ELEMENTOS DE ESTADISTICA AL SERVICIO DE LA EVALUACION DEL RENDIMIENTO

FLORALBA CANO DE BECERRA
Universidad Nacional de Colombia

Cuando un profesor tiene a su cargo una clase, un cursillo o una cátedra es su deber planear su contenido, atender a su desarrollo lógico y metodológico, idear las ayudas didácticas tendientes a presentar lo que desea enseñar en tal forma que asegure el mayor nivel de asimilación y, por supuesto, de manera ineludible, evaluar ese nivel de asimilación para determinar si ha alcanzado el grado deseado, porque haciéndolo, obtiene entre otras cosas, una medida objetiva de su propia capacidad para transmitir conocimientos.

Con frecuencia, la evaluación del rendimiento escolar o académico se lleva a cabo con instrumentos que a "juicio" del profesor le permiten obtener una "medida" del conocimiento alcanzado por el alumno, sin que ello lo lleve a preguntarse sobre el grado de confianza que tal medida le brinda. Las características de Validez y Confiabilidad se adscriben, hasta ahora, como indispensables sólo en los instrumentos estandarizados ya que tales conceptos no están lo suficientemente generalizados, como para que sean motivo de preocupación de un profesor común y corriente, cuando emite una calificación indicativa del caudal de conocimientos de un alumno.

Tal es la situación y tan primitiva, que la calificación que emite es el producto de transformaciones lineales, simplísimas, de una medida obtenida contando el número de respuestas correctas, cuyo patrón de referencia es "lo que el profesor piensa que el alumno *debe* saber", sin tener en cuenta "lo que realmente *pudo* aprender" o sin pensar, si no sería prudente, que la evaluación tuviera en cuenta estos dos patrones de referencia, o, ¿porqué uno y no otro? Las operaciones matemáticas que aquél tipo de evaluaciones conlleva se satisfacen con un razonamiento equiparable al de un niño de 5 años cuando reúne objetos y los redistribuye en grupos y, a pesar, de que a partir del primer año de primaria, el nivel del razonamiento que implica aprender lo que el profesor enseña, se torna elaborado y complejo, el profesor utiliza un estilo operacional de agrupación y repartición concretos para evaluar lo aprendido. Haciendo una analogía, es como si el médico, al hacer un examen físico, diera el dato sobre la estatura del paciente en "dedos" o en cuántas veces su palma de la mano cabe en el sitio donde lo ha colocado para hacerle el examen.

Este artículo está, entonces, encaminado a hacer algunas reflexiones sobre la evaluación del aprovechamiento dicente, a la vez que dar ejemplos prácticos de cómo resolver problemas es-

pecíficos con la ayuda de estadísticas elementales y para el caso de evaluaciones con exámenes objetivos; indicaciones precisas de cómo elaborar exámenes objetivos de conocimientos no son propósito aquí, puesto que sobre el particular hay suficiente literatura y de fácil acceso para profanos en cuestiones de psicometría.

¿Cómo y cuándo evaluar?

En un sentido general, siempre que el profesor emite juicios sobre un alumno, lo está evaluando; no importa si el juicio se traduce o no en una apreciación cuantitativa; sin embargo, cuando se piensa en evaluación, se hace en términos cuantitativos y se entiende como sistemática: así se toma cuando se piensa en exámenes de tipo tradicional o de los llamados objetivos. Uno y otro tipo tiene sus ventajas y sería aconsejable utilizarlos ambos, precisamente para controlar, entre sí, los vicios derivados de la utilización de un único tipo de examen. Si para el examen tradicional es indispensable que se piense en preguntas que no tengan sino una sola interpretación, porque de lo contrario se podrían obtener una serie de respuestas correctas que bien pudieran no referirse al conocimiento que se desea explorar en particular, para el examen objetivo, realizado con preguntas de escogencia múltiple, es necesario que entre las posibilidades de respuesta esté la "más acertada", y a juicio de varios peritos en la materia que se desea evaluar.

Hacer un examen objetivo no es fácil ni puede improvisarse ni elaborarse sin conocer las reglas de construcción: es preferible un examen tradicional si no se tienen los conocimientos técnicos o el tiempo para elaborar uno objetivo. Un examen tradicional mal elaborado es, en muchos casos, más útil que uno objetivo defectuoso, porque la respuesta libre en un examen tradicional da mayores indicios que el objetivo elaborado sin conocer los principios que rigen su construcción, ya que éste puede dar, simplemente, puntajes o calificaciones producto de la adivinación o del simple azar. Por otro lado, evaluaciones hechas con un solo tipo de exámenes son arriesgadas.

Hay ciertas indicaciones que pueden ayudar al respecto:

1. Preferir el examen tradicional para contenidos de poca extensión y los objetivos, para los contenidos mayores.

2. Evaluaciones frecuentes son preferibles a una única; aconsejable, un examen por cada unidad lógica dentro del contenido general.

3. Los exámenes de diagnóstico, aquellos que se realizan sin previo aviso y con el fin de averiguar qué contenidos se han asimilado o requieren mayores explicaciones por considerarlos básicos en la adquisición de nuevos, deben tener peso mínimo en la evaluación final porque las distribuciones de estos exámenes tienden a comportarse como las de azar. Sin embargo, no deben ignorarse porque son un incentivo del rendimiento y control en el desarrollo metodológico del programa.

4. Los exámenes objetivos deben tener como mínimo 25 preguntas, sobre todo cuando no se trata de preguntas probadas con anterioridad; debe recordarse que exámenes relativamente largos son más confiables que los cortos. Como regla general, puede estimarse que el número de preguntas será tal, que provea margen suficiente para medir el rendimiento del mejor de los alumnos. Si al construír un examen de 25 preguntas, se encuentra que un alumno lo ha contestado todo correctamente, se tendrá evidencia suficiente de haberse quedado "corto" en la medida; no se sabe, para una situación particular, cuál hubiera sido el nivel máximo de rendimiento.

5. Los exámenes de pregunta única y los de preguntas dependientes, aquellas en los que la respuesta de una de las

preguntas es conocimiento esencial y sin el cual no se puede resolver correctamente la siguiente, deben evitarse.

6. Exámenes que arrojen distribuciones normales no son aconsejables para evaluar conocimientos. Las de distribuciones asimétricas negativas, que discriminan mejor entre los alumnos de bajo rendimiento, deben preferirse para los exámenes finales. Las distribuciones asimétricas positivas, para los primeros exámenes de un curso, éstos, son los que discriminan mejor entre los alumnos de buen aprovechamiento.

¿Cómo escoger las mejores preguntas?

Esto se refiere, específicamente, a preguntas para construír exámenes objetivos. Alguna vez, un profesor tiene que comenzar a construír su primer examen objetivo; la primera vez que lo haga para una materia, las preguntas tendrán como única referencia su propio criterio y juicio, siempre y cuando se trate de una materia de la cual, sea él el único profesor; cuando hubiere más de un profesor, el examen debe construírse con el criterio de más de un profesor.

Una vez que un examen objetivo se haya aplicado por primera vez, ya se tienen indicios objetivos sobre: cómo se entendieron o interpretaron las preguntas, cuáles fueron contestadas por la mayoría, cuáles por ninguna persona, cuales por la minoría de los mejores, cuáles por los peores y cuáles por los mejores solamente.

Sin necesidad de hacer cálculos estadísticos, utilizando el sentido común y un gráfico para apreciar cómo se comportaron las preguntas, un profesor puede escoger las mejores preguntas y tendrá mayores seguridades de hacer mejores escogencias, que aquellas hechas al "ojo de la espontaneidad".

Los siguientes pasos pueden ayudar a escoger las mejores preguntas:

1. Después de corregir los exámenes, se ordenarán en forma descendente de acuerdo con el número de respuestas correctas.

2. En un papel cuadriculado se construye un tablero de doble entrada de tal manera que sobre el eje X se pueda disponer de una casilla para cada una de las preguntas y en el eje Y una para cada uno de los individuos.

3. Se anotan las contestaciones buenas de cada individuo, con una barrita vertical, en la casilla correspondiente a cada pregunta; es decir, se construye una matriz de respuestas correctas.

4. Con las sumas verticales se obtendrá el número de veces que ha sido correctamente contestada cada pregunta. Observando toda la matriz, se tendrá información de si las preguntas tienden a ser bien contestadas por los alumnos que obtuvieron altos puntajes, por los de bajos puntajes, por casi todos o por muy pocos.

5. Con las sumas horizontales se obtendrá el número de respuestas buenas para cada individuo. La suma de totales verticales y totales horizontales deben ser iguales, si la matriz ha sido bien levantada.

6. Se divide la matriz en tres grupos iguales y se indica en ella. El tercio superior constituirá el grupo *alto* y el tercio inferior, el *bajo*. El tercio medio no se tiene en cuenta. Así, se prepara el material para escoger las preguntas por un procedimiento que utiliza la comparación de grupos extremos; en éste caso, el tercio superior vs. tercio inferior. Este procedimiento será tanto más seguro cuantos más individuos intervengan en la matriz de puntajes; si se tiene información sobre grupos muy pequeños (de 15 individuos), se acopiará el material de grupos diferentes, siempre y cuando guarden condiciones iguales. Para situaciones habituales en una escuela con grupos cercanos a los 50 individuos, ya puede pensarse en hacer un análisis por grupos extremos.

GRAFICO No 1.

MATRIZ DE RESPUESTAS CORRECTAS

	PREGUNTAS											
	1	2	3	4	5	6	7	8	9	10	PB	
1	1	1	1	1	1	1	1	1		1	9	ALTO
2	1	1	1		1	1	1	1		1	8	
3	1	1	1		1	1	1	1		1	8	
4	1	1			1	1	1	1		1	7	
5	1		1		1	1	1	1		1	7	
6	1		1		1	1	1	1		1	7	
7	1		1		1		1	1			5	
8	1	1	1		1					1	5	
9	1				1		1	1		1	5	
10	1	1			1	1	1				5	
11	1		1			1				1	4	BAJO
12	1		1		1		1				4	
13	1		1					1		1	4	
14	1	1						1			3	
15								1	1		2	
	14	7	10	1	11	8	10	11	1	10	83	

INDIVIDUOS

7. Se descartan para futuras aplicaciones las siguientes preguntas:

a) Aquellas que han sido contestadas por casi todos los individuos, por ejemplo la pregunta número 1, a menos que se desee incluírla como medio de introducir, con confianza, al alumno en el examen; en éste caso, debe tenerse en cuenta que es como si el examen tuviera tantas preguntas menos cuantas de éstas se incluyen, para efectos de la fiabilidad de la medida.

b) Aquellas que han sido contestadas por menos del 10% del grupo total, a menos que se desee construír exámenes muy difíciles. Es bueno recordar que exámenes de dificultades extremas disminuyen las probalidades de ser contestadas correctamente, porque se sabe que aumentan las contestaciones por adivi-

nación. Es muy posible que las contestaciones a las preguntas 4 y 9 de la matriz de ejemplo se deban al azar.

c) Aquellas que han sido contestadas por igual proporción de individuos en el grupo *alto* que en el *bajo*. Ejemplo, preguntas 1 y 3. Este fenómeno se define todavía más y tiene mayor significación a medida que se trata con grupos numerosos.

d) Las contestadas por más individuos del grupo *bajo* que del *alto*. Ejemplo, pregunta 9.

Las preguntas que tienen las anteriores características no son deseables porque no discriminan los alumnos de alto de los de bajo rendimiento, puesto que han sido contestadas por todos, por "ninguno" o por igual número de altos que de bajos.

8. Se prefieren preguntas que han sido contestadas correctamente por mayor número de "altos" que de "bajos". Ejemplo, preguntas 2, 5, 6, 7 y 10. Cuyo orden de bondad es: 6, 2, 7, 10 y 5.

9. Se reordena la matriz de respuestas correctas de tal manera que las preguntas aparezcan ordenadas de las más fáciles (contestadas por la mayoría) a las más difíciles.

10. Se acumula, para futuras aplicaciones, las preguntas buenas, las que puedan mejorarse o las que se desee ensayar nuevamente. Se conserva, también,

GRAFICO No 2.

MATRIZ DE RESPUESTAS CORRECTAS.
(Reordenada)

		PREGUNTAS											
		1	5	8	7	10	3	6	2	4	9	PB	
INDIVIDUOS	1	1	1	1	1	1	1	1	1	1		9	ALTO
	2	1	1	1	1	1	1	1	1			8	
	3	1	1	1	1	1	1	1	1			8	
	4	1	1	1	1	1		1	1			7	
	5	1	1	1	1	1	1	1				7	
	6	1	1	1	1	1	1	1				7	
	7	1	1	1	1		1					5	
	8	1	1			1	1		1			5	
	9	1	1	1	1	1						5	
	10	1	1		1			1	1			5	
	11	1				1	1	1				4	BAJO
	12	1	1		1		1					4	
	13	1		1		1	1					4	
	14	1		1					1			3	
	15			1							1	2	
		14	11	11	10	10	10	8	7	1	1	83	
		X	V	?	V	V	X	V	V	X	X		

información de su comportamiento en diferentes exámenes y en diferentes grupos; por lo cual es necesario asegurarse sobre la estricta seguridad de los exámenes para evitar su conocimiento previo por parte de los estudiantes. En esta forma se construye un Banco de Preguntas que en lo sucesivo proveerá material para construír exámenes confiables, para los propósitos que se tengan en mente y de las características deseables.

11. Se incluyen en aplicaciones futuras aquellas preguntas con características no deseables después de una aplicación, para lo cual pueda hallarse explicación lógica y en esta forma asegurarse de si se está en lo cierto al sospechar que "hubiera podido servir si determinadas condiciones se presentan".

Se entiende que los pasos que aquí se han esbozado están dirigidos a profesores sin conocimientos de psicometría y que no desean o no encuentran justificado dedicarle demasiados cálculos a la tarea de la evaluación y teniendo en cuenta que solo se ha enfocado la bondad de la pregunta desde el punto de vista de la elección de la alternativa correcta en preguntas de selección múltiple y para el caso de que el comportamiento de las respuestas, en las alternativas no correctas, sea el deseado.

El comportamiento deseado en cuanto a las alternativas no correctas es precisamente el contrario al de la alternativa correcta. Se espera que las alternativas no correctas sean preferidas por los alumnos clasificados en el grupo *bajo (en la matriz de puntajes)*; son inocuas las que han sido escogidas por igual número de "altos" que de "bajos", por ninguna persona o solamente por los "altos". Cuando una pregunta de selección múltiple tiene alternativas inocuas, es como si tuviera tantas alternativas menos y por lo tanto aumentarán las posibilidades de contestaciones al azar para tal pregunta. Esto es necesario tenerlo en cuenta si se desea establecer el número de preguntas que pueden contestarse correctamente por azar en un examen objetivo construído con preguntas de selección múltiple.

¿Cómo tomar decisiones en un caso particular?

Se tomará como ejemplo un examen objetivo de 30 preguntas, con un valor de 15/50, aplicado a un grupo de 45 alumnos, construído con preguntas ya probadas y cuya distribución de puntajes fue la siguiente:

De la observación de la distribución de puntajes se puede, ya, sacar algunas conclusiones:

1. El examen diferenció claramente dos grupos: uno alto (A) con Modo = 18 y uno bajo (B) con Modo = 13.

2. Si el examen tenía 30 puntos y el mayor puntaje fue de 19, el instrumento de medida tuvo suficiente margen de seguridad para establecer el rendimiento máximo.

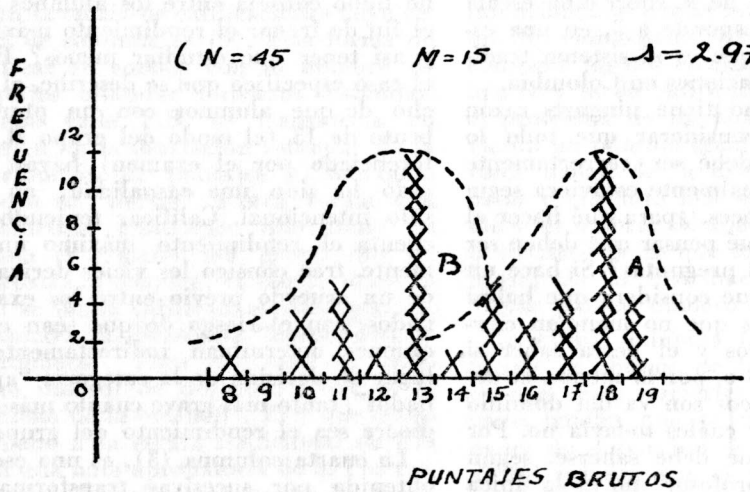
3. El rendimiento promedio del grupo (M = 15) está más cerca del Modo del grupo B que del Modo del grupo A.

4. Abstrayendo una distribución unimodal, el examen discriminó mejor dentro de los alumnos de bajo rendimiento; ésta condición se cumple aún para cada uno de los grupos que arrojó el examen. Esto es, diferencian mejor los alumnos menos buenos del grupo alto y los peores del grupo bajo. En un nivel práctico, se puede pensar en que son pocas las probabilidades de equivocarse al considerar que quienes tuvieron puntajes bajos tienen un nivel de aprovechamiento bajo. Se deberá pensar y "más de dos veces" la decisión de aproximar, por ejemplo, a puntajes superiores las calificaciones iguales o inferiores a 12, para el ejemplo que se describe.

5. El umbral de decisión para la categoría "pasaron" puede establecerse en 18 puntos, en 15 puntos o en 13 puntos.

GRAFICO No. 3

DISTRIBUCION DE PUNTAJES PARA UN EXAMEN OBJETIVO



A continuación se presentan sucesivas transformaciones del mismo puntaje bruto, conservando la escala de calificación (0 a 15), pero variando el procedimiento; se analizan las consecuencias y el patrón de referencia para cada procedimiento.

Transformaciones de un mismo puntaje bruto

(Conservando la escala y variando el procedimiento)

TABLA NUMERO 1

P.B.	(1)	(2)	(3)	(4)	(5)	(6)
19	9,5	15	13	15	13,5	12
18	9	14	12	13,5	12	12
17	8,5	13	11	13,5	12	12
16	8	12	10	12	10,5	9
15	7,5	11	9	10,5	10,5	9
14	7	10	8	9	9	9
13	6,5	9	7	7,5	9	9
12	6	8	6	6	7,5	6
11	5,5	7	5	6	7,5	6
10	5	6	4	4,5	6	6
9	4,5	5	3	3	4,5	3
8	4	4	2	3	4,5	3

En la primera columna aparece el número de respuestas correctamente dadas, o puntaje bruto (P.B.). El número de puntos buenos, por supuesto, no tiene ningún significado de por sí porque se trata de calificar en una escala de 0 a 15, ya que el examen tiene un valor de 15/50.

La segunda columna (1), es la primera escala de calificaciones obtenida por una simple transformación lineal: el puntaje bruto se ha dividido por 2. Cada dos puntos bien contestados tienen un valor de un punto sobre la escala de calificaciones de 0 a 15. El patrón de referencia utilizado es "lo que el profesor considera que el alumno debe saber". Es la manera más usual de calificar, pero también la más inconsecuente. En el caso de que alguien hubiera contestado correctamente todos los 30 puntos, el profesor no podría contestarse la pregunta: ¿ese alguien, conoce toda la materia?, ¿qué es lo que no sabe?; realmente, sabe más de lo que considero, ¿es el rendimiento ópti-

mo?; ¿cuál sería el rendimiento óptimo y cuál el mínimo? Tampoco podría saber cuál sería el rendimiento promedio del grupo, ni si solamente merecen aprobar 15 de los 45 alumnos, puesto que una calificación de 9, sobre una escala de 0 a 15, corresponde a 3, en una escala de 0 a 5, que es el sistema tradicional de calificaciones en Colombia.

El profesor no tiene ninguna razón objetiva para considerar que todo lo que pregunta debe ser correctamente contestado; si realmente estuviera seguro de ello, entonces, ¿para qué hacer el examen y por qué pensar que deben ser 30 y no 40 o 20 preguntas? Si hace un examen es porque considera que habrá algunos alumnos que no manejan ciertos conocimientos y el desea saber si está en lo cierto o, por lo menos, si determinados tópicos son ya del dominio del estudiante y cuáles todavía no. Por lo tanto, "lo que debe saberse, según el criterio del profesor" no es la única referencia para dar una nota apreciativa de lo que el alumno sabe, a menos que el profesor esté seguro de que no existe ningún tipo de interferencia en el proceso de aprendizaje, lo cual es muy arriesgado asegurar.

La tercera columna (2), es también una escala obtenida por una tranformación lineal; del puntaje bruto se ha restado 4, porque esa es la diferencia que hay del punto mayor de la escala de calificaciones al puntaje máximo alcanzado. Si ésta es la razón, no se entiende por qué se hizo un examen de 30 y no de 15 puntos. Este tipo de transformaciones es muy usual y algunos profesores lo llaman "calificar por curva". Si la transformación se hace sumando, se añade a cada puntaje bruto un número arbitrario de puntos. Este número se determina, buscando el complemento del puntaje bruto máximo alcanzado, para obtener el punto máximo de la escala sobre la cual se desea calificar. El patrón de referencia sería en este caso el rendimiento máximo alcanzado; es el mejor alumno el que marca la pauta de calificación. Se diría que es el extremo opuesto al caso anterior (1).

Si fuera el rendimiento real máximo, no se tendrían mayores objeciones; sin embargo ¿cómo estar seguros, de que no hubo conseja entre los alumnos con el fin de frenar el rendimiento máximo y así tener que estudiar menos? Para el caso específico que se describe, el hecho de que alumnos con un puntaje bruto de 13 (el modo del grupo B, diferenciado por el examen) hayan pasado ha sido una casualidad; no ha sido intencional. Calificar teniendo en cuenta el rendimiento máximo únicamente, trae consigo los vicios derivados de un acuerdo previo entre los examinados, con el riesgo de que sean ellos quienes determinan indirectamente el lugar de decisión de la categoría "aprobados", tanto más grave cuanto más mediocre sea el rendimiento del grupo.

La cuarta columna (3), es una escala obtenida por sucesivas transformaciones lineales, donde cada puntaje bruto se ha disminuido en 15, el promedio de rendimiento, y dividido por 2,97, la desviación estandar. En otras palabras, la distancia que hay de cada puntaje el promedio, se ha expresado en unidades estandar o sea tomando como unidad la estadística que indica el poder de discriminación del instrumento de medida (qué tan diferentes son entre sí respecto del rendimiento que pretende evaluar el examen, los alumnos que lo han abordado). Para obtener las calificaciones sobre una escala de 0 a 15, se hace una segunda y doble transformación lineal multiplicando cada nota, producto de la primera transformación, por 3, la desviación parámetro arbitrario en la escala de 0 a 15 y, sumándole 9, el promedio parámetro arbitrario en la escala de 0 a 15. Un ejemplo aclarará lo anterior:

Para un puntaje bruto de 19, se tiene que está a 4 puntos del promedio del grupo que es 15; 4 se divide por 2,97 que es la desviación estandar (s) para obtener 1,32; 1,32 se multiplica por 3

que es la desviación estandar arbitraria fijada y éste producto se suma a 9, el promedio arbitrario fijado para la escala de calificaciones lo cual da, aproximando, una calificación de 13 que es la nota correspondiente a un P.B. de 19 en la escala de calificaciones cuya cabeza de columna es (3). Esta forma de calificar se conoce con el nombre de puntajes estandar derivados. Estandar, porque cada P.B. se transforma en función de la desviación estandar de la distribución de puntajes y, derivada, porque se "deriva" a otra escala con parámetros arbitrarios que tiene en cuenta el tamaño del grupo que ha abordado el examen, la escala sobre la cual se desea calificar, el sitio que determinará la dicotomía "aprobados vs. reprobados". Para este caso se tomó como umbral de decisión 9 que es la mediana de la escala 0 a 15 y que transformado a la escala tradicional de 0 a 5 da una nota aprobatoria de 3. Por eso, el promedio arbitrario que se ha tomado es 9. Se tomó una desviación estandar arbitraria de 3, porque se espera que para una distribución con un N = 50, el rango total de dispersión sea de aproximadamente 4 desviaciones (4, pp. 81) y por lo tanto los límites del rango total de la escala de calificaciones (0 a 15), se pueden obtener restando y sumando a 9, el producto de multiplicar 3. La desviación estandar arbitraria, por 2, el número de desviaciones por encima o por debajo del promedio que se espera teóricamente, para una distribución de N = 50.

Este sistema de calificaciones tiene en cuenta las características del instrumento en sí y el rendimiento promedio del grupo, de tal manera que el umbral de aprobación lo marca el rendimiento promedio del grupo, muchísimo más difícil de "estabilizar" previamente y de común acuerdo por los estudiantes.

Puede observarse que tienen una calificación de 9, aquellos que obtuvieron una nota bruta de 15, que es el promedio de rendimiento del grupo y por lo cual aprueba, aproximadamente, el 50% del grupo. En cuanto al umbral de decisión de la categoría "aprobados", la diferencia con la escala (2) es que, en ésta, el umbral resultó de la casualidad y en la (3), de la intencionalidad. Lo intencional se puede manejar, lo casual no. Sin embargo tal y como se hicieron coincidir las dos escalas, la de puntajes brutos y la de calificación, no resolvió el problema de los alumnos que obtuvieron un puntaje igual o mayor que el valor modal de la distribución de las bajas y que está más cerca del promedio —el umbral de aprobación para ésta escala— de lo que lo está el puntaje modal de la distribución de los altos (véase gráfico número 3).

La quinta columna (4), es una escala que se obtiene de la representación en una recta numérica de cinco intervalos iguales que se expresan en unidades de desviación de la distribución, de tal manera que el promedio sea el centro del intervalo central y logrando que a cada gráfica le correspondan dos coordenadas, una para la escala de puntajes brutos y otra para la escala de calificaciones tal y como se representa a continuación.

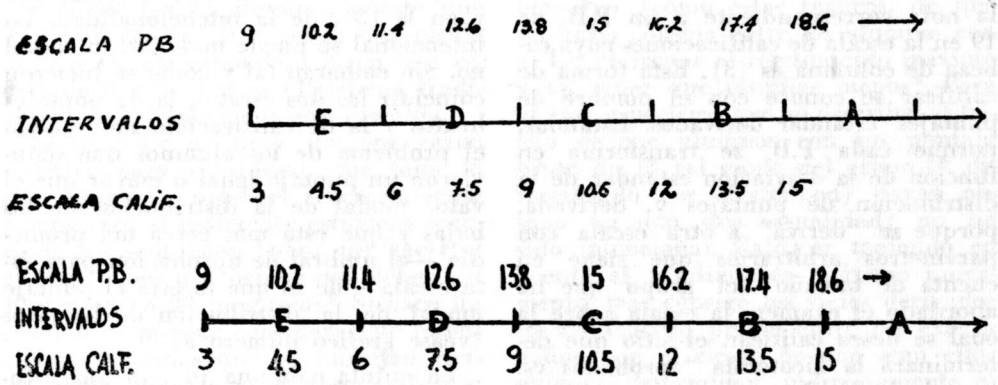
Si se quiere establecer más de cinco categorías de calificación, no solamente se asignarán coordenadas a las gráficas que separan los cinco intervalos, sino a los puntos medios de éstos.

¿Cómo se obtiene el valor para cada intervalo? Se ha dicho que para una distribución de N = 50, se espera un rango total de 4 desviaciones.

Si se deseara hacer solamente cuatro intervalos, cada intervalo tendría una amplitud de una desviación estandar; como no se desea hacer 4 intervalos sino 5, se dividirá la amplitud total, 4, por 5, los intervalos deseados, lo cual da 0,8 unidades de desviación. La desviación estandar de la distribución es de 2,97 que se aproxima a 3 para facilitar los cálculos; de tal manera que el valor de cada intervalo expresado en

GRAFICO No. 4.

DOS ESCALAS EN UNA MISMA RECTA NUMERICA — PARA N = 50



unidades de desviación estandar es de 0,8 x 3 = 2,4.

Como se desea hacer coincidir el punto medio del intervalo central con el rendimiento promedio del grupo que es de 15, se toma 15 como la coordenada de la gráfica central; las dos primeras gráficas laterales y centrales, respecto de la gráfica central, se obtendrán sumando y restando a 15, la mitad del valor del intervalo 2,4; entonces, se tienen las gráficas 13,8 y 16,2 en la escala de puntajes brutos. La gráfica que separa el primer intervalo tendrá como primer coordenada 18,6, (16,2 + 2,4), y la coordenada de la gráfica que separa el último intervalo, será 11,4, (13,8 — 2,4); la gráfica que cierra el último intervalo tendrá como coordenada 9, (11,4 — 2,4). El primer intervalo se deja abierto. Así, se han obtenido hasta ahora cinco puntos en la escala de puntajes brutos, que pueden hacerse coincidir con otras tantas coordenadas en la escala de calificaciones. Se pueden determinar puntos medios en los intervalos y así se obtendrán cuatro gráficas más, de tal manera que puede tenerse un total de 9 categorías para calificar. Esta escala tiene en cuenta el rendimiento promedio del grupo y el máximo rendimiento. Sin embargo, se observa que hay calificaciones demasiado extremas. Los alumnos de alto rendimiento fueron mejorados en gran medida a la vez que los de bajo rendimiento quedaron desmejorados, también, en gran medida. La distribución así obtenida jala por los extremos y en cuanto al grupo en total no se ha obtenido mejoría; sigue aprobando aproximadamente el 50% del grupo. Los mejores del grupo bajo no se han tenido en cuenta [1].

La sexta columna (5), corresponde a una escala que guarda los mismos principios de construcción de la anterior, sólo que se ha corregido la influencia que el tamaño del grupo tiene sobre la calificación. El argumento es: si en lugar de haber examinado 45 alumnos se hubieran examinado 500, ¿qué características hubiera tenido la superposición de las dos escalas —la de puntajes brutos y la de calificaciones—? ¿Se pueden obtener mayores beneficios si se

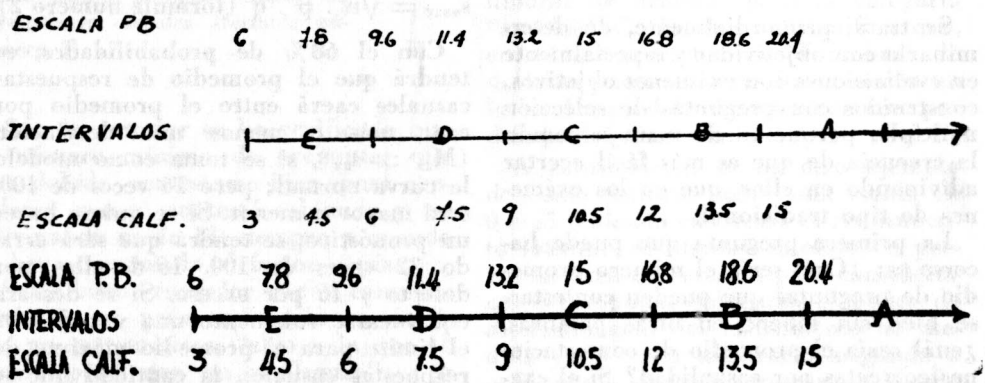
[1] El valor del intervalo en la escala de calificaciones es 3 y la calificación aprobatoria se ha hecho coincidir con el límite inferior del intervalo central.

aislan los efectos del sesgo? En otras palabras, si la escala de calificaciones se levanta para un N de 500 se tiene mayor fiabilidad de la medida porque los valores de la variable (lo que pretende medir el examen) se presentan en forma más continua, se eliminan los descensos y ascensos abruptos, muy posiblemente la distribución no sea bimodal sino unimodal y se conserve la asimetría negativa si todos los 500 estudiantes, en conjunto, mantienen las características de los 45 examinados y las circunstancias de aplicación se mantienen.

Con un supuesto de N = 500, se espera un rango de dispersión total de aproximadamente 6 desviaciones. Si se desea hacer sobre la recta numérica 5 intervalos, cada intervalo vale 6/5 = 1,2 desviaciones. Como la desviación para la distribución de puntajes ha sido aproximada a 3, el valor de cada intervalo expresado en unidades de desviación estandar es 1,2 × 3 = 3,6. A la gráfica central del intervalo central, se le ha dado como coordenada 15, siendo las coordenadas de las gráficas del intervalo central 13,2 y 16,8. Como se puede observar, tanto la escala de calificaciones como el punto medio de la escala de puntajes brutos han conservado la misma posición relativa, sólo se han cambiado las coordenadas para el resto de las gráficas en la escala de puntajes brutos, tal como se muestra a continuación.

GRAFICO No. 5

DOS ESCALAS EN UNA MISMA RECTA NUMERICA — PARA N = 500



¿Qué se ha obtenido? El umbral de decisión para la categoría "aprobaron" ha incluido el puntaje modal de la distribución de los alumnos de bajo rendimiento. Es el centro de la distribución el que ha jalado los puntajes y se ha corregido el distanciamiento excesivo de los puntajes extremos. La calificación correspondiente al mejor de los alumnos se acerca el rendimiento promedio y lo mismo sucede con el peor de los alumnos: mientras el mejor ha bajado, el peor ha subido. Es una escala que comparte las bondades de la (3) y de la (4), corrige sus vicios, controla las exigencias del profesor, tiene en cuenta el rendimiento promedio del grupo, la característica de la distribución real, la de N = 45 y la esperada, la de N = 500. De ésta manera se con-

trola también las consejas previas entre los alumnos, ya que todos han sido tratados por igual.

La séptima columna (6), es una escala que se ha obtenido de la anterior, pero en la cual se han dejado menos categorías de calificación. Esto se hace cuando hay razones suficientes para pensar que no hay necesidad de discriminar finamente; por ejemplo, si se encuentra que quienes obtuvieron una nota de 13.5 no tienen un rendimiento sensiblemente diferente y mejor que quienes obtuvieron una nota de 12. Se ha tomado como calificación que representa todo el intervalo, la coordenada que corresponde al límite inferior de cada intervalo; sin embargo, puede tomarse la coordenada del punto medio o las del límite superior de cada intervalo, cuando se observe, para casos específicos, su conveniencia.

¿Cómo determinar un nivel mínimo de conocimientos?

Se trata, primordialmente, de determinarlo con objetividad y especialmente en evaluaciones con exámenes objetivos, construídos con preguntas de selección múltiple, porque es de manejo común la creencia de que es más fácil acertar adivinando en ellos, que en los exámenes de tipo tradicional.

La primera pregunta que puede hacerse es: ¿Cuál sería el número promedio de preguntas que pueden contestarse bien sin saber? en otras palabras, ¿cuál sería el promedio de contestaciones correctas por casualidad? Si el examen está construído con preguntas de selección múltiple con 4 alternativas, se tiene que la probabilidad de escoger cualquiera de ellas por casualidad es de $\frac{1}{4}$; esta condición, por supuesto, es válida para la alternativa correcta. Si el examen tiene 30 preguntas el promedio de contestaciones por casualidad será, entonces, $(\frac{1}{4})$ 30.

Si se llama *p* la probabilidad de escoger por azar la alternativa correcta y *n*, el número de preguntas que contiene cada examen, se tiene que:

$$M_{azar} = p \cdot n \quad \text{(fórmula número 1)}$$

La segunda pregunta por hacerse es: ¿Qué tanto se dispersaría el número posible de contestaciones correctas, respecto del valor central obtenido anteriormente? En otras palabras, ¿cuál sería la desviación estandar esperada por casualidad? Tenemos que la varianza para cada pregunta es el producto de la probabilidad de escoger la alternativa correcta (p), por, la probabilidad en contra de tal escogencia (q). Si $p + q = 1$ (ya que 1 representa la probabilidad máxima, puesto que si todas las alternativas fueran correctas, la probabilidad de acertar sería 4/4), $q = 1 - p$.

Si el examen tiene 30 preguntas, la varianza total será entonces, $n \cdot p \cdot q$. Como lo que interesa es la desviación estandar, se tiene que:

$$s_{azar} = \sqrt{n \cdot p \cdot q} \quad \text{(fórmula número 2)}$$

Con el 68% de probabilidades, se tendrá que el promedio de respuestas casuales caerá entre el promedio por azar más o menos una desviación ($M_{az} \mp s_{az}$), si se toma como modelo la curva normal; pero 16 veces de 100, será mayor o menor. Si se quiere hacer un pronóstico, se tendrá que será errado 32 veces de 100, 16 de ellas por defecto y 16 por exceso. Si se deseara equivocarse solamente una vez de 100 el límite para el promedio máximo de respuestas casuales, la cantidad que se suma a M_{az}, será el resultado de multiplicar s_{az} por el valor, en unidades de desviación estandar, que en la curva normal separa el 1% extremo de la distribución, ese valor es *2,32*. Si se desea correr el riesgo de equivocarse 5 veces de 100, se tomará el valor en unidades de desviación estandar, que separa el 5% extremo de la distribución normal, que es *1,64* (4, Tabla B, pp. 569), *(3,* pp. 305 a 309).

Si se toma M_{az}, como el promedio de contestaciones correctas por azar, s_{az} la desviación estandar para una distribución de azar y Z el valor en unidades de desviación estandar que separan el 1% o el 5% extremo de la distribución normal, se tiene que el límite para el promedio máximo de respuestas casuales, con probabilidades de equivocarse 1 vez en 100 es:

$$M_{az} + (s_{az} \cdot Z); \quad M_{az} + (s_{az} \cdot 2{,}32)$$

(fórmula número 3)

Se tomará el examen del ejemplo anterior para aclarar prácticamente lo dicho:

Número de puntos del examen ...	n = 30
Alternativas de cada pregunta ...	4
Probabilidad de acertar por azar en cada pregunta	p = .25
M_{az} (fórmula número 1)	7,5
s_{az} (fórmula número 2)	2,37
Límite para el promedio máximo de respuestas casuales (fórmula número 3)	13

Con lo anterior se ha verificado que el número mínimo de preguntas que han debido contestarse bien, para considerar que se contestó sabiendo y no adivinando es de 13, siempre y cuando se haya controlado el factor copia.

Sería bueno recordar que el concepto de límite es, para expresarlo en términos no matemáticos, "algo a lo que se tiende pero que no se alcanza"; por ejemplo, el límite de un polígono regular es el círculo; sinembargo un polígono regular de infinito número de lados no podrá ser nunca un círculo, sin perder las características de polígono.

El examen que se ha venido tomando como ejemplo, fue aplicado a un grupo de alumnos al iniciar el curso, esto es, a personas que no tenían conocimientos sobre lo que el examen pretendía evaluar y, que por lo tanto, deberían escoger la respuesta correcta guiados por factores casuales y no por el conocimiento. La distribución que se obtuvo fue la siguiente, siendo prudente anotar que dos de las cuatro personas que obtuvieron un puntaje de 13 estaban repitiendo el curso, las otras dos habían asistido al curso anterior en calidad de oyentes, 2 con puntajes en 12 y una con puntajes en 11 estaban en condiciones similares. (Ver gráfico 6, página siguiente).

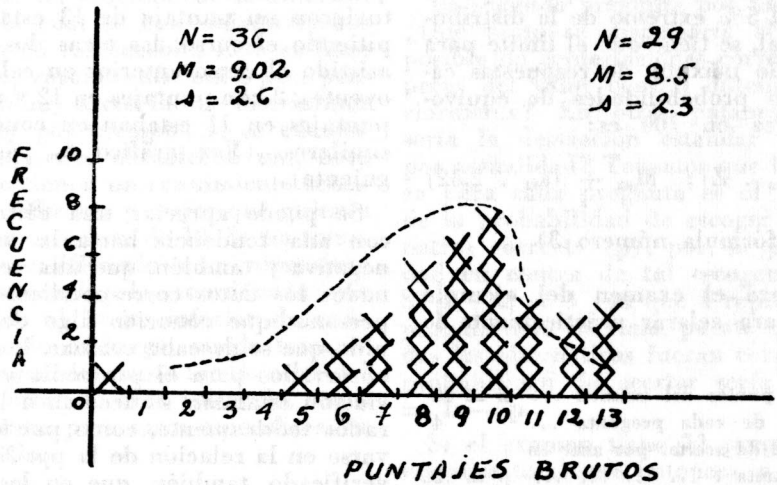
Se puede apreciar una distribución con una tendencia hacia la asimetría negativa y también, que una vez eliminados los datos correspondientes a las personas que conocían algo del contenido que se deseaba evaluar, los valores observados para el promedio y la desviación estandar, se acercan a los esperados teóricamente, como puede observarse en la relación de la pp. 23. Se ha verificado, también, que en las escalas (5) y (6) de la Tabla número 1, el nivel mínimo de conocimientos marcó el umbral de decisión para la categoría "aprobados"; lo que no significa que, en otras circunstancias, el umbral de decisión pueda correrse hacia arriba en la escala; no deberá nunca correrse por debajo del nivel mínimo.

A continuación se da otro ejemplo de cómo calificar sobre una escala de 0 a 15 teniendo en cuenta el rendimiento máximo, el número mínimo de preguntas bien contestadas para considerar que se responde sabiendo, el poder de discriminación del instrumento de medida y un valor arbitrario asignado a cada punto de cada examen, cuidando de conservar abierta en el extremo superior la escala de puntajes brutos.

Valor del examen	15/50
Número de puntas en el examen .	60
Valor arbitrario de cada punto en la escala de calificaciones	.25
Número de alternativas en las preguntas	4
Puntaje máximo alcanzado	40

GRAFICO No 6

DISTRIBUCION DE RESPUESTAS CASUALES



NOTA: A la izquierda aparecen las estadísticas del grupo total en la segunda aplicación y a la derecha, descontados los personas que algo sabían.

Puntaje mínimo obtenido	20
Promedio de rendimiento (M) ...	30,8
Desviación estandar observada (s) .	5,43
Promedio de respuestas correctas casuales (M_{az})	15
Desviación estandar de la distribución de azar (s_{az})	3,35
Límite para el promedio máximo de contestaciones casuales y correctas	23

Los datos anteriores sirven para determinar el umbral de la categoría "aprobados". Si no se desea que el punto (en la escala de puntajes brutos) que marca el conocimiento mínimo coincida con la calificación aprobatoria en la escala de calificaciones, ésta puede hacerse coincidir con el punto que marca, en la escala de puntajes brutos, el puntaje indicativo del conocimiento mínimo adicionado en el valor de la desviación estandar observada y se toma éste punto como umbral de aprobación en la escala de puntajes brutos. Para el caso es de 28 (23 +5,43 = 28, valor redondeado).

Para construír la escala se escriben los puntajes brutos posibles de obtener y para facilitar la conversión de puntajes brutos a valores en la escala de calificaciones, se coloca al frente del umbral de la categoría aprobados en la escala de P.B., la calificación aprobatoria y hacia arriba y hacia abajo incrementos o decrementos iguales a .25, tal como lo muestra la Tabla número 2.

Escala de conversión de puntajes brutos a calificaciones en una escala de 0 a 15.

TABLA NUMERO 2

P.B.	Calf.	P.B.	Calf.	P.B.	Calf.
52	15	39	11,75	26	8,50
51	14,75	38	11,50	25	8,25
50	14,50	37	11,25	24	8
49	14,25	36	11	23	7,75
48	14	35	10,75	22	7,50
47	13,75	34	10,50	21	7,25
46	13,50	33	10,25	20	7
45	13,25	32	10	19	6,75
44	13	31	9,75	18	6,50
43	12,75	30	9,50	17	6,25
42	12,50	29	9,25	16	6
41	12,25	28	9	15	5,75
40	12	27	8,75	14	5,50

¿Cómo saber si el rendimiento promedio observado en un grupo, puede atribuirse al conocimiento?

Puede pensarse que si el promedio de rendimiento está sobre el límite del promedio máximo de contestaciones casuales, ya es atribuíble al conocimiento. Pero puede verificarse sin necesidad de calcular tal límite, utilizando la dócima estadística *t* (t de student) para calcular el intervalo de confianza para el promedio verdadero (M') (6, pp. 213-216).

Lo que se desea verificar es si establecido el intervalo de confianza para el promedio verdadero (M') a partir del promedio observado (M), el intervalo incluye o no el promedio esperado por azar (M_{az}). Mientras más distante se halle M del M_{az}, menos probabilidades habrá de que M_{az} caiga en el intervalo de confianza de M'. Tanto más cerca de M_{az} esté M, tanto más probable será que $M = M_{az}$ y que por lo tanto, M' sea producto del azar. Para el caso que se desea verificar:

$$t = \frac{M - M_{az}}{ES}$$ (fórmula Nº 4), donde

$$ES = \frac{s}{\sqrt{n}}$$ (fórmula Nº 5) (6, pp. 211).

Se verificará para el mismo examen en tres condiciones diferentes:

Condición A: cuando se aplicó a un grupo de 45 alumnos al final de un curso, gráfico número 3.

Condición B: cuando se aplicó a un grupo al principio del curso, incluyendo personas que algo sabían, gráfico número 6.

Condición C: la misma situación de B, pero eliminando las personas que no cumplían la condición de absoluto desconocimiento de lo que se deseaba avaluar, gráfico número 6.

Para las tres situaciones se toman como constantes M_{az} puesto que se trata del mismo examen, la hipótesis de nulidad (H_0), puesto que se desea verificar el mismo fenómeno en las tres situaciones y el nivel de significación (NS) porque el riesgo de decisión está determinado para el mismo instrumento. En el cuadro siguiente se resumen las estadísticas para cada una de las situaciones.

Resumen de estadísticas, para las condiciones A, B, y C, necesarias para verificar la hipótesis de nulidad $M = M_{az}$.

TABLA NUMERO 3

Estadísticas variables	A	B	C
N	45	36	29
M	15	9,02	8,5
s	2,97	2,6	2,3
ES	.44	.43	.43
gi	44	35	28
t	16,93 (1)	3,5 (1)	2,32 (2)
Umbral de decisión	2,69	2,7	2,76

Estadísticas constantes	A	B	C
M_{az}	7,5	7,5	7,5
NS	1%	1%	1%
H_0 :	El rendimiento promedio está determinado por el azar, por lo tanto $M = M_{az}$		

Notas: (1) significativa; (2) No significativa.

Como se puede apreciar en la Tabla número 3, la distancia a la cual cae M de M_{az}, es tan grande, que expresada en unidades de ES sobrepasa el valor que en la distribución t, para el correspondiente grado de libertad, contiene el 99% de la distribución, en las situaciones A y B; no es así para la situación C.

Teniendo en cuenta lo anterior, se rechaza la hipótesis de nulidad (H_O) para las condiciones A y B, mientras que se verifica para la condición C; se formula la hipótesis alternativa H_1 para las condiciones A y B solamente. Por lo tanto se puede concluír que en las condiciones A y B el factor responsable del comportamiento de los datos fue diferente del azar; si se ha controlado el factor copia puede concluirse que hubo aprendizaje.

Se pensará que la conclusión anterior no es cierta para la situación B, puesto que la mayoría de los individuos que abordaron el examen (29 de 36), no tenían conocimiento ya que no habían iniciado el curso que se deseaba evaluar; pero debe tenerse en cuenta que se trata de una comparación de promedios, y, el promedio, es una estadística que se afecta por valores extremos y que en este caso eran 6, esos valores (ver gráfico número 6: los que se han excluído de la curva). Sin embargo, puede apreciarse en la Tabla número 3 que el valor absoluto de *t* es muchísimo mayor en A que en B. Esta situación sirve tambien para ejemplificar las consecuencias que, a nivel de decisión, causa la presencia de un sesgo en la muestra; en este caso se pudo controlar porque se conocía y permitió depurar la situación para producir la condición C.

Queda por verificar si establecido el intervalo de confianza para M' a partir de M, con el 99% de confianza (con posibilidades de fallar 1 vez de 100), el intervalo incluye M_{az}.

Si se toma el umbral de decisión en la distribución t, como UD_t (ver valores t, 4, pp. 580, Tabla D), ya que los demás símbolos se conocen, el promedio verdadero M' puede estimarse (6, pp. 213):

$$M' = M \mp (UD_t \cdot ES) \quad \text{(fórmula Nº 6)}$$

Resumen de estadísticas para calcular el intervalo de confianza para M' en las condiciones A, B y C

TABLA NUMERO 4

Estadísticas	A	B	C
M	15	9,02	8,5
UD_t	2,69	2,7	2,76
ES	.44	.43	.43
Límite superior del intervalo de confianza	16,19	10,18	9,69
Límite inferior del intervalo de confianza	13,81	7,86	7,31
M_{az}	7,5	7,5	7,5
Decisión	El intervalo de confianza para M', *no* incluye M_{az}		*si* lo incluye

Como puede observarse en la tabla precedente, el intervalo de confianza al estimar el promedio verdadero, manteniendo las condiciones A, B y C, sólo lo incluye en esta última que es la condición depurada de donde se ha asegurado la intervención del azar.

Con estas reflexiones sobre la evaluación sólo se ha querido demostrar que, relativamente, toma poco tiempo preocuparse seriamente por ella y que los exámenes objetivos brindan más confiables indicios, si están bien construídos, para hacerla controladamente.

La tarea de evaluar no puede reducirse, simplemente, a producir una calificación, sino que le es inherente la certeza de que ella represente, con cierto margen de exactitud, el verdadero nivel de conocimientos del alumno.

BIBLIOGRAFIA

1. ADKINS, D. *Elaboración de tests. Desarrollo e interpretación de tests de aprovechamiento.* México: F. Trillas, 1968.
2. BROWN, F. G. *Principles of Educational and Psychological Testing.* Illinois: The Dryden Press, 1970.
3. DUBOIS, Ph. *An introduction to Psychological Statistics.* New York: Harper & Row, 1965.
4. GUILFOD, J. P. *Fundamental Statistics in Psychological and Education.* New York: McGraw Hill, 1965.
5. MAGNUSSON, D. *Teoría de los tests.* México: F. Trillas, 1969.
6. YOUNG, R. & VELDMAN, D. *Introducción a la estadística aplicada a las ciencias de la conducta.* México: F. Trillas, 1968.