MODELO PARA EL ANÁLISIS DE REACTIVOS OBJETIVOS POR COMPUTADORA

Model for computer aided analysis of items

ABSTRACT

An original model that justifies the traditional item analysis is presented here, using some properties of the Rasch Model. It is shown that a good yardstick is needed for a good measure and the characteristics of a good scale are provided. Some false postulates contained in the traditional item analysis via difficulty and discrimination are studied and some new theorems are proposed; they justify the possibility to have a rational model based on the two parameters already mentioned. The model in this paper includes the Degree of Difficulty and the Discrimination Power, showing that the domain of permissible values of an item is a convex triangular domain in the plane DD-DP. The weaknesses of the traditional discrimination norms are shown and a new Norm of Discrimination and the Discrimination Ratio used in the KALT program are offered herein. This model is the only one possible to build from the available data, if no further hypothesis about the persons or the items are included.

RESUMEN

Se presenta un modelo original que fundamenta el análisis tradicional de reactivos, complementado en el modelo de Rasch. En este trabajo se muestra que para realizar una buena medida se tiene la necesidad de contar con una buena escala y se dan las características que debe incluir la escala de un instrumento de evaluación. Se aclaran algunas falacias contenidas en el análisis tradicional de reactivos basado en la dificultad y la discriminación y se postulan los teoremas que justifican la existencia de un modelo racionalmente construido. El modelo presentado en este trabajo incluye el Grado de Dificultad y el poder de discriminación, demostrando que el dominio de valores permisibles de los reactivos es de forma triangular. Se demuestra la debilidad de las normas discriminativas tradicionales y se ofrece la Norma del Modelo de KALT y la Relación Discriminativa. Este modelo es el único posible de construir a partir de la información disponible, sin incluir hipótesis adicionales sobre la población o el instrumento, permitiendo estudiar la calidad de la medida de manera unívoca.

KEYWORDS / PALABRAS CLAVE

Calificación por computadora. Dificultad. Discriminación. Escala. IRT. Lógito. Modelo de KALT. Norma discriminativa. Rasch.

INTRODUCCION

El problema básico y general para todo evaluador es el saber, con algún nivel de precisión, la calidad intrínseca que tiene la prueba empleada y, en particular, cada uno de los reactivos o ítems empleados.

De un punto de vista general se pueden distinguir dos grandes corrientes que permiten estudiar la calidad de un reactivo. La primera es la corriente tradicional, fundamentada en el análisis de dos parámetros: la dificultad y la discriminación del reactivo. La segunda, enfocada por la corriente de la Teoría de la Respuesta al ítem (IRT por sus iniciales en inglés, o TRI en español), donde a su vez se engloban en las referencias a dos tendencias: el análisis de Rasch, (denominado modelo de un parámetro) y los modelos de dos y tres parámetros.

Desde la forma de agrupar ambas corrientes se tienen problemas, debido a que tienen implícitas connotaciones no necesariamente correctas. Este es el caso del análisis de Rasch que, en realidad, no forma parte de la corriente de la IRT, pero atendiendo a que esta última generó modelos probabilistas de 1 a 3 parámetros de manera "muy parecida", cuando menos a la vista de sus autores, de lo que hace el análisis probabilista de Rasch, decidieron incluirlo en su IRT como un modelo de un parámetro. Nada más injusto para varios de los seguidores de Rasch, quienes se defienden continuamente de esta clasificación, no solamente por incluirlos en esta corriente, sino también porque se le identifica como modelo de un parámetro. Resulta evidente que, de acuerdo con

la IRT, el modelo más completo y preciso, el que se dice que ajusta mejor a los resultados es el de tres parámetros, siendo el de un parámetro (y por lo tanto para ellos, el modelo de Rasch) un modelo muy burdo e ineficiente que no ajusta con los datos. No es motivo de este trabajo identificar los problemas conceptuales contenidos en el IRT y en el modelo de tres parámetros, para ello pueden consultarse con provecho los trabajos de Wright. Por lo dicho aquí, en este trabajo se identificará el análisis de Rasch como una corriente distinta de la Teoría de la Respuesta al ítem.

En lo que respecta al análisis tradicional, de entrada se le está dando la terrible asignación de "tradicional" que lleva consigo una alta carga de "obsoleto", "viejo", "no actualizado". En efecto puede afirmarse que el análisis tradicional tiene un alto contenido de elementos incorrectamente planteados. Pero se sataniza como inconveniente todo análisis que contiene a la dificultad y a la discriminación como parámetros de trabajo, cuando lo único que hay que hacer es establecer el modelo correcto que utiliza a ambos.

Se encuentra el evaluador en una encrucijada muy compleja. Si se sigue una corriente debería reportar sus resultados en lógitos, siendo poco claro el significado del lógito para las autoridades, los estudiantes y el público en general. Si se reporta en términos de los tres parámetros se va a enfrentar con problemas conceptuales insalvables y con puntajes en escala logarítmica de difícil interpretación. Por último, si se reporta los resultados en los términos de grados de dificultad y valores de discriminación en términos de correlaciones o de valores de X2, se enfrentará con resultados "manipulables" y de muy baja o nula utilidad.

Una solución que se sigue muy comúnmente es adquirir o desarrollar un sistema de cómputo que haga todos los cálculos y que facilite la tarea de calificación y análisis de las pruebas. Generalmente se puede percibir que los usuarios de los programas "confían" en los resultados que emiten porque fueron obtenidos por la computadora. Pero no hay que olvidar que la computadora solamente es una herramienta que facilita los cálculos que han sido programados siguiendo algún esquema de cálculo o algoritmo.

Tanto en el caso del análisis por el método tradicional como en el de la IRT o el análisis de Rasch, es indispensable contar con un modelo. No se trata de repetir cálculos que se realizan a mano, o decisiones tomadas "por inspección", sino de realizar procedimientos justificados por medio de un modelo.

La pertinencia y necesidad de los modelos puede no ser evidente para un evaluador, ya que su tarea cotidiana la realiza "por inspección", atendiendo a la gran cantidad de datos y a la necesidad de obtener resultados rápidamente. Generalmente los procesos mentales que se efectúan a la "inspección" son poco sistematizables, dependen del criterio de cada persona y, por lo mismo, se prestan a múltiples justificaciones y "permisos" que se otorgan al evaluador para agilizar su tarea.

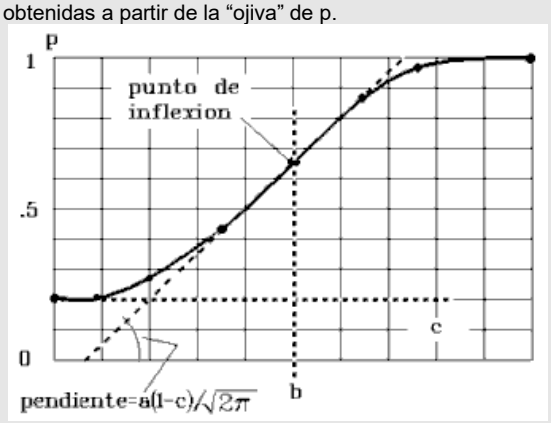
La Teoría de la Respuesta al Ítem se fundamenta en hipótesis "fuertes" basadas en la probabilidad de respuesta de una pregunta y de una persona. La teoría puede emplear uno o más parámetros para estimar dicha probabilidad, aunque es demostrable que, a partir de la información disponible, no pueden obtenerse más de dos parámetros.

Modelo de tres parámetros:

p = c + (1 – c) / (1 + $e^{-1.7a(\phi-b)}$ **)**

donde Ø representa el rasgo medido (habilidad, capacidad, conocimiento)

p representa la probabilidad de respuesta al rasgo de n el ítem y por la persona.
a, b, c son parámetros de interpolación relacionados con la *dificultad* "b", la *discriminación* "a" y la *adivinación sistemática* "c" obtenidas a partir de la "ojiva" de p.



Modelo de Rasch
El modelo de Rasch es un modelo probabilista que hace lineal e independientes la medida de la persona y la medida del ítem, con la expresión:

$$p = e^{(Bn-Di)} / (1 + e^{(Bn-Di)})$$

donde
p representa la probabilidad de respuesta
Bn es un estimador de la medida de la persona n
Di es un estimador de la medida de la dificultad del ítem i
El modelo de Rasch se expresa en lógitos (logaritmo natural del momio), de su nombre en inglés LOGIT = LOG ODD RATIO. El lógito es la unidad de medida dada por:
Lógito = log (pni/qni) = log (pni/(1-pni))
Siendo log el logaritmo natural del momio pni/qni, donde pni es la probabilidad de respuesta de la persona n al ítem i.

La sistematización de los procesos de inspección con ayuda de la computadora se reduce a comparaciones simples, ajustando lo más posible al criterio del evaluador; también se hace la aplicación de pruebas de hipótesis estadísticas en el mejor de los casos. Ambos procedimientos contienen un mismo problema: no necesariamente son modelos para evaluación educativa. Desgraciadamente se siguen tomando, con ayuda de la computadora, las mismas decisiones injustificadas y permisivas, ya que el programa trata de emular la decisión del evaluador.

Este es uno de los problemas del esquema "tradicional" y el mismo que contiene el modelo de la IRT: buscar que los modelos ajusten a los datos. Esta aproximación no sigue el método científico, pero sí es una forma de trabajo común en todo el mundo cuando no se sabe qué otra cosa mejor hacer.

Si se trabaja con el modelo de Rasch o con un modelo especialmente diseñado para la dificultad y la discriminación, los problemas se clarifican y el evaluador puede tomar algunas decisiones más sensatas. Un modelo diseñado para la dificultad y la discriminación se encuentra integrado al sistema KALT para calificación de reactivos objetivos, en vista de ello se le denominará como "Modelo de KALT".

El mensaje que se pretende transmitir en este trabajo es que si el evaluador desea tomar sus decisiones exclusivamente en términos de la información disponible, entonces no debe hacer intervenir hipótesis adicionales que no se puedan deducir de los datos disponibles. Estas hipótesis adicionales, por mas “razonables” que se consideren, no son sino maquillajes para los datos y pueden conducir a cualquier lugar, atendiendo a la corriente del evaluador y a las hipótesis que haga.

1. El problema de la medida. Uso de Banco de Reactivos.

Todos los autores y los evaluadores en la práctica conocen que el problema de la medida forma parte de las primeras dificultades que deben resolverse. A pesar de que el problema de base, la forma de trabajar la medición difiere entre evaluadores. Posiblemente sea uno de los aspectos en que todos hablan de lo mismo sin estar de acuerdo en lo que hablan, dando diferentes interpretaciones a una misma cosa.

Hay varios problemas relacionados con la medida y que a su vez definen el proceso de evaluación:
a) Definir la escala

b) Comparar a las personas respecto a la escala

c) Ubicar a las personas respecto a la escala

d) Emitir juicios de valores respecto a las personas

e) Emitir juicios de valor respecto del instrumento

En la generalidad de los casos se publican e informan los datos relativos al punto (c): son las calificaciones escolares, las puntuaciones para ingreso a las escuelas, las puntuaciones promedio de una institución para compararse con las de otras escuelas, etc. Por añadidura, toda persona que recibe o analiza los datos emitidos, puede con toda facilidad pasar al punto (d), emitiendo juicios de valor respecto a las personas sobre todo si se trabaja en una escala en base 10, para la cual es costumbre en nuestro medio reportar puntuaciones desde la escuela primaria.

Dice Ruiz Massieu: "*Resultó dramático conocer que de los alumnos admitidos en bachillerato en el decenio 1976-1985, la calificación promedio de corte en una escuela de diez fue de 3.85 y que si la Universidad hubiera aceptado sólo a quienes obtuvieron 6 o más de calificación, sólo hubiera admitido en promedio al 7.6% de los aspirantes. En el mismo tenor se encuentra la admisión a estudios profesionales, en donde el promedio de calificación de corte en el mismo lapso es de 4.56*."

Cuando se reportan las cifras es claro que se tiene un grave problema de deficiencia educativa o un problema en la escala. Nadie sabe, por el simple reporte de promedios en base 10, si el instrumento y la escala fueron los adecuados para medir a la población. ¿Cómo es posible que los alumnos que sustentan el examen que reporta el autor obtengan del orden de 4 en base 10, si se trata de egresados de ciclo anterior? ¿Cómo es que pudieron egresar? La evaluación está inmersa en problemas no solamente de evaluación *per se*, sino que incluye aspectos sociales, económicos y políticos que hacen difícil la elaboración de conclusiones si no se quiere uno meter en aguas profundas.
El problema es complejo y sencillo al mismo tiempo. Bastaría con conocer cual es la escala de medida. No se trata de una escala del 0 al 10, que no dice nada, sino de la verdadera escala (la del instrumento para ser más preciso) utilizada para medir a las personas.

La definición de la escala es, por lo tanto, el primer problema a resolver. En KALT se reportan los valores en varias escalas: porcentual, percentilar, normalizada y, en la próxima versión en lógitos. En Rasch se usa la escala de lógitos. Lo que hay en común entre ambos modelos es que antes de establecer la escala, el evaluador puede (y, de hecho debería siempre) eliminar los reactivos ineficientes que afectan la precisión del instrumento de medida. Es muy común que este proceso no se realice, sino que se dejan los efectos negativos de los reactivos mal diseñados dentro de la calificación o puntaje de la persona, afectando automáticamente la calidad de la escala.

La escala en lógitos tiene la ventaja de ser una escala que hace lineal la proporción entre un puntaje y el grado de dominio que refleja. En la escala porcentual la diferencia que tienen dos personas en su grado de dominio no es lineal. Por ejemplo: Una persona obtuvo 30% de puntuación y otra 40% de puntuación, su "distancia" es de 10%, lo mismo se tiene para las otras dos personas que obtuvieron 70% y 80%; sin embargo el grado de dominio necesario para pasar de 30% a 40% es muy distinto que el de pasar de 70% al 80%. En una escala de lógitos la diferencia en el primer caso es de 0.442 lógitos y en el segundo es de 0.539 lógitos, mostrando claramente que la escala porcentual no es lineal.

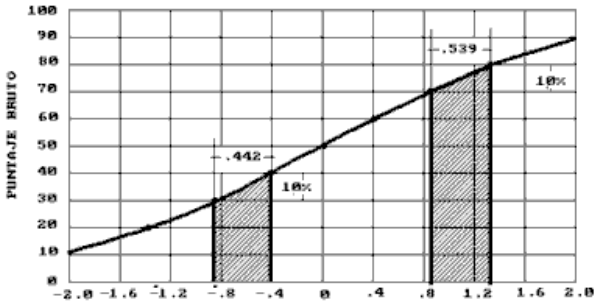


FIG.1 ESCALA EN LOGITOS

Puede decirse que al no conocer la escala del instrumento de medida con el cual reporta Ruiz Massieu, resulta imposible afirmar que la situación de los aspirantes a la UNAM es "desesperada". No es suficiente "sobrentender" que el 10 es dominio total y el 0 es dominio nulo. Dominio total ¿de qué? Dominio nulo ¿de qué? Se hace la hipótesis de que el cuestionario es un reflejo de los contenidos, habilidades, niveles cognoscitivos, etc. que se desean medir. Para poder afirmar tal cosa debe garantizarse que el muestreo es realizado rigurosamente. Es muy fácil demostrar lo difícil (o imposible) que es conseguir un muestreo riguroso en un cuestionario dado.

En resumen se tienen estos elementos de una escala, para los cuales se presenta un ejemplo de calor.

ELEMENTOS PARA DETALLAR UN INSTRUMENTO DE MEDIDA

a) La especificación de la escala	Temperaturas
b) La unidad de medida	Grados Celsius
c) El rango de medidas	36 oC - 40 oC para uso humano
	0 oC – 100 oC para agua
	100 oC – 300 oC para pasteles
	800 oC – 1400 oC para cerámica
d) La precisión del instrumento	1 / 2 oC para uso humano
	1 / 4 oC para agua
	10 oC para pasteles
	50 oC para cerámica artística
	0.1 oC para cerámica industrial de alta precisión
e) Las condiciones de uso	Al ambiente para uso humano
	Al nivel del mar para agua
	Al ambiente para pasteles
	A una velocidad de 10 oC/min para cerámica artística
	En atmósfera controlada para cerámica industrial de alta precisión

Si se usa el termómetro para agua (con rango de 0 a 100) y se reportan las temperaturas corporales de varias personas, se podría estar muy desilusionado de que el ser humano no alcance el 100% de temperaturas del termómetro: No alcanzaría el 100% de dominio, dicho en términos de pruebas de conocimientos. Pero esto carece de sentido, ya que lo esperado es que el ser humano se encuentre alrededor de los 37ºC, que es la temperatura "normal", valores por arriba o por abajo de este número indican anormalidades que pueden ser hasta peligrosas.

Con el ejemplo de las temperaturas puede apreciarse que es necesario saber cual es el valor "normalmente esperado" y poder medir diferencias respecto a dicho valor. Cuando al aplicar un examen se tiene que el valor esperado es "10" (recuérdese que 10 es el máximo), la escala está, automáticamente, mal construida. No es de extrañar que se reporten resultados de 3.85 y se juzguen como pésimos.

Una vez comentado esto y regresando al tema del trabajo, puede plantearse un examen por medio de una herramienta de cómputo. Si la computadora se programa para que emule los pasos realizados por el hombre para construir su cuestionario, no hay duda de que se generarán pruebas mal construidas. El programa generador de pruebas a partir de un Banco de Reactivos deberá contemplar la posibilidad de resolver los elementos indicados: Especificar la escala, establecer la unidad de medida, establecer el rango de medida, la precisión del instrumento y las condiciones de uso.

El CENEVAL está haciendo un gran esfuerzo por cambiar el enfoque de la presentación de los resultados de una prueba aplicada, con objeto de que no se hagan juicios fáciles a partir de los puntajes brutos obtenidos de la aplicación. En los próximos meses el sistema para gestión del Banco de Reactivos estará funcionando y permitirá que los exámenes sean eficientemente preparados, tomando en cuenta los elementos para detallar un instrumento de medida.

2. El problema de la dificultad

El segundo problema fundamental está relacionado también con el instrumento. En particular con el rango de aplicación y la precisión del instrumento. Resulta difícil, si no imposible, que una persona alcance los 100ºC de temperatura corporal. Pero no es posible juzgar mal a las personas por dicha incapacidad, ni juzgar bien al instrumento de manera independiente de lo que está midiendo: el termómetro para el agua no está hecho para el rango humano.

Lo que dice Dorothy Adkins es muy significativo del error conceptual que se puede tener respecto al instrumento de medida: "*Las opciones de los expertos en pruebas mentales todavía difieren en el sentido de si todos los reactivos deberían ser aproximadamente del 50% de dificultad, o si se debería preferir el tener reactivos con un rango bastante amplio de dificultad y con un promedio de 50*". Al término del párrafo concluye: "*Así, pues, el 50% de nivel de dificultad para un solo reactivo es óptimo si todas las otras condiciones permanecen constantes. En el medio educativo, sin embargo, cuando menos algunos pocos reactivos significativamente más fáciles y otros pocos significativamente más difíciles que los del nivel del 50% de dificultad, son incluidos por lo regular, con la intención de motivar un poco a los estudiantes más malos y de desafiar a los mejores*."

Hay varios aspectos a resaltar en estas afirmaciones: Afirmar que el 50% de dificultad es óptimo; mencionar que las otras condiciones permanecen constantes; incluir algunos pocos reactivos más fáciles para motivar a los malos e incluir algunos pocos reactivos más difíciles para desafiar a los mejores. La escala de medida no se plantea en ninguno de estos casos de manera correcta. En particular dice Dorothy Adkins "si todas las otras condiciones permanecen constantes", revisando cuidadosamente el texto, no se aclara cuales deben ser las otras condiciones que deben permanecer constantes. Pero dejando de lado este detalle, a continuación se refutarán estas afirmaciones tradicionales para muchos evaluadores.

2.1 El rango de dificultad y el rango de la escala.

Si se diseña un termómetro que mida exactamente 37ºC para uso humano, se tiene el problema de que sólo se sabrá que algunas personas tienen exactamente esta temperatura, pero que muchas más estarán por arriba o por abajo de 37ºC, sin posibilidades de saber con precisión su temperatura corporal. Un termómetro de este tipo resulta muy inconveniente.

Hay instrumentos de medida diseñados para tomar decisiones dicótomas: "pasa-no-pasa", aceptado-rechazado" Si tal es el propósito de un cuestionario debe diseñarse en esa dirección. Pero la generalidad es que los cuestionarios se diseñen para poder tomar decisiones correctivas a partir de los resultados que se obtienen, para poder ubicar mejor a las personas dentro de un continuo de habilidades o conocimientos. El instrumento debe, pues, ser pensado con otras características.

En particular, igual que el caso del termómetro, deben considerarse reactivos de toda la gama de dificultades posibles, cubriendo todo el rango de medida. Si el rango debe estar entre 36 y 40 ºC y la

precisión deseada es de 1/2º, entonces se debe "graduar" el termómetro como sigue: 36, 36.5, 37, 37.5... 40. En el caso de un instrumento de medición de conocimientos la graduación deberá contener dificultades bajas y altas. Lo ideal sería contar con dificultades graduadas, por ejemplo, de 1 en 1, para poder correr todo el dominio entre 1 y 99 de dificultad.

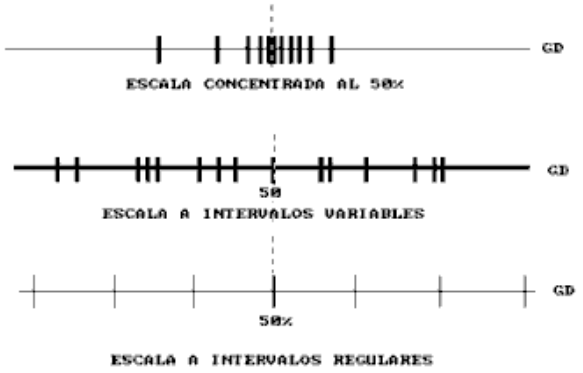


FIG.2. DEFINICION DE LA ESCALA

Obsérvese que **el objetivo de incluir reactivos de diferentes dificultades es disponer de una escala bien graduada,** no "motivar" a los malos y "retar" a los mejores. **Se trata de contar con una escala precisa y útil que permita identificar la posición de cada persona de la mejor manera posible. Esto permite ubicar sin lugar a dudas el grado de conocimientos o habilidades que dispone una persona, desde el que tiene pocos conocimientos o habilidades hasta la más apta de las personas**.

Es claro que habrá pocas personas de muy bajo dominio, pocas en el rango más alto y la gran mayoría se ubicarán en zonas centrales, sobre todo si la población se distribuye de manera normal. Esta distribución normal de las personas es independiente de la graduación de la escala. Al igual que la temperatura corporal el termómetro tiene una escala graduada uniformemente, pero las temperaturas de las personas se distribuyen de manera normal.

Es un error común pensar que la población se distribuye de manera normal porque las preguntas o reactivos se prepararon con dificultades que se distribuyen de manera normal. Recuérdese que son dos cosas diferentes el instrumento y la población, pudiendo distribuirse también en forma diferente.

En conclusión: **las pruebas deben diseñarse con reactivos graduados en dificultad lo más uniformemente posible, cubriendo todo el rango deseado de la escala**.

El **Grado de Dificultad** se define por medio de este cociente:

$$GD(\%) = \frac{\text{Suma de respuestas correctas}}{\text{Total de personas que responen}} 100 \qquad [1]$$

En símbolos se acostumbra escribir como p al porcentaje de aciertos, en este caso se escribirá p(G) el porcentaje de aciertos de la población G, por lo que:

$$GD(\%) = \frac{p(G)}{G} x100 = px100 \qquad [2]$$

A su vez, se define la Medida del Reactivo por la expresión:

Medida = □ = log [(1-p(G))/p(G)] [3]

Las unidades de la Medida son lógitos (el lógito es el logaritmo natural del monto).

A su vez □ = Bn – Di [4]

siendo Bn la medida de la persona n y Di la medida del reactivo i.

Puede establecerse una equivalencia entre GD y la medida:

GD(%) = 1/N [1/(e□ -1)] x 100 [5]

que corresponde con el modelo de Rasch.

El Grado de Dificultad es un valor cuyo estimador más probable es p(G) o simplemente p. Esto quiere decir que se trata de un índice muy estable, sobre todo cuando se está usando un mismo instrumento en poblaciones "razonablemente" similares. Por ello la medida 0 es asimismo un valor estable en diferentes aplicaciones y, de hecho, el análisis de Rasch postula que 0 es independiente de la población y del instrumento.

2.2 La dificultad óptima

La otra afirmación es muy común es que la dificultad óptima ocurre para 50%. La discusión presentada anteriormente ya hizo ver que no debe planearse un examen con reactivos al 50%, puesto que esto llevaría consigo un deficiente instrumento de medida.

Adicionalmente debe aclararse que no hay forma de afirmar que la dificultad óptima es de 50%. El error se comete al afirmar tal cosa parte del comportamiento de la varianza de la distribución binomial. A partir de un reactivo con p porcentaje de aciertos y q porcentaje de errores, tal que p+q es igual a 1, se tiene que la varianza es el producto pq, que corresponde con el número de casos que se puede discriminar en una población.

De este modo, si se dan valores a p y q, la varianza 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 entre 0 y 1 (su complemento q es decreciente), y el producto pq toma valores que algunos autores identifican con una distribución normal (sin serlo, como se muestra más adelante), ascendente de 0 a 0.5 y descendente de 0.5 a 1. El máximo ocurre para p=0.5 (50%), con un valor de discriminación posibles de .25 (25%).

Partiendo de esta distribución los evaluadores afirman que la óptima dificultad es por lo tanto en p=0.5 (50%), sin percatarse que es la varianza la que obtiene el óptimo al 50%, pero no puede afirmarse nada sobre la dificultad.

Si desde el punto de vista de la lógica es evidente la falacia del razonamiento de Adkins y de otros autores (Tristán-2), también es fácil la demostración del error desde el punto de vista matemático.

Defínase la función varianza:

$$s2=pq=p-p2 \qquad [6]$$

Se trata de una parábola (no representa por lo tanto una distribución normal) cuyo vértice se encuentra en (0.5, .25), con su concavidad hacia abajo.

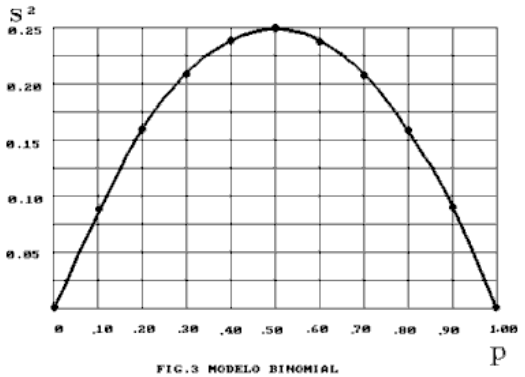


FIG.3 MODELO BINOMIAL

Por construcción de la parábola se sabe que el máximo ocurre para 0.5 y vale 0.25, lo cual también puede demostrarse por la condición necesaria de extremo. Matemáticamente se obtiene que el máximo de la varianza ocurre para p=0.5, pero así como puede afirmarse que el óptimo s2 es 0.25, nada permite afirmar que p es óptimo en 0.5.

De hecho no hay una dificultad "óptima" general. El óptimo depende de los propósitos de la evaluación, de las características de la población, etc. El óptimo de la temperatura corporal es 37ºC no 50ºC, la temperatura óptima para hornear el pastel depende del tipo y tamaño del pastel deseado, no se encuentra a 200ºC simplemente por ser el valor medio del rango de 100ºC a 300ºC en el horno de la estufa.

Esta es una de las razones por las que el "análisis tradicional" falla, debido a que se hace una hipótesis adicional respecto a la escala al propiciar mediciones al 50% de dificultad de manera injustificada.

3. El problema de la discriminación

El segundo parámetro de importancia es la discriminación. En el "análisis tradicional" este parámetro se ha prestado a todo tipo de hipótesis sin fundamento. De hecho en muchos libros ni siquiera aparece una forma razonable de calcular la discriminación existiendo numerosas definiciones directas, así como otras indirectas en términos de pruebas de hipótesis igualmente mal planteadas.

Revisando de nuevo el libro de Adkins, se tiene esta sugerencia para el cálculo de la discriminación. "*Un procedimiento útil cuando el número de casos es 100 o más requiere que primero dividamos al grupo en dos mitades, inmediatamente se hará uso de la gráfica para la computación de la correlación tetracórica... la mayoría de las autoridades en el campo del análisis de reactivos probablemente no consideraría que el coeficiente tetracórico, con sus presupuestos un tanto restrictivos, es la mejor medida que se puede utilizar, pero con todo, se la utiliza con mucha frecuencia precisamente por su facilidad de computación*." Mas adelante dice: "*Seguramente que resulta ahora claro que la sola inspección del número de sujetos que pasen los reactivos de los grupos superior e inferior, revelará de inmediato si la correlación entre el reactivo y el criterio es positiva o cercana a cero o negativa.*

Para ciertos propósitos, tal inspección por sí sola será suficiente, evitando al profesor el trabajo de convertir a porcentajes los números de cada grupo-criterio, que han pasado el reactivo; tener que leer los coeficientes tetracóricos de la gráfica y registrarlos."

De nuevo se tienen varios problemas en esto. En primer lugar no aparece una definición simbólica acerca de lo que se entiende por discriminación. En vez de ello se explica un procedimiento tipo "caja negra" donde debe usarse una gráfica para la correlación tetracórica, cualquier cosa que esto quiera decir. Curiosamente afirma que algunas autoridades no están de acuerdo en el uso de la correlación tetracórica por sus presupuestos restrictivos que, desde luego, tampoco aclara en que consisten. Para completar el cuadro presenta unas conclusiones acerca de un procedimiento que permitiría no usar esta correlación tetracórica, consistente en la inspección visual del número de personas que contestan en los grupos inferior y superior. Todo este panorama es muy débil.

La correlación tetracórica efectivamente tiene un supuesto muy restrictivo. Está relacionada con un patrón de cotejo para respuestas esperadas por azar. No se discutirá aquí todo el conjunto de implicaciones que tiene esta hipótesis, pero es necesario insistir que **los reactivos se redactan justamente para no ser contestados por azar. Existe un patrón de respuesta que difiere substancialmente del azar y que ubica claramente porque hay una respuesta correcta y otras opciones que no lo son. En el caso de azar no hay respuesta correcta**. Haciendo de lado este hecho, la correlación tetracórica es una prueba débil, con la cual los evaluadores pretenden afirmar que el reactivo "discrimina" cuando en realidad sólo está comparando contra un patrón de respuestas azarosas. El uso de las gráficas era necesario hace años que no había computadora, pero ahora se puede programar el algoritmo de la correlación tetracórica en caso de desearse.

El procedimiento de inspección visual es igualmente débil de soportar. Si no hay un modelo matemático de por medio, las comparaciones entre los grupos superior e inferior quedan al buen criterio del evaluador, mismo que puede cambiar de un año para otro, incluir efecto de "halo" y ser diferente entre evaluadores. **Ningún proceso de toma de decisiones en evaluación debería hacerse "por inspección visual"**. Esta es otra de las fallas del "análisis tradicional".

El **Poder de Discriminación PD** tiene una definición matemática simple. Sea G la población, que se divide de acuerdo con la mediana en dos subgrupos GS (Grupo superior) y GI (Grupo inferior) y defínanse las siguientes expresiones:

G = GS + GI	Población	[7]
p(G) = p(GS) + p(GI)	Porcentaje de aciertos	[8]
PD(%) = (p(GS) - P(GI)/Gx100	Poder de discriminación	[9]

La discriminación identifica por lo tanto las diferencias entre los grupos superior e inferior. Se espera que las personas del GS respondan siempre mejor que las del GI (discriminación positiva), en caso contrario se tendría una discriminación negativa que indicaría un reactivo ineficiente, pudiendo ser confuso para las personas del GS, o que propicia la respuesta por azar, entre muchas otras causas más.

Algunos evaluadores prefieren dividir al grupo en dos partes separadas por los cuartiles, definiendo el grupo superior como el 25% de personas de más alto porcentaje de aciertos y como grupo inferior al más bajo 25%. Al tomar las colas de la distribución se pretende eliminar el efecto que tienen las personas que están al centro, cerca de la mediana. Al reducir la población se tiene una prueba necesaria pero no suficiente, con esto se quiere decir que esta prueba sólo es útil cuando la discriminación resulta negativa (si es que llega a ocurrir), porque en generalidad de casos el resultado es positivo y no permite probar nada. Este problema se demuestra en el teorema 3 y el corolario 2, presentados en el apartado siguiente.

Las dos formas de definir a la discriminación (dividida en la mediana por cuartiles) pueden incluirse en programas de cómputo. De hecho KALT contiene ambas opciones que pueden manejarse a elección del usuario, aunque sólo se recomienda la división por la mediana.

4. El problema de la relación entre la dificultad y la discriminación.

Ninguna referencia presenta relación alguna entre dificultad y discriminación. En realidad, tal y como se definen ambos parámetros, existe una relación funcional entre ciertos valores de ellos.

La deducción siguiente parte de la tabla de contingencias de un reactivo, resumida en su forma 2 x 2, con valores normalizados en porcentaje. No

se hace el análisis con valores nominales, ya que es fácil demostrar que el usar valores nominales automáticamente hace que las pruebas de hipótesis fallen en su ámbito de aplicación al hacer intervenir el tamaño de la población. No es propósito de este trabajo hacer la demostración correspondiente, pero puede demostrarse que pruebas como X2, diferencia de medias, etc., fallan sistemáticamente al hacer intervenir a la población en valores nominales, gracias a ello los evaluadores de grandes poblaciones pueden llega a "demostrar" que su instrumento de evaluación es bueno estadísticamente.

Regresando a la tabla 2 X 2 con datos normalizados a 100, se observa que es una tabla de un solo grado de libertad. Esto quiere decir que una vez conocidos los valores marginales, al disponer de uno de los elementos de la tabla se deducen automáticamente los demás. En este caso el grado de libertad se tendría, por comodidad, en el porcentaje de respuestas correctas del Grupo Superior p(GS). Se define entonces un espacio vectorial sobre los números reales, formado por los porcentajes de aciertos y fallas de la población, con una operación entre ellos (suma) y una aplicación externa (multiplicación por cualquier número real). La dimensión del espacio vectorial formado por los porcentajes de aciertos y fallas de la población es de dimensión 1, siendo la BASE formada por un solo valor.

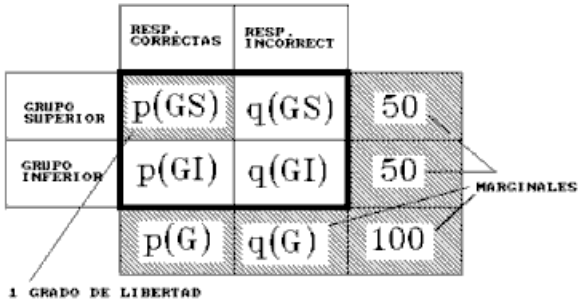


FIG.4 TABLA DE CONTINGENCIAS 2 X 2

Se tiene además que dicha tabla 2 x 2 más los subtotales marginales conducen a una nueva tabla (que se denominará aquí hipertabla) de 3 x 3, con dos grados de libertad. Los dos grados de libertad de un reactivo serían, por ejemplo, el propio Grado de Dificultad GD y, de nuevo, el porcentaje de respuestas correctas del Grupo Superior p(GS). De esta tabla no pueden generarse más hipertablas, quedando por lo tanto que estos dos grados de libertad forman una BASE para el espacio vectorial formado por el conjunto de porcentajes de respuestas totales y de los grupos superior e inferior. Dicha BASE tiene por lo tanto una dimensión de 2.

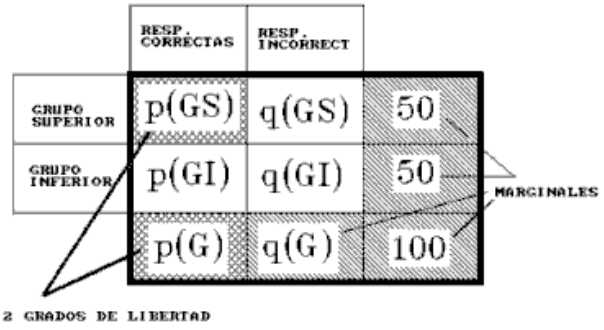


FIG.5 HIPERTABLA 3 X 3

La hipertabla se construye dado un porcentaje de aciertos totales p(G) y un porcentaje de aciertos del grupo superior p(GS). Cualquier otro parámetro (**grado de dificultad, poder de discriminación** u otro que se desee), diferente de P(G) y p(GS) será dependiente de los dos datos que forman la BASE.

La BASE puede establecerse con cualquier pareja de valores independientes dentro del espacio vectorial de resultados del reactivo. Obsérvese que las definiciones de **Grado de Dificultad y Poder de Discriminación** pueden realizarse por incluir exclusivamente las operaciones permitidas en el espacio vectorial.

Esto quiere decir que **a partir de la información disponible de un reactivo no se pueden establecer más de 2 parámetros independientes.** Esto desconcertará, desde luego, a los evaluadores que usan el modelo de tres parámetros que no se han percatado que no es posible obtener 3 parámetros independientes: solamente dos son independientes, el tercero por lo tanto dependerá de los dos primeros y será entonces redundante. Esto puede desconcertar adicionalmente porque uno de los postulados del modelo de tres parámetros es que son independientes.

No obstante, con objeto de demostrar sus afirmaciones, involucran algunas hipótesis muy liberales, siendo la más importante de ellas el concepto de la "adivinación sistemática" como tercer parámetro que, como se puede apreciar de este análisis, carece de sentido.

A partir de lo anterior se establecen cuatro Teoremas:

Teorema 1

De la información disponible de un reactivo, el número máximo de parámetros independientes que puede tenerse es 2.

Este teorema ya fue demostrado en los párrafos anteriores. La importancia del teorema 1 estriba en que si se desean obtener más parámetros para un reactivo deberán, por fuerza, hacerse intervenir hipótesis adicionales. Lo interesante de esto es que nada justifica la inclusión de hipótesis adicionales en el análisis.

Corolario 1

Dado un reactivo, toda la información que contiene la tabla de contingencias 2 x 2 se puede representar en un plano.

Este corolario es evidente a la luz del álgebra lineal y no requiere demostración.

Teorema 2

Dada la información completa de un reactivo (a partir de la tabla de contingencias que incluye todas las opciones del reactivo), el número máximo para parámetros necesarios y suficientes para el análisis es 2.

Este teorema no se demuestra aquí, pero puede comprenderse fácilmente si se recuerda que solamente existe una respuesta correcta y las demás opciones son distractoras. Resulta irrelevante para el estudio del reactivo que se trate de 2, 3 o más opciones distractoras, ya que tanto el Grado de Dificultad como el Poder de discriminación están referidos a la respuesta correcta. Todos los distractores pueden englobarse en una sola opción nueva que es la "respuesta incorrecta", es decir, se dicotomiza el comportamiento del reactivo, lo cual conduce a la explicación del teorema 2.

Teorema 3

Dada la información de un reactivo, si los datos se agrupan en m subgrupos de personas y las respuestas se dicotomizan, entonces la dimensión del espacio vectorial que definen es m.

Este teorema se demuestra estableciendo la hipertabla del reactivo organizado en m subgrupos y determinando los grados de libertad de la tabla. Su demostración es muy sencilla y no se hace aquí (fig.6).

Corolario 2

Dado un reactivo definido para una dimensión m, el número de parámetros independientes es m.

Esto es importante porque al establecer más subgrupos, se pueden obtener más parámetros sin hipótesis adicionales. Por ejemplo, al dividir a la población en tres subgrupos el evaluador deberá establece tres parámetros independientes para su estudio, con lo que la dificultad y la discriminación se vuelven insuficientes.
Cuando se establece que la discriminación se debe realizar con los subgrupos definidos en términos de las cuartiles y solamente se trabaja con la dificultad y la discriminación, el corolario 2 indica que el estudio es insuficiente: no se pueden tomar decisiones tan sólo con los dos parámetros "tradicionales".

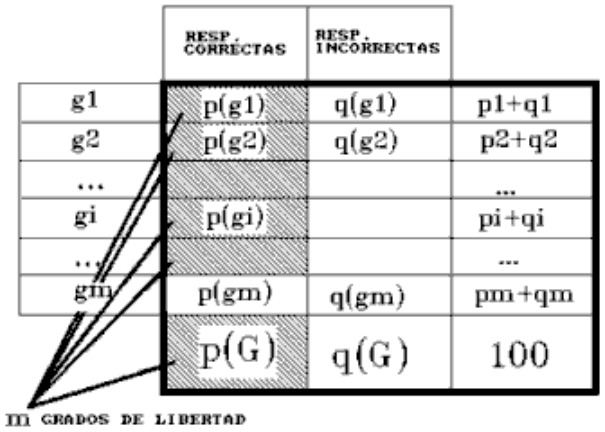


FIG.6 HIPERTABLA DE m SUBGRUPOS

Teorema 4

Dado un reactivo y dos parámetros independientes, la condición necesaria y suficiente para poder tomar decisiones es que la tabla sea de dimensión 2.

La demostración no se presenta aquí, sería motivo de otro trabajo, pero brinda la plena seguridad al evaluador que desea trabajar con dos parámetros que basta con establecer el análisis de la tabla de contingencias 2 x 2 para poder emitir juicios suficientes sobre el reactivo.

Como puede verse, la importancia de estos teoremas radica en la posibilidad de hacer análisis suficientes y completos a partir de la información disponible y sólo de dicha información, sin incurrir en el planteamiento de hipótesis adicionales que, como demuestra el teorema 2, son injustificadas. El teorema 4 en particular, tranquilizará a los evaluadores para mostrarles que lo que han estado haciendo de manera "tradicional" puede no ser tan malo después de todo, siempre y cuando se disponga de un modelo.

Resulta altamente conveniente, y hasta indispensable, que el criterio de partición de la población sea la mediana, ya que permite trabajar con la tabla de contingencias 2 x 2 y, por lo tanto, con el espacio vectorial de aciertos del reactivo.

Gracias al corolario 1 puede pensarse en dibujar los dos parámetros del reactivo en el plano donde se define la BASE. Como se recordará del álgebra lineal, una vez identificada la dimensión del espacio vectorial basta con establecer dos vectores independientes para la BASE, pero no hay

compromiso en la elección de dichos vectores, es decir, puede elegirse la pareja de valores como mejor convenga a los fines del estudio.

Se elige el porcentaje de aciertos totales, es decir, el **Grado de Dificultad,** como primer parámetro. El segundo parámetro puede elegirse de nuevo como el porcentaje de respuestas correctas o aciertos del Grupo Superior, pero esto puede resultar poco útil para estudios posteriores, por ello aquí se elige el Poder de Discriminación como segundo parámetro.

No es posible elegir a la varianza pq como segundo parámetro porque conduciría a un análisis no lineal, la operación pq no está definida en el espacio vectorial de los porcentajes del reactivo.

De este modo se establece el plano **Grado de Dificultad-Poder de Discriminación** que puede representarse como se muestra en la fig. 7 (Tristán-1).

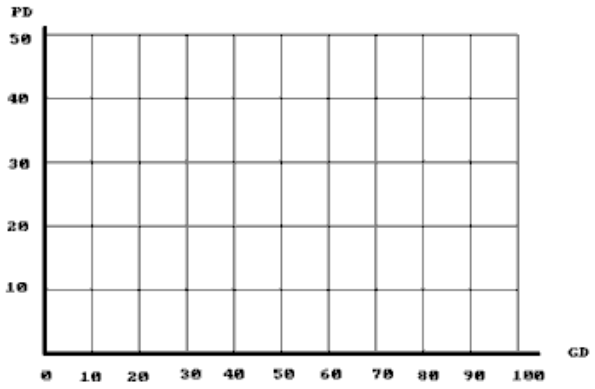


FIG.7 PLANO DIMENSION 2, DIFICULTAD-DISCRIMINACION (SOLO DISCRIMINACION POSITIVA)

Todos los posibles valores de los dos parámetros elegidos de un reactivo caen en algún lugar de este plano. Puede demostrarse además que hay un dominio de valores permisibles para los parámetros de un reactivo. Para ello obsérvese que el **Grado de Dificultad** fluctúa entre 0 100% (nadie contesta - todos contestan correctamente), por su parte el **Poder de Discriminación** varía entre -50 y +50 (discriminación pésima - discriminación óptima).

Si se calculan los valores de la máxima discriminación posible dado una dificultad de un reactivo, se tiene esta tabla:

GD	0	10	20	30	40	50	60	70	80	90	100
GS(máximo)	0	10	20	30	40	50	50	50	50	50	50
G (mínimo)	0	0	0	0	0	0	10	20	30	40	50
PD(máximo)	0	10	20	30	40	50	40	30	20	10	0

Al representar el Poder de Discriminación máximo contra el grado de dificultad, se tiene que el dominio de valores permisibles para GD y PD está limitado por dos rectas, como se muestra en la figura 8, obteniéndose un dominio de forma de triángulo.

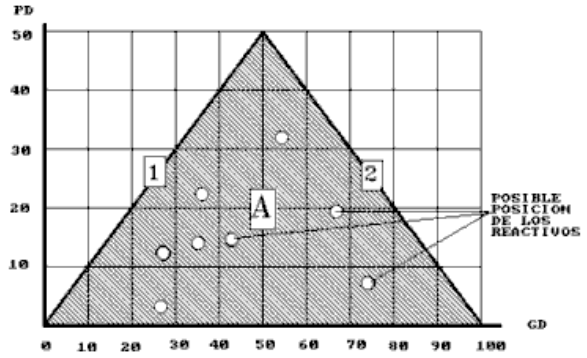


FIG.8 DOMINIO DE VALORES PERMISIBLES

Las ecuaciones de las rectas 1 y 2 se determinan fácilmente (Tristán -3). Se observa que, de nueva cuenta, el óptimo de discriminación ocurre para una dificultad de 50%, sin haber ningún óptimo para el **Grado de Dificultad.** Se denomina **Dominio A** a la zona triangular limitada por las rectas 1 y 2, y el eje horizontal con PD=0. Se tiene que los reactivos cumplen estas propiedades:

1- TODOS los reactivos que discriminan positivamente se encuentran en A

2. Ningún reactivo puede discriminar más alto que el valor limitado por las rectas 1 y 2.

3. La mínima discriminación positiva es 0

Del **Dominio A** puede concluirse que los reactivos deseables óptimos NO se encuentran al 50% de dificultad, sino se encuentran para cualquier Grado de Dificultad dada directamente sobre las rectas 1 ó 2. Pero sería igualmente ilusorio esperar que todos los reactivos de una prueba discriminen en forma óptima. Tampoco es suficiente decir que basta con que discriminen por arriba de cero.

De ello se plantea el problema que consiste en establecer los límites inferiores de aceptación para

el **Poder de Discriminación** de los reactivos, a partir de la información disponible.

Antes de pasar a presentar una solución al problema, debe comentarse que existe un **Dominio A',** simétrico de **A** respecto al eje PD=0, que corresponde con los reactivos que discriminan negativamente. Este **Dominio A'** carece de interés, ya que considera que los reactivos que discriminan negativamente son inconvenientes para la evaluación. El dominio completo de valores permisibles es la unión de los dominios A y A', de forma de rombo, simétrico respecto al eje horizontal PD=0 y respecto al eje vertical GD=50. Este dominio es dibujado por **KALT** como resultado de la corrida y se presenta en las referencias (Tristán).

En resumen, se ha mostrado aquí que se puede definir una BASE de dimensión 2 para los porcentajes de aciertos contenidos en la hipertabla 3 x 3, con la cual pueden caracterizarse los reactivos **exclusivamente a partir de la información disponible, sin incluir hipótesis adicionales.** Se mostró que a partir de la información disponible se puede localizar unívocamente cualquier reactivo en el plano GD-PD, definiéndose un dominio triangular para los reactivos que discriminan positivamente (un rombo cuando se incluyen también los reactivos que discriminan negativamente), cuyo límite superior corresponde a la máxima discriminación posible que, a su vez, es dependiente del **Grado de Dificultad** del reactivo.

5. El problema de la norma discriminativa.

Ya se ha dicho que se espera que los reactivos discriminen positivamente y que no hay un valor óptimo para la dificultad, sino que deben plantearse reactivos en toda la gama de dificultades para disponer de un buen instrumento de medida. Sin embargo no solamente se espera que la discriminación sea positiva sino también que sea alta. Una discriminación de 5% será suficiente? ¿o tal vez es preferible que los reactivos discriminen 10%?

Defínase la **Exigencia** para un **Grado de Dificultad** dado, a la relación entre la norma y la cantidad de personas que contestan el reactivo, en porcentaje:

Exigencia (GD%) = ND / GD x 100 [10]

con ND la **Norma Discriminativa** y GD el **Grado de Dificultad.**

Con este parámetro se juzgarán a continuación algunas normas "tradicionales".

5.1 Normas "tradicionales"

Existen recomendaciones de algunos autores para establecer una norma mínima del 10% del número de personas. En este caso se trabaja con poblaciones normalizadas a 100%, por lo que se tendría una norma de fácil aplicación del 10%. Se trata de una recomendación empírica pero nada permite demostrar este valor.

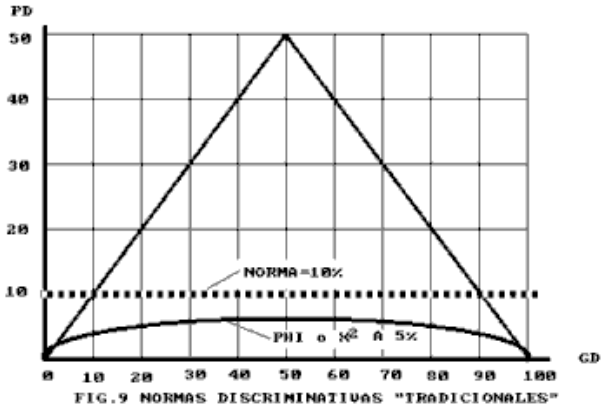


FIG.9 NORMAS DISCRIMINATIVAS "TRADICIONALES"

Esta norma corta a las rectas 1 y 2 del **Dominio A** en los puntos GD=10 y GD=90. De acuerdo con esta norma, todos los reactivos deberían discriminar por arriba del 10%. El rango de dificultades aceptables es entre 10 y 90. Ahora supónganse dos reactivos, uno de 20% de **Grado de Dificultad** y otro del 80%; a ambos se les pide que discriminen por lo menos 10%, esto quiere decir que la **Exigencia** de la norma es, para cada caso:
Exigencia (20%) = 10/20 x 100 = 50% [11]

Exigencia (80%) = 10/80 x 100 = 12.5% [12]

Este resultado es paradójico: para un reactivo difícil, donde hay menos personas que contestan, la **Exigencia** que es más alta que en el caso de reactivo fácil donde hay más gente que contesta. La norma propuesta no es sensible al número real de personas que contestan correctamente, sino solamente al número de personas que forman la población. Una norma como esta es muy exigente para los reactivos difíciles y benévola para los reactivos fáciles. La lógica de construcción de reactivos diría lo contrario: debería contarse con

una prueba suficientemente justa tanto para fáciles como para difíciles.

Una segunda forma de definir la norma discriminativa es por medio de X2, uso de la correlación tetracórica, prueba de diferencia de medias, PHI, Gamma, etc. Puede demostrarse que este tipo de pruebas conduce a formas similares a la siguiente expresión (Tristán-4):

$$ND2/C1 + (GD-50)2/502 = 1 \quad [13]$$

con

$$C1^2 = N1(\%)/200 \quad [13a]$$

ND es la norma discriminativa dada por la prueba de hipótesis.

Esta expresión se deduce en otro documento (Tristán-5). Por el momento puede observarse que se trata de la ecuación de una elipse con centro en GD=50, PD=0, que pasa por GD=0 y GD=100, y cuyo máximo valor ocurre en GD=50 con el semieje C1. El valor de C1 depende del nivel de significación n1(%) que se establezca para la prueba de hipótesis con X2 de 1 grado de libertad. Para una significación dada, pueden encontrarse los límites de dificultad G1 y G2, que son las intersecciones de la elipse con las rectas 1 y 2 respectivamente.

Por ejemplo para un nivel de significación del 5% n1(5%)=3.84 y se tiene la norma:

$$ND = 0.13856\ [GD(100-GD)]1/2 \quad [14]$$

La máxima discriminación que exige esta norma es de 6.92% y la dificultad aceptable se encuentra en el rango de 1.88 a 98.12. **Con una prueba de este tipo se puede decir que prácticamente cualquier reactivo será aceptable.**

Los evaluadores nacionales y extranjeros no se han percatado que el uso de las pruebas de hipótesis para definirla norma discriminativa conduce a este tipo de comportamiento poco exigente, ya que resulta muy conveniente poder afirmar que el conjunto de reactivos pasa la prueba de hipótesis.

Obsérvese que esta norma es muy más benévola que la del 10% presentada anteriormente y, como agravante, es creciente en exigencia para los reactivos de dificultad inferior al 50% (lado difícil) y reduce su exigencia conforme el reactivos e hace más fácil.

5.2 La Norma Discriminativa en el Modelo de KALT

Aunque no se demuestra en este documento, sino en las referencias (Tristán). Se tratará aquí de establecer una lógica para la idea de la **Norma.**

A partir de la **Exigencia (GD%)** si se establece que la Norma debe ser igualmente exigente sin importar el **Grado de Dificultad** del reactivo, se tendría que debería ser una constante, pudiendo escribirse:

$$\text{Exigencia (GD\%)} = \text{Norma} / GD \times 100 = k \quad [15]$$

siendo k una constante

De aquí se deduce una fórmula sencilla para la **Norma Discriminativa**:

$$ND = k/100\ GD \quad [16]$$

Esto indica que una exigencia igualmente justa conduce a una norma que es lineal dependiente del
Grado de Dificultad.

Sin demostrar la expresión de la **Norma de KALT,** se tiene que **KALT** exhibe:

$$ND = 0.3\ GD \quad [17]$$

k = 30%, la **Norma Discriminativa** es el 30% del **Grado de Dificultad.**

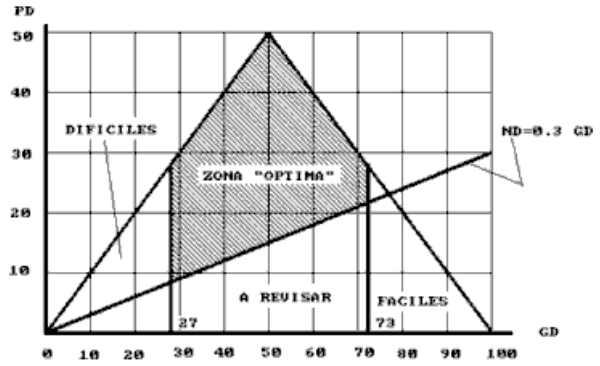


FIG.10 NORMA DISCRIMINATIVA MODELO DE KALT

Al deducir este valor se encontró asimismo el rango de aplicación que es de 27 a 73 para el **Grado de Dificultad,** esta deducción queda fuera del propósito de este documento. Resulta mucho más complicada que la explicación propuesta aquí en términos de la **Exigencia (GD%)** si se trata de

una aproximación lineal a dos parábolas que definen la **Norma Discriminativa.** Fuera de este intervalo no se puede demostrar la relación lineal. Sólo debe apuntarse aquí que esta deducción se hace por un procedimiento ajeno al uso de la distribución binomial y el intervalo tiene cierta semejanza simplemente por casualidad. Algunas personas han tratado de relacionar estos valores con los intervalos de confianza de la distribución binomial aplicada a reactivos de 5 opciones, pero debe insistirse que no hay relación alguna con dicha distribución.

Para fines prácticos, la **Norma Discriminativa** se extiende en **KALT** hasta cortar a la recta 2 y se amplía el rango de dificultades definiendo los reactivos difíciles (por abajo de 27%) y fáciles (por arriba de 73%), mismos que deben ser revisados por el evaluador.

Esta norma es la más racional, desde el punto de vista de la **Exigencia (GD%)** además de ser la más exigente respecto a las otras mencionadas. Los reactivos que pasan la **Norma de KALT** pueden catalogarse como reactivos "impecables", los que no satisfacen deben ser revisados en sus opiniones y modificarse en consecuencia.

5.3 La relación discriminativa.

Por último se establece una relación entre el **Poder de Discriminación** real del reactivo y el valor proporcionado por la **Norma Discriminativa** denominada **Relación Discriminativa RD:**

RD = PD / ND [18]

La **Relación Discriminativa** tiene esta interpretación:

RD > 1 Óptimos, discriminan por arriba de la Norma

RD = 1 Correctos, discriminan exactamente en la Norma

RD < 1 A revisar, discriminan por abajo de la Norma

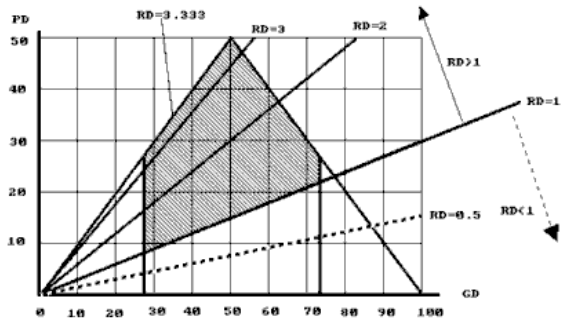


FIG.11 RELACION DISCRIMINATIVA

Con el uso de RD basta, ahora si, con una inspección visual para identificar los reactivos que deben ser revisados: aquellos cuya RD sea menor que 1. Aquí cabe el criterio adicional del evaluador para aceptar reactivos "ligeramente" abajo de la Norma. La máxima discriminación está identificada por la recta 1, donde RD=3.3333 (ó 10/3); ningún reactivo puede tener una **Relación Discriminativa** superior a 3.3333.

6. Modelos Y KALT

Muchos otros aspectos pueden ser comentados respecto al modelo "tradicional", mismo que no serán tratados en este trabajo. Por lo pronto obsérvese que la construcción de un modelo a partir de la información disponible en el reactivo resulta fundamental para que un programa de cómputo lleve a buen término un análisis de reactivos.

Aquí se ha presentado sólo parte del **Modelo de KALT,** mostrando la racionalidad de su fundamentación que hace que sea una herramienta mucho más allá de un simple programa que calcula "buenas", "malas" y saca promedios.

Es necesario disponer de modelos antes de tomar decisiones. Ninguno de los dictámenes de **KALT** se emite por inspección. De hecho el modelo se está reforzando y complementando cada vez más en función de las sugerencias o pedidos de los clientes y usuarios.

Como ya se apuntó anteriormente, se trabaja en adaptar la medida en lógitos y se desarrollará el modelo de discriminación en esta métrica, lo cual representará un avance significativo ya que el modelo de **Rasch** no considera explícitamente el uso de la discriminación del reactivo.

Además del modelo presentado en este trabajo, se incluyen modelos para el dictamen de los distractores, modelo para verificar la adivinación local (presente en un reactivo dado o en una persona dada), las respuestas inesperadas y las características a través de un análisis de varianza.

KALT se ha utilizado desde 1976 en diferentes medios, incluyendo la UNAM, universidades de varios estados de la República Mexicana y, recientemente, en el CENEVAL, con inmejorables resultados.

Se incluyen unos ejemplos de dictámenes de análisis de reactivos que emite **KALT** con base en

este modelo (anexo 1). Como puede observarse el reporte incluye la "nube" de reactivos dentro del Dominio A, del plano Grado de Dificultad-Poder de Discriminación, así como un reporte detallado reactivo por reactivo donde se asientan el Grado de Dificultad, el Poder de Discriminación, la Norma Discriminativa y la Relación Discriminativa. A partir de estos valores resulta muy cómodo emitir un dictamen de los reactivos.

El reporte **KALT** incluye una serie adicional de índices y parámetros estadísticos que no se explicaron en este documento y que son motivo de otros estudios.

CONCLUSIONES

Resulta imprescindible contar con modelos para la evaluación. Los análisis por "inspección", por "evidencia empírica", por "hipótesis razonables", conducen a conclusiones de índole práctica, pero no sustentables a menos que se disponga de un modelo que las justifique. El modelo presentado aquí es el mejor posible que puede construirse a partir de la información dada sin incluir hipótesis adicionales.

El modelo mejora y justifica el "análisis tradicional", puede aproximarse al modelo de **Rasch** y permite dictaminar la calidad de los reactivos de manera unívoca. El modelo forma parte de la fundamentación de **KALT** utilizando en la práctica con buenos resultados.

REFERENCIAS Y BIBLIOGRAFIA

Adkins W.D. "ELABORACION DE TESTS", Ed. Trillas, México 1983, pp 103 y sigs

Diederich P.B. "SHORT-CUT STATISTICS FOR TEACHER MADE TESTS",Educational Testing Service, Princeton, N.J. 1960.

Guilford J.P. y Fruchter B. " FUNDAMENTAL STATISTICS IN PSYCHOLOGY AND EDUCATION", 5A. ED. McGraw Hill Kogakusha, 1973

Landsheere G. "INTRODUCTION A LA RECHERCHE IN EDUCATION" 4a. ed. Armand Colin Bourrelier, París, 1976

Lord F.M. "APPLICATIONS OF ITEM RESPONSE THEORY TO PRACTICAL TESTING PROBLEMS", Lawrence Erlbaum Assoc. Nueva Jersey, 1980.

Magnusson D. "TEORIA DE LOS TESTS", Ed. Trillas, México, 1966.

Quesada R y Co. "EVALUACION DEL APROVECHAMIENTO ESCOLAR", Comisión de Nuevos Métodos de Enseñanza, UNAM, 2a. versión, 1975.

Rasch G. "PROBABILISTIC MODELS FOR SOME INTELLIGENCE AND ATTAINMENT TEST", Mesa Press Chicago, 1980.

Ruiz Massieu M. "EL CAMBIO EN LA UNIVERSIDAD", Universidad Nacional Autónoma de México, 2a. Ed. 1987, pp 20 y sigs.

Stockton F. "ESTADISTICAS APLICADAS A LAS PRUEBAS DE RENDIMIENTO ESCOLAR", UNAM. Comisión de Nuevos Métodos de Enseñanza 1976.

Thondike R.L. y Hagen E. " TESTS Y TECNICAS DE MEDICION EN PSICOLOGIA Y EDUCACION", Ed. Trillas México, 1975.

Tristán L.A. y González V. F. "SISTEMA DE CALIFICACION DE EXAMENES UTILIZANDO LA COMPUTADORA". Semanario de la Facultad de Ingeniería, UNAM, Año VIII, N.8, 14 dic. 1977.

Tristán L.A.-1 "MODELO DE LA EVALUACION PARA LA FACULTA DE INGENIERIA", UNAM, México, 1976-1977.

Tristán L.A.-2 "RELACIONES ENTRE GRADOS DE DIFICULTAD Y DISCRIMINACION. PRIMERA PARTE ESTUDIO DEL GRADO DE DIFICULTAD", Noticias ICI. México, E -10, 8 de marzo de 1995.

Tristán L.A.-3 "RELACIONES ENTRE GRADO DE DIFICULTAD Y DISCRIMINACION. SEGUNDA PARTE. ESTUDIO DE LA DISCRIMINACION", Noticias ICI, México E-12, 8 de marzo de 1995.

Tristán L.A.-4 "RELACIONES ENTRE GRADO DE DIFICULTAD Y DISCRIMINACION. TERCERA PARTE, EL DOMINIO DE DISCRIMINACION "Noticias ICI. México E-12, 11 de marzo de 1995.

Tristán L.A.-5 "PRUEBA DE HIPOTESIS PARA LA DISCRIMINACION. DEFINICION DE LA NORMA DISCRIMINATIVA DE UN REACTIVO. Noticias ICI México, E-15, 20 de marzo de 1995.

Wright B. "IRT IN THE 1990S: WHICH MODELS WORK BEST?, Rasch Measurement Transactions Vo. 6 N. 1, Primavera 1992, AERA pp 196-200.

Wright B.D., Stone M.H: "BEST TEST DESIGN", Mesa Press, Chicago, 1979.

ANEXO 1

KALT- RESUMEN TECNICO DEL CUESTIONARIO Pág 1

02-24-2005/20:11:30

PRUEBA DE DIAGNOSTICO
INSTITUCION: INSTITUTO BOLIVAR DE CORDOBA
FECHA: 20 DE ENERO DE 2005

RESULTADOS DE CALIDAD DEL CUESTIONARIO

Número de reactivos : 96 Reactivos con discriminación positiva: 91

GRADO DE DIFICULTAD (%)	
MEDIA:	54.81
POSITIVO PROMEDIO:	55.36
POSITIVO MINIMO:	17.94
POSITIVO MAXIMO:	96.51

PODER DE DISCRIMINACION (%)	
MEDIA:	7.40
POSITIVO PROMEDIO:	8.00
POSITIVO MINIMO:	0.44
POSITIVO MAXIMO:	17.62

RELACION DISCRIMINATIVA (%)	
MEDIA:	0.53
POSITIVO PROMEDIO:	0.57
POSITIVO MINIMO:	0.03
POSITIVO MAXIMO:	1.46

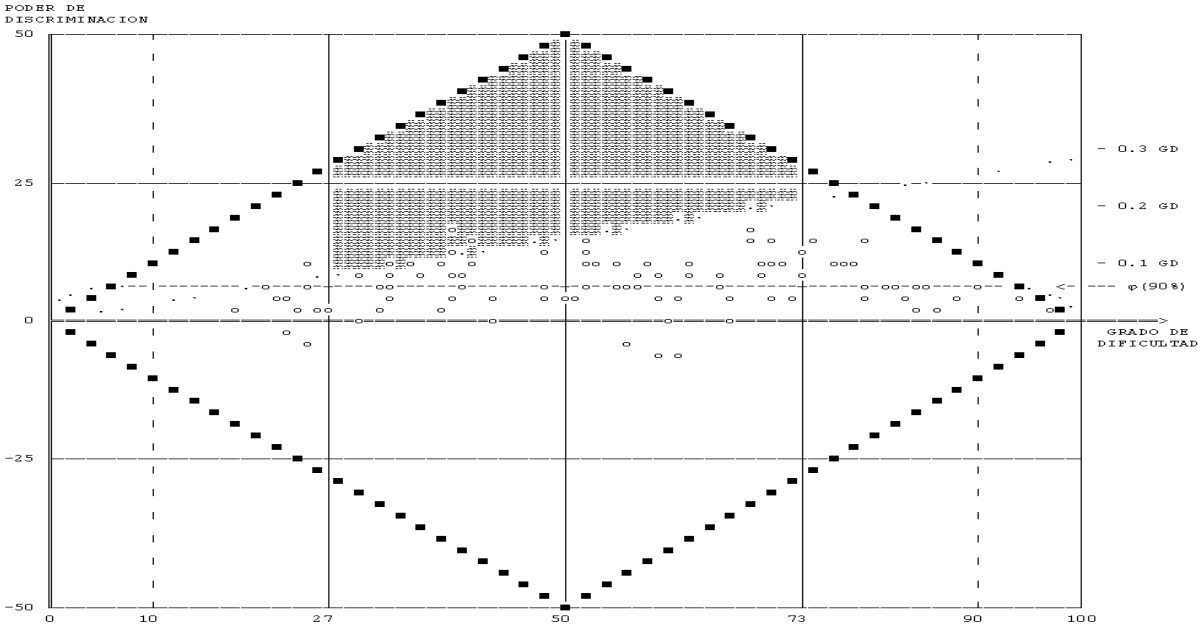
COEFICIENTE DE CONTINGENCIA (%)	
MEDIA:	16.10
POSITIVO PROMEDIO:	16.52
POSITIVO MINIMO:	0.25
POSITIVO MAXIMO:	33.71

CORRELACION DE ATRIBUTOS (%)	
MEDIA:	16.47
POSITIVO PROMEDIO:	16.91
POSITIVO MINIMO:	0.25
POSITIVO MAXIMO:	35.81

FACTOR KALT/MAX.DEPENDENCIA (%)	
MEDIA:	10.80
POSITIVO PROMEDIO:	11.47
POSITIVO MINIMO:	0.00
POSITIVO MAXIMO:	100.00

CONFIABILIDAD DEL INSTRUMENTO (α) : 0.73170

DIAGRAMA DISCRIMINACION/DIFICULTAD



NOTAS:
A) Los reactivos se ubican en el plano PD/GD con un círculo (o)
B) Se indican los extremos de las rectas al 10% de GD, 20% de GD y 30% de GD como referencia. La recta a 30% GD se identifica con un punto (·)
C) Se señala la posición aproximada de los valores de φ, al 90% de confianza (-)
D) Los reactivos arriba de la recta al 30% de GD y entre 27 y 73 de dificultad (zona sombreada) se consideran totalmente aceptables para KALT.
E) El evaluador podr modificar el criterio apoy ndose en las rectas de referencia indicadas en este diagrama

ANALISIS DE REACTIVOS (1) - CRITERIO 1 - VERSION 1 - TEMA 1 Pag. 1

REACTIVO	GRADO DE DIFICULTAD	DIFICULTAD MODIFICADA	PODER DE DISCRIMINACION	NORMA DISCRIMINATIVA	RELACON DISCRIMINATIVA	DICTAMEN DIFICULTAD	SOBRE DISRIMINACION	RESUMEN DEL DICTAMEN
1	17.00	-3.75	9.00	5.10	1.76	DIFICIL		√
2	43.00	28.75	23.00	12.90	1.78			√
3	8.00	-15.00	2.00	2.40	0.83	MUY DIFICIL	< NORMA	MUY DUDOSO
4	7.00	-16.25	5.00	2.10	2.38	MUY DIFICIL		A REVISAR
5	11.11	-11.11	7.07	3.33	2.12	DIFICIL		√
6	19.00	-1.25	13.00	5.70	2.28	DIFICIL		√
7	5.00	-18.75	3.00	1.50	2.00	MUY DIFICIL		A REVISAR
8	12.37	-9.54	10.31	3.71	2.78	DIFICIL		√
9	13.00	-8.75	-1.00	3.90	-0.26	DIFICIL	MUY MAL	DESECHABLE
10	40.00	25.00	-14.00	12.00	-1.17		MUY MAL	DESECHABLE
11	7.00	-16.25	5.00	2.10	2.38	MUY DIFICIL		A REVISAR
12	39.18	23.97	18.56	11.75	1.58			√

ANEXO 2.
COMENTARIOS SOBRE EL MODELO PRESENTADO

Con el objeto de hacer esta presentación lo más clara posible, se incurrió en un abuso de lenguaje desde el punto de vista de álgebra lineal, aunque sin perder generalidad.

El interés de los teoremas de la parte 4 se centra en el número de parámetros independientes necesarios para representar a la base del espacio vectorial. Cuando se trabaja con la hipertabla 3 x 3 el espacio vectorial está definido por el elemento genérico siguiente:

$$v = \begin{bmatrix} b & N/2-b & N/2 \\ a-b & N/2-(a-b) & N/2 \\ a & N-a & N \end{bmatrix} \quad [a.2.1]$$

con N = número total de personas que contestan el ítem

a = número de personas que contestan correctamente el ítem

b = número de personas que contestan correctamente en el grupo superior

Se observa que con 3 parámetros se construye cualquier hipertabla, pudiendo definirse una base cualquiera con 3 hipertablas linealmente independientes, como pueden ser las siguientes:

$$u1 = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0.5 & 0 & 0.5 \\ 1 & 0 & 1 \end{bmatrix} \quad [a.2.2]$$

$$u2 = \begin{bmatrix} 0 & 0.5 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0 & 1 & 1 \end{bmatrix} \quad [a.2.3]$$

$$u3 = \begin{bmatrix} 0.5 & 0 & 0.5 \\ 0 & 0.5 & 0.5 \\ 0.5 & 0.5 & 1 \end{bmatrix} \quad [a.2.4]$$

Estas tres hipertablas son linealmente independientes y se escogieron arbitrariamente de forma que asemejan lo más posible a una base canónica.

Dada esta base, la hipertabla v se escribe como combinación lineal de u1, u2 y u3 como sigue:

$$v= k\ u\ 1 + i\ u2 + m\ u\ 3 \quad [a.2.5]$$

siendo k, l, m escalares reales sobre los que se estableció el espacio vectorial.

Resolviendo la ecuación [a.2.5] se llega a:

$$k= 2a - 2b \quad [a.2.6]$$

$$l= N - 2b \quad [a.2.7]$$

$$m= 4b - 2^a \quad [a.2.8]$$

Hasta este momento se tiene que el espacio vectoriales de dimensión 3. Si se trabaja ahora con la hipertabla normalizada a 100, como se presenta en el documento, es decir, trasformando todos los valores de v por medio del escalar 100/N, se normaliza la hipertabla denominada Vn, que se expresa:

$$Vn= \begin{bmatrix} B & 50-B & 50 \\ A-B & 50-(A-B) & 50 \\ A & 100-A & 100 \end{bmatrix} \quad [a.2.9]$$

para Vn se tiene la solución de los escalares:

$$k= 2A - 2B \quad [a.2.10]$$

$$l = 100 - 2B \quad [a.2.11]$$

$$m= 4B - 2^a \quad [a.2.12]$$

con lo cual solamente se requieren dos parámetros independientes, en este caso A y B, para los datos normalizados. Con esto se demuestra la independencia propuesta en la parte 4 y puede continuarse con los teoremas planteados.

El espacio formado por las hipertablas normalizadas del tipo Vn en un espacio afín del espacio vectorial que definen las hipertablas del tipo v. Puede trabajarse en el espacio vectorial o en el espacio afín, sin perder generalidad.

Es motivo de otro trabajo presentar las ventajas de disponer de un análisis de reactivos en términos de las hipertablas estocásticas que forman la base, pudiendo ser las tres representadas en [a.2.2], [a.2.3] y [a.2.4] o un conjunto diferente.