UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA
SECRETARÍA DE APOYO A LA DOCENCIA

ANÁLISIS DEL PRIMER EXAMEN FINAL
ÁLGEBRA
SEMESTRE 2008-1

COORDINACIÓN DE EVALUACIÓN EDUCATIVA

ABRIL DE 2008

PRESENTACIÓN

Se presenta aquí un trabajo de análisis cuantitativo del examen final de Álgebra aplicado a alumnos inscritos en la asignatura en el semestre 2008-1.

El trabajo se ubica en el marco de la revisión de los procedimientos de evaluación del aprendizaje de la División de Ciencias Básicas de la Facultad de Ingeniería.

Luego de la introducción, en este texto, se presentan los resultados, conclusiones y sugerencias. Y al final se sitúan las tablas y figuras.

OBJETO

El instrumento, objeto de análisis, se compone de cuatro tipos de examen, que contienen 15 reactivos de respuesta estructurada con cuatro opciones. Se trata del Primer Examen Final de Álgebra fechado el 4 de diciembre de 2007.

Las fuentes de información son el archivo *final_Algebra*, de *WinZip* y el *Expediente del Primer Examen Final de Álgebra 2008-1* elaborados por el Departamento de Álgebra de la Coordinación de Matemáticas.

INDICADORES Y CRITERIOS

Se procedió al análisis de los reactivos en particular y de los exámenes en general.

De los reactivos se analiza:

- Grado de dificultad
- Poder de discriminación
- Frecuencia de elección de distractores

De los exámenes en su conjunto se analiza:

- Confiabilidad
- Equivalencia
- Validez predictiva

Grado de Dificultad (GD)

La dificultad de un reactivo se concibe como la proporción de estudiantes que responden correctamente a un ítem de una prueba. Se trata de un índice inverso: un grado de dificultad bajo implica un reactivo difícil, mientras que un grado de dificultad alto implica un reactivo fácil.

Para calcular la dificultad de un reactivo, simplemente se divide el número de alumnos que contestó correctamente el reactivo entre el número total de alumnos que contestó el reactivo. Respecto a los criterios para calificar a los reactivos según su grado de dificultad, no existe un criterio uniforme (Tristán, 1995); a manera ejemplo considérese los siguientes términos:

GD	Calificación
0.80 o más	Muy fácil
Entre 0.61 y 0.80	Fácil
Entre 0.41 y 0.60	Medio
Entre 0.21 y 0.40	Difícil
0.20 o menos	Muy difícil

Poder de Discriminación (PD)

El poder de discriminación es la capacidad que tiene un reactivo de diferenciar a los alumnos según la competencia que se está midiendo a través de todo el examen.

Para su obtención se comienza por identificar dos subconjuntos de alumnos, el "grupo superior" (el 27% con las puntuaciones más altas en el examen) y el “grupo inferior” (el 27% con las puntuaciones más bajas en el examen); y de ahí para cada reactivo se obtiene la diferencia de la proporción de alumnos del grupo superior que lo contestó acertadamente menos la proporción de alumnos del grupo inferior que también lo contestó acertadamente.

Tampoco en este caso existen criterios únicos; optaremos por el siguiente (Ebel y Fribie, 1986):

PD	Calificación
Superior a 0.40	Muy buenos reactivos
Entre 0.31 y 0.40	Buenos reactivos
Entre 0.21 y 0.30	Regulares, deben mejorarse
Inferior a 0.21	Deficientes, deben descartarse

Confiabilidad

La confiabilidad se define como el grado de consistencia de las mediciones que arroja la prueba o examen en su conjunto. Una buena confiabilidad estaría dada por una alta correlación entre las puntuaciones que resultaran de aplicar la misma prueba o examen en dos ocasiones consecutivas a las mismas personas.

Para efectos prácticos en este trabajo se ha acudido a la obtención del "Alfa de Crombach", que opera con las matrices de correlación de las puntuaciones de los reactivos, produciendo un coeficiente unitario. Un excelente coeficiente es 0.80, difícil de lograr en exámenes de rendimiento académico.

Equivalencia

La equivalencia está referida a la posibilidad de que dos o más pruebas o exámenes miden lo mismo y con el mismo nivel de dificultad: si es el caso se habla de pruebas paralelas.

Para determinar el grado de equivalencia entre dos exámenes es preciso aplicarlos bajo las mismas condiciones e igualando al máximo el nivel de competencia de los examinados.

En este estudio procedió a efectuar pruebas de hipótesis, basadas en la distribución *t de student* para determinar si las diferencias son significativas considerando un nivel de $p = 0.05$.

Validez

Se dice que un instrumento de medición es válido en la medida en que mide lo que pretende medir. Según el criterio de validación, existen diversas clases de validez. En este caso se analizará la validez predictiva, contrastando el resultado en el examen final con la calificación final de la asignatura y con el número de aprobadas en el primer semestre.

En los estudios realizados en este campo, véase por ejemplo Garritz y cols (1996) o Backohoff y Tirado (2000), a los coeficientes inferiores a 0.35 se les considera indicativos de una correlación baja, a los que fluctúan entre 0.35 y 0.45 se les denomina correlación moderada y a los superiores a 0.45 se les considera signo de una correlación elevada.

RESULTADOS

El número de alumnos reportados en las fuentes de información disponibles, ascendió a 767, los que se distribuyeron de la siguiente manera:

	Examen T1 (A o C)	Examen T2 (B o D)	
Matutino	211	196	407
Vespertino	182	178	360
	393	374	766

Se presenta en la Tabla 1 la distribución de alumnos por grupo. Se observa que son 47 grupos, lo que da un promedio de 16.3 alumnos por grupo.

Comportamiento de los reactivos

Los resultados de los reactivos en cuanto a grado de dificultad, poder de discriminación y frecuencia de elección de los distractores, se presentan en las tablas 2 a 4 y Figura 1 del Anexo.

En cuanto al grado de dificultad, se observa una distribución aceptablemente proporcionada, con una moda situada en el intervalo de 0.50 a 0.60 y una frecuencia descendente en dirección a los extremos.

En cuanto al poder de discriminación, se tiene que más del 90% de los reactivos son entre buenos y muy buenos; siendo que únicamente uno puede ser calificado como deficiente, tal es el caso del reactivo 11 del Examen Tipo B

Los distractores de máxima frecuencia - o sea, el que más fue seleccionado en cada caso- presentan un rango de elección de 4.7% a 40.1%.

Comportamiento de los exámenes: confiabilidad

La confiabilidad de los exámenes, según los coeficientes *alpha* obtenidos son diferentes, observándose una excelente confiabilidad en el Tipo A, media en los Tipo B y C y baja en el Tipo D, según puede verse en las tablas 5 a 9.

Comportamiento de los exámenes: equivalencia

Considerando como variable dependiente el número de aciertos, se podría concluir que no existe diferencia significativa ($p < 0.05$) entre los exámenes A y B como tampoco entre los exámenes C y D, por lo que podría declararse en ambos casos la equivalencia (tablas 9 y 10).

No obstante la probabilidad asociada a la diferencia entre los exámenes A y B es tal ($p = 0.059$) que se procedió a examinar con más detalle los resultados, encontrándose que al situar como variable dependiente la calificación obtenida en el examen, la diferencia entre ambos sí es significativa ($p = 0.047$). (Ver análisis complementario en tablas 11 a 13)

Así mismo, considerando como unidad de análisis los grupos de más de 10 alumnos del turno matutino y asignando a cada grupo un signo dependiendo de en que tipo de examen presentó mejor promedio, se concluye –al contrastar con la distribución binomial - que la diferencia entre los grupos es significativa ($p = 0.008$) (tablas 14 y 15).

Comportamiento de los exámenes: validez predictiva

En cuanto a la validez predictiva de los exámenes, utilizando como criterios las calificaciones de acta y el número de aprobadas en el primer semestre, los coeficientes de correlación revelan una validez elevada en los exámenes Tipo A, B y C y una entre moderada y baja en el examen Tipo D (tablas 16 y 17)

En las figuras 2 y 3 se ilustra la relación entre los resultados del examen final en su conjunto y el comportamiento correspondiente en los dos criterios considerados.

CONCLUSIONES

El Primer Examen Final de Álgebra del 4 de diciembre de 2007 es un examen compuesto por reactivos con una muy aceptable distribución en cuanto a grados de dificultad y una gran mayoría de reactivos con un adecuado poder de discriminación.

Se puede concluir que el examen Tipo A posee los mejores atributos como son una elevada confiabilidad y comprobada validez de predicción. Los exámenes Tipo B y C, aunque en menor grado, definitivamente también poseen buenos atributos en cuanto a confiabilidad y validez de predicción. El examen Tipo D, en cambio, se muestra deficiente en dichos renglones.

Debe concluirse también que las evidencias aportadas mediante el presente permiten confirmar la equivalencia entre los exámenes Tipo C y D pero no plenamente entre los exámenes Tipo A y B. Es por tanto indicado reparar en los procedimientos para garantizar la equivalencia entre los diversos tipos de examen.

SUGERENCIAS

Se recomienda efectuar análisis para ponderar los efectos en el grado de dificultad que ocasionan las diferentes variaciones que se hacen para elaborar reactivos “paralelos”. Es necesario también determinar la equivalencia entre los exámenes del turno matutino y del turno matutino.

Se sugiere también emprender otra clase de análisis, como puede ser de validez de contenido, pero en especial de análisis de los patrones de respuesta –observando las frecuencias de elección de los distractores- con el fin de detectar errores comunes de los estudiantes, lo que permite ajustar actividades y/o materiales de enseñanza.

Hay que reconocer, por último, que la experiencia y dedicación que regularmente se aplica en estas áreas a los exámenes comunes, por parte de los grupos académicos responsables, le confiere a la evaluación correspondiente atributos muy estimables, como son la claridad, la consistencia y la seguridad.
¡ Enhorabuena !

REFERENCIAS

Garrtiz, A y cols. *Antecedentes escolares y avances en la educación superior*. Asociación Nacional de Universidades e Institutos de Educación Superior. Temas de Hoy en la Educación Superior No 14 México, D.F. 1996

Backhoff Escudero, E., Larrazolo Reyna, N. y Rosas Morales, M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXCOBA) *Revista Electrónica de Investigación Educativa.* 2 (1), 2000.

Backhoff Escudero, E., Tirado Segura, F y Larrazolo Reyna, N. Ponderación de reactivos para mejorar la validez de una prueba de ingreso a la universidad. *Revista Electrónica de Investigación Educativa*. 3 (1), 2001.

Ebel, R.L. y Frisbie, D.A. *Essentials of educational measurement*. Englewood Cliffs, N.J.: Printice Hall, 1986

Lafourcade Análisis de Items Cap. 10, 1975

Tristán, A. Relaciones entre grado de dificultad y discriminación (1) (Primera parte: estudio del grado de dificultad) *Colección de Noticias ICI sobre Evaluación Educativa*. S.L.P México, 1995