UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA
DIVISIÓN DE CIENCIAS BÁSICAS

SECRETARÍA GENERAL
COORDINACIÓN DE EVALUACIÓN EDUCATIVA
FEBRERO DE 2002

EVALUACIÓN DEL EXAMEN DIAGNÓSTICO DE LA FACULTAD DE INGENIERÍA A TRAVÉS DE SUS INDICADORES DE NIVEL DE DIFICULTAD, GRADO DE DISCRIMINACIÓN, CONFIABILIDAD Y VALIDEZ (*)

INFORME PARA LA COMISIÓN DE VINCULACIÓN DE LA FACULTAD DE INGENIERÍA CON LA ESCUELA NACIONAL PREPARATORIA Y EL COLEGIO DE CIENCIAS Y HUMANIDADES

PRESENTACIÓN

Se presentan los resultados de la evaluación cuantitativa del examen diagnóstico aplicado a los alumnos de nuevo ingreso a la Facultad de Ingeniería de la Generación 2002. Este trabajo está inscrito en el marco de la *Comisión de Vinculación de la Facultad Ingeniería con la Escuela Nacional Preparatoria y el Colegio de Ciencias y Humanidades*, la que ha contribuido de manera fundamental en la elaboración del examen diagnóstico de la Facultad de Ingeniería, en los últimos dos años.

PROPÓSITO

El propósito de esta evaluación es observar el comportamiento del examen diagnóstico, a través del análisis cuantitativo de sus reactivos, en particular y en conjunto; para reconocer algunos de sus aciertos y limitaciones, y así proponer medidas para mejorarlo.

(*) Este trabajo ha sido realizado por la Coordinación de Evaluación Educativa de la Facultad de Ingeniería, con la colaboración del profesor Agustín Arreguín Rojas, del Colegio de Ciencias y Humanidades e integrante de la Comisión de Vinculación, y las valiosas aportaciones de Marlene Flores Mares y Alma Favila González, estudiantes de la Facultad de Psicología de la UNAM

OBJETO

El instrumento, objeto de análisis, se compone de dos tipos de examen, Tipo 1 (E-T1) y Tipo 2 (E-T2), que contienen 50 reactivos de respuesta estructurada con cinco opciones de respuesta. Los primeros 30 reactivos son del área de matemáticas, los siguientes 15 de física y los 5 restantes de química.

Los dos tipos de examen reúnen reactivos elaborados por profesores de las tres dependencias que integran la *Comisión de Vinculación*. Los reactivos se elaboraron con base en una tabla de especificaciones, que integra contenidos de los programas del bachillerato necesarios para iniciar los estudios de ingeniería.

La fuente de información para este trabajo es el archivo de resultados del examen diagnóstico aplicado a la Generación 2002 de la jefatura de la División de Ciencias Básicas; este archivo presenta las respuestas y puntuaciones de 1,794 alumnos (934 del E-T1 y 860 del E-T2) en los 50 reactivos correspondientes.

INDICADORES Y CRITERIOS

Para este análisis, se procedió a la obtención de los indicadores del comportamiento de los reactivos y del examen en su conjunto. En cuanto a indicadores de los reactivos se obtuvieron dos: grado de dificultad y poder de discriminación. En cuanto a indicadores del examen en su conjunto se obtuvieron también dos: confiabilidad y validez.

Grado de Dificultad

La dificultad de un reactivo se concibe como la proporción de personas 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 personas que contestó correctamente el reactivo entre el número total de personas que contestó el reactivo. Identificaremos a esta proporción como GD (grado de dificultad)

Respecto a los criterios para calificar a los reactivos de un examen según su grado de dificultad, no existe un criterio uniforme, pues mientras autores como Backhoff y cols. (2000) califican de "muy difíciles" a los reactivos con GD < 0.30, Lafourcade (1975) califica así a los de GD < 0.15.

---

* Tristán (1995) señala: "En rigor debería llamarse "grado de facilidad"; sin embargo la costumbre se ha hecho regla y el nombre no se cambia, aunque haya personas que estén pugnando porque dicha costumbre se modifique. Una vez aclarado el término, su interpretación no debe causar problema"

Para el presente análisis convenimos por la siguientes calificaciones:

GD	Calificación
Superior a 0.80	Muy fácil
Entre 0.51 y 0.80	Fácil
Entre 0.20 y 0.50	Difícil
Inferior a 0.20	Muy difícil

Poder de Discriminación

El poder de discriminación es la capacidad que tiene un reactivo de diferenciar a los estudiantes en términos de la competencia propia que se está midiendo a través de todo el examen. Identificaremos este indicador como PD (poder de discriminación)

Para obtenerlo se comienza por identificar dos subconjuntos de alumnos, uno, el "grupo superior" compuesto por el 27% que obtuvo las puntuaciones más altas en el examen, y el otro, "grupo inferior", compuesto por el 27% que obtuvo las puntuaciones más bajas. En seguida se obtiene la diferencia del número de alumnos del grupo superior que contestó acertadamente el primer reactivo menos el número de alumnos del grupo inferior que contestó acertadamente ese mismo reactivo; y finalmente, esta diferencia que se divide entre el número de alumnos del grupo más numeroso.

De esta manera se procede reactivo por reactivo, bajo el supuesto de que quien obtuvo una puntuación alta en toda la prueba tiene mayor probabilidad de contestar correctamente el reactivo. Así, entre más alto es su PD, el reactivo diferenciará mejor a los estudiantes con altas y bajas calificaciones.

En cuanto a los criterios para calificar los índices de discriminación, en este trabajo optaremos por los siguientes (Ebel y Fribie, 1986):

PD	Calificación
Superior a 0.40	Muy buenos reactivos
Entre 0.30 y 0.39	Buenos reactivos
Entre 0.20 y 0.29	Regulares, deben mejorarse
Inferior a 0.20	Deficientes, deben revisarse acuciosamente o 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 los 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 cada uno de los reactivos, produciendo un coeficiente unitario. Un excelente coeficiente es 0.85, difícil de lograr en exámenes de rendimiento académico.

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, se habla de diversas clases de validez. En nuestro caso contamos con dos criterios de validación: el promedio de bachillerato y los resultados en el primer examen parcial colegiado de cursos propedéuticos.

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 considera indicativos de una correlación moderada, y a los superiores a 0.45 de una correlación elevada.

RESULTADOS

Grado de Dificultad

De los 100 reactivos que componen ambos tipos de examen (ver Anexo), se tiene que ninguno presentó un grado de dificultad GD > 0.80 y que casi la cuarta parte presentó un GD < 0.20, en particular los siguientes:

Examen Tipo 1: R39, R44, R25, R16, R17, R1, R45, R18, R33, R37, R42, R35 Y R48
Examen Tipo 2: R40, R16, R32, R19, R25, R41, R22, R13, R45, R12

En la Tabla 1 se presentan las medias de los valores GD de los reactivos de las tres áreas en los dos tipos del examen. Lo que más llama la atención en los resultados de esta tabla, sin duda, es

Tabla 1 Medias del grado de dificultad (GD) por área y tipo del examen diagnóstico 2002

Área	Reactivos	Media
Matemáticas	1 a 30	0.32
Física	31 a 45	0.25
Química	46 a 50	0.36
Examen Tipo 1	1 a 50	0.30
Matemáticas	1 a 30	0.34
Física	31 a 45	0.35
Química	46 a 50	0.36
Examen Tipo 2	1 a 50	0.34

que mientras en el E-T1 la dificultad fue de 0.30, en el E-T2 fue de 0.34, observándose esta diferencia de manera muy pronunciada en el área de Física, en donde en el E-T1 la dificultad fue de 0.25 y en el E-T2 de 0.35

Para analizar esta situación con detalle, procedimos a aplicar directamente a los promedios de calificación de los alumnos de los dos grupos, la prueba estadística "*t*" *de Student,* en su aproximación para grupos independientes, encontrándose los resultados que se presentan en la Tabla 2, en donde se observan diferencias significativas ($p < .001$) entre E-T1 y E-T2, tanto en el promedio general ($t = -6.112$), como en el promedio del área de física. ($t = -14.135$). Dado estos resultados, definitivamente no se comprueba la equivalencia entre los dos tipos de examen.

Tabla 2. Resultados de la comparación de los dos tipos de examen a partir de la aplicación de la prueba T de Student a la distribución de los promedios obtenidos por los alumnos en el examen en su conjunto y en las distintas áreas

Área	E-T1 (N = 934) Media	D.S.	E-T2 (N = 860) Media	D.S.	t =	g.l.	p
Matemáticas	3.216	1.684	3.389	1.635	-2.200	1786.9	= 0.028
Física	2.469	1.509	3.471	1.493	-14.135	1782.7	< 0.001
Química	3.593	2.362	3.560	2.466	0.286	1764.3	= 0.775
General	3.030	1.410	3.431	1.366	-6.112	1787.4	< 0.001

Posteriormente a la realizaciòn de este análisis se encontró que el reactivo 37 del E-T1 fue mal calificado (se definió como correcta la opción "3", siendo que la correcta es la opción "5"). La falta de equivalencia, sin embargo permanece, puesto que luego de rectificar este error, se obtiene una media de 2.690 en el área de Física del E-T1 y de 3.096 en el promedio general del E-T1, las que comparadas con las medias correspondientes del E-T2 continúan reportando diferencias significativas: $t = -11.032$ ($p < 0.001$) respecto a Física y $t = -5.123$ ($p < 0.001$) respecto al promedio general.

Poder de Discriminación

Al aplicar los criterios establecidos a los índices de discriminación de los 100 reactivos que componen ambos tipos de examen, se tiene la siguiente distribución: 35 reactivos son muy buenos (17 del E-T1 y 18 del E-T2), 24 reactivos son buenos (13 del E-T1 1 y 11 del E-T2), 17 reactivos son regulares (9 del E-T1 y 8 del ET-2) y 24 reactivos son deficientes (11 del E-T1 y 13 del E-T2), De los 35 muy buenos en su poder de discriminación, hay 14 cuyo índice PD es aún superior a 0.50, estos son:

Examen Tipo 1: R26, R50, R28, R12, R3, R5 y R4
Examen Tipo 2: R6, R26, R8, R18, R4 y R2

Los 24 reactivos con PD inferior a 0.20 son:

Examen Tipo 1: R33, R35, R42, R17, R40, R36, (R37), R48, R20, R25 y R31
Examen Tipo 2: R22, R40, R27, R33, R48, R35, R12, R25, R36, R11, R13, R41 y R32,

En la Tabla 3 se presentan las medias de los valores PD de los reactivos de las tres áreas en los dos tipos del examen. Se observa que la media de los índices de los reactivos de los dos tipos de examen es 0.33, la que sin lugar a duda es buena. Se observa también que el PD de los reactivos de matemáticas asciende a 0.37 en el E-T2 y 0.38 en el E-T1.

Tabla 3 Medias y desviaciones estándar del poder de discriminación (PD) por área y tipo del examen diagnóstico 2002.

Área	Reactivos	Media	Desviación estándar
Matemáticas	1 a 30	0.38	0.14
Física	31 a 45	0.25	0.11
Química	46 a 50	0.32	0.16
Tipo 1	1 a 50	0.33	0.15
Matemáticas	1 a 30	0.37	0.16
Física	31 a 45	0.24	0.13
Química	46 a 50	0.33	0.13
Tipo 2	1 a 50	0.33	0.15

En los resultados de esta tabla se puede ver que el número de reactivos por área, aunque determinante no lo es todo, puesto que así como los 30 reactivos de matemáticas discriminan mejor que los restantes, los 5 de química discriminan mejor que los 15 de física, lo que se observa básicamente igual en los dos tipos de examen.

Para terminar la presentación de resultados de análisis cuantitativo de los reactivos, cabe identificar los 13 reactivos que combinan un alto grado de dificultad (GD < 0.20) y un bajo poder de discriminación (PD < 0.20), tales son (*):

Examen Tipo 1: R17, R25, R33, R35, R42 y R48
Examen Tipo 2: R12, R13, R22, R25, R32, R40 y R41

_____

* En esta relación se ha suprimido el R37 del E-T1 por la razón antes señalada.

Confiabilidad

En cuanto al análisis de la confiabilidad de los exámenes en su conjunto, los resultados son positivos en lo general, habiéndose obtenido un coeficiente Alfa de Crombach de 0.82 para el examen Tipo 1 y de 0.80 para el examen Tipo 2. De las tres áreas, a los resultados de matemáticas se les confiere una alta confiabilidad (Alfa = 0.79, en ambos tipos de examen), no así los de física (Alfa = 0.58 en E-T1 y Alfa = 0.48 en E-T2) y química (Alfa = 0.36 en E-T1 y Alfa = 0.40 en E-T2), en donde el reducido número de reactivos limita la confiabilidad de sus resultados.

Validez

Para al análisis de la validez procedimos a estimar los coeficientes de correlación de los resultados del examen diagnóstico con los promedios de bachillerato (lo que indaga sobre la validez concurrente) y con los promedios del primer examen parcial de cursos propedéuticos (lo que indaga sobre la validez predictiva), obteniéndose los coeficientes que se muestran en las Tabla 4 para el E-T1 y la Tabla 5 para el E-T2.

Hay que aclarar que el promedio del bachillerato no es el oficial sino el que reporta el alumno en el *Cuestionario Sociodemográfico y de Antecedentes Escolares* antes de comenzar los cursos. Hay que considerar también que el número de alumnos, en este caso. se reduce a 530 y 432 en los E-T1 y E-T2 respectivamente, dado que son los que están asignados a cursos propedéuticos y realizaron el primer examen parcial colegiado.

Tabla 4. Matriz de correlación ("r" de Pearson) entre las áreas del examen diagnóstico (E-T1), el promedio de bachillerato y los resultados del primer examen parcial colegiado de los cursos propedéuticos (N =530 alumnos)

	Diag. Mat.	Diag. Fis.	Diag. Qui.	Diag. Gral.	Bach. Prom.	Props. 1er par.
Diag. Mat.	1.000	0.498	0.389	0.941	0.280	0.305
Diag. Fis.		1.000	0.328	0.732	0.109	0.130
Diag. Qui.			1.000	0.551	0.197	0.243
Diag. Gral.				1.000	0.269	0.334
Bach. Prom.					1.000	0.306
Props. 1er par.						1.000

Tablas 5. Matriz de correlación ("r" de Pearson) entre las áreas del examen diagnóstico (E-T2), el promedio de bachillerato y los resultados del primer examen parcial colegiado de los cursos propedéuticos (N = 432 alumnos)

	Diag. Mat.	Diag. Fis.	Diag. Qui.	Diag. Gral.	Bach. Prom.	Props. 1er par.
Diag. Mat.	1.000	0.438	0.365	0.932	0.310	0.423
Diag. Fis.		1.000	0.279	0.697	0.152	0.075
Diag. Qui.			1.000	0.539	0.197	0.165
Diag. Gral.				1.000	0.295	0.382
Bach. Prom.					1.000	0.318
Props. 1er par.						1.000

Estos coeficientes indican que ambos tipos de examen se correlacionan escasamente con el promedio de bachillerato (r = 0.27 y r = 0.29) y un poco más con los resultados de los alumnos en el primer examen parcial (r = 0.33 y r = 0.38). Se destaca aquí la correlación (r = 0.423) entre los promedios de calificaciones de matemáticas del E-T2 y los resultados del primer examen parcial de los cursos propedéuticos, lo que apunta a favor de su validez predictiva.

CONCLUSIONES

En resumen, el examen diagnóstico aplicado a los alumnos de nuevo ingreso Generación 2002, es un examen compuesto por una mayoría de reactivos de alto grado de dificultad y buen poder de discriminación; el examen en su conjunto presenta propiedades que permiten afirmar la confiabilidad y en menor grado la validez de sus resultados. Hay que agregar que la confiabilidad, uno de los mejores atributos de este examen, se reduce en esta ocasión, dado que la equivalencia entre el E-T1 y E-T2 no se ha confirmado.

Al concluir este estudio, observamos varios puntos que conviene revisar en nuestro examen.

Comencemos por su dificultad. Es cierto que la dificultad de un reactivo es relativa. De hecho, el grado de dificultad de un reactivo, definido en los términos en que se ha definido en este trabajo, depende casi por completo del repertorio de cada alumno, o mejor dicho, de cada grupo de alumnos, al que se le aplique dicho reactivo.

Es muy probable que la mayoría de los reactivos que en este análisis reportan un alto grado de dificultad, sean reactivos no solamente bien estructurados sino también altamente representativos de lo que se desea evaluar, o sea, contenidos expresos de los programas de bachillerato, necesarios para el estudio de las carreras de ingeniería.

La dificultad, así entendida, es una noción limitada, pero su limitación no impide afirmar que un conjunto de reactivos aplicado a un grupo representativo de una población, sea más (o menos) difícil que otro conjunto de reactivos aplicados a otro grupo representativo de esa misma población, como ocurrió en este caso.

Es necesario asegurar la equivalencia no solamente entre los dos tipos de examen que se aplican a los alumnos de una misma generación, sino los que se aplican de un año a otro. Evidentemente, los resultados desiguales de esta aplicación no indican nada respecto a los resultados de aplicaciones de años anteriores.

Hay que reconocer, sin presunciones, que el conocimiento y la dedicación que tradicionalmente se aplica a este examen, por parte de los grupos académicos responsables, brinda una elevada dosis de seguridad y de confianza. Lo deseable es complementar esta práctica con medidas y acciones sistemáticas para incrementar la estabilidad del instrumento.

Estas medidas, de hecho, han comenzado a marchar bajo el impulso de la *Comisión de Vinculación*, nos referimos a la elaboración de la *tabla de especificaciones* y *el banco de reactivos*, cuya estructura debe consolidarse, sus funciones y usos extenderse, y sus contenidos revisarse.

Finalmente, en la revisión de contenidos a evaluar sería deseable considerar contenidos con énfasis en habilidades, no sólo en conocimientos, así como incluir contenidos *necesarios* para los estudios de ingeniería, que probablemente estén más explícitos en los programas de educación básica y media básica, que en los de educación media superior.

Lo anterior podría llevar a un instrumento mejor calibrado y a un diagnóstico más preciso del repertorio escolar de nuestros estudiantes.

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

A N E X O

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
FACULTAD DE INGENIERÍA
SECRETARÍA GENERAL
COORDINACIÓN DE EVALUACIÓN EDUCATIVA

EVALUACIÓN DEL EXAMEN DIAGNÓSTICO DE LA FACULTAD DE INGENIERÍA A TRAVÉS DE SUS INDICADORES DE NIVEL DE DIFICULTAD, GRADO DE DISCRIMINACIÓN, CONFIABILIDAD Y VALIDEZ

INFORME PARA LA COMISIÓN DE VINCULACIÓN DE LA FACULTAD DE INGENIERÍA CON LA ESCUELA NACIONAL PREPARATORIA Y EL COLEGIO DE CIENCIAS Y HUMANIDADES

FEBRERO DE 2002