PROPUESTA PARA EL DISEÑO DE EXÁMENES SUMARIOS PARA EL NIVEL MEDIO SUPERIOR

LUIS ALBERTO ÁLVAREZ ALDACO, ANA ISABEL MIRAMONTES BUSH Y
MARTA TORRES INFANTE

Introducción

Algunos investigadores en didáctica y diseñadores de planes de estudios coinciden en la necesidad de innovar los procedimientos para efectuar las evaluaciones de los contenidos curriculares, sin embargo, poco se ha podido corroborar en las reformas educativas de diversos países (Solomon y Aikenhead, 1994).

La evaluación de contenidos puede convertirse en un obstáculo, sobre todo cuando los profesores desarrollan estos contenidos en las aulas, ya que son pocos los profesores están inmersos en los diferentes procesos que se realizan y sobre todo no están al día de cuales son los temas a evaluar y que no saben cómo evaluarlos (Bell, Lederman y Abd-El-Khalick, 2000).

Algunos especialistas en educación mencionan que debe existir coherencia entre la evaluación de los estudiantes y los contenidos curriculares (Hofstein, Aikenhead y Riquarts, 1988), para lo cual, señalan la necesidad de buscar alternativas a la evaluación tradicional, mediante instrumentos válidos y nuevos criterios de evaluación específicamente diseñados para la evaluación.

Esto refleja la necesidad de que las instituciones del nivel medio superior cuenten con un instrumento que cumpla con las expectativas de ser de alto impacto y que permita monitorear la calidad del aprendizaje de los contenidos curriculares.

En el país se carece de este tipo de instrumentos a gran escala, lo cual se confirma con los mencionado por la funcionaria de la SEP (Vázquez, 2007) “Se hace necesario contar con una evaluación que refleje de manera clara el aprendizaje del alumno, a partir del cual se puedan tomar medidas enfocadas a mejorar el proceso educativo”.

En este contexto, durante la última década se ha observado un creciente interés por la evaluación educativa y en particular por la evaluación de la calidad de la educación (Martínez y Colaboradores, 2000).

En diversos documentos que forman parte de la literatura reciente relativa a la investigación educativa nacional, se comenta la falta de una cultura de la evaluación generalizada y una escasa investigación sobre la evaluación del aprendizaje Jackson (1993). Lo anterior resulta extraño si se considera la importancia que representa el enorme subsistema de educación media superior como insumo para la Universidad y los diferentes posgrados.

El hecho de contar con un curriculum nacional para el nivel medio superior, como los señala (Nitko, 1994), fue diseñado para hacer racional el proceso educativo, resulta claro que también debe ser la base racional sobre la que se sustente la evaluación educativa. Sin embargo, no es posible utilizar una sola estrategia en una actividad tan delicada como es la evaluación, por lo cual, la propuesta se complementa con la metodología para la construcción de tests criteriales propuesta por Popham (1990); por el esquema metodológico aportado por Ronald (1984) para la construcción de tests criteriales así como con los planteamientos psicométricos propuestos por Kellaghan y Greaney para el caso de países en desarrollo (1992; 1995).

Cabe señalar que, hasta hace relativamente poco tiempo, la mayor parte de los exámenes de gran escala que se elaboraban eran de tipo normativo. Por ello, presentan un

nivel de desarrollo psicométrico superior al que tienen los criteriales, mismos que empezaron a desarrollarse en los años setenta (Popham, 1990)

Por la dimensión y por el poderoso impacto social que tienen sobre las vidas de los alumnos, profesores, padres, directivos escolares, autoridades educativas y la sociedad en general, los tests de gran escala plantean condiciones especiales que determinan que tanto su elaboración, como su aplicación y evaluación, deben ajustarse a rigurosos estándares de calidad tales como la definición de su uso y cobertura, la exhibición de evidencias de validez y confiabilidad, el uso de procedimientos estandarizados para la administración, calificación e interpretación de resultados, entre otros (Rudner, 1993; Joint Committee on Testing Practices, 1994; Popham, 1990).

Por ejemplo, el National Center for Research on Evaluation, Standards and Student Testing (CRESST, 1994-b), en Estados Unidos, desarrolló criterios para revisar la calidad técnica de las evaluaciones que apelan a la complejidad cognitiva (pensamiento crítico, solución de problemas y razonamiento), la calidad del contenido (contenido que represente un reto y sea importante), significatividad (que las tareas evaluativas valgan la pena y que el estudiante entienda su valor), propiedad del lenguaje (que resulte claro y al nivel del estudiante), transferencia y generabilidad (que permite generalizaciones válidas respecto a la habilidad para realizar otras tareas), justicia (que no da cabida a factores irrelevantes para el aprendizaje o no pretendidos y califica con equidad), confiabilidad (se considera que las respuestas a las preguntas representan consistentemente lo que el alumno sabe) y consecuencias pretendidas (que tiene los efectos deseados).

En el caso de nuestro país, hasta muy recientemente, no existían antecedentes relacionados con el establecimiento de estándares nacionales o estatales de evaluación o de

sistemas de captación de información como podrían ser los exámenes nacionales (Martínez, 1993).

Objetivos

Utilizar una metodología que permita diseñar exámenes sumarios de tipo criterial (EXSUBACH) de gran escala, para evaluar el aprendizaje que logran los estudiantes del nivel medio superior en Baja California, iniciar el proceso de validación del instrumento, empleando para ello una muestra de los planteles del CBBC y establecer la aplicación del instrumento a gran escala, como un mecanismo permanente para monitorear la calidad del aprendizaje en el nivel medio superior.

Método

La metodología propuesta presenta las siguientes etapas:

Sujetos: a) selección de los coordinadores de cada una de las etapas del diseño de exámenes sumarios; b) selección de tres maestros por asignatura para el diseño del examen sumario distribuido en tres etapas (18 profesores por etapa, en total 54 profesores) y c) población de alumnos (23,000 alumno) correspondiente al ciclo 2006-2 de los 19 planteles oficiales del Colegio de Bachilleres del Estado de Baja California

Instrumentos: a) manual para el diseño de exámenes sumarios prediseñado; b) programa oficial de estudio de las asignaturas implicadas; c) paquete grafico Visio 2003, diccionario electrónico técnico WordNet, traductor Thesauros, Office 2003, base de datos SQL; d) software para el análisis de índices prediseñado y e) paquete estadístico SPSS V. 13.

Procedimiento: Una vez definida la metodología general para construir y pilotear los exámenes sumarios, se procedió a formar un comité coordinador del examen, con funciones

de diseño general, capacitación, piloteo de instrumentos, análisis de datos y control de calidad, así como elaboración de materiales e informes. Para cada una de las etapas se nombró a un coordinador, el cual fue capacitado con anticipación, siendo el responsable durante el proceso de la etapa.

La primera etapa consistió en la selección y capacitación del un comité diseñador del examen, integrado por tres especialistas para cada una de las asignaturas (18 maestros por etapa), quienes fueron seleccionados a partir de su experiencia en la materia y una alta responsabilidad para el trabajo en equipo, mismos que se encargaron de los aspectos relativos al análisis curricular, en la elaboración de redes de contenido retícular (Robledo, Ledesma y Alvarado, 1983) y en el diseño de especificaciones de ítem, a fin de contar con un grupo de especialistas entrenados para construir el instrumento.

En la segunda etapa se formó el grupo de trabajo del comité diseñador de especificaciones, siendo los encargados de estructurar cada una de las especificaciones solicitadas por el primer comité (18 profesores), quienes diseñaron el total de especificaciones marcadas con un promedio de 65 especificaciones y 70 ítems por asignatura.

Para la tercera etapa se formo el comité elaborador de los ítems (18 profesores), quienes se encargaron del diseño de los ítems para cada una de las especificaciones, de tal manera que se estructuraron tres diferentes versiones con ítems equivalentes (190 ítems en promedio para cada asignatura).

Cuarta etapa, se efectuó una aplicación piloto a una muestra de estudiantes de cada una de las asignaturas trabajadas, para tal evento, se realizó una selección de maestros distintos a los diseñadores para reducir el sesgo, posteriormente, la información se vació en una base de datos para realizar el análisis psicométrico utilizando: el índice de dificultad,

análisis de los distractores, índice de discriminación y el coeficiente de correlación. Posteriormente se volvió a trabajar con el equipo encargado del diseño de ítem, realizando los ajustes necesarios y la estructuración de la versión final paras su aplicación a gran escala.

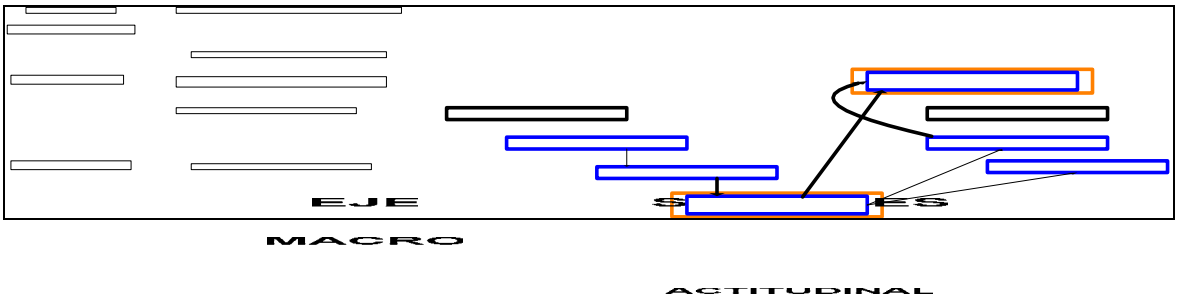
Quinta etapa, posterior a la aplicación, se realizo el análisis global de la versión utilizada con el objetivo de identificar y depurar el instrumento a través de los índices y criterios para pruebas a gran escala.

Análisis de resultados

El análisis curricular efectuado permitió hacer explícito el dominio de resultados de aprendizaje que establece el curriculum de cada una de las asignaturas del tronco común.

Los productos obtenidos en la primera etapa fueron: diseño de la retícula del contenido a evaluar (ver ejemplo de matemáticas III en la Figura 1), diseño de la tabla de especificaciones y tabla de justificaciones, el análisis consistió en representar gráficamente los contenidos de cada una de las asignaturas del tronco común, por eje curricular y grado educativo, haciendo explícitas las relaciones de servicio entre los contenidos. Para efectuar la estructuración se utilizó como base el programa de cada asignatura, pues en ella los contenidos presentaban ya un cierto nivel de organización. En la tabla de especificaciones se indican los contenidos que serán evaluados a través del instrumento.

Fig. 1. Muestra de la estructura de una retícula para matemáticas III



En la segunda etapa se diseñaron las especificaciones (ver ejemplo de matemáticas III en la Figura 2) para cada contenido a evaluar y el diseño de un ítem muestra.

Para tal evento, la información se vació en un formato, el cual agrupaba la información: datos de identificación del contenido a evaluar, comentario aclaratorio del contenido a evaluar, atributos relevantes de los estímulos que se presentarán al estudiante y el reactivo muestra. Este documento fue la base para la elaboración de los ítems correspondientes para cada una de las especificaciones.

Figura 2. Muestra de una especificación para matemáticas III

1. Datos de identificación del contenido a evaluar Curso: Matemáticas III Unidad: Conceptos Básicos de Geometría Analítica Tema: Subtema: P.1.1. Representación gráfica de puntos y segmentos rectilíneos dirigidos y no dirigidos en el plano.

2. Comentario aclaratorio acerca del sentido evaluativo del contenido: Es un contenido esencial ya que su dominio es fundamental para temas subsecuentes como

3. Atributos relevantes de los estímulos que se presentarán a los estudiantes. 3.1 Instrucciones para responder al reactivo: 3.2 Base del reactivo: La base del ítem presentará la definición o el concepto de un segmento rectilíneo dirigido y no dirigido y/o se presentarán las coordenadas de tres puntos y se solicitará al evaluado que los localice en una gráfica formalizada de puntos en un plano cartesiano o se presentara' una grafica con tres puntos, solicitando al examinado identifique sus coordenadas Se utilizarán letras mayúsculas para cada punto.

La tercera etapa consistió en la elaboración de ítems con base en el manual de especificaciones y en el entrenamiento recibido, el Comité Elaborador desarrolló un conjunto de 180 ítems para la prueba de conformidad con las normas, a fin de propiciar su validez.

Durante este procedimiento se presentaron diferentes tipos de interacciones entre los miembros de los comités y el coordinador, siendo las más comunes las que tenían por objeto hacer aclaraciones y retroalimentar las actividades. Al finalizar la elaboración, los responsables entregaron los ítems que diseñaron.

En general, la elaboración de los ítems duró de dos a tres semanas y en los casos en que los reactivos contenían dibujos, se prolongó una semana más.

Los reactivos se analizaron a la luz de las especificaciones elaboradas, se procedió a probarlos empíricamente ante una muestra de alumnos, una vez aplicados los modelos de examen, el comité coordinador procedió a capturar los resultados y a efectuar un análisis de ítems y de confiabilidad de los modelos, para estimar su calidad técnica. La información se registró en una base de datos, posteriormente se obtuvieron índices de dificultad apropiada para cada ítem y modelo mediante tres procedimientos: a) Se calculó el índice de dificultad o valor p del reactivo. Es decir, la proporción de examinados que contestaron correctamente el ítem; b) Se obtuvo el índice de discriminación o valor D del ítem. Es decir, la diferencia entre p para el grupo alto (27%) y p para el grupo bajo (27%) y c) Se estimo la calidad de los puntajes del test, mediante el coeficiente de consistencia interna de cada modelo de examen; es decir, el coeficiente alfa de Cronbach o kuder-Richardson 20, que puede ser considerado como la correlación promedio obtenida de todas las posibles estimaciones de confiabilidad, mediante división por mitades, de los reactivos de un modelo de examen (Popham, 1990, p. 133)

Una vez calibrados los ítems de los tres modelos, se aplico una de las versiones al total de la población (23,000 alumnos) del CBBC, posterior a su aplicación, se efectuó el análisis estadísticos referido a los diferentes índices (dificultad, discriminación, distractores e índice de confiabilidad (ver ejemplo de matemáticas III en la Tabla 1).

Tabla I. Análisis psicométrico del examen sumario de matemáticas III

INDICE DE DIFICULTAD	0.590
INDICE DE DISCRIMINACIÓN	0.311
ÍNDICE DE CONFIABILIDAD	0.811
TOTAL DE ÍTEMS	50
TOTAL DE ALUMNOS	23,000

Conclusiones

Los resultados obtenidos, nos permiten concluir que la propuesta metodología para el diseño de exámenes sumarios de tipo criterial es valida y confiable, ya que después de cuatro años de su implementación al interior del CBBC, ha permitido conocer el logro obtenido de los alumnos del nivel medio superior de una forma tangible.

Los productos generados en cada etapa, muestran su calidad estructural, ya que cada uno de estos fueron sometidos a una detallada revisión de contenido.

Se realizó un análisis psicométrico contra las especificaciones de ítems correspondientes para garantizar la calidad y validez de los ítems.

Los ítems no presentaron estereotipos étnicos o de género, los cual no favoreció a ningún grupo sobre otros.

El índice de dificultad (0.590) obtenido para matemáticas III, fue aceptable ya que es mayor que 0.05 y menor que 0.95; un índice de discriminación (0.311) es adecuado ya

que es mayor de 0.2; el índice de confiabilidad para este caso requería ser mayor o igual que 0.75, y se obtuvo una de 0.811.

Los distractores se ajustaron a un valor de 0.25, índices que son marcados por el (CRESST, 1994-b) y el comité internacional de estándares para pruebas de gran escala.

Bibliografía

Bell, R. L., Lederman, N. G. y Abd-El-Khalick F. (2000). Developing and acting upon one's conception of the nature of science: A follow-up study. *Journal of Research in Science Teaching, 37* (6), 563-581.

CRESST. (1994 b). Assessment Profile-State Summary. Evaluation Comment. National Center for Research on Evaluation, Standards and Student Testing.

Hofstein, A., Aikenhead, G. y Riquarts, K. (1988). Discussions over STS at the Fourth IOSTE Symposium. *International Journal of Science Education, 10* (4), 357-366.

Jackson, M., D. (1993). Desarrollo del Español como Primera Lengua. En II Congreso Nacional de Investigación Educativa. La Investigación Educativa en los Ochenta, Perspectiva para los Noventa. Estados del Conocimiento, cuaderno 9.

Joint Committee on Testing Practices. (1994). Code of Fair Testing Practices in Education. American Psychological Association.

Kellaghan, M. y Greaney, P. para el caso de países en desarrollo (1992; 1995). "Meliá, J. L. (2001) Teoría de la Fiabilidad y la Validez. Valencia: Cristóbal Serrano. www.uv.es/psicometría".

Martínez, F. (1993). Evaluación del Aprendizaje. Presentación del Estado del Conocimiento. En II Congreso Nacional de Investigación Educativa. Síntesis del Congreso Nacional Temático. D.F.

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

Nitko, A. J. (1994). A model for developing curriculum-driven criterion-referenced and norm-referenced national examinations for certification and selection of students. A paper

Popham, J. (1990). Modern Educational Measurement. A Practitioner's Perspective. MA. Allyn and Bacon. presented the 2nd International Conference on Educational Evaluation and Assessment of the Association for the Study of Educational Evaluation in Southern Africa, Pretoria, South Africa, July.

Robledo, J. M., Ledezma, R. y Alvarado, J. F. (1983). Reticulación: una estrategia para la elaboración de programas de estudio. UNAM. Facultad de Psicología. Tesis para obtener el grado de licenciatura.

Ronald, B. (1984). Test Evaluation. Disponible en Gopher ERIC/AE. 12/ 93.

Rudner L. (1993). Test Evaluation. Disponible en Gopher ERIC/AE. 12/ 93.

Solomon, I. y Aikenhead, R. (1994). La evaluación del aprendizaje. Tendencias y reflexión crítica. Revista Cubana de Educación Superior 2000, XX (1):47-62

Vázquez, M. j. (2007). Comunicado de prensa de la Republica, Febrero, 15, 2007 http://www.sep.gob.mx/.