactivos y dejar pasar un tiempo para que se pueda apreciar al ítem de una manera crítica. Además, reconoció el trabajo de la profesora de español quien a pesar de que no dominaba el contenido de la prueba, sí los apoyó en la revisión ortográfica y de redacción.

Finalmente, la instructora cerró el curso agradeciendo la participación de los profesores, así como por darse la oportunidad para reflexionar sobre los exámenes, la importancia que tienen, el impacto en la vida de los examinados, los derechos de éstos últimos y del esfuerzo que se está haciendo para tener pruebas justas en el

aspecto sociocultural, cognitivo y curricular. Por último reconoció la labor que realizaron en el módulo, y cómo se dieron el tiempo de reflexionar y analizar sobre sus exámenes para mejorarlos.

Anexo No. 10

Resultados del análisis de los ítems mediante el software del IIDE-UABC

Resultado del ítem análisis de los 15 reactivos revisados y modificados

	Ítem	Dificultad	Discriminación
	1	0.63	0.53
	2	0.46	0.41
	3	0.27	0.44
	4	0.55	0.6
	5	0.42	0.61
	6	0.29	0.56
	7	0.25	0.31
	8	0.22	0.4
	9	0.43	0.35
	10	0.76	0.32
	11	0.73	0.31
	12	0.61	0.44
	13	0.57	0.3
	14	0.36	0.44
	15	0.39	0.4
Promedio total:		0.46	0.43
Total Pregunt:		15	
Total alumnos:		344	
Confiabilidad:		0.53	

# de Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
Correctas	D	B	C	C	B	A	D	A	A	B	C	D	C	B	C
Total A	27	135	123	79	58	99	93	75	148	22	40	70	53	125	122
Total B	39	159	80	40	143	47	60	191	39	261	29	38	22	125	32
Total C	60	22	91	189	27	150	104	47	65	43	252	25	197	48	134
Total D	217	26	49	34	115	46	86	30	92	17	21	209	72	45	55
Total no contesto	1	2	1	2	1	2	1	1	0	1	2	2	0	1	1
% A	0.08	0.39	0.36	0.23	0.17	0.29	0.27	0.22	0.43	0.06	0.12	0.2	0.15	0.36	0.36
% B	0.11	0.46	0.23	0.12	0.42	0.14	0.17	0.56	0.11	0.76	0.08	0.11	0.06	0.36	0.09
% C	0.17	0.06	0.27	0.55	0.08	0.44	0.3	0.14	0.19	0.13	0.73	0.07	0.57	0.14	0.39
% D	0.63	0.08	0.14	0.1	0.33	0.13	0.25	0.09	0.27	0.05	0.06	0.61	0.21	0.13	0.16
% no contesto	0	0.01	0	0.01	0	0.01	0	0	0	0	0.01	0.01	0	0	0
Dificultad	0.63	0.46	0.27	0.55	0.42	0.29	0.25	0.22	0.43	0.76	0.73	0.61	0.57	0.36	0.39
Discriminación	0.53	0.41	0.44	0.6	0.61	0.56	0.31	0.4	0.35	0.32	0.31	0.44	0.3	0.44	0.4

Resultados del ítem análisis de 17 que no fueron revisados

	Ítem	Dificultad	Discriminación
1	16	0.38	0.39
2	17	0.42	0.44
3	18	0.35	0.16*
4	19	0.47	0.42
5	20	0.45	0.3
6	21	0.62	0.3
7	22	0.18	0.23
8	23	0.55	0.32
9	24	0.46	0.47
10	25	0.42	0.35
11	26	0.43	0.35
12	27	0.26	0.33
13	28	0.42	0.51
14	29	0.26	0.25
15	30	0.24	0.29
16	31	0.37	0.37
17	32	0.23	0.2
Promedio total:		0.38	0.34
Total Pregunt:		17	
Total alumnos:		344	
Confiabilidad:		0.33	

# de Reactivo	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Reactivo	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Correctas	A	B	A	A	C	D	A	B	B	C	D	A	C	B	D	A	D
Total A	132	77	120	163	58	62	61	82	60	56	43	88	109	132	40	128	63
Total B	78	143	72	22	53	28	208	189	158	115	43	102	53	88	76	57	87
Total C	73	47	125	66	155	42	22	22	72	143	109	78	143	63	145	86	111
Total D	61	77	22	91	76	212	49	49	54	27	148	76	38	59	83	71	78
total no contesto	0	0	5	2	2	0	4	2	0	3	1	0	1	2	0	2	5
% A	0.38	0.22	0.35	0.47	0.17	0.18	0.18	0.24	0.17	0.16	0.13	0.26	0.32	0.38	0.12	0.37	0.18
% B	0.23	0.42	0.21	0.06	0.15	0.08	0.61	0.55	0.46	0.33	0.13	0.3	0.15	0.26	0.22	0.17	0.25
% C	0.21	0.14	0.36	0.19	0.45	0.12	0.06	0.06	0.21	0.42	0.32	0.23	0.42	0.18	0.42	0.25	0.32
% D	0.18	0.22	0.06	0.27	0.22	0.62	0.14	0.14	0.16	0.08	0.43	0.22	0.11	0.17	0.24	0.21	0.23
% no contesto	0	0	0.02	0.01	0.01	0	0.01	0.01	0	0.01	0	0	0	0.01	0	0.01	0.02
Dificultad	0.38	0.42	0.35	0.47	0.45	0.62	0.18	0.55	0.46	0.42	0.43	0.26	0.42	0.26	0.24	0.37	0.23
Discriminación	0.39	0.44	0.16*	0.42	0.3	0.3	0.23	0.32	0.47	0.35	0.35	0.33	0.51	0.25	0.29	0.37	0.2

Anexo No. 11
Resultados del análisis de ítems
mediante el ITEMAN

Anexos

15

MicroCAT (tm) Testing System Page 1
 copyright (c) 1982, 1984, 1986, 1988, 1993 by Assessment Systems Corporation
 Item and Test Analysis Program -- ITEMAN (tm) Version 3.50

Item and Test Analysis Program -- ITEMAN (tm) Version 3.50

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics				
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser. Key
1	0-1	.63	.49	.40	A	.08	.12	.02	-.17
					B	.11	.21	.02	-.23
					C	.17	.26	.07	-.19
					D	.63	.40	.88	.40
					other	.00	.00	.00	-.06
2	0-2	.46	.46	.37	A	.39	.51	.26	-.22
					B	.46	.25	.71	.37
					C	.06	.09	.00	-.14
					D	.08	.14	.02	-.16
					other	.01	.00	.00	.00
3	0-3	.26	.42	.42	A	.36	.42	.28	-.13
					B	.23	.35	.12	-.25
					C	.26	.10	.52	.42
					D	.14	.14	.08	-.05
					other	.00	.00	.00	-.02
4	0-4	.55	.57	.48	A	.23	.32	.07	-.22
					B	.12	.22	.01	-.28
					C	.55	.33	.89	.48
					D	.10	.12	.02	-.17
					other	.01	.00	.00	-.06
5	0-5	.42	.58	.50	A	.17	.16	.14	-.07
					B	.42	.18	.76	.50
					C	.08	.12	.01	-.12
					D	.33	.55	.09	-.40
					other	.00	.00	.00	-.02
6	0-6	.29	.55	.54	A	.29	.08	.63	.54
					B	.14	.18	.09	-.13
					C	.44	.57	.20	-.33
					D	.13	.17	.07	-.11
					other	.01	.00	.00	-.03
7	0-7	.25	.29	.27	A	.27	.32	.20	-.12
					B	.17	.24	.09	-.14
					C	.30	.35	.32	-.03
					D	.25	.10	.39	.27
					other	.00	.00	.00	.00
8	0-8	.22	.38	.43	A	.22	.07	.45	.43
					B	.56	.58	.47	-.13
					C	.14	.19	.04	-.16
					D	.09	.14	.03	-.19
					other	.00	.00	.00	-.06

Page 1

Anexos

9	0-9	.43	.35	.33	15					
					A	.43	.24	.59	.33	*
					B	.11	.18	.09	-.13	
					C	.19	.26	.11	-.19	
					D	.27	.33	.21	-.11	
					other	.00	.00	.00		
10	0-10	.76	.28	.29	A	.06	.09	.02	-.14	*
					B	.76	.62	.89	.29	
					C	.13	.19	.05	-.18	
					D	.05	.10	.03	-.13	
					other	.00	.00	.00	-.04	
11	0-11	.73	.31	.28	A	.12	.20	.03	-.20	*
					B	.08	.11	.03	-.11	
					C	.73	.59	.91	.28	
					D	.06	.09	.02	-.11	
					other	.01	.00	.00	.02	
12	0-12	.61	.41	.35	A	.21	.31	.13	-.20	*
					B	.11	.21	.03	-.22	
					C	.07	.10	.05	-.10	
					D	.61	.38	.79	.35	
					other	.00	.00	.00	.02	
13	0-13	.57	.29	.30	A	.15	.22	.07	-.18	*
					B	.07	.12	.05	-.13	
					C	.57	.42	.71	.30	
					D	.21	.24	.17	-.13	
					other	.00	.00	.00		
14	0-14	.36	.43	.32	A	.36	.35	.28	-.03	*
					B	.36	.18	.61	.32	
					C	.14	.20	.08	-.12	
					D	.13	.27	.02	-.28	
					other	.00	.00	.00	-.06	
15	0-15	.39	.40	.33	A	.35	.45	.27	-.14	*
					B	.09	.15	.06	-.14	
					C	.39	.16	.56	.33	
					D	.16	.22	.11	-.13	
					other	.01	.00	.00	-.08	

There were 344 examinees in the data file.

scale statistics

scale:	0
N of Items	15
N of Examinees	344
Mean	6.927
Variance	6.963
Std. Dev.	2.639
Skew	0.390
Kurtosis	-0.220
Minimum	1.000
Maximum	14.000
Median	7.000
Alpha	0.563
SEM	1.744

15

Mean P 0.462
Mean Item-Tot. 0.375
Mean Biserial 0.488
Max Score (Low) 5
N (Low Group) 113
Min Score (High) 9
N (High Group) 95

SCALE # 0

Score Distribution Table

Number Correct	Freq- uency	Cum Freq	PR	PCT	
1	2	2	1	1	#
2	5	7	2	1	#
3	22	29	8	6	#####
4	37	66	19	11	#####
5	47	113	33	14	+#####
6	48	161	47	14	#####
7	47	208	60	14	#####
8	41	249	72	12	#####
9	42	291	85	12	#####
10	19	310	90	6	+#####
11	14	324	94	4	####
12	10	334	97	3	###
13	5	339	99	1	#
14	5	344	99	1	#
15	0	344	99	0	+

-----+-----+-----+-----+-----+
5 10 15 20 25
Percentage of Examinees

Anexos

17
Item and Test Analysis Program -- ITEMAN (tm) Version 3.50

Seq. No.	Scale -Item	Item Statistics			Alternative Statistics					Key
		Prop. Correct	Disc. Index	Point Biser.	Alt.	Prop. Total	Endorsing Low	High	Point Biser.	
1	0-1	.38	.33	.32	A	.38	.25	.58	.32	*
					B	.23	.23	.17	-.06	
					C	.21	.30	.14	-.14	
					D	.18	.22	.10	-.19	
					other	.00	.00	.00		
2	0-2	.42	.38	.37	A	.22	.30	.09	-.22	*
					B	.42	.27	.64	.37	
					C	.14	.20	.12	-.11	
					D	.22	.24	.14	-.13	
					other	.00	.00	.00		
3	0-3	.35	.15	.18	A	.35	.28	.43	.18	*
					B	.21	.25	.16	-.08	
					C	.36	.38	.36	-.09	
					D	.06	.07	.05	-.02	
					other	.01	.00	.00	-.05	
4	0-4	.47	.39	.37	A	.47	.29	.68	.37	*
					B	.06	.07	.07	-.02	
					C	.19	.25	.11	-.17	
					D	.26	.38	.14	-.24	
					other	.01	.00	.00	-.06	
5	0-5	.45	.26	.25	A	.17	.24	.12	-.13	*
					B	.15	.20	.15	-.06	
					C	.45	.30	.56	.25	
					D	.22	.27	.16	-.14	
					other	.01	.00	.00	.05	
6	0-6	.62	.27	.26	A	.18	.21	.14	-.10	*
					B	.08	.14	.03	-.18	
					C	.12	.20	.10	-.11	
					D	.62	.45	.73	.26	
					other	.00	.00	.00		
7	0-7	.18	.19	.29	A	.18	.11	.30	.29	*
					B	.60	.70	.44	-.23	
					C	.06	.05	.06	-.02	
					D	.14	.13	.18	.02	
					other	.01	.00	.00	.01	
8	0-8	.55	.26	.24	A	.24	.33	.19	-.14	*
					B	.55	.40	.66	.24	
					C	.06	.09	.03	-.13	
					D	.14	.18	.11	-.09	
					other	.01	.00	.00	.05	
9	0-9	.46	.45	.38	A	.17	.25	.06	-.21	*
					B	.46	.25	.70	.38	
					C	.21	.26	.15	-.11	
					D	.16	.24	.09	-.17	
					other	.00	.00	.00		
10	0-10	.42	.32	.32	A	.16	.21	.10	-.15	

Page 1

Anexos

					17							
					B	.33	.38	.25	-.14			
					C	.42	.27	.59	.32	*		
					D	.08	.13	.05	-.14			
					other	.01	.00	.00	-.02			
11	0-11	.43	.28	.26	A	.13	.15	.11	-.04			
					B	.13	.16	.09	-.07			
					C	.32	.41	.23	-.19			
					D	.43	.29	.57	.26	*		
					other	.00	.00	.00	.01			
12	0-12	.26	.35	.36	A	.26	.10	.45	.36	*		
					B	.30	.41	.19	-.21			
					C	.23	.25	.18	-.06			
					D	.22	.23	.18	-.09			
					other	.00	.00	.00				
13	0-13	.42	.52	.42	A	.32	.44	.13	-.27			
					B	.15	.21	.08	-.10			
					C	.42	.19	.71	.42	*		
					D	.11	.16	.08	-.11			
					other	.00	.00	.00	-.10			
14	0-14	.26	.21	.18	A	.38	.36	.34	-.02			
					B	.26	.20	.41	.18	*		
					C	.18	.23	.13	-.05			
					D	.17	.21	.12	-.13			
					other	.01	.00	.00	-.02			
15	0-15	.24	.27	.31	A	.12	.19	.06	-.17			
					B	.22	.27	.15	-.15			
					C	.42	.41	.39	-.03			
					D	.24	.14	.41	.31	*		
					other	.00	.00	.00				
16	0-16	.37	.33	.30	A	.37	.21	.54	.30	*		
					B	.17	.18	.14	-.04			
					C	.25	.29	.22	-.11			
					D	.21	.32	.10	-.20			
					other	.01	.00	.00	.00			
17	0-17	.23	.22	.24	A	.18	.21	.12	-.10			
					B	.25	.28	.20	-.11			
					C	.32	.36	.30	-.03			
					D	.23	.14	.36	.24	*		
					other	.01	.00	.00	-.00			

There were 344 examinees in the data file.

Scale Statistics

Scale:	0
N of Items	17
N of Examinees	344
Mean	6.488
Variance	5.628
Std. Dev.	2.372
Skew	0.342
Kurtosis	0.103
Minimum	0.000

17

Maximum 15.000
 Median 6.000
 Alpha 0.348
 SEM 1.916
 Mean P 0.382
 Mean Item-Tot. 0.295
 Mean Biserial 0.384
 Max Score (Low) 5
 N (Low Group) 128
 Min Score (High) 8
 N (High Group) 106

SCALE # 0

Score Distribution Table

Number Correct	Freq- uency	Cum Freq	PR	PCT	
0	1	1	1	0	+
1	2	3	1	1	#
2	9	12	3	3	###
3	17	29	8	5	#####
4	37	66	19	11	#####
5	62	128	37	18	+#####
6	54	182	53	16	#####
7	56	238	69	16	#####
8	37	275	80	11	#####
9	35	310	90	10	#####
10	10	320	93	3	+###
11	16	336	98	5	#####
12	6	342	99	2	##
13	1	343	99	0	
14	0	343	99	0	
15	1	344	99	0	+
16	0	344	99	0	
17	0	344	99	0	

-----+-----+-----+-----+-----+
 5 10 15 20 25
 Percentage of Examinees