Evaluación de la educación
Crítica a un modelo

Miguel Ángel Rosado Chauvet[1]

Resumen

En 1994 se crea el Centro Nacional de Evaluación, A.C. (CENEVAL) por acuerdo y con subsidio de la Secretaría de Educación Pública (SEP) y de la Asociación Nacional de Universidades e Institutos de Educación Superior (ANUIES), utilizando para los análisis de ítemes y calificación de sustentantes el modelo KALT® del Dr. Agustín Tristán López.

Este modelo asume una serie se supuestos que no comprueba y que se contraponen con varios de los principios de medición. No obstante, por el apoyo político recibido se ha impuesto en diversas instituciones viciando los procesos de evaluación

En el análisis que se presenta se incluyen algunas de las críticas que se han señalado en diferentes momentos, iniciando por las mencionadas al autor del modelo y a algunasde las personas e instituciones de quienes recibe apoyo plítico e ingresos económicos.

El autor divide en grupos superior e inferior a los sustentantes para realizar al análisis de ítemes, distribuyendo al azar aproximadamente un 20% de los casos de la clase mediana y no explica de que se vale para obligar a que el grupo "inferior" no pueda responder un solo acierto hasta que el grupo "superior" termine con todas sus respuestas.

Asimismo asume que existe una correlación prácticamente perfecta entre el valor del Grado de Dificultad y el Poder de Discriminación cuando en su modelo la mitad ascendente presenta correlación negativa perfecta quedando en resumen una ausencia de correlación.

[1] Docente-Investigador de tiempo completo adscrito a la Licenciatura en Administración de la Universidad Autónoma Metropolitana–Unidad Iztapalapa.

Por otra parte, menciona que no se han publicado límites críticos para la dificultad de los ítemes; cunado el CENEVAL publicó los límites que habían sido utilizados en el Instituto Politécnico Nacional y en la Universidad Autónoma Metropolitana por el autor de este artículo y que concuerdan con los límites logísticos del Modelo Rasch, y señala que debe tomarse toda la amplitud de difcultades para, después, quitar el 54% de esta amplitud, evitando realizar análisis sobre los extremos de 27% que es donde se encuentran las respuestas de estudiantes que, a pesar de ser ítemes muy dificiles, son capaces de acertar y en el extremo contrario las respuesta de estudiantes que, a pesar de ser ítemes muy fáciles , no aciertan a ellos.

En 1994 se crea el Centro Nacional de Evaluación, A. C. (CENEVAL) por acuerdo y con subsidio de la Secretaría de Educación Pública (SEP) y de la Asociación Nacional de Universidades e Institutos de Educación Superior (ANUIES) y desde entonces se ha encargado de aplicar pruebas de ingreso al Nivel Medio Superior (Bachillerato), al Nivel Superior (Licenciatura) y al Nivel de Posgrado, así como para exámenes de egreso de Licenciatura y otras pruebas especiales.

Al comienzo de su empresa debió elegir un modelo de evaluación de los ítemes y las pruebas, tanto como desarrollar una serie de procedimientos en apoyo a los procesos de admisión de alumnos en los niveles mencionados.

Para la admisión de estudiantes se optó por calificaciones normalizadas con Media de 1000 y Desviación Estándar de 100 y se utilizó una distribución teórica que cubriera hasta tres desviaciones negativas y tres desviaciones positivas, utilizando la siguiente fórmula[2]:

CNE = 6X%+700
donde
X% = cada uno de los porcentajes de acierto de los sustentantes
6 = desviaciones permitidas en la distribución
700 = valor mínimo de la distribución

[2] Propuesta por el Dr. Miguel Rosado a solicitud del Dr. Agustín Tristán y del Dr. Eduardo Backoff.

Para el análisis de la prueba y sus ítemes se optó por el modelo KALT®[3] que es del que se ocupa el presente estudio, aclarando mediante una serie de notas las discrepancias que el autor de este artículo tiene con el autor del modelo.

Análisis crítico

Cálculo de la dificultad y la discriminación (*Noticias ICI E-10* a E-15)[4]

El modelo KALT se origina de una distribución donde grafica el Grado de Dificultad (GD) en el eje de X y el Poder de Discriminación (PD) en el eje de Y.

Las rectas 1 y 2 tienen por ecuaciones:

a) Máx PD = GD [RECTA 1]

b) Máx PD = 100–GD [RECTA 2]

Indicadores

Los procedimientos y valores típicos del modelo corresponden básicamente a la Dificultad y la Discriminación de los ítemes, obteniéndose como aportación del modelo el criterio utilizado para la Norma Discriminativa (ND) y la Relación Discriminativa (RD) derivadas de los anteriores, de acuerdo con lo siguiente:

El **Grado de Dificultad** (GD) se obtiene dividiendo el número de aciertos del grupo total, entre el número de casos del mismo. El rango de la dificultad aceptada está entre 26.795 y 73.205, (≅ 27↔73).

Fórmula: GD = A ÷N×100
Criterio: 27 ≤ GD ≤ 73

El **Poder de Discriminación** (PD) se obtiene dividiendo al grupo total por la mediana de respuestas al valor de referencia (suma de aciertos), formando dos grupos: Grupo Superior (GS) con los casos que tuvie-

[3] Diseñado por el Dr. Agustín Tristán.

[4] Publicaciones del Dr. Agustín Tristán a través de su empresa ICI en Mariano Jiménez 1839, Col. Balcones del Valle, CP 78280, San Luis Potosí, S.L.P.

ron la mayor cantidad de aciertos y Grupo Inferior (GI) con los que tuvieron una menor cantidad de aciertos.

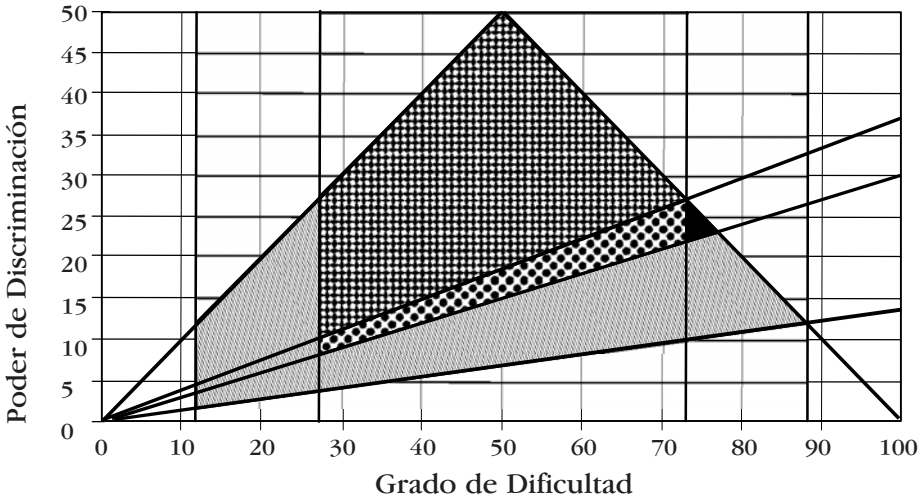
PD = [(GS̃-GI) +N]×100

La **Norma Discriminativa** (ND) se obtiene multiplicando por 0.30 el Grado de Dificultad (GD) hasta llegar al punto de incidencia de 76.92 entre la línea descendente de discriminación y la línea ascendente de la norma, a partir de ahí será 100-GD.

ND = 0.30×GD ≤ 76.92; ND = 10õ-GD > 76.92

La **Relación Discriminativa** (RD) se obtiene al dividir PD entre ND, y se fija en el modelo original en una norma de 1.00 o mayor.

Fórmula: PD ÷ ND
Criterio: ND ≥ 1.00



SA	0	10	20	30	40	50	50	50	50	50	50
IA	0	0	0	0	0	0	10	20	30	40	50
TA	0	10	20	30	40	50	60	70	80	90	100
GD	0.00	10.00	20.00	30.00	40.00	50.00	60.00	70.00	80.00	90.00	100.00
PD	0.00	10.00	20.00	30.00	40.00	50.00	40.00	30.00	20.00	10.00	0.00

[...]"

Nota 1. *Se observa en el modelo que, para que incida el límite superior del Grado de Dificultad, debería multiplicarse GD por 0.37 en lugar de 0.30, quedando el valor crítico del Poder de Discriminación sobre una línea de ascenso de 0 a 37. Por tener un rango de discriminación entre 0 y 50 el 0.37 equivale a 0.74 en una distribución con un rango entre 0 y 100 (ver línea ascendente superior en la gráfica).*
En sentido inverso, si se quiere mantener en 0.30 la ponderación sobre GD, se tienen que abrir los límites del rango del GD a 23 y 77, en lugar del 27 y 73 mencionados en el modelo (ver línea ascendente intermedia en la gráfica).
La correlación (incluyendo la correlación Gamma) tiene, para grados de libertad de 100 y una probabilidad de error de 0.05, un valor crítico de 0.196 (≅0.20) aceptado por el CENEVAL. Dado que KALT sólo tiene un rango de discriminación de 0 a 50, el 0.30 que sugiere equivale a una norma de 0.60 sobre 100, cuando se trabaja con un rango de dificultad de 23↔77, lo cual triplica los valores críticos aceptables para la correlación. Esto hace extremadamente exigente el modelo.
En sentido inverso, si asumimos que el 30% es sobre un 100%, para el modelo KALT que sólo llega al 50% le correspondería un valor de ponderación sobre GD de 0.15, lo cual obliga a abrir el rango de dificultad a 13 y 87, quedando ya muy próximo a los valores de los demás modelos (12-88) (ver línea ascendente inferior de la gráfica).
De hecho, si tomamos el valor de correlación en r=0.196 para 100, debería ser un valor de r=0.098 para el 50 del modelo Kalt, ajustándose el rango del Grado de Dificultad a 9↔91, quedando aun más amplio que el de los demás modelos con los cuales se ha comparado.

En la publicación *Noticias ICI*[5] *E-10* del 8 de marzo de 1995, "Relaciones entre grado de dificultad y discriminación (1) (Primera parte: Estudio del grado de dificultad)", se menciona:

"Aunque ambos conceptos son manejados en los textos y artículos relacionados con evaluación, ninguna referencia hasta la fecha ha inclui-

[5] Publicadas por ICI en Mariano Jiménez 1839, Col. Balcones del Valle, CP 78280, San Luis Potosí, S.L.P. At´n. Guadalupe Carrión FAX (48) 15 48 48 y Tel. (48) 20 37 88.

do este modelo que los relaciona. Este es el modelo original de KALT (Tristán L., 1976)"

Nota 2. *No existe ninguna referencia, porque de hecho no hay relación entre la dificultad y la discriminación de un reactivo.*

SA	0	10	20	30	40	50	50	50	50	50	50	**r**
IA	0	0	0	0	0	0	10	20	30	40	50	
TA	0	10	20	30	40	50	60	70	80	90	100	
Dif	0.00	0.10	0.20	0.30	0.40	0.50	0.60	0.70	0.80	0.90	1.00	**0.00**
Dis	0.00	0.33	0.50	0.65	0.82	1.00	0.82	0.65	0.50	0.33	0.00	
GD	0.00	10.00	20.00	30.00	40.00	50.00	60.00	70.00	80.00	90.00	100.00	**0.00**
PD	0.00	10.00	20.00	30.00	40.00	50.00	40.00	30.00	20.00	10.00	0.00	

Nota 3. *Tomando en cuenta los razonamientos planteados, la correlación entre los Indices de Dificultad y de Discriminación, calculados mediante el coeficiente Fi, no guardan ninguna relación (r = 0.00), pero tampoco se relacionan el Grado de Dificultad y el Poder de Discriminación de KALT (r = 0.00).*

En un estudio de cinco años, con muestras de 10,000 casos por versión y por año, las correlaciones entre Grado de Dificultad (GD) y Poder de Discriminación (PD) de los 128 ítemes de cada versión fueron los siguientes:

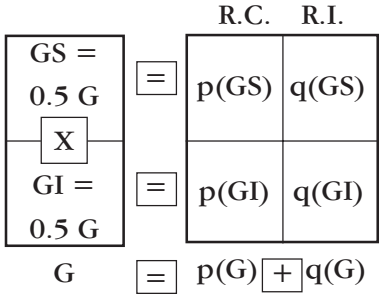
Versión	M1996	M1997	M1998	M1999	M2000
1	0.110	0.168	0.137	0.303	0.186
2	0.087	0.254	-0.099	0.428	0.279

Se puede observar que las correlaciones tienden a ser bajas, comparadas con el supuesto del modelo como correlación positiva perfecta (r=1.000), y en un caso (M1998/V2) no sólo no es positiva perfecta, sino que resulta negativa. Esto descalifica la insistencia del autor de tomar, como valor crítico del Poder de Discriminación, una proporción del Grado de Dificultad.

"1. Antecedentes. Supóngase una población G a la que se aplica un cuestionario objetivo. De manera arbitraria se divide al grupo en dos subgrupos de acuerdo con la mediana X: por arriba de X se denomina Grupo Superior GS y por abajo Grupo Inferior GI. Cada subgrupo contiene, por lo tanto un 50% de personas de G.

La mediana divide al grupo en dos partes iguales (salvo un individuo cuando la población es impar). La mediana no corresponde necesariamente con la media, excepto en casos particulares.

En la *Noticia ICI E-07* (Cálculo del poder de discriminación de varias muestras) se demuestra que la división de grupo no es aditiva para subpoblaciones, requiriéndose el cálculo de la mediana poblacional en cada caso.



La división de los subgrupos es arbitraria, pudiendo definirse otras formas (como son los cuartiles, por ejemplo), sin embargo la manera que aquí se presenta, permite una interpretación clara y sistemática de la relación entre los dos parámetros considerados en el modelo."

Nota 4. *De hecho, la mediana sí divide al grupo en dos partes iguales, pero la clase mediana contiene una gran cantidad de casos. En las 6 versiones de las pruebas aplicadas en el Metropolitano 2001 se obtuvo una muestra aleatoria de 1200 casos (600 mujeres y 600 hombres), con las siguientes frecuencias y porcentajes en la clase mediana (Clase Med.):*

Como podemos constatar, cuando tomamos 24 reactivos la cantidad de individuos que cae en la clase mediana queda entre 6.58% y 8.58% y cuando tomamos 10 reactivos la cantidad de individuos que cae en la clase mediana queda entre 15.50% y

21.58%. Una aclaración que proporciona el autor del modelo para apoyar su aseveración es que los individuos que caen en la clase mediana se dividen al azar en los Grupos Superior e Inferior[6]. Sin embargo, hay que tener en cuenta que la división entre Grupo Superior y Grupo Inferior se realiza utilizando el total del tema, por lo que la proporción de acierto o error en cada reactivo se ve afectada por ese porcentaje de manera aleatoria.

	Ver	Esp	His	Geo	Civ	Núm	Mat	Fís	Quí	Bio
N	1200	1200	1200	1200	1200	1200	1200	1200	1200	1200
Núm. de R	24	10	10	10	10	24	10	10	10	10
Clase Med. V1	7.08	19.42	21.33	17.92	21.17	7.25	17.58	18.25	17.75	19.00
Clase Med. V1	7.33	15.67	21.00	19.50	18.42	7.83	19.00	18.92	18.75	18.17
Clase Med. V1	7.33	19.42	19.92	18.25	19.17	7.75	17.50	19.08	19.08	21.42
Clase Med. V1	7.33	18.25	20.08	19.17	20.25	8.50	18.42	17.25	20.92	17.92
Clase Med. V1	6.58	19.67	19.92	19.25	21.50	8.42	15.50	18.67	21.58	16.33
Clase Med. V1	8.58	18.75	19.58	16.50	17.92	7.67	18.67	19.92	21.50	16.50
Ajuste	44	111	122	111	118	47	107	112	120	109

Existen otras posibilidades de dividir los casos: 1. pasar la clase mediana al Grupo Inferior, 2. pasar la clase mediana al Grupo Superior, 3. excluir la clase mediana y, 4. tomar los cuartiles 1 y 4 para el análisis. No obstante, en cualquiera de estos casos se desnivela la proporción del 50% para cada uno de los grupos, repercutiendo en los valores del Grado de Dificultad y del Poder de Discriminación, además en el punto 4 tendríamos nuevamente el problema de la clase cuatrilar en lugar de la clase mediana. En los casos de pasar la clase mediana a otro grupo estaríamos haciendo un ajuste promedio de 44 ó 47 casos en los temas de 24 reactivos, entre 107 y 122 casos en los temas de 10 reactivos en cada grupo afectando el análisis de los reactivos. Si se excluye la clase mediana se excluiría la cantidad de casos mencionados arriba.

[6] Comunicación personal del autor del modelo.

Además, por su forma de cálculo, el modelo KALT es sensible a cualquier diferencia en las proporciones que no sean de GS = 50% y GI = 50%.

"A su vez, para cada reactivo o pregunta, cada subgrupo se divide en dos, dependiendo de que las personas respondan correctamente R.C. o incorrectamente R.I., dando lugar a una tabla de 2´2, donde p(GS) es la frecuencia o número de respuestas correctas de las personas que se encuentran en el Grupo Superior, q(GS) es la frecuencia o número de respuestas incorrectas de las personas que se encuentran en el Grupo Superior, y de igual manera se definen p(GI) y q(GI) para el Grupo Inferior.

2. Grado de Dificultad. Se define por medio del cociente:

$$GD(\%) = \frac{\text{Suma de respuestas correctas}}{\text{Total de personas que responden}} \text{ X } 100$$

El Grado de Dificultad varía entre estos valores:

$$GD(\%) = \frac{p(G)}{G} \text{ x } 100 = p \text{ x } 100$$

0 = reactivo imposible de contestar, ninguna respuesta correcta.
100 = reactivo facilísimo, todas las personas contestan correctamente.

Entre ambos extremos hay toda la gama de números posibles.

En general las pruebas deben medir TODO EL RANGO DE DIFICULTADES, con objeto de disponer de un instrumento de medida que permita identificar a los individuos de mayor dominio de los de menor dominio. Contrariamente, algunos autores sostienen que un buen instrumento es aquél que tiene reactivos centrados en dificultad. Sin embargo, cuando el instrumento de medida se centra, lo único que se logra es que no podrá distinguirse entre los individuos extremos."

Nota 5. *No se trata de un instrumento que tiene reactivos centrados en dificultad, sino de un instrumento cuyo promedio de dificultad en*

los reactivos esté próximo al 50%, pero que abarquen desde un mínimo hasta un máximo aceptables de dificultad. De hecho no se hablaría de los límites que citan Stockton y Diederich como mínimo y máximo aceptables de dificultad (citados más abajo); no obstante, hay dos formas de cálculo de los valores extremos derivadas por Rosado a partir de la prueba binomial con aproximación a z (diseñada para la Dirección de Evaluación del Instituto Politécnico Nacional en el proceso de modificación del examen de ingreso al nivel medio superior) y de la fórmula general de x (diseñada en la Universidad Autónoma Metropolitana para comparación con la anterior, y presentada a la Dirección de Evaluación del CENEVAL, como extremos máximos aceptables).

Para el límite inferior externo (LIE):

$$Z_{LIE} = \frac{N - z\sqrt{N(k-1)}}{K} - \frac{1}{2}$$

$$\chi_{LIE}^2 = \frac{N}{K} - z\sqrt{\frac{N}{K}}$$

Para el límite superior externo será $N\text{-}z_{LIE}$ y $N\text{-}\chi^2_{LIE}$ prefiriéndose trabajar con porcentajes, como en el caso del modelo KALT. En ambos casos se tiene en cuenta el número de opciones del reactivo (k), dando las fórmulas como límites para 100 casos y 5 opciones 11.66↔88.34 (≅12↔88) la primera y 11.23↔88.77 (≅11↔89) la segunda.

		z		χ^2	
N	**k**	**LIE**	**LSE**	**LIE**	**LSE**
100	**5**	11.66	88.34	11.23	88.77
100	**4**	16.01	83.99	15.20	84.80
100	**3**	23.59	76.41	22.02	77.98
100	**2**	39.70	60.30	36.14	63.86

Podemos observar que el número de opciones (k) afecta los límites externos inferior (LIE) y superior (LSE), pero también notamos que el ejemplo de Diederich aplicado a 100 casos con 5 opciones (10↔90) queda muy próximo a ≅12↔88 y ≅11↔89. Asimismo, el modelo logístico de Rasch sugiere como máximo aceptable el rango –2.00 ↔ +2.00, representando los valores de los lógitos ln(11.92/88.08) y ln(88.08/11.92) respectivamente.

"Una vez comprendido y aceptado que los instrumentos de medida deben abarcar todo el rango de dificultades, se tiene que NO PUEDE JUSTIFICARSE POR LA SOLA DEFINICION DEL GRADO DE DIFICULTAD QUE LA MEDIA DE DIFICULTADES SE DEBE UBICAR AL 50%, NI QUE EL OPTIMO VALOR DE DIFICULTAD DEBA SER PARA GD=50%.

El modelo presentado aquí NO SOSTIENE que GD es óptimo si vale 50%, ya que caería en esta argumentación a todas luces errónea y absurda. Debe insistirse en que sólo los propósitos de la evaluación y la reducción del error en la medida conducen a un valor "óptimo" para cada caso, población y aplicación."

Nota 6. *Extraña la insistencia de que "deben abarcar todo el rango de dificultades" y después elimine con su procedimiento el 54% de la distribución (de 0 a 27 y de 73 a 100), quedando sólo los valores que el mismo autor hace consciente al mencionar que "cuando el instrumento de medida se centra, lo único que se logra es que no podrá distinguirse entre los individuos extremos".*

Verifique que la cita (mayúsculas subrayadas) se refiere al promedio de dificultades, no a la dificultad de todos y cada uno de los reactivos del tema. El valor óptimo de dificultad se da en los términos explicados abajo.

La "dificultad óptima", en realidad, dependerá del propósito de la evaluación, del tipo de medida a realizar, de las características de la población, etc.

No olvidemos que el "propósito de la evaluación" consiste en emitir un juicio justo, nutriéndose de hechos y datos lo más objetivo y estable posibles, incluyendo cualquier tipo de observación y forma de medición, como la que nos ocupa. Además, un ítem cuyo error no pueda ser explicado ni por el rango de azar puede deberse a un

error propositivo que corresponda a fines personales del que responde, o a conductas de copia sobre versiones diferentes; asimismo, un acierto que se ubique por arriba del rango de azar de los ítemes fáciles puede deberse a que el aprendizaje medido corresponde a conocimientos rebasados por el nivel del grupo al que se le aplica. En cualquiera de estos casos la medición estará fuera de las características de la población.

Se dice que "el Grado de Dificultad es óptimo si vale el 50%, siempre y cuando el 50% de aciertos corresponda al Grupo Superior". El supuesto de KALT no puede verificarse en este sentido porque el valor máximo admitido es de 50% por una diferencia de procedimiento. Veamos este fundamento utilizando Fi:

	R.C.	R.I.	TOTAL
GS	50	0	50
GI	0	50	50
TOTAL	50	50	100

Nota 7. *Esta distribución da una correlación Fi de $r_\phi = 1.000$ que es la máxima posible. De hecho es un parteaguas determinista para concluir una óptima discriminación. En el caso de los modelos logísticos esto correspondería a una vertical sobre el eje de X y paralela al eje de Y, donde los valores negativos corresponderían a GI y los positivos a GS.*

Suponiendo que un item es acertado por el 60% del grupo superior y por el 40% del grupo inferior, cuando son 50% y 50% los casos para los grupos la Razón Discriminativa (RD) acepta al ítem (1.50 > 1.00), pero con la misma proporción de aciertos al pasar la clase mediana al grupo superior la RD lo rechaza (0.78 < 1.00) y al pasar la clase mediana al grupo inferior la RD también lo rechaza (0.00 < 1.00).

Sin embargo, puede verificarse que ni Fi ni Gamma se alteran, manteniendo el mismo valor correlacional al mantener también la misma proporción de aciertos, con independencia de los casos que tenga cada grupo.

	p	q	R.C.	R.I.	Tot.	GD	ND	PD	RD	Fi	G
GS	0.60	0.40	30	20	50						
GI	0.40	0.60	20	30	50						
Tot.			50	50	100	50.00	15.00	10.00	1.50	0.20	0.20

	p	q	R.C.	R.I.	Tot.	GD	ND	PD	RD	Fi	G
GS	0.60	0.40	36	24	60						
GI	0.40	0.60	16	34	40						
Tot.			52	48	100	52.00	15.60	20.00	0.78	0.20	0.20

	p	q	R.C.	R.I.	Tot.	GD	ND	PD	RD	Fi	G
GS	0.60	0.40	24	16	40						
GI	0.40	0.60	24	36	60						
Tot.			48	52	100	48.00	14.40	0.00	0.00	0.20	0.20

Tomemos ahora el ejemplo anterior donde el modelo KALT "NO SOSTIENE que GD es óptimo si vale 50%":

	p	q	R.C.	R.I.	Tot.	GD	ND	PD	RD	Fi	G
GS	1.00	0.00	50	0	50						
GI	0.00	1.00	0	50	50						
Tot.			50	50	100	50.00	15.00	50.00	3.33	1.00	1.00

Nótese que para las correlaciones Fi y Gamma resulta una correlación positiva perfecta.

3. "**El modelo binomial.** Bajo el concepto de la distribución binomial, se sabe que el número de casos que se pueden discriminar en una población está dado por el producto pq.

De este modo, se tiene que al dar valores a p y q el número de discriminaciones posibles puede tabularse como sigue:

p	0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1
q	1	.9	.8	.7	.6	.5	.4	.3	.2	.1	0
pq	0	.09	.16	.21	.24	.25	.24	.21	.16.	.09	0

Se observa que p es creciente, mientras que q es decreciente, sin embargo pq toma valores que se asemejan a una distribución normal (ascendente hasta 0.5 y descendente a continuación), con máximo en 0.5, dando un número de discriminaciones posibles de 0.25.

Partiendo de esta distribución, algunos autores sugieren que el 50% de dificultad es el óptimo deseable para un reactivo, independientemente de los propósitos de la evaluación."

Nota 8. *Este razonamiento es totalmente válido, pero hay que tomar en cuenta que p y q no se refieren al Grupo Superior y al Grupo Inferior, sino a los aciertos y errores de cada uno de los grupos.*
Obsérvese además que este razonamiento refuerza lo que KALT desea desvirtuar, ya que en p = .5 y q = .5 es donde hay un máximo de discriminación.

En la publicación *Noticias ICI*[7] *E-12* del 11 de marzo de 1995 "Relaciones entre grado de dificultad y discriminación (3) (Tercera parte: El dominio de discriminación)" se menciona:

4. Dominio de aceptacion de reactivos. La deducción de la norma discriminativa que se hace en otra Noticia ICI, permite determinar el rango de aceptación de los reactivos. En KALT se sigue este criterio:

Norma discriminativa:
N = 0.3 GD (3.1)
Intervalo de dificultad:
27 <= GD <= 73 (3.2)

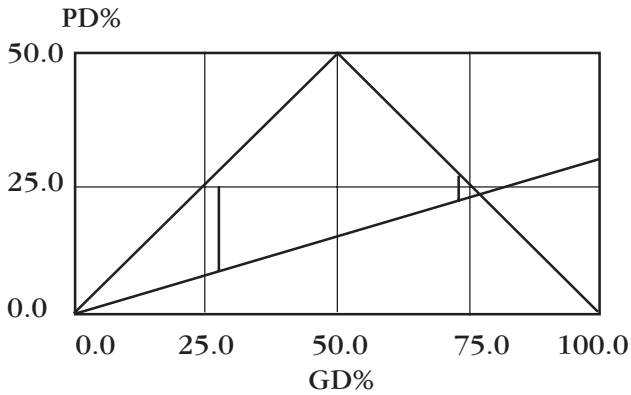
Fuera del intervalo propuesto deben hacerse algunas consideraciones de tipo práctico, con objeto de poder analizar los reactivos de un

[7] Publicadas por ICI en Mariano Jiménez 1839, Col. Balcones del Valle, CP 78280, San Luis Potosí, S.L.P. At´n. Guadalupe Carrión FAX (48) 15 48 48 y Tel. (48) 20 37 88.

cuestionario. La combinación de la norma (3.1), los límites fijados por (3.2) y las rectas 1 y 2 del dominio A, permiten establecer el subdominio ND de reactivos aceptables.

5. Relación discriminativa. Para poder comparar fácilmente el Poder de Discriminación y la Norma Discriminativa, se introduce la Relación Discriminativa, dada por:

$$RD = PD / ND \qquad (3.3)$$



Los reactivos que discriminan por lo menos igual que la norma tiene una RD >= 1. Es deseable que todos los reactivos de un cuestionario tengan valores de RD superiores a la unidad. Pero será responsabilidad del evaluador aceptar reactivos cuya RD no llegue a 1, cuando se considere que tal libertad no afecte a los propósitos de la evaluación o desvirtúe el cuestionario. Con esto se quiere decir que pueden aceptarse ALGUNOS reactivos con RD inferior a 1, pero seguramente no serán todos los reactivos los que estén en este caso.

6. Valores fuera del subdominio ND. El último problema planteado es el de establecer un criterio para los valores externos al dominio ND. Se divide el problema en dos partes:

Para GD > 73. Se trata de reactivos fáciles, que no se consideran "aceptables" de acuerdo con lo especificado en (3.1) y (3.2). Sin embargo pueden ser estudiados observando que si se prolonga la recta (3.1), se tiene una intersección con la recta 2 de dominio A en GD = 76.92. Para

fines prácticos puede continuarse con la misma recta hasta dicho punto de intersección. A partir de ahí, el valor que rige NO ES LA RECTA (3.1), SINO LA RECTA 2. Recuérdese que las rectas 1 y 2 representan los valores máximos posibles de discriminación, por ello los valores de la recta (3.1) más allá de GD = 76.92 no son aplicables.

Nota 9. *Si el rango admisible de dificultad se encuentra entre 27↔73, ¿cómo se explica ahora que "más allá de GD = 76.92 no son aplicables", cuando debiera ser para el modelo un máximo de GD = 73?*

Este es el mismo criterio empleado en KALT: Se pide que los reactivos discriminen por arriba de la recta (3.1) hasta GD = 76.92. A partir de ahí se exige que discriminen "AL MAXIMO", de acuerdo con la recta (3.2). Este criterio es muy exigente para los reactivos fáciles, pero de manera práctica se ha visto que funciona. **Ya se ha dicho cómo las pruebas tienden a aceptar a los reactivos fáciles, por lo que exigir que RD=1 representa una exigencia fundamental que hace la diferencia entre TODOS los criterios de la literatura y el que incluye KALT. Los evaluadores que empleen este criterio pueden estar seguros de estar trabajando al máximo de calidad posible de un reactivo.**

Se demuestra fácilmente que para GD>76.92 el valor máximo posible de RD es 1.

No hay un límite superior claro para los reactivos fáciles. Se puede definir arbitrariamente que los reactivos con GD>90% son extremadamente "fáciles".

Nota 10. *Con los criterios probabilísticos mencionados con anterioridad, tanto para los modelos clásicos como para los modelos logísticos, queda precisado cuál debe ser el límite superior admisible.*

De hecho, no permiten sacar grandes conclusiones porque el producto pq en este caso sólo permite discriminar un 9% de casos (ver *Noticia ICI E-10*, recuadro "El Modelo Binomial")

Nota 11. *No corresponde a la discriminación de "un 9% de casos", sino del 9% de aciertos del total de casos.*

En resumen para este caso:
N=0.3GD para 73 < GD <= 76.92 (3.4)
N=100-GD para 76.92 < GD <= 100 (3.5)
Reactivos fáciles para 73 < GD <= 90
Reactivos muy fáciles para GD > 90

Para GD < 27. Se trata de reactivos difíciles que las pruebas de hipótesis tienden a rechazar. Desde el punto de vista de la evaluación, los reactivos difíciles que discriminan positivamente deberán ser incluidos en los cuestionarios.

Durante las etapas de la prueba de la norma (3.1) se propusieron múltiples maneras de analizar los reactivos difíciles. No existe un tratamiento claro de este tipo de reactivos, por lo que cualquier modelo que se establezca siempre estará sujeto a discusiones. Para fines prácticos se ha optado por seguir la misma norma (3.1).

En este caso no hay problema de intersección con la recta 1, ya que ambas cortan en el origen. En cuanto al valor aceptable de dificultad se maneja el mismo criterio de limitar un 10% como se hizo en los reactivos fáciles. En este caso se puede decir que los reactivos con GD<10% son muy difíciles, aunque la norma seguirá siendo la expresión (3.1).

Nota 12. *Para este otro extremo también se han propuesto con anterioridad los límites probabilísticos aceptables, tanto para modelos clásicos como para logísticos.*

N=0.3GD para 0 <= GD < 27 (3.6)
Reactivos difíciles para 10 <= GD < 27
Reactivos muy difíciles para 0 <= GD < 10

7. Trazado del dominio de reactivos aceptables y la norma discriminativa. Con todo lo anterior puede dibujarse el dominio completo en el diagrama de DISCRIMINACION/DIFICULTAD (Figura 2), como se presenta en la salida de KALT, donde se identifican las diversas zonas presentadas hasta ahora. Obsérvese que en este diagrama se han marcado los valores al 0.3 GD, 0.2 GD y 0.1 GD que pueden servir como referencia al evaluador para análisis especiales, con objeto de conocer qué

tan lejos está la Norma Discriminativa, hasta qué valores puede aceptar en un cuestionario, estimar la eficiencia del instrumento, etc.

Cuando los reactivos discriman negativamente

En la misma Figura 2 se incluye el dominio del cuarto cuadrante, donde caen los reactivos que discriminan negativamente. Como ya se indicó en la *Noticia ICI E-11*, carece de interés estudiar las propiedades de la zona negativa, sin embargo es importante presentarla ya que los reactivos "desechables", cuya discriminación es negativa, caen en dicha zona.

Esto es debido entre otras causas a:

Reactivos con base u opciones mal planteadas, de manera que confunde a los individuos.

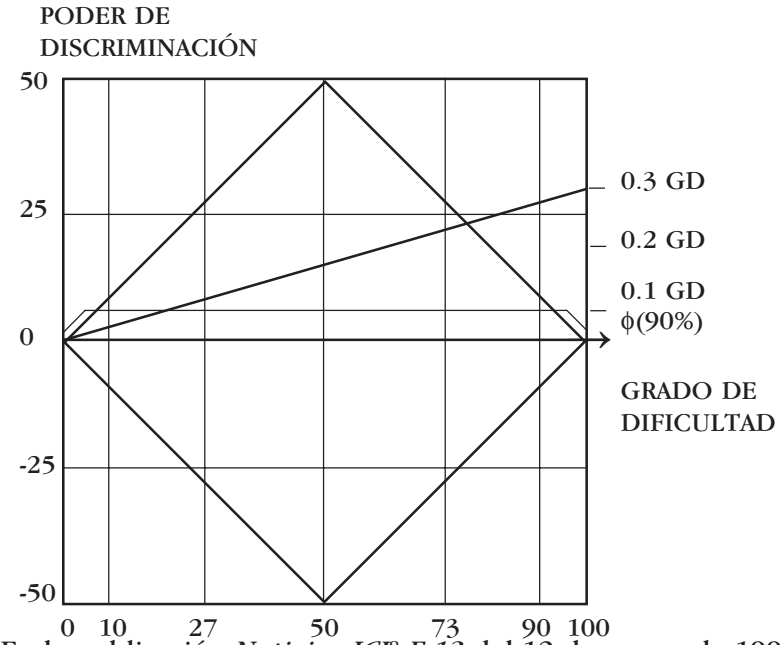
Cuestionario con muchos reactivos ineficientes (con opciones mal planteadas), que los convierte en reactivos de 2 ó 3 opciones, modificando la esperanza matemática de respuesta. Esto generalmente ocurre en reactivos muy fáciles.

Reactivos sin respuesta correcta.

Cuando un reactivo discrimina negativamente se considera "desechable", dado que está engañando la medición o, francamente, no sirve para los propósitos de la medición.

Las *Noticias ICI E-10 a E-13* están basadas en: A. Tristán López. *Modelo de evaluación para la Facultad de Ingeniería*, UNAM, 1977.

Nota 13. *Nótese que en el diagrama de KALT la correlación Fi (r_ϕ), anotada como ϕ, queda prácticamente sobre el eje de X por el error conceptual de comparación de los valores de KALT (–50↔+50), contrastados con los valores de r_ϕ (–1.00↔+1.00). Ya hemos mencionado anteriormente el requisito de mantener el mismo tipo de medida para realizar este contraste: o los porcentajes de KALT se toman como proporciones (–0.50 ↔+0.50), o los valores de Fi se toman como porcentajes (–100 ↔+100), de no ser así, la comparación resulta irracional.*



En la publicación *Noticias ICI*[8] *E-13* del 12 de marzo de 1995 Relaciones entre grado de dificultad y discriminación (4) (Cuarta parte: Norma discriminativa) se menciona:

8. Deducción de la Norma Discriminativa por el principio del trabajo virtual mínimo. *En la Noticia ICI E-11* (Estudio de la discriminación), se discutió la necesidad de decidir una norma y se comentó cómo el subdominio B cumple ciertas propiedades que incluyen:

Simetría del subdominio respecto a GD = 50%

Los criterios de discriminación mínima dados por pruebas de hipótesis o criterios, tienden sistemáticamente a

b.1) rechazar los reactivos difíciles (inferiores a GD = 50%)
b.2) aceptar los reactivos fáciles (por arriba de GD = 50%)

[8] Publicadas por ICI en Mariano Jiménez 1839, Col. Balcones del Valle, CP 78280, San Luis Potosí, S.L.P. At´n. Guadalupe Carrión FAX (48) 15 48 48 y Tel. (48) 20 37 88.

Nota 14. *Esta diferenciación en aceptación y rechazo de reactivos difíciles y fáciles está dada por el error conceptual en el uso de la correlación Gamma. Presentamos un ejemplo de un item difícil GD = 30 < 50 y un item fácil GD = 70 > 50 de acuerdo con lo que menciona KALT:*

Reactivo difícil

	%	R.C.	R.I.	Tot.	GD	PD	RD	G
GS	40	20	30	50				
GI	20	10	40	50				
G	30	30	70	100	30	10.00	1.11	0.20

Reactivo fácil

	%	R.C.	R.I.	Tot.	GD	PD	RD	G
GS	80	40	10	50				
GI	60	30	20	50				
G	70	70	30	100	70	10	0.48	0.20

Nota 15. *La diferencia en porcentaje de acierto entre los grupos superior e inferior es la misma (40 - 20 = 20; 80 - 60 = 20). KALT favorece al ítem difícil (RD = 1.11 > 1.00) y desfavorece al fácil (RD = 0.48 < 1.00), cuando Gamma califica igual a ambos (G = 0.20).*

Desde el punto de vista de la estadística puede afirmarse que para las pruebas de hipótesis no hay diferencia entre un reactivo de 25% y otro de 75% de Grado de Dificultad.

Con objeto de reducir o eliminar el defecto inherente a las pruebas de hipótesis, se planteó la necesidad de definir una norma discriminativa que sea congruente con la lógica de la evaluación, partiendo de estos postulados:

Nota 16. *Ya se ha establecido la baja correlación entre la discriminación y la dificultad de un ítem, por lo que tomar un valor crítico para la discriminación en función de un porcentaje de la dificultad*

resulta irracional. Más aún cuando el cálculo de la discriminación contiene un error conceptual en su manejo, haciéndolo de la mitad de su valor natural, es decir, que cualquiera de los postulados que se mencionan a continuación arrastran estos sesgos.

A.1.- La norma discriminativa debe ser una función creciente del Grado de Dificultad. De preferencia debe emitir valores equiparables para todo el rango de Dificultades, de acuerdo con el número de personas que responden.

Justificación: Evitar que el criterio rechace, de manera sistemática, a los fáciles, como es el caso con las pruebas de hipótesis.

A.2.- La norma discriminativa debe permitir identificar el rango de dificultad en donde tiene validez.

Justificación: Conseguir límites para el Grado de Dificultad, ya que no hay valores óptimos del dominio.

A.3.- La norma debe ser suficientemente exigente para que permita identificar las características de calidad de los reactivos y suficientemente flexible para poder manejarla e interpretarla en la evaluación.

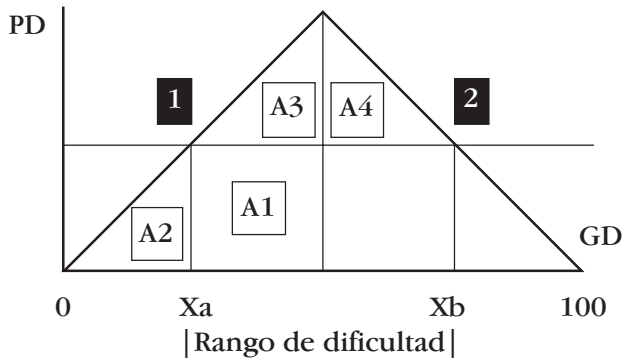
Justificación: Las otras normas propuestas en la literatura no tienen una interpretación clara, haciendo muy difícil su modificación y aplicación en casos particulares. Dicho de otro modo: La norma debe ser una expresión de uso sencillo.

A este respecto, si se desea manejar la norma discriminativa propuesta por una prueba de hipótesis, (elipse E) se tiene la expresión explícita:

$$ND = C_1\sqrt{\frac{50^2 - (GD - 50^2)}{50}} \qquad (4.1)$$

Esta fórmula que es general para toda prueba de hipótesis, tiene el inconveniente de ser de uso poco práctico. Ya se han discutido los problemas que tiene implícita la prueba de hipótesis.

El proceso original para establecer la norma discriminativa empleada en KALT (que después fue depurado y justificado por otros medios y pruebas en exámenes reales), la define a partir del concepto de trabajo virtual mínimo, que en términos generales parte de establecer un área básica bajo las rectas 1 y 2 que limitan al dominio A (lo cual evita aceptar los extremos de dicho dominio con discriminación nula) y el resto del área bajo dicha curva conducirá a una sobretasa para la norma en términos de Grado de Dificultad. De manera gráfica, y para simplificar esta deducción se presenta [...].



Definición de áreas para la norma discriminativa

Los elementos del modelo son:

El área básica de la norma discriminativa está definida por A1 y A2 (norma mínima).

La recta que limita por arriba a A2 y que intersecta a la recta 1, permite definir Xa

La recta que limita por arriba a A2, al intersectar a la recta 2 permite definir Xb

El resto de la norma (sobretasa sobre la norma mínima) será k veces las áreas A3 y A4.

Para determinar k se impone que la discriminación cumpla que el área que se obtenga brinde como resultado la discriminación que posee un reactivo de Grado de Dificultad GD cualquiera que este sea:

$$\frac{A1 + A2}{\text{ÁREA BÁSICA}} + \frac{k(A3 + A4)}{\text{FRACCIÓN DE LAS RESTANTES}} = 100GD \quad (4.2)$$

(El coeficiente 100 de GD permite dimensionar el Grado de Dificultad respecto al primer miembro).

Observaciones respecto a las curvas [...]

1.- Se muestra que para los diversos valores de k se obtienen los límites del rango de Dificultad, simétricos respecto a 50%.

2.- Las curvas Na y Nb tienen un comportamiento inverso entre sí: Na decrece, mientras que Nb crece, para el mismo rango de valores de k.

3.- Solamente hay un valor de k para el cual Na = Nb. Esto ocurre donde se cruzan Na y Nb, para k = 3.1547. En este valor se tiene el rango de Dificultad que cumple con los postulados A1 y A2.

Así el intervalo de dificultad es:

$$26.795 <= GD(\%) <= 73.205 \quad (4.7)$$
para k = 3.1547 y Na = Nb = 36.60254 %

De (4.7) se tiene que el rango de dificultades fluctúa aproximadamente entre 27 y 73% (en números cerrados). Resulta muy interesante verificar que por caminos completamente diferentes, se tiene que bajo el Modelo de Rasch el intervalo limitado por +/– 1 lógito corresponde de manera muy cercana con este valor."

Nota 17. *Hay que recordar que, si bien este es un límite óptimo para Rasch, donde la linealidad del modelo logístico es virtualmente completa, se admiten valores que van de -2.00 a +2.00 lógitos, donde no se justifica el rechazo por desajustes en los extremos de la distribución.*

Para $26.795 <= GD <= 50$

$$N = -21.13 + 0.01(157.735GD - 1.577GD^2) \quad (4.8)$$

Para $50 <= GD <= 73.205$

$$N = 57.735 - 0.01(157.735GD - 1.577GD^2) \qquad (4.9)$$

Las expresiones (4.8) y (4.9) corresponden a 2 trozos de parábolas continuas en GD = 50%, cuyo rango de validez está plenamente identificado para cada una de ellas, fuera de ese rango no tienen sentido. Las curvas (4.8) y (4.9), las rectas Xa = 26.795 y Xb = 73.205, junto con las rectas 1 y 2 del dominio A, definen el subdominio V que se estaba buscando, donde se encuentran los reactivos aceptables en dificultad y discriminación.

9. Formas alternativas y complementarias para la Norma de Discriminación. Las expresiones (4.8) y (4.9) brindan la forma más exacta de la Norma Discriminativa de acuerdo con los postulados propuestos en la parte 4 de esta *Noticia*, con excepción de su facilidad de uso.

Si bien estas expresiones pueden calcularse muy fácilmente con ayuda de la computadora, la experiencia en las primeras aplicaciones de esta norma mostraron que:

Los evaluadores desean una norma de "fácil aplicación". En este caso la norma está definida en dos intervalos diferentes, por medio de dos ecuaciones que, además, requieren de elevar al cuadrado y hacer varias operaciones. Los evaluadores generalmente rechazan este nivel de "complejidad".

La norma (4.8) y (4.9) es rigurosa. Los evaluadores desean disponer de una norma "flexible", con objeto de bajar o ajustar el grado de rigor de acuerdo con los propósitos de la evaluación.

La norma (4.8) y (4.9) es válida para GD en el intervalo (26.795, 73.205) ó (27, 73) en números cerrados. Los evaluadores desean saber qué hacer con los reactivos "indeseables" que quedan fuera de este intervalo.

Para atender estos problemas se hacen los siguientes trabajos.

9.1. Interpolación Lineal. Se observa que la recta que pasa por los extremos de validez de (4.8) y (4.9) tiene por ecuación:

$$N = 0.3660254\ GD$$

Esta recta que pasa por el origen, cruza por el punto de continuidad entre (4.8) y (4.9). Se observa que esta recta de interpolación reduce ligeramente el rigor de la curva (4.9) en los reactivos del lado "difícil" (menor de GD = 50%) y aumenta ligeramente el rigor de la curva (9) en los reactivos el lado "fácil" (mayores de GD = 50%).

Nota 18. *Ya habíamos anotado esta diferencia con respecto a los supuestos del modelo, mencionando que de acuerdo con KALT la Norma Discriminativa no debería ser de ND = 0.30GD, sino del 0.37GD que se admite en este momento.*

9.2. Ajuste por Mínimos Cuadrados. Otra forma alterna se obtiene por medio de una recta de ajuste de mínimos cuadrados. La ecuación que se obtiene es:

$$N = 1.9776538 + 0.3264939\ GD \qquad (4.11)$$

que tiene una correlación de 0.9673338.

Esta recta de ajuste ya no pasa por el origen y es ligeramente menos rigurosa que la interpolación (4.10). Al cruzar a las parábolas (4.8) y (4.9) se aprecia que compensa de manera diferente en los intervalos de reactivos "fáciles" y "difíciles".

9.3. Forma alterna general. Por lo anterior, se puede ver que una forma "razonablemente" cercana a la norma (4.8) y (4.9) puede estar dada por una recta de muy fácil aplicación del tipo

$$N = A + B \times GD \qquad (4.12)$$

Ya se vio que A varía entre (0 y 1.97) y B es del orden de (0.326 a 0.366).

10. Aplicación de la norma. La norma (4.8) y (4.9) y su forma alternativa (4.12) fue probada ampliamente en la Facultad de Ingeniería de la UNAM durante 1976 y 1977. De acuerdo con las recomendaciones de los profesores y responsables de la evaluación se vio de manera práctica que la norma es un poco rigurosa, por lo que se sugirió que no era substancial incluir el valor de A para fines prácticos, ya que las diferencias

que se tenían no eran muy relevantes, pero en cambio hacían ligeramente más "complicada" la aplicación de la fórmula.

Por lo anterior, se sugirió emplear la fórmula donde se reduce en un 20% la Norma Discriminativa:

$$N = B \times GD = 0.8 \times 0.3660254GD = 0.293GD \qquad (4.13)$$

en el intervalo (26.795, 73.205).

Finalmente se optó por reducir los números de (4.13) a valores más sencillos, con lo que queda:

Norma Discriminativa:

$$N = 0.3GD \qquad (4.14)$$

Intervalo de Dificultad:

$$27 <= GD <= 73 \qquad (4.15)$$

Este es el modelo original KALT. Con objeto de dejar al evaluador la posibilidad de adaptar los valores de esta norma, se ha dejado en KALT la expresión (4.12), permitiendo al evaluador especificar los valores de A y B a su conveniencia. Por omisión se usan las fórmulas (4.14) y (4.15), que son las que demostraron su utilidad en la práctica y que, hasta la fecha, incluyendo las aplicaciones realizadas en el CENEVAL, siguen demostrando que son de uso sencillo y muy "estables" entre las diversas aplicaciones.

Uso de la Distribución Binomial para calcular el Intervalo de Dificultad

Con el objeto de comparar los resultados se usará la distribución binomial (Ver *Noticia ICI E-10*). Para un cuestionario de opción múltiple de cinco opciones se tiene una esperanza matemática de 20%. Si se busca el intervalo de valores que discriminen por lo menos al 0.2 de acuerdo con la distribución binomial se tiene:

$$p+q=1 \qquad pq=0.2$$

Combinando ambas ecuaciones se llega a:

$$p^2 + p + 0.2 = 0$$

ecuación de segundo grado que se resuelve para

$$p1 = 0.7236 \text{ y } p2 = 0.2764$$

Estos serán los límites de dificultad aceptables para un reactivo de 5 opciones. Para 4 opciones se tiene

p1 = 0.75 y p2= 0.25

Generalmente en la literatura sólo se reportan valores de reactivos de 5 opciones. Esta aproximación brinda solamente intervalos de dificultad, pero no informa acerca de la norma mínima, para ello habría que hacer una prueba de hipótesis de la diferencia de medias, con los problemas ya anotados en la *Noticia ICI E-11.*

Nota 19. *Se debe recordar que en las fórmulas que ofrece el modelo Siseval®[9] (ver Nota 5) la k se refiere a las categorías de respuesta del ítem, por lo que se incluyen los casos de cualquier número de categorías incluyendo "elegida o no" o "elegida entre 2, 3, 4, 5 o más categorías".*

Es interesante observar, como ya se ha hecho antes, que el criterio de trabajo virtual máximo, brinda un rango de dificultades similar a otros modelos, informando al mismo tiempo, del valor de la norma discriminativa y resolviendo las necesidades planteadas en los postulados iniciales."

11. Conclusiones

Se mostró que los criterios estadísticos fallan respecto a los propósitos de la evaluación, por lo que resultó obligatorio un criterio determinista de fácil manejo e interpretación para poder trabajar con diversas poblaciones e instrumentos y poder hacer estudios sistemáticos referidos a esta norma precisa.

Se demuestra la obtención de la Norma Discriminativa y el dominio de aceptación de los reactivos.

Se hizo ver que la norma rigurosa dada por las expresiones (4.8) y (4.9) fue ajustada en la práctica, reduciéndola un 20% para facilitar el trabajo de los evaluadores.

---

[9] Publicadas por ICI en Mariano Jiménez 1839, Col. Balcones del Valle, CP 78280, San Luis Potosí, S.L.P. At´n. Guadalupe Carrión FAX (48) 15 48 48 y Tel. (48) 20 37 88.

Por último se proporcionó una expresión general (4.12) que puede ser manejada muy fácilmente por los evaluadores, pero a los que debe recordarse que debe emplearse con las debidas precauciones, ya que puede conducir a criterios indefinidos, incorrectos o que hagan difíciles las comparaciones entre cuestionarios y poblaciones."

En la publicación *Noticias ICI E-15* del 20 de marzo de 1995 "Prueba de hipótesis para la discriminación (Definición de la Norma Discriminativa de un reactivo)", se menciona:

1. Definición de variables y valores. Dado el cuadro de valores observados:

	R.C.	**R.I.**	**TOTAL**
GS	a	b	NA
GI	c	d	NB
TOTAL	N1	N2	N

Donde GS=grupo superior, GI=grupo inferior, las respuestas se clasifican en R.C. (respuesta correcta) y R.I. (respuesta incorrecta) y las frecuencias observadas en cada caso son a, b, c y d.

Se pueden obtener los valores de las frecuencias en cada casilla, a partir del conocimiento del Grado de Dificultad (GD) y la Discriminación (D) del reactivo como sigue:

$$GD = \frac{a + c}{N} \times 100$$

$$PD = \frac{a + c}{N} \times 100$$

Nota 20. *Con anterioridad ya se ha mencionado que la fórmula del Poder de Discriminación no corresponde a la anterior, sino a:*

$$PD = \frac{a}{n_a} - \frac{c}{n_c}$$

Dado el error conceptual en el manejo de la correlación Gamma, cualquier derivación que se realice a partir de este falso supuesto acarreará el sesgo del cálculo.

se observa:

$$NI = GD \times \frac{N}{100}$$

$$N2 = N - NI = N - GD \times \frac{N}{100} = N \left[\frac{100 - GD}{100}\right]$$

2. Modelo de hipótesis. Se dice que un reactivo discrimina correctamente si su proporción de aciertos en el grupo superior es significativamente mayor que la proporción de aciertos en el grupo inferior.

Puede realizarse una prueba χ^2 que se expresa de modo general:

$$\chi^2 = \Sigma \frac{(o - e)2}{e}$$

Siendo "o" las frecuencias observadas y "e" las esperadas.
Para una tabla de 2 x 2 se reduce a:

$$\chi^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)} = \frac{N(ad - bc)^2}{N1\ N2\ NA\ NB}$$

Nota 21. *No deja de llamarnos la atención que después de tratar de invalidar los modelos estadísticos en repetidas ocasiones, ahora el autor de KALT trate de usar la prueba de Ji Cuadrada para sustentar*

la Norma Discriminativa. En todo caso la aceptación de esta fórmula implicaría la aceptación de la correlación Fi, por ser modelos afines de estadística: $\chi^2 = Nr_{\phi}^2$. *Compárese la similitud de las fórmulas cuando se trabaja con tablas de contingencia de 2×2.*

$$r_{\varphi} = \frac{(ad - bc)}{\sqrt{(a+b)(c+d)(a+c)(b+d)}} = \frac{(ad - bc)}{\sqrt{N1\ N2\ NA\ NB}}$$

Con 1 grado de libertad, para la prueba de hipótesis:
Hipótesis Nula H_0: a = c
Hipótesis alternativa: a > c
con frecuencias esperadas bajo el criterio de datos independientes:

$$\chi^2 = \frac{N(a(N/2 - c) - (N/2 - a)c)^2}{N1\ N2\ NA\ NB}$$

con un nivel de significación

3. Desarrollo. Se substituyen en el modelo las diversas variables:
Al rechazar H_0, se tiene:

$$\frac{200(PD)^2}{GD(100 - GD)} \geq nl\ (\%)$$

Con n1(%) el valor de χ^2 para un nivel de significación dado % y se resuelve para PD:

$$PD \geq \sqrt{\frac{GD(100 - GD)}{200}\ nl(\%)}$$

Esto indica que la discriminación tiene que cumplir con la desigualdad para que pueda rechazarse H_0.

Al mínimo valor de D se le llamará NORMA DISCRIMINATIVA, quedando:

$$ND = \sqrt{\frac{GD(100 - GD)}{200}}\ nl(\%)$$

[...]"

En una reunión del Consejo Asesor Externo del CENEVAL se solicitó que se comparara la calificación de los ítemes mediante varios modelos, con el objeto de verificar la equivalencia de dictamen. Si los ítemes son los mismos y los sujetos que responden son también los mismos, debe haber equivalencia entre la calificación de los ítemes con diferentes modelos bajo los principios de replicación directa (investigación cuantitativa) y de triangulación (investigación cualitativa).

Para estos fines de equivalencia de dictamen se presenta un análisis realizado sobre la comparación entre los modelos Kalt, Bigsteps, XCalibre y Siseval. Con el propósito de verificar las equivalencias de calificación entre varios modelos, se compararon los resultados de cuatro modelos en 5 años con dos versiones por año, haciendo un total de 1280 ítemes, con las respuestas de una muestra aleatoria simple de 10,000 casos por año y versión.

Se hace la aclaración de que "equivalencia de calificación" implica que ambos modelos hayan calificado igual al ítem, independientemente de que haya sido aceptado o rechazado.

	Calificados igual								Max - Min = 29.77
	K	**RM**	**RZ**	**XL**	**Res**	**XC**	**XT**	**S**	**MODELOS**
	914	1233	1211	1190	1270	1229	1223	1245	**Aceptados en todos**
K		913	893	924	914	915	915	911	**Kalt**
RM	71.33		1192	1199	1225	1236	1236	1234	**Rasch Medio Cuadrático**
RZ	69.77	93.13		1187	1207	1204	1202	1196	**Rasch Estandarizado**
XL	72.19	93.67	92.73		1190	1211	1217	1203	**X Calibre Tres Parámetros**
Res	71.41	95.70	94.30	92.97		1227	1221	1245	**X Calibre Residual**
XC	71.48	96.72	94.06	94.61	95.86		1274	1232	**X Calibre Clásico**
XT	71.48	96.56	93.91	95.08	95.39	99.53		1234	**X Calibre rasgo latente**
S	71.17	96.41	93.44	93.98	97.27	96.25	96.41		**Siseval**

$27 \leq GD \leq 73$; $RD \geq 1.00$

Estándares típicos de los modelos

En este caso se utilizaron los indicadores básicos con sus valores críticos naturales, de acuerdo con los modelos, pudiéndose observar que el modelo Kalt resulta ser rígido en extremo.

Se puede verificar que las distancias entre las calificaciones por binas resultan tener una diferencia de 29.77%. La combinación de modelos que difieren más en la calificación de los ítemes es Kalt y Rasch estandarizado, calificando igual al 69.77% y en la que difieren menos es XCalibre Clásico y XCalibre con rasgo latente, coincidiendo en el 99.53%. Se hace notar que, sin tomar en cuenta a Kalt, la diferencia máxima entre los demás modelos se encuentra entre Rasch estandarizado–XCalibre Logístico (92.73), y XCalibre Clásico–XCalibre con rasgo latente (99.53), por un total de 99.53–92.73 = 6.80, quedando todos los modelos en una igualdad de calificación en un mínimo de 92.73 que resulta bastante aceptable.

Se toman ahora los valores críticos extremos de Rasch y Siseval para verificar si se logra que los modelos coincidan en el 90% como mínimo:

	Calificados igual								Max - Min = 8.98
	K	**RM**	**RZ**	**XL**	**Res**	**XC**	**XT**	**S**	**MODELOS**
	1217	1233	1211	1190	1270	1229	1223	1245	**Aceptados en todos**
K		1184	1162	1159	1213	1186	1186	1208	**Kalt**
RM	92.50		1192	1199	1225	1238	1236	1234	**Rasch Medio Cuadrático**
RZ	90.78	93.13		1187	1207	1204	1202	1196	**Rasch Estandarizado**
XL	90.55	93.67	92.73		1190	1211	1217	1203	**X Calibre Tres Parámetros**
Res	94.77	95.70	94.30	92.97		1227	1221	1245	**X Calibre Residual**
XC	92.66	96.72	94.06	94.61	95.86		1274	1232	**X Calibre Clásico**
XT	92.66	96.56	93.91	95.08	95.39	99.53		1234	**X Calibre rasgo latente**
S	94.38	96.41	93.44	93.98	97.27	96.25	96.41		**Siseval**

$12 \leq GD \leq 88$; $RD \geq 0.45$

Valores con los que Kalt se iguala a los estándares de Rasch, XCalibre y Siseval

Se puede verificar, por último, que las distancias entre las calificaciones por binas resultan tener una diferencia de 8.98%. La combinación de modelos que difieren más en la calificación de los ítemes es Kalt y XCalibre Logístico, calificando igual al 90.55% y en la que difieren menos es XCalibre Clásico y XCalibre Logístico, coincidiendo en el 99.53%.

Un asunto no resuelto cabalmente por el Consejo Técnico tiene que ver con el hecho de que los límites naturales del modelo Kalt, establecen un Grado de Dificultad entre ≅ 27 (26.795) y su complementario ≅ 73 (73.205), y una Razón Discriminativa igual o mayor que 1.00, sin que haya coincidencia en el punto de incidencia de la pendiente de descenso de discriminación y la línea de ascenso del 30% del Grado de Dificultad sugerido como discriminación por el modelo. El valor correspondiente a la incidencia de las líneas de ascenso y descenso mencionadas sería de 36.6%, lo que implicaría una altísima exigencia en el nivel de discriminación de los reactivos. Por este motivo, el Consejo Técnico originalmente optó por situar los límites del Grado de Dificultad entre 23 (≅23.08) y 77 (≅76.92), a fin de lograr incidencia entre las líneas de ascenso y descenso mencionadas, para una Razón Discriminativa del 1.00 o mayor.

Sin embargo, posteriormente, con el fin de dar una mayor flexibilidad al proceso de construcción del banco de reactivos del examen, recomendó flexibilizar el criterio de discriminación, estableciendo como valor mínimo de la Razón Discriminativa, el 0.70 el cual corresponde al 21% del Grado de Dificultad en vez del 30% sugerido por el modelo. Esto plantea la cuestión de si deben modificarse los límites del Grado de Dificultad, con el fin de hacerlos congruentes con el punto de incidencia de la pendiente de discriminación que marca el modelo Kalt.

En los programas de Rasch para ordenador se reportan dos valores (*infit*, promedio ponderado de los residuos estandarizados y *outfit*, residuo cuadrático medio) para cada uno de los procedimientos de cálculo del ajuste (medio cuadrático y estandarizado), en los que se sugieren como límites aceptables de ajuste:

Medio cuadrático: $0.80 \leq MC \leq 1.20$
Estandarizado: $-2.00 \leq Z \leq +2.00$

Los valores máximos admitidos por el modelo de Rasch[10], más allá de los cuales la información llega a ser irrelevante y confusa, corresponden a:

$12 \leq GD \leq 88$ (en porcentaje)
$-1.992 \leq D \leq +1.992$ (en lógitos)

Los valores se han redondeado a $-2.00 \leq D \leq +2.00$ que corresponden a valores entre $11.92 \leq GD \leq 88.08$ (en porcentajes).

Se ha tratado de igualar a Kalt con los demás modelos, pero pudiera ser óptimo igualar a los demás modelos con Kalt. Sin embargo, debemos tomar en cuenta que los modelos pueden estar sustentados en dos criterios básicos con respecto al azar para determinar si un ítem debe o no ser aceptado:

Incluirlo, asumiendo que el azar pueda favorecer a los sustentantes de mayor rendimiento al tener dudas con respecto a una menor cantidad de las opciones de respuesta, y teniendo información sobre los casos extremos donde se ubican sustentantes que sólo teniendo un buen manejo del tema pueden responder acertadamente. Esto implica un mayor rango de dificultades al contener la mínima dificultad de la variabilidad aceptada en el límite inferior (LIE) y el máximo en el límite superior (LSE), tomándose los límites externos del dominio.

Excluirlo, evitando cualquier sesgo debido al azar con el riesgo de no tomar en cuenta a los sustentantes que realmente responden con acierto un item en extremo difícil, por dominar ampliamente el tema. Esto implica un menor rango de dificultades al contener la máxima dificultad de la variabilidad aceptada en el límite inferior (LII) y la mínima en el superior (LSI), tomándose los límites internos del dominio.

Si mantenemos $\mathbf{N} = 100$ como valores de porcentaje, $\mathbf{k}$ como número de opciones y $\mathbf{z}$ como valor probabilístico con un error máximo de $\alpha = 0.05$, tenemos que[11]:

[10] En *Apendix 2: Diagnosing misfit* del manual de Linacre & Wrigth se mencionan como ítemes "mudos" los que se encuentran por debajo de MnSq < 0.80 o Stdzd < -ð2.00 y "ruidosos" los que se encuentran por arriba de MnSq > 1.20 o Stdzd > +ð2.00

[11] Fórmulas derivadas por el Dr. Miguel Angel Rosado.

$$Z_{LIE} = \frac{N - z\sqrt{N(K-1)}}{k} - \frac{1}{2} \qquad Z_{LSE} = \frac{N(K-1) + z\sqrt{N(K-1)}}{k} + \frac{1}{2}$$

Modelo Siseval (Binomial con aproximación a z, ejemplo con 5 opciones).

LIE = 11.66 ≅ 12 = (((Ñ – (z*RAIZ(N*(k̃-1))))/k̃) – (1/2))
LSE = 88.34 ≅ 88 = ((((N*(k̃–1)) + (z*RAIZ(N*(k̃-1))))/k) +(1/2))

$$\chi^2_{LIE} = \frac{N}{k} - Z\sqrt{\frac{N}{k}} \qquad \chi^2_{LSE} = N - \frac{N}{k} + Z\sqrt{\frac{N}{k}}$$

Modelo Siseval (Ji Cuadrada, ejemplo con 5 opciones)

LIE = 11.23 ≅ 11 = (N/k)˜ (z*RAIZ(N/k))
LSE = 88.77 ≅ 89 = Ñ (N/K) z*RAIZ(N/k))

			z				χ^2			
N	**k**	**z**	**LIE**	**LII**	**LSI**	**LSE**	**LIE**	**LII**	**LSI**	**LSE**
100	**5**	**1.96**	**12**	28	72	**88**	**11**	29	71	**89**
100	**4**	**1.96**	16	34	66	84	15	35	65	85
100	**3**	**1.96**	24	43	57	76	22	45	55	78
100	**2**	**1.96**	40	**60**	**40**	60	36	**64**	**36**	64

Se puede observar que a medida que decrece el número de opciones se aproximan los límites los valores entre sí cuando se utilizan los límites externos, ocurriendo lo contrario si se utilizan los límites internos hasta llegar a ser incongruentes, quedando el límite inferior como superior y viceversa cuando se trata de dos opciones. Este punto refuerza la idea de utilizar los límites externos para cualquiera de los casos.

El valor crítico de la discriminación para cualquier caso estaría dado por $z/\sqrt{N} = 1.96/\sqrt{100} = 0.196 \cong 0.20$ (aceptado por el CENEVAL), mediante el coeficiente de correlación Fi (r_ϕ).

En la parte complementaria se espera un porcentaje máximo de respuestas del 80% con un Ítem de cinco opciones, oscilando entre 71.66% y 88.34%. Por arriba del 88.34% corresponde a un Ítem en extremo fácil para la población a la que se le aplica, o errores tales en los distractores que permiten sólo elegir la clave por exclusión.

Ahora, observemos lo que ocurre en un ítem de dos opciones (k=2). Prácticamente se tiene un porcentaje de acertar o fallar al 50% con una oscilación entre 39.70% y 60.30%, donde sólo deberían aceptarse los ítemes que se responden por arriba del 60.30% si se excluye el azar, o por arriba del 39.70% si se incluye. Cualquier porcentaje de respuestas por debajo del 39.70% deberá rechazarse por no responder a las expectativas de respuesta con respecto al modelo utilizado. Los modelos que excluyen el azar también deberán rechazar el rango entre 39.70% y 60.30% aceptando sólo los que tengan un mínimo de acierto del 60.30% o mayor.

Resumiendo los aspectos principales, tenemos que:

En la base del modelo el autor asume que los aciertos del grupo superior (SA) crecen de 0 a 50, mientras que los aciertos del grupo inferior (IA) permanecen en 0. A partir de 50 el grupo superior permanece en 50 creciendo los aciertos del grupo inferior hasta llegar a 50. Sin embargo, en ningún momento explica por qué el grupo inferior no puede acertar ni por azar hasta que el grupo superior llega a su total de respuestas posible. Si el total de aciertos fuera 10 podría ser que la relación fuera SA = 10, IA = 0, pero también podría ser SA = 9, IA = 1, SA = 8, IA = 2, etc. donde la relación no es lineal sino matricial.

No existe una relación directa entre el Grado de Dificultad y el Poder de Discriminación, que autorice el uso de un porcentaje del primero como valor crítico del segundo. De acuerdo con el modelo en la primera mitad (dificultad de 0 a 50) la correlación es positiva perfecta y en la segunda mitad (dificultad de 50 a 100) la correlación es negativa perfecta, quedando para el total una ausencia de correlación.

Para que se cumpla que “la mediana divide en dos partes iguales (salvo un individuo cuando la población es impar)”, por omisión el modelo Kalt asigna al azar entre un 8 y un 9% de los casos en secciones de 10

ítemes y cerca del 19% en secciones de 24 ítemes que corresponden a la clase mediana.

Si el autor del modelo Kalt acepta que "En general las pruebas deben medir TODO EL RANGO DE DIFICULTADES,...", no se explica por qué el rango de dificultad que utiliza rechaza un 54% de información, aceptando sólo dificultades con un rango entre 27 y 73%.

De hecho ningún modelo "SOSTIENE que GD es óptimo si vale 50%", sino que el promedio de dificultades debe estar cercano al 50%. Se puede verificar por qué se dice que el óptimo valor de dificultad es 50% si tomamos en cuenta que este porcentaje de acierto "corresponda al grupo de mejor rendimiento" que no ha sido tomado en cuenta por el autor del modelo.

El modelo Kalt es sensible a la variación de casos en los grupos superior (GS) e inferior (GI), dejando esta variación al azar sin que esta desigualdad afecte ni a la correlación Fi ni a la correlación Gamma.

Es curioso que el autor del modelo Kalt mencione que "si se prolonga la recta (3.1), se tiene una intersección con la recta 2 de dominio A en GD = 76.92" (GD≅77) y no asocie este dato con los valores que asigna al rango del Grado de Dificultad (27↔73), debiendo ser, de acuerdo con el autor, un rango de 23↔77, para que realmente exista una incidencia.

El autor del modelo presenta un error de equivalencias al tratar de comparar porcentajes con valores de correlación. Si la manera como trata de calcular la correlación Gamma la multiplica por 100 para obtener el Poder de Discriminación, también debe multiplicar por 100 el valor de la correlación Fi para tener datos comparables, en cuyo caso Fi duplicaría el valor máximo del modelo.

Llama la atención que después de mencionar "el defecto inherente a las pruebas de hipótesis", utilice la χ^2 para justificar el Poder de Discriminación y la Norma Discriminativa, rechazando la correlación Fi (r_ϕ) que tiene como referente de valores críticos al modelo que utiliza en dicha justificación, donde $\chi_2 = Nr_\phi^2$.

Existe en el autor del modelo un error conceptual en el cálculo de la discriminación al utilizar $(a–c)/N \times 100$, en vez de usar $((a/n_a)–(cn_c)) \times 100$, teniendo como efecto que el modelo total presente un rango de discriminación entre –50 y +50 en vez de utilizar un rango entre –100 y +100, al multiplicar los valores de Gamma por 100.

Bibliografía

Adkins W., D. (1983). *Elaboración de tests,* Trillas, México.

Diederich, P. B. (1960). *Short-cut statistics for teacher made tests*, Educational Testing Service, Princeton, N. J.

Donlon, Thomas F. [Ed.] (1984). *The College Board Technical Handbook for the Scholastic Aptitude Test and Achievement Tests,* College Entrance Examination Board, New York.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment,* Mesa Press, Chicago.

Stockton, F. (1976). *Estadísticas aplicadas a las pruebas de rendimiento escolar,* Comisión de Nuevos Métodos de Enseñanza, UNAM México.

Tristán, L. A. (8 de marzo de 1995). "Relaciones entre grado de dificultad y discriminación. Primera parte: Estudio del grado de dificultad", *Noticias ICI E-10*, Ingeniería Computarizada Integral, San Luis Potosí.

Tristán, L. A. (11 de marzo de 1995). "Relaciones entre grado de dificultad y discriminación. Segunda parte: Estudio de la discriminación". *Noticias ICI E-12*, Ingeniería Computarizada Integral, San Luis Potosí.

Tristán, L. A. (11 de marzo de 1995). "Relaciones entre grado de dificultad y discriminación. Tercera parte: El dominio de discriminación", *Noticias ICI E-12,* Ingeniería Computarizada Integral, San Luis Potosí.

Tristán, L. A. (20 de marzo de 1995). "Prueba de hipótesis para la discriminación definitiva de un reactivo", *Noticias ICI E-15*, Ingeniería Computarizada Integral, San Luis Potosí.

Vidal, R.; Leyva, Y.; Tristán, A. y Martínez Rizo, F. (2000). *Manual Técnico del CENEVAL,* CENEVAL, México.

Wright, B. D. y Stone, M. H. (1998). *Diseño de mejores pruebas utilizando la técnica de Rasch,* CENEVAL, México.

http://www.ieesa-kalt.com/articulo1 ka.html