RIDU REVISTA DIGITAL DE INVESTIGACIÓN EN DOCENCIA UNIVERSITARIA
ISNN 2223-2516

ARTÍCULO METODOLÓGICO

Relación entre los índices de dificultad y discriminación

Relationship between the Difficulty and Discrimination Indices

Relação entre os índices de dificuldade e discriminação

Luis Leoncio Hurtado Mondoñedo*

https://orcid.org/0000-0001-5373-556X

Centro Preuniversitario, Universidad de Ingeniería y Tecnología, Lima – Perú

Recibido: 15/12/17 **Revisado:** 19/03/18 **Aceptado:** 08/05/18 **Publicado:** 30/06/18

Resumen. Dos de los principales índices usados al hacer el análisis psicométrico de una prueba de rendimiento son el índice de dificultad y el índice de discriminación. Estos índices se convierten en indicadores de la calidad de una prueba en la medida que se encuentren dentro de rangos aceptables. Esto trae dos consecuencias, en primer lugar la determinación de la fórmula para el cálculo de los índices y en segundo lugar la interpretación de los mismos según determinadas normas. El presente trabajo muestra la forma como se determinan y relacionan estos índices, así como la manera en que influye la norma de discriminación en la valoración de la prueba. Se plantean recomendaciones sobre el uso de ambos índices a partir del análisis realizado.

Palabras clave: *índice de dificultad, índice de discriminación, confiabilidad, validez, medición educacional.*

Abstract. Two of the main indices used when doing a psychometric analysis of a performance test are the index of difficulty and the index of discrimination. These indices become indicators of the quality of a test as long as they are within acceptable ranges. This has two consequences, first the determination of the formula for the calculation of the indices, and secondly, the interpretation of them according to certain standards. This work shows the way how these indices are determined and related, as well as the way in which the discrimination rule influences the valuation of the test. Recommendations are also proposed for the use of both indices based on the analysis performed.

Palabras clave: *difficulty index, discrimination index, reliability, validity, educational measurement.*

CITAR COMO: Hurtado Mondoñedo, L. L. (2018). Relación entre los índices de dificultad y discriminación. *Revista Digital de Investigación en Docencia Universitaria, 12*(1),273-300. doi: http://dx.doi.org/10.19083/ridu.12.614

***Correo de correspondencia:** lhurtadom@utec.edu.pe

Resumo. Dois dos principais índices utilizados na realização da análise psicométrica de um teste de desempenho são o índice de dificuldade e o índice de discriminação. Esses índices tornam-se indicadores da qualidade de um teste, desde que estejam dentro dos limites aceitáveis. Isto tem duas consequências, em primeiro lugar a determinação da fórmula para calcular os índices e em segundo a interpretação deles de acordo com certos padrões. O presente trabalho mostra a maneira como esses índices são determinados e relacionados, bem como a maneira pela qual a norma de discriminação influencia a avaliação do teste. São feitas recomendações sobre o uso de ambos os índices com base na análise realizada.

Palavras-chave: índice de dificuldade, índice de discriminação, confiabilidade, validade, medida educacional.

El análisis psicométrico de las preguntas que conforman una prueba de rendimiento incluye el cálculo de índices que permiten caracterizarlas (Mejía, 2005). Dos de los principales índices son el índice de dificultad y el índice de discriminación. El índice de dificultad de una pregunta, como su nombre sugiere, está dado por la expresión numérica de la dificultad que representó para los examinados contestar la pregunta. El índice de discriminación de una pregunta separa, distingue, diferencia, selecciona entre los examinados de mayor y menor rendimiento en la prueba. La calidad de una prueba de rendimiento es explicada, en gran medida, por arrojar índices que se encuentren dentro de rangos aceptables. Frente a esto cabe preguntarse ¿cómo se calculan estos índices? ¿Existe una relación entre los índices de dificultad y discriminación?, ¿qué criterio determina que los índices están dentro de rangos aceptables? El presente trabajo muestra la forma cómo se determinan y relacionan estos índices, así como la manera en que influye la norma de discriminación en la valoración de la calidad de la prueba.

Objetivos

a. Determinar la relación entre los índices de dificultad y de discriminación de las preguntas de una prueba.

b. Mostrar la influencia de la norma de discriminación en la valoración de la calidad de una prueba.

MARCO TEÓRICO

Patrón de respuestas.

En una prueba de rendimiento, los índices de dificultad y discriminación de las preguntas son determinados a partir del patrón de respuestas dadas por los examinados en las preguntas

de dicha prueba. Tomando el marco de referencia de Guttman, las personas son ordenadas de acuerdo con su puntaje total, el cual es un índice de su habilidad, y los ítems son ordenados por su *score* total (Andrich, 2008). Las respuestas de cada persona a cada uno de los ítems son los datos brutos con los que empezamos (Wright & Stone, 1979). Independientemente del número de alternativas que tenga la pregunta, si estas son del tipo dicotómicas, el examinado solo tendrá dos opciones de respuesta: correcta o incorrecta. Por cada examinado registramos sus respuestas a cada una de las preguntas de la prueba. Convenimos en usar el 1 para codificar el acierto y el 0 para la falla. El patrón de respuestas se construye como una tabla donde la primera columna se registra la identificación del examinado y en las siguientes el código de respuesta (1 o 0) a cada una de las preguntas. El patrón de respuestas permite determinar el puntaje del examinado y el número de aciertos en cada pregunta. El puntaje se encuentra contando la cantidad de aciertos que tuvo el examinado en todas las preguntas de la prueba; es decir, contando el número de códigos 1 en la fila de cada examinado. El número de aciertos en cada pregunta se puede encontrar contando el número de examinados que la respondieron correctamente; es decir, contando el número de códigos 1 en la columna de cada pregunta. A partir de los puntajes podemos elaborar un ranking de los examinados, ordenarlos y distinguir entre aquellos de mayor y menor rendimiento. El conteo de aciertos y la distinción de los grupos son necesarios para determinar los índices mencionados. En la tabla 1 se muestra el ranking y el número de aciertos, construido a partir del patrón de respuestas de los examinados de nuestro ejemplo (ver Tabla 1).

Tabla 1
Ejemplo de ranking y el número de aciertos de un grupo de examinados

PATRÓN DE RESPUESTAS Y RANKING DE LAS PUNTUACIONES OM		P1	P2	P3	P4	Puntaje
1	Examinado 6	1	1	1	1	4
2	Examinado 3	1	1	1	0	3
3	Examinado 8	1	1	1	0	3
4	Examinado 2	1	0	1	1	3
5	Examinado 1	1	1	0	0	2
6	Examinado 4	1	0	1	0	2
7	Examinado 5	0	0	1	1	2
8	Examinado 7	1	0	0	0	1
	Número de aciertos	**7**	**4**	**6**	**3**	

Índice de dificultad.

La pregunta P1 tuvo el mayor número de aciertos (7), mientras que P4 tuvo el menor número de aciertos (3). En tanto P1 fue respondida correctamente por más examinados que P4 ella resultó más fácil para el grupo. Diremos entonces que, para este grupo de examinados, P4 es más "difícil" que P1. Para determinar una medida del grado en que la pregunta ha sido difícil para un grupo comparemos el número de examinados que no la respondieron correctamente con respecto al ideal. Llamaremos índice de dificultad ($IDif$) a la comparación del número de fallas $N - C$ de la pregunta con el número total de examinados *N*.

$$IDif = \frac{N - C}{N}$$

Descomponiendo esta fracción tenemos: $IDif = \frac{N}{N} - \frac{C}{N}$

La primera fracción equivale a 1 y la segunda corresponde al llamado índice de facilidad.

$$IDif = 1 - Ifac$$

Esta relación muestra que la dificultad y la facilidad de una pregunta son conceptos excluyentes y complementarios con respecto a la unidad. Fácil y Difícil son adjetivos polares. En un grupo de examinados, una pregunta será más fácil si ella es acertada por un mayor número de sujetos, y por tanto será más difícil cuanto más sujetos la fallen. Si se quiere analizar la dificultad de una pregunta no se deberían contar el número de aciertos, sino el número de fallas. Es frecuente encontrar en la literatura especializada nombrar como "índice de dificultad" o "grado de dificultad" a la relación entre el número de aciertos y el total de examinados (Canales, 2005; García-Cueto, 2005; Gronlund, 1999; Tristán, 2001). De acuerdo con esta definición cuanto mayor sea este índice, mayor será el número de aciertos y por tanto más fácil la pregunta, lo que resulta algo contrario a la dificultad. Desde un punto de vista puramente semántico es más exacto denominar índice de facilidad (García-Cueto, 2005) a la relación entre el número de aciertos y el total de examinados, tal como lo hemos considerado anteriormente. En este trabajo hacemos la distinción señalando que al hablar de índice de dificultad nos referimos a la expresión numérica del grado en el que una pregunta resulta difícil de responder correctamente para el grupo al cual se aplica y viene dada por:

$$IDif = 1 - \frac{C}{N}$$

Donde C representa el número de aciertos en la pregunta y N el número de examinados (ver Tabla 2).

Tabla 2
Índices de dificultad calculados

	P1	P2	P3	P4
Número de aciertos (C)	7	4	6	3
N	8	8	8	8
IDif = 1 - C/N	0.125	0.5	0.25	0.625

El índice de dificultad $(\ IDif\ )$ solo puede tomar valores dentro de un intervalo. Si una pregunta es contestada correctamente por todos los examinados tendremos que el número de aciertos (C) será igual al número de examinados $(\ N\ )$, en este caso $C = N$ y por tanto el índice de dificultad será de $IDif = 1 - \frac{N}{N} = 0$. Por otro lado, si una pregunta no es contestada correctamente por ninguno de los examinados tendremos que el número de aciertos será igual a cero $(\ C = 0\ )$ y por tanto el índice de dificultad será de $IDif = 1 - \frac{0}{N} = 1$

En general, dado que $0 \le C \le N$, entonces $0 \le \frac{C}{N} \le 1$ de donde se desprende que $-1 \le \frac{C}{N} \le 0$ y por tanto $0 \le 1 - \frac{C}{N} \le 1$ De esta forma el índice de dificultad puede tomar valores entre 0 y 1, incluidos estos. Cuanto mayor el índice de dificultad, mayor es la dificultad de la pregunta.

$$0 \le IDif \le 1$$

Índice de discriminación.

Para Bazán (2000) "la discriminación de una pregunta se mide por el grado en que la pregunta ayuda a ampliar las diferencias estimadas entre los que obtuvieron un puntaje total de la prueba relativamente alto de los que obtuvieron un puntaje relativamente bajo" (p.6). Así el índice de discriminación es la expresión numérica de la medida en que una pregunta separa a los examinados de más alto rendimiento de los de más bajo rendimiento. Estos grupos, a los que denominaremos grupo superior (GS) y grupo inferior (GI), son determinados utilizando como punto de corte la mediana de los puntajes, la misma que para nuestro ejemplo resulta

2.5 (ver Tabla 1). El GS estará formado por los examinados con puntajes mayores que 2.5 y el GI por los examinados con puntajes menores que 2.5. En las Tablas 3 y 4 se presentan los puntajes y el patrón de respuestas de los sujetos pertenecientes a cada uno de los grupos.

Tabla 3
Patrón de respuestas del grupo superior

		P1	P2	P3	P4	Puntaje
1	Examinado 6	1	1	1	1	4
2	Examinado 3	1	1	1	0	3
3	Examinado 8	1	1	1	0	3
4	Examinado 2	1	0	1	1	3
	Número de aciertos	4	3	4	2	

Tabla 4
Patrón de respuestas del grupo inferior

		P1	P2	P3	P4	Puntaje
5	Examinado 1	1	1	0	0	2
6	Examinado 4	1	0	1	0	2
7	Examinado 5	0	0	1	1	2
8	Examinado 7	1	0	0	0	1
	Número de aciertos	3	1	2	1	

Dado que no es posible observar directamente el verdadero nivel que tienen los examinados en el tema tratado en la prueba, este debe ser inferido. Aunque los puntajes de una prueba constituyen una medida ordinal, tradicionalmente estos suelen ser el estimador del nivel del dominio de la persona examinada. En ocasiones la razón del número de aciertos y el número de preguntas de la prueba es usado como indicador. De forma similar, una medida del grado de dominio de un grupo, frente a una determinada pregunta, está dada por la razón del número de aciertos al número de examinados que conforman dicho grupo. Cuanto más aciertos en un grupo, más homogéneo será el grupo en el dominio del tema tratado en la pregunta. Cuanto más aciertos en el grupo, mayor será la razón del número de aciertos al número de examinados en el grupo y por tanto más homogéneo el grupo. Una pregunta que pretenda diferenciar a los examinados de mayor rendimiento de los de menor rendimiento tendría que

comparar la razón del número de aciertos al número de examinados en cada grupo. Pero esta comparación debe ser por exceso, cuanto mayor sea la diferencia entre la razón del número de aciertos al número de examinados del GS y GI, mayor será la medida de la discriminación. En la teoría tradicional del test, la alta discriminación es interpretada como una característica deseable de un ítem y un indicador clave de la calidad del ítem (Masters, 1988).

Consideremos un grupo de N examinados, el GS y el GI establecidos a partir de la mediana tendrían $N/2$ examinados cada uno. Si representemos con C_S el número de aciertos del GS y con C_i el número de aciertos del GI, la medida en que se discrimina el GS del GI está dada por la diferencia $C_S\ /\ (N/2) - C_i\ /\ (N/2)$. Llamamos a esta diferencia índice de discriminación, lo representaremos por $IDisc$ y se calcula a partir de:

$$IDisc = \frac{C_S - C_i}{N\ /\ 2}$$

Tabla 5
Índices de discriminación calculados

	P1	P2	P3	P4
Número de aciertos *GS* (C_S)	4	3	4	2
Número de aciertos *GI* (C_i)	3	1	2	1
N	8	8	8	8
***IDisc* = (C_s -C_i)/(N/2)**	0.25	0.5	0.5	0.25

El índice de discriminación $(IDisc)$ solo puede tomar valores dentro de un intervalo. Si una pregunta es contestada correctamente por todos los examinados del GS y ninguno del GI tendremos $C_s = N/2$ y $C_i = 0$, el índice de discriminación será $IDisc = \frac{(0\ \ N/2 - 0)}{(N/2)} = 1.$

Por otro lado, si una pregunta no es contestada correctamente por ninguno de los examinados del GS pero contestada correctamente por todos los examinados del GI tendremos $C_s = 0$ y $C_i = N/2$ que nos daría un índice $IDisc = (0 - N/2\ ) / (N/2) = -1.$ Un caso intermedio se presentaría cuando la pregunta no es contestada por ninguno de los examinados de ambos grupos $(C_s = 0$ y $C_i = 0)$. lo cual arroja un índice $IDisc = (0 - 0\ ) / (N/2) = 0.$ De esta forma el índice de discriminación puede tomar valores entre -1 y 1, incluidos estos.

$$-1 \leq IDisc \leq 1$$

Podemos decir que una pregunta con $IDisc = -1$ discrimina totalmente, mientras que en el otro extremo, una pregunta con $IDisc = 1$ discrimina erróneamente. Si lo que se busca es ver la medida en que se diferencian los examinados en cuanto a un mayor dominio del GS con respecto al GI, un $IDisc$ negativo indicaría un mayor dominio del GI que el GS, situación que no es admisible, ya que contradice el sentido del índice.

La forma como se ha calculado y descrito el índice de discriminación está basada en grupos extremos. Esta es una forma más sencilla de determinar la discriminación que otros índices de discriminación. Una forma frecuente de calcular este índice es por medio de la correlación biserial pregunta-prueba. Una pregunta discrimina en forma adecuada en una prueba si ella sirve para diferenciar, distinguir, distanciar entre los sujetos con puntajes mayores y los sujetos con puntajes menores. La consecuencia inmediata es que cuando la pregunta discrimina adecuadamente, existirá una correlación positiva entre las puntuaciones de la pregunta y de la prueba (García-Cueto, 2005). De esta forma, el índice de discriminación es un índice estadístico que describe el grado en que una pregunta es coherente con las demás preguntas que buscan discriminar entre las personas (Andrich, 2008). El tipo de correlación que se utilice dependerá de las características de medida que tengan las preguntas y la prueba, como por ejemplo si ambas variables son dicotómicas, dicotomizadas, continuas o una combinación de ellas. En general, cuanto mayor sea esta correlación, mayor es la discriminación, aunque hay algunas excepciones en las que no se espera una discriminación alta. Para Garret (1966) la conveniencia de otros métodos para calcular el índice de discriminación "... se juzga según el grado hasta el cual son capaces de dar resultados que se aproximen a los obtenidos mediante la correlación biserial" (p. 403).

El problema de la dificultad óptima

Las preguntas demasiado fáciles o aquellas que resulten demasiado difíciles darían lugar a distribuciones asimétricas con respecto al porcentaje de aciertos y fallas en ellas. Si la pregunta es fácil presenta una fuerte asimetría negativa y la asimetría es positiva cuando la pregunta es más difícil, siendo totalmente simétrica si su índice de dificultad es de 0.5 (García-Cueto, 2005). En un estudio experimental de la distribución de los puntajes de una prueba con los valores de dificultad de los ítems, Ebel (1977) señala: "... los resultados de esta investigación ratifican la recomendación de preferir preguntas de dificultad intermedia al construir pruebas de rendimiento" (p.491). En la misma línea, García-Cueto (2005) sostiene que "... en general se conseguirán los mejores resultados en las evaluaciones cuando la mayoría de los ítems sea de dificultad media" (p.60). Una posición contraria es la formulada por Tristán (2001): "... es conveniente disponer de reactivos en toda la gama de dificultades y no solamente reactivos centrados al 50% de dificultad con objeto de poder medir el dominio de cada persona con

precisión" (p.7). Tristán toma como ejemplo la temperatura corporal medida en un termómetro para agua, debidamente graduado en un rango de 0 a 100, donde la temperatura óptima no se encuentra a los 50°C. De igual manera "... el objetivo de incluir reactivos con diferentes niveles de dificultades es disponer de una escala bien graduada"[1] de ahí la importancia de considerar en la prueba preguntas con toda la gama de dificultades. Las evaluaciones referidas a normas -como las pruebas de admisión- permiten clasificar a los examinados, ordenándolos en una escala común y diferenciando entre grupos de alto y bajo rendimiento. En las evaluaciones de aula -por lo general- el docente no busca que su prueba de rendimiento se convierta en una escala bien graduada, lo más probable es que su rango cambie al aplicarla en distintos grupos. El bajo número de examinados a los que usualmente se aplican las pruebas suele ser la mayor limitación para obtener escalas precisas. De igual manera, el limitado número de examinados y de preguntas en las pruebas de aula, provoca un mayor error en el cálculo de la confiabilidad. Por otro lado, no debe ser una condición que la distribución de los puntajes sea simétrica. Los rendimientos de los alumnos en la prueba no deben seguir necesariamente una distribución normal. Esto es más probable cuando el número de examinados es grande y están presentes una serie de rasgos diferenciales. Delgado (2004) refiere la coincidencia de Bloom, Hastings y Madaus con De Landsheere "... en que la curva normal es la distribución más apropiada para la actividad casual, mientras que la educación tiene un propósito intencional" (p.165). En este trabajo buscaremos recoger las dos posiciones anteriores. Por un lado, determinar un intervalo para el índice de dificultad en un entorno cercano a la dificultad media y, por otro lado, que en dicho intervalo ellos se encuentren distribuidos como en una escala bien graduada. Gráficamente, ambos enfoques se pueden representar en lo que proponemos como región óptima.

MÉTODO

Empezaremos analizando el caso de los GS y GI determinados a partir de la mediana, con esto se podrá definir una región de valores admisibles para los índices. Cada uno de los índices, por separado, solo puede tomar valores dentro de un intervalo de valores, pero también analizados en conjunto como un par ordenado, debe ubicarse dentro de una región en un plano bidimensional. Definir matemáticamente esta región y con ello determinar la relación entre los $IDif$ e $IDisc$ será nuestra primera tarea. Una vez definida buscaremos aquella parte de esta región donde es deseable que se encuentren los índices, dada una norma de discriminación. A continuación buscaremos optimizar esta región de forma tal que ella contenga el mayor número de pares ordenados de la forma $(IDif, IDisc)$ asociados con las preguntas de la prueba de rendimiento. Finalmente haremos una generalización de estas regiones para grupos de n sujetos donde $n < N$.

1 Tristán (1995, 2001, 2006) desarrolla estas ideas al presentar las bases del modelo Kalt.

Dificultad y discriminación

Por lo general se menciona la existencia de una relación entre el $IDif$ y el $IDisc$ aunque no se indica nada más de lo que se puede desprender en los valores críticos. Una pregunta con dificultad 0 e 1, tiene discriminación 0; y una pregunta con dificultad de 0.5 tiene discriminación 1. Buscaremos profundizar en esta relación. Tomando los resultados presentados en las tablas anteriores podemos resumir los índices de las cuatro preguntas de nuestro ejemplo en la siguiente tabla:

Tabla 6
Índices de dificultad y discriminación calculados

	P1	P2	P3	P4
IDif	0.125	0.500	0.250	0.625
IDisc	0.250	0.500	0.500	0.250

Estos índices, obtenidos posaplicación de la prueba, permiten describir las preguntas y analizar los resultados obtenidos. Este análisis incluye la caracterización de acuerdo al grado de dificultad de la pregunta, la caracterización de acuerdo al grado en que ella discrimina y la determinación de la calidad de la prueba a partir de la ubicación de estos índices en un intervalo de valores deseables. Existe una relación entre el $IDif$ y el $IDisc$. El conjunto de valores posibles que toma uno está relacionado con el valor que toma el otro. Una comparación de los índices mostrados en la tabla 5 con una determinada norma de discriminación o dificultad nos haría aceptar alguna de las preguntas y revisar (o desechar) otras. En tanto que uno de los objetivos de una prueba de rendimiento referida a normas es la de ordenar a los examinados en función del dominio del tema que mida la prueba, el poder de discriminación de una pregunta es de la mayor importancia. Una buena pregunta tiene que ser acertada por una proporción mayor de los sujetos con mayor puntuación en la prueba que aquellos con puntuaciones más bajas (García-Cueto, 2005).

Región de valores admisibles

El presente trabajo se ha desarrollado tomando las bases del modelo Kalt[2], sin embargo hemos basado nuestro análisis en el número de aciertos y no en porcentajes como se presenta en Kalt. Esto nos permitirá, por un lado, simular los casos extremos e intermedios del patrón de respuestas de los examinados con una sencilla presentación y, por otro lado, formular

2 El modelo Kalt es usado para el análisis de preguntas por computadoras por el IEIA de México.

matemáticamente las que llamaremos región de valores admisibles, región normada y región óptima. Empezaremos tomando el caso de los grupos GS y GI definidos a partir de la mediana.

Tabla 7
Índices de dificultad y discriminación calculados

	CASO A	CASO B	CASO C	CASO D
C_s	N/2	N/2	0	0
C_i	N/2	0	N/2	0
$C = C_s + C_i$	N	N/2	N/2	0
$C_s - C_i$	0	N/2	-N/2	0
N	N	N	N	N
$IDif$	0	0.5	0.5	1
$IDisc$	0	1	-1	0

El GS estará formado por los sujetos con puntajes mayores a la mediana y el GI por los sujetos con puntajes menores a la mediana. Al no haber grupo intermedio, los sujetos del GS y GI completan el total de examinados, razón por la cual el número total de aciertos en cada pregunta será igual a la suma de los aciertos en el GS y GI para dicha pregunta, es decir $C=C_s + C_i$.

De acuerdo a esto podemos distinguir cuatro casos extremos:

Caso A: Aciertan todos los examinados: $C_s = N/2,\ C_i = N/2$ y $C = N$.

Caso B: Aciertan solo los sujetos del GS: $C_s = N/2,\ C_i = 0$ y $C = N/2$.

Caso C: Aciertan solo los sujetos del GI: $C_s = 0,\ C_i = N/2$ y $C = N/2$.

Caso D: Fallan todos los examinados: $C_s = 0,\ C_i = 0$ y $C = 0$.

En la tabla 8 se muestran los $IDif$ e $IDisc$ para los casos señalados anteriormente.

Con el fin de explicar más fácilmente este punto supondremos el caso de 80 examinados ($N=80$). Los GS y GI tendrían 40 examinados cada uno y por tanto el número de aciertos C_s o C_i no podrían ser mayores que 40. El supuesto teórico es que en el GS se encuentran los examinados de más alto rendimiento y en el GI los examinados de más bajo rendimiento. El comportamiento ideal en una pregunta que discrimine a estos grupos es que la acierten todos los del GS y la fallen todos los del GI. El comportamiento erróneo se daría si la fallan todos los del GS y la aciertan todos los del GI. Entre estos dos comportamientos - ideal y erróneo - se pueden encontrar otros fijando el comportamiento de uno de los grupos y haciendo variable el comportamiento del otro.

Tabla 8

Casos AB. Comportamiento ideal del GS ($C_s = 40$) y comportamiento variable del GI.

	CASO AB1	CASO AB2	CASO AB3	CASO AB4	CASO AB5
C_s	40	40	40	40	40
C_i	0	10	20	30	40
$C = C_s + C_i$	40	50	60	70	80
$C_s - C_i$	40	30	20	10	0
N	80	80	80	80	80
$IDif = 1 - C/N$	0.5	0.375	0.25	0.125	0
$IDisc = (Cs - Ci)/(N/2)$	1	0.75	0.5	0.25	0
Intervalos		$0. \leq IDif \leq 0.5$		$0 \leq IDisc \leq 1$	

Tabla 9

Casos BD. Comportamiento ideal del GI ($C_i = 0$) y comportamiento variable del GS.

	CASO BD1	CASO BD2	CASO BD3	CASO BD4	CASO BD5
C_s	0	10	20	30	40
C_i	0	0	0	0	0
$C = C_s + C_i$	0	10	20	30	40
$C_s - C_i$	0	10	20	30	40
N	80	80	80	80	80
$IDif = 1 - C/N$	1	0.875	0.75	0.625	0.5
$IDisc = (Cs - Ci)/(N/2)$	0	0.25	0.5	0.75	1
Intervalos		$0.5 \leq IDif \leq 1$		$0 \leq IDisc \leq 1$	

Tabla 10

Casos DC. Comportamiento erróneo del GS ($C_s = 0$) y comportamiento variable del GI

	CASO DC1	CASO DC2	CASO DC3	CASO DC4	CASO DC5
C_s	0	0	0	0	0
C_i	0	10	20	30	40
$C = C_s + C_i$	0	10	20	30	40
$C_s - C_i$	0	-10	-20	-30	-40
N	80	80	80	80	80
$IDif = 1 - C/N$	1	0.875	0.75	0.625	0.5
$IDisc = (Cs - Ci)/(N/2)$	0	-0.25	-0.5	-0.75	-1
Intervalos		$0.5 \leq IDif \leq 1$		$-1 \leq IDisc \leq 1$	

Tabla 11
Casos CA.Comportamiento erróneo del GI ($C_i = 40$) y comportamiento variable del GS.

	CASO CA1	CASO CA2	CASO CA3	CASO CA4	CASO CA5
C_s	0	10	20	30	40
C_i	40	40	40	40	40
$C = C_s + C_i$	40	50	60	70	80
$C_s - C_i$	-40	-30	-20	-10	0
N	80	80	80	80	80
$IDif = 1 - C/N$	0.5	0.375	0.25	0.125	0
$IDisc = (Cs - Ci)/(N/2)$	-1	-0.75	-0.5	-0.25	0
Intervalos		$0 \leq IDif \leq 0.5$		$-1 \leq IDisc \leq 0$	

Los casos anteriormente descritos nos permiten tener un conjunto de posibilidades para el *IDif* y el *IDisc*. La tabla 12 muestra los *IDif* y *IDisc* ordenados en forma decreciente a la dificultad.

Tabla 12
*Casos e **IDif** y **IDisc** ordenados por dificultad*

CASO	*IDif*	*IDisc*
BD1	1.000	0.000
DC1	1.000	0.000
BD2	0.875	0.250
DC2	0.875	-0.250
BD3	0.750	0.500
DC3	0.750	-0.500
BD4	0.625	0.750
DC4	0.625	-0.750
AB1	0.500	1.000
BD5	0.500	1.000
CA1	0.500	-1.000
DC5	0.500	-1.000
AB2	0.375	0.750
CA2	0.375	-0.750
AB3	0.250	0.500
CA3	0.250	-0.500
AB4	0.125	0.250
CA4	0.125	-0.250
AB5	0.000	0.000
CA5	0.000	0.000

Formaremos pares ordenados de la forma $(IDif,IDisc)$ para cada uno de los casos considerados y los llevaremos a un plano bidimensional $(IDif_{VS}\ IDisc)$. En el eje horizontal se han considerado los valores del $IDif$ y en el eje vertical los valores del $IDisc$. La figura 1 muestra el conjunto de puntos de la forma $(IDif,IDisc)$ para cada uno de los casos señalados en la tabla 12. La unión de los puntos marcaría el contorno de un rombo donde se ubicarían las posibilidades extremas de los puntos cuyas coordenadas están dadas por los pares ordenados de la forma $(IDif, IDisc)$. Así, por ejemplo, una posibilidad extrema ocurriría en una pregunta donde aciertan los 40 examinados del GS y solo 6 del GI por lo que el punto $(0.575,0.850)$ asociado con esta pregunta formaría parte del contorno del rombo. En su interior se ubicarían las posibilidades no extremas. Por ejemplo, si aciertan 37 examinados del GS y 9

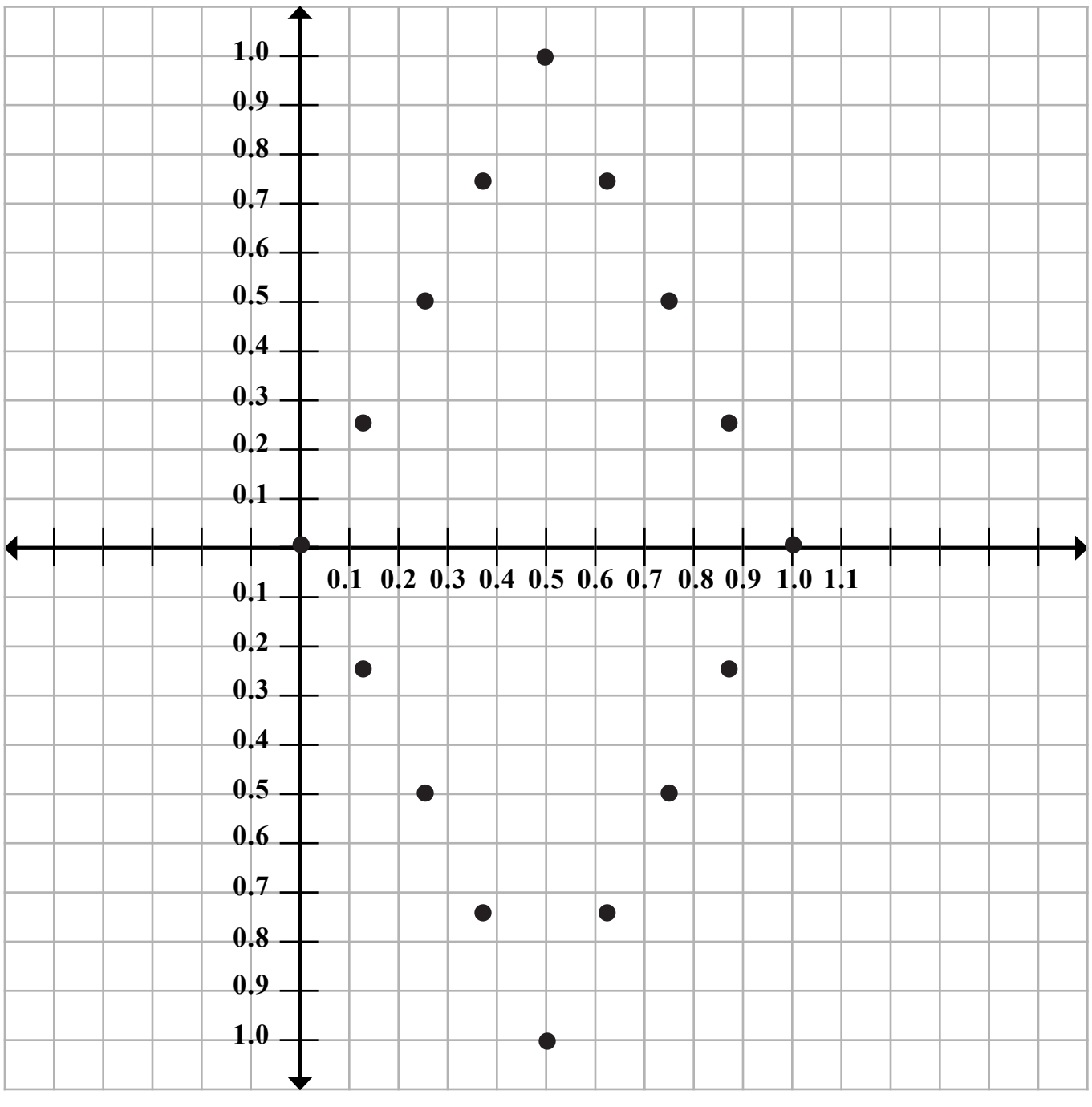

Figura 1. Distribución de puntos $(IDif,IDisc)$ para los casos de la tabla 12

del GI, tendríamos el punto (0.575,0.700) que se encuentra por debajo del punto (0.575,0.850) perteneciente al contorno del rombo. De esta forma, tanto el perímetro como el interior del rombo contendrían el conjunto de todas las posibilidades de pares ordenados de la forma (*IDif, IDisc)* por cada una de las preguntas de la prueba de rendimiento.

Si nombramos los vértices del rombo como P(0.0,0.0), Q(0.5,1.0), R(1.0,0.0) y S(0.5,-1.0) bajo la forma como han sido definidos los índices de dificultad y discriminación en este trabajo, el par ordenado con primera componente *IDif* y segunda componente *IDisc* debe pertenecer a la región limitada por el rombo PQRS tal como se muestra en la figura 2. Dado que no es admisible un *IDisc* negativo, del rombo de la región anterior debería considerarse solo los casos con *IDisc* no negativos, es decir solo los casos AB y BD.

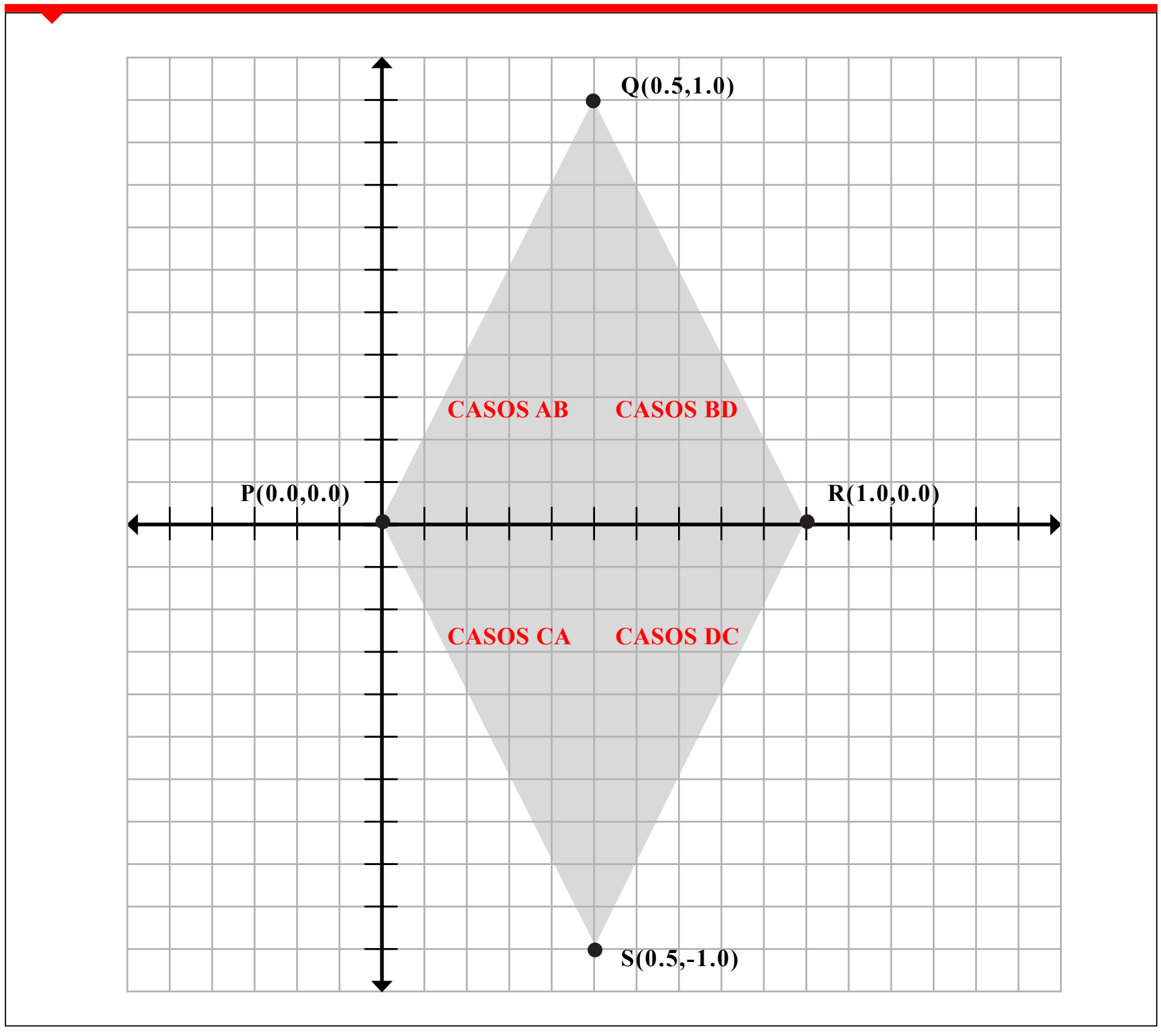

Figura 2. Región formada por el conjunto de puntos (*IDif,IDisc)*

Esto nos haría redefinir nuestra región como la limitada por el triángulo PQR del primer cuadrante donde se encontrarían los valores admisibles del $IDisc$ tal como se presenta en la figura 3. Decimos entonces que por cada una de las preguntas de una prueba de rendimiento los puntos de la forma ($IDif,IDisc$) asociados a ellas deberán pertenecer a la región triangular limitada por los puntos $(0.0,0.0)$, $(0.5,1.0)$ y $(1.0,0.0)$ la misma que denominamos región de valores admisibles.

Formulación matemática de la región de valores admisibles

La región triangular PQR se convertirá en nuestra región de interés y ella puede ser definida matemáticamente por medio de desigualdades. Cuando una desigualdad contiene solo dos variables, el conjunto solución está representado gráficamente por medio espacio del plano cartesiano. El medio espacio de una desigualdad del tipo $y \leq ax+b$ está compuesto por la línea recta correspondiente y todos los puntos situados debajo de ella. Si la inecuación es del tipo $y \geq ax+b$ el medio espacio incluye la recta y todos los puntos situados por encima de ella. Tomaremos al $IDif$ como la variable x y al $IDisc$ como la variable y y buscaremos primero las ecuaciones de las rectas que contienen a los segmentos PQ y QR que son dos de los límites de nuestra región triangular. Usaremos la forma punto-pendiente.

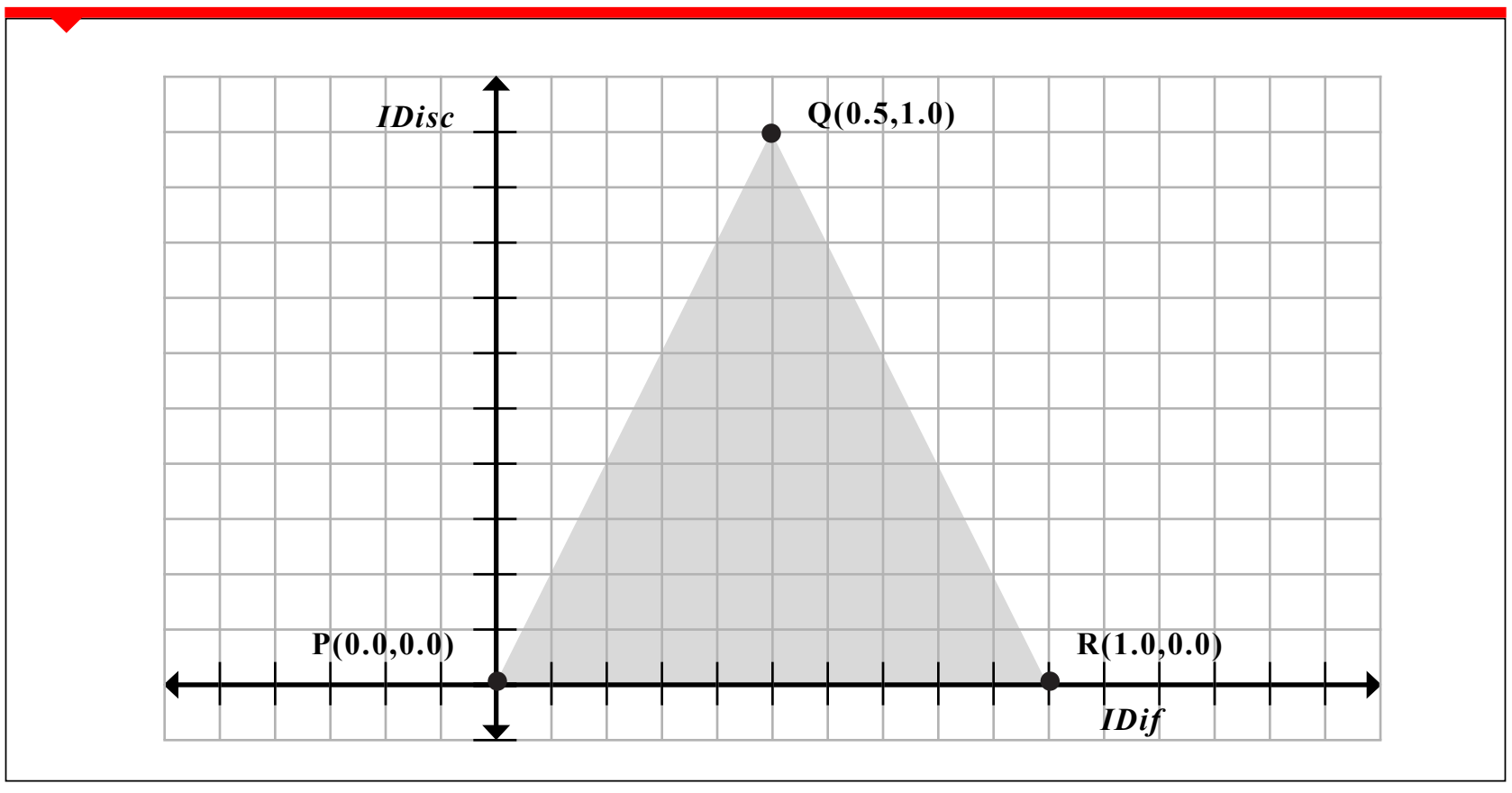

Figura 3. Región de valores admisibles

El segmento PQ está limitado por los puntos $P(0,0)$ y $Q(0.5,1)$ por lo cual su pendiente está dada por $\frac{1-0}{0.5-0} = 2$. Tomando como punto de paso $P(0,0)$ y la pendiente $m=2$,

planteamos la ecuación de la recta que contiene a PQ:

$$L_{PQ}: y - 0 = 2(x - 0)$$
$$y = 2x$$

El segmento QR está limitado por los puntos $Q(0.5,1)$ y $R(1,0)$ por lo cual su pendiente está dada por $\frac{0-1}{1-0.5} = -2$. Tomando como punto de paso al punto $R(1,0)$y la pendiente $m=-2$,

planteamos la ecuación de la recta que contiene a QR:

$$L_{PQ}: y - 0 = -2(x - 1)$$
$$y = -2x + 2$$

Las ecuaciones halladas corresponden a dos de las rectas que limitan la región triangular PQR. La tercera es la que contiene al segmento PR, es decir el eje horizontal, el cual tiene por ecuación $y = 0$ ($L_{PQ}: y = 0$) A partir de estas ecuaciones podemos definir la región triangular PQR como la intersección de tres medio-espacios:

Medio-espacio 1:

Compuesto por los puntos de la recta L_{PQ} y todos los situados debajo de ella: $y \leq 2x$

Medio-espacio 2:

Compuesto por los puntos de la recta L_{QR} y todos los situados debajo de ella: $y \leq -2x + 2$

Medio-espacio 3:

Compuesto por los puntos de la recta L_{PR} y todos los situados encima de ella: $y \geq 0$

Esto es,

$$\text{Región } PQR \begin{cases} y \leq 2x \\ y \leq 2 - 2x \\ y \geq 0 \end{cases}$$

También podemos redefinir la región triangular como la unión de dos triángulos rectángulos. Si llamamos T al punto medio del segmento PR, tenemos que la región PQR es la unión de las regiones triangulares PTQ y RTQ. Estas dos últimas quedan definidas restringiendo convenientemente el intervalo de la variable x. Tenemos:

Región PTQ: $y \geq 0;\ y \leq 2x;\ 0 \leq x \leq 0.5$

Región RTQ: $y \leq 0;\ y \leq 2 - 2x;\ 0.5 \leq x \leq 1$

La región PQR tiene tres puntos extremos. El punto P con coordenadas $x=0$, $y=0$ el punto Q con

coordenadas $x = 0.5,\ y = 1$; y el punto R con coordenadas $x=1,\ y=0$. O en forma equivalente:

$$\text{Región } PQR = \text{Región } PTQ \ \smile\ \text{Región } RTQ$$

$$\text{Región } PQR \begin{cases} 0 \le y \le 2x & si,\ 0\ x\ 0.5 \\ 0 \le y \le 2 - 2x & si,\ 0.5\ x\ 1 \end{cases}$$

Dado que x e y corresponden al $IDif$ y al $IDisc$ respectivamente, la región PQR corresponde al conjunto de todas las posibilidades de pares ordenados de la forma $(IDif,\ IDisc)$ para una pregunta dada. Llamaremos a esta zona como REGIÓN DE VALORES ADMISIBLES (RVA) de los índices de todas las preguntas de la prueba de rendimiento que se analiza.

$$(IDif,\ IDisc)\ \varepsilon\ RVA \leftrightarrow \begin{cases} 0 \le IDisc \le\ 2IDif & si, 0 \le\ IDif \le 0.5 \\ 0 \le IDisc \le\ 2-2IDif & si, 0.5 \le IDif \le 1 \end{cases}$$

A partir de la RVA podemos formular matemáticamente la relación entre el $IDif$ y el $IDisc$ cuando los grupos han sido determinados a partir de la mediana.

$$IDisc = \begin{cases} 0 & si,\ IDif\ =\ 0 \\ 1 & si,\ IDif\ =\ 0.5 \\ 0 & si,\ IDif\ =\ 1 \end{cases}$$

$$0 \le IDisc \le \begin{cases} 2\ IDif & si,\ 0\ \le\ I\ Dif\ \le 0.5 \\ 2 - 2\ IDif & si,\ 0.5\ \le\ I\ Dif\ \le 1 \end{cases}$$

Esto significa que el intervalo de valores que puede tomar el $IDisc$ está determinado por el valor que toma el $IDif$. Así, por ejemplo, si tenemos una pregunta con $IDif\ =\ 0.35$, su índice de discriminación debe estar entre 0 y 2 x 0.35; es decir: $0 \le IDisc \le 0.70$; si para otra pregunta, el índice de dificultad es de 0.75 su índice de discriminación debe estar entre 0 y $2 - 2$ x 0.75; es decir: $0 \le IDisc \le 0.50$

Región normada

La región triangular PQR descrita en el punto anterior nos muestra la RVA; es decir la región del plano $IDisc$ vs $IDif$ donde se van a distribuir los índices de las preguntas. Esta región es la región teórica, pero no necesariamente la deseable. Podemos establecer un valor para el $IDisc$ por encima del cual se deberían encontrar los índices de discriminación de las preguntas. Llamaremos a este *valor norma de discriminación*, y lo representaremos por $IDisc_{norma}\ =\ k$, donde k es una constante positiva menor que 1. Por otro lado, se sabe que cuanto más cerca de 0.5 están los índices de dificultad de las preguntas, mayor confiabilidad para la prueba y mejor distribución de los puntajes. Deberá existir un intervalo de valores para el índice de dificultad que posibilite esta mayor confiabilidad sin sacrificar la discriminación entre

los grupos; es decir, deberá existir un intervalo cerrado donde se encuentren los índices de dificultad deseables. Si definimos el extremo inferior del intervalo con el valor D , y dada la simetría de la región de valores admisibles, el valor superior del intervalo resultará igual $1-D$. Llamaremos a D *valor norma de dificultad* y lo representaremos por $IDif_{norma} = D$.

A partir de la RVA trazaremos una recta horizontal correspondiente al valor norma de discriminación. Todas las preguntas cuyos puntos pertenezcan a la RVA y se encuentren en o por encima de la recta horizontal discriminarán de acuerdo a la norma. En la figura 4 esto queda representado por la recta horizontal $IDisc = k$. Al mismo tiempo se ha definido el intervalo cerrado $[D, 1-D]$ donde se encuentran los valores normados para el índice de dificultad.

Todas las preguntas cuyos puntos pertenezcan a la RVA y se encuentren entre las rectas verticales trazadas por los extremos del intervalo tendrán una dificultad deseable. De esta forma las preguntas cuyos puntos con coordenadas $(IDif, IDisc)$ se encuentre en la RVA, entre las rectas verticales $IDif = D$ y $IDif = 1-D$ y en o por encima de la recta horizontal $IDisc = k$. pertenecerán a la región normada.

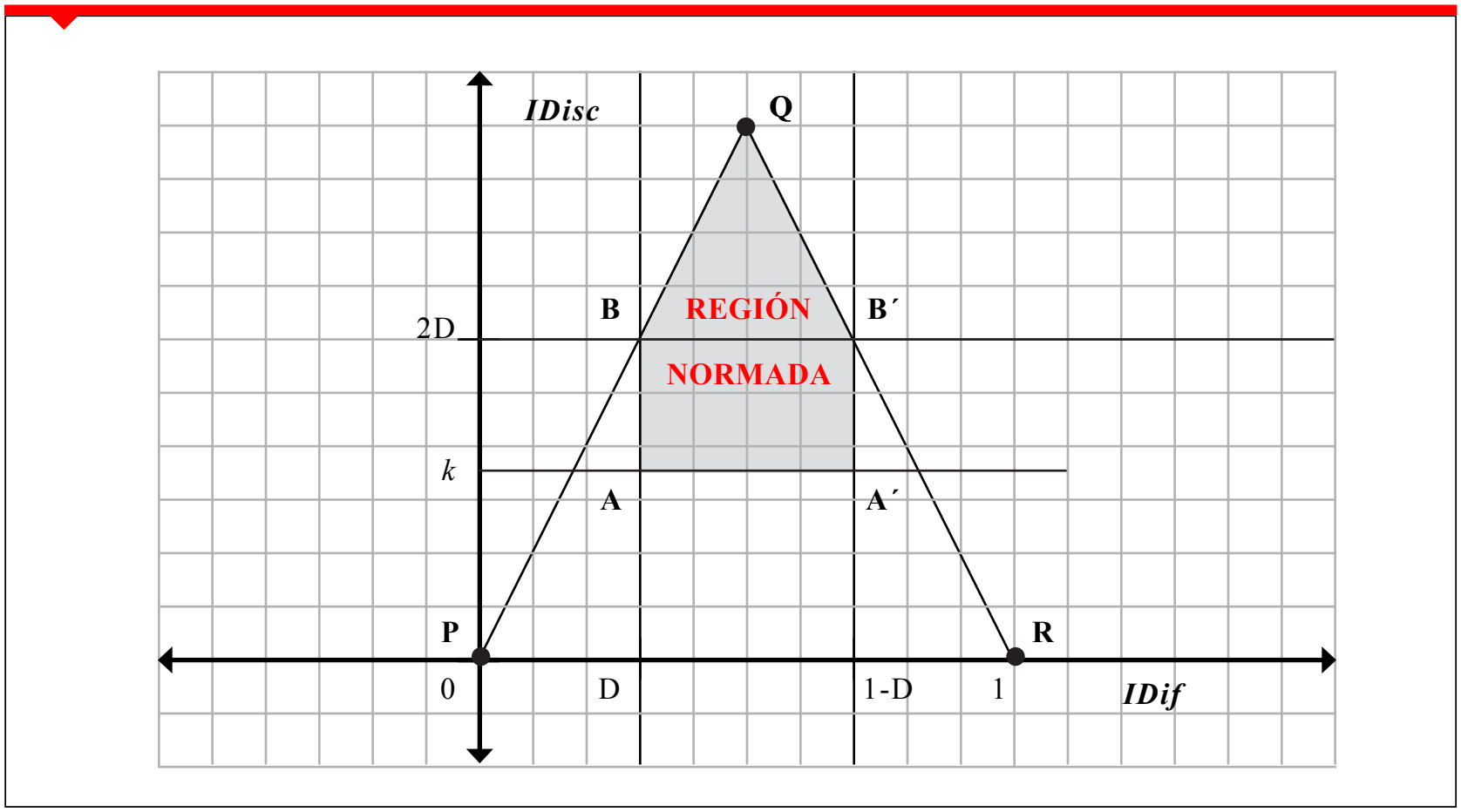

Figura 4. Región normada para el caso: GS = GI = N/2

La región normada corresponde al interior y contorno del polígono $ABQB'A'$. Las coordenadas de los vértices del polígono están dadas por $A(D,k)$; $B(D,2D)$; $Q(0.5,1.0)$; y $B'(1-D,2D)$ y $A'(1-D,k)$. Los puntos B y B' se determinan reemplazando $x=D$ y $x=1-D$ en las ecuaciones de las rectas L_{PQ} y L_{QR} respectivamente.

Región óptima

Cuantos más puntos estén dentro de la región normada, mejor será el comportamiento de las preguntas y por tanto mejor construida está la prueba de rendimiento. Estos puntos deben distribuirse cubriendo el intervalo de dificultad deseable y por encima del valor norma de discriminación cubriendo una parte del área de la región poligonal. Un mayor número de puntos con esta dispersión, cumpliendo ambas normas –dificultad y discriminación–, generarían una región óptima determinada por la mayor área posible que puede tomar el polígono $ABQB'A'$. Nos encontramos frente a un problema de optimización, cuya función objetivo es la función área del polígono $ABQB'A'$.

Con la ayuda de la figura 4 podemos modelar la función área del polígono.

Área del polígono $ABQB\acute{A}'$ = Área del rectángulo $ABQB'A'$ + Área del triángulo BQB'

Área del polígono $ABQB\acute{A}' = (1-2D)(2D-k) + \frac{1}{2}(1-2D)(1-2D)$

Área del polígono $ABQB'A' = -2D^2 + 2kD - k + 0.5$

Al analizar una prueba comparamos los índices de las preguntas con algún criterio clasificatorio. Si aceptamos una determinada norma de discriminación, el valor de k es constante, por tanto el área del polígono $ABQB'A'$ queda expresada como una función del valor norma de dificultad D. Representaremos con S el área del polígono $ABQB\acute{A}'$, con lo cual tenemos:

$$S(D) = -2D^2 + 2kD - k + 0.5$$

Para maximizar esta función debemos primero encontrar su valor crítico, es decir aquel valor de D que maximiza la función $S(D)$. Para ello buscamos la primera derivada de la función y la igualamos a cero.

$$S(D) = -4D + 2k$$
$$-4D + 2k = 0$$
$$D = \frac{k}{2}$$

La segunda derivada de la función $S(D)$ está dada por $S''(D) = -4$. Por lo que para $D = \frac{k}{2}$ tenemos que $S'' = \left(\frac{k}{2}\right) < 0$, esto muestra que la función $S(D)$ alcanza su máximo valor cuando $D = \frac{k}{2}$. Dicho de otra forma, para un valor norma de dificultad $D = \frac{k}{2}$ el área de la región deseable es lo mayor posible. Con $D = \frac{k}{2}$ podemos encontrar las coordenadas de los vértices del polígono, así tenemos $A\left(\frac{k}{2}, k\right)$; $B\left(\frac{k}{2}, k\right)$; $Q\,(0.5, 1.0)$; $B'\left(1-\frac{k}{2}, k\right)$ y $A'\left(1-\frac{k}{2}, k\right)$.

Nótese que tanto los puntos A y B como los puntos A' y B' tienen las mismas coordenadas para el valor $D = \frac{k}{2}$ que hace óptima la región poligonal. En otras palabras, el óptimo ocurre cuando la poligonal se convierte en la región triangular BQB'.

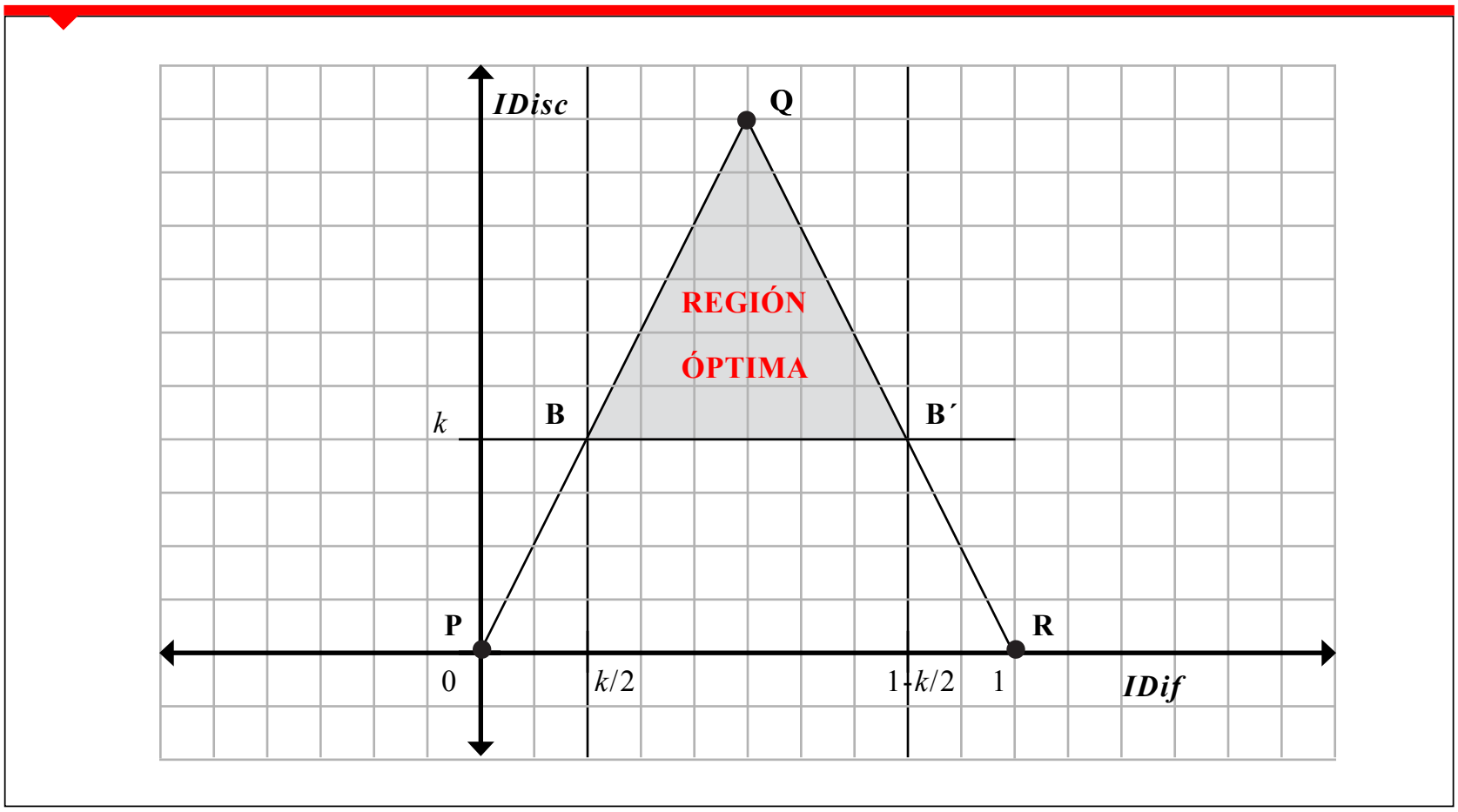

Figura 5. Región óptima para el caso: GS = GI = N/2 y norma de discriminación k

Esto significa que una prueba de rendimiento cuyas preguntas discriminen igual o por encima de un valor norma de discriminación k, los índices de dificultad deben pertenecer al intervalo $\left(\frac{k}{2}, 1-\frac{k}{2}\right)$ para que el mayor número de preguntas cuyas coordenadas $(IDif, IDisc)$. se encuentren dentro de la región óptima. Como consecuencia, una prueba de rendimiento será de mayor calidad en la medida que contenga mayor número de preguntas en la región óptima. No debe sorprender el hecho de que la región óptima haya resultado un triángulo. Al buscar un intervalo para el índice de dificultad en un entorno cercano al valor $IDif = 0.5$, debíamos cubrir la región con puntos distribuidos horizontalmente y cada vez más cercanos al 0.5. Al mismo tiempo, al buscar una discriminación por encima de un valor $IDisc = k$ debíamos cubrir la región con puntos distribuidos verticalmente tendiendo a 1. La distribución más

plausible para este doble comportamiento es la triangular. En la región óptima podemos encontrar puntos que se distribuyen en forma horizontal tendiendo a 0.5 y verticalmente por encima de k tendiendo a 1.

Índices para grupos de "n" sujetos (n<N/2)

Para calcular el $IDif$ se contó el número de fallas del total de examinados, lo cual es independiente de la forma en que se hayan separado los grupos superior e inferior. Por el contrario, el cálculo del $IDisc$ depende de la forma como ellos son separados. El establecimiento del punto de corte de los grupos no es uniforme entre los evaluadores. En la primera parte de este trabajo se ha considerado la mediana de los puntajes como punto de corte de los GS y GI. En ocasiones, cuando se tienen grandes grupos de examinados, se utiliza los cuartiles o los deciles. Se ha demostrado que el poder discriminativo de una pregunta se determina de manera más exacta si los grupos se basan en el 27% superior e inferior, en lugar de cualquier otro porcentaje de la distribución (Garret, 1966). Aunque esto es lo óptimo, Ebel (1977) señala que "... no son en realidad mucho mejores que los del grupo del 25 o 33%" (p. 476). Una de las razones que se dan para no escoger la mediana es el hecho de separar mejor los grupos y que ellos no se vean afectados por los puntajes de los examinados de rendimiento medio. Considerando que el GS y GI están formados por n sujetos cada uno, el índice de discriminación se calcularía según

$$IDisc = \frac{C_s - C_i}{n}$$

Dado que cada uno de los grupos tendrá n sujetos, la diferencia de los aciertos será dividida por n y no por la mitad del total de examinados $\frac{N}{2}$ como lo hicimos para el corte a partir de la mediana. El cálculo del $IDif$ no se ve afectado por esta separación, ya que su cálculo depende del número total de aciertos y no de la suma de los aciertos de cada grupo. Es importante hacer esta distinción ya que lo contrario, es decir usar para el cálculo de $IDif$ la relación $1-\frac{C_s + C_i}{N}$, nos daría un intervalo para el índice de dificultad $\left[1-\frac{2n}{N}, 1\right]$ distinto al teórico de $[0,1]$. A partir de la información anterior construimos la tabla 13 que incluye los casos extremos. Descartando los casos III y IV por ser contrarios al sentido del índice de discriminación tenemos los siguientes puntos críticos:

$$A_I\ (0,0);\ B_I\left(\frac{n}{N}, 1\right)\ ;\ C_I\left(\frac{2n}{N}, 0\right)\ ;\ D_I\left(1\ \frac{2n}{N},\ 0\right)\ ;\ E_I\left(1\ \frac{n}{N}, 1\right)\ ;\ F_I\ (1,0)$$

que al ser graficados definen una región trapezoidal (RVA) tal como se muestra en la figura 6.

Tabla 13

	CASO 1	CASO 2	CASO 3	CASO 4	CASO 5	CASO 6	CASO 7	CASO 8
Cs	*N*	*N*	0	0	*n*	*N*	0	0
Ci	0	0	*n*	*N*	*n*	*N*	0	0
Cs - Ci	*N*	*N*	*-n*	*-n*	0	0	0	0
C	*N*	*N-n*	*n*	*N-n*	*2n*	*N*	*N-2n*	0
N	*N*	*N*	*N*	*N*	*N*	*N*	*N*	*N*
IDif	1-*n*/*N*	*n*/*N*	1-*n*/*N*	*n*/*N*	1-2n/N	0	2*n*/*N*	1
IDisc	1	1	-1	-1	0	0	0	0

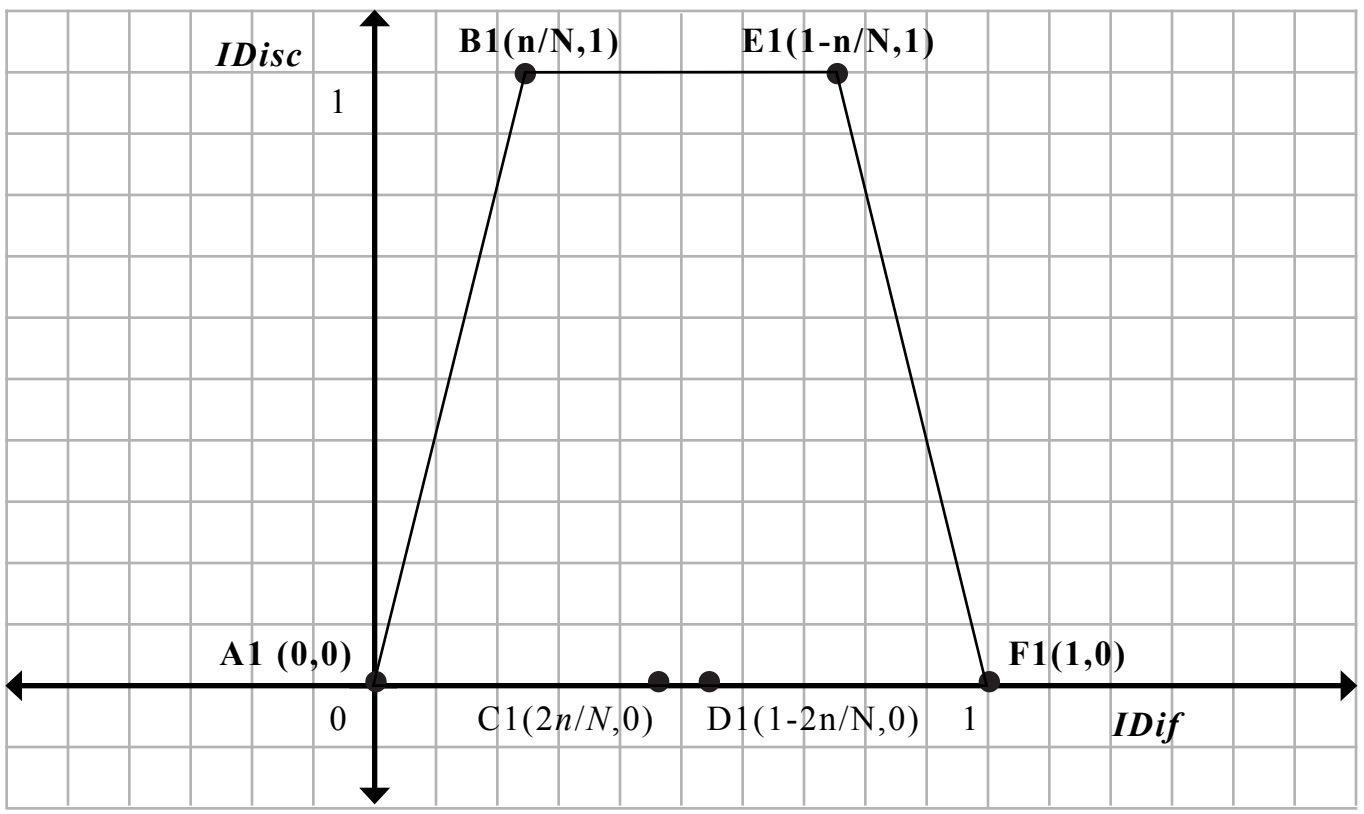

Figura 6. Región de valores admisibles para el caso: GS = GI = n (n < N/2)

Región de valores admisibles para grupos de ¨n¨ sujetos (n<N/2)

Siguiendo similar metodología que para los grupos formados a partir de la mediana se puede demostrar que La REGIÓN DE VALORES ADMISIBLES (RVA) para grupos con n sujetos, con $n < \frac{N}{2}$,queda determinada por:

$$(IDif\,,\ IDisc)\ \in \text{RVA} \leftrightarrow \begin{cases} 0 \le IDisc \le \frac{N}{n}\,(IDif) & si,\ 0 \le IDif \le \frac{n}{N} \\ 0 \le IDisc \le 1 & si,\ \frac{n}{N} \le IDif \le 1\,\frac{n}{N} \\ 0 \le IDisc \le \frac{N}{n} - \frac{N}{n}(IDif) & si,\ 1 - \frac{n}{N} \le IDif \le 1 \end{cases}$$

Si la prueba de rendimiento ha sido aplicada a un grupo numeroso de N examinados, se recomienda dividirlos en tres grupos GS, GM (grupo medio) y GI donde los grupos extremos estén formados por n sujetos cada uno ($n < \frac{N}{2}$). En este caso tenemos una región limitada por un trapecio y no por un triángulo como en los grupos con $\frac{N}{2}$ sujetos cada uno.

Adviértase que los puntos C_1 y D_1 no son vértices del trapecio, solo son puntos incluidos en el segmento $A_1 F_1$. De la figura 6 notamos que los puntos $B_1\left(\frac{n}{N},1\right)$ y $E_1\left(1\ \frac{n}{N},1\right)$ se encuentran a una distancia $d(B_1E_1)=1-\frac{2n}{N}$ uno del otro. Llevando al límite, cuando n tienda a 0 la distancia sería entre los puntos sería 1.

$$\lim_{n\to 0} d(B_1E_1) = \lim_{n\to 0}\left(1-\frac{2n}{N}\right) = 1$$

Este caso límite (e imposible) ocurriría cuando $n = 0$ convirtiendo la región trapezoidal en la región rectangular dada por $[0,1]\ .\ [0,1]$, que corresponde a los rangos teóricos de los índices y no mostraría ningún distinción entre los grupos. El caso contrario ocurriría cuando n tienda a $\frac{N}{2}$, ya que ahora la distancia entre los puntos sería 0.

$$\lim_{n\to \frac{N}{2}} d(B_1E_1) = \lim_{n\to \frac{N}{2}}\left(1-\frac{2n}{N}\right) = 0$$

Según esto último B_1 y E_1 coincidirían en un solo punto, aquel con coordenadas $(0.5,1.0)$. De esta forma la región trapezoidal $A_1B_1E_1F_1$ se convertiría en la región triangular del caso estudiado anteriormente, es decir de los grupos separados a partir de la mediana. Igualmente se puede demostrar que, en el caso límite, los puntos C_1 y D_1 coincidirían en el punto medio del segmento A_1F_1, es decir en el punto de coordenadas $(0.5,0.0)$ Cuanto más se aleje n por debajo del valor de $\frac{N}{2}$ los puntos B_1 y E_1 estarían más distantes y por tanto mayor área para la región de valores admisibles.

Región normada

A partir de la RVA trazaremos una recta horizontal correspondiente al valor norma de discriminación $IDisc = k$ y dos rectas verticales trazadas por los extremos del intervalo de dificultad deseable $[D,1-D]$. Todas las preguntas cuyos puntos pertenezcan a la RVA y se encuentren en o por encima de la recta horizontal y entre las rectas verticales pertenecerán a

la región normada. En la figura 7 se muestra esta región poligonal $P_1Q_1B_1E_1Q_1'P'_1$. Dado que k es constante se puede demostrar que para estas condiciones el área S de la región normada es función de D y está dada por:

$$S\ (\mathrm{D}) = -\frac{N}{n}\ D^2\ +\ 2kD\ +\ 1-k-\frac{n}{N}$$

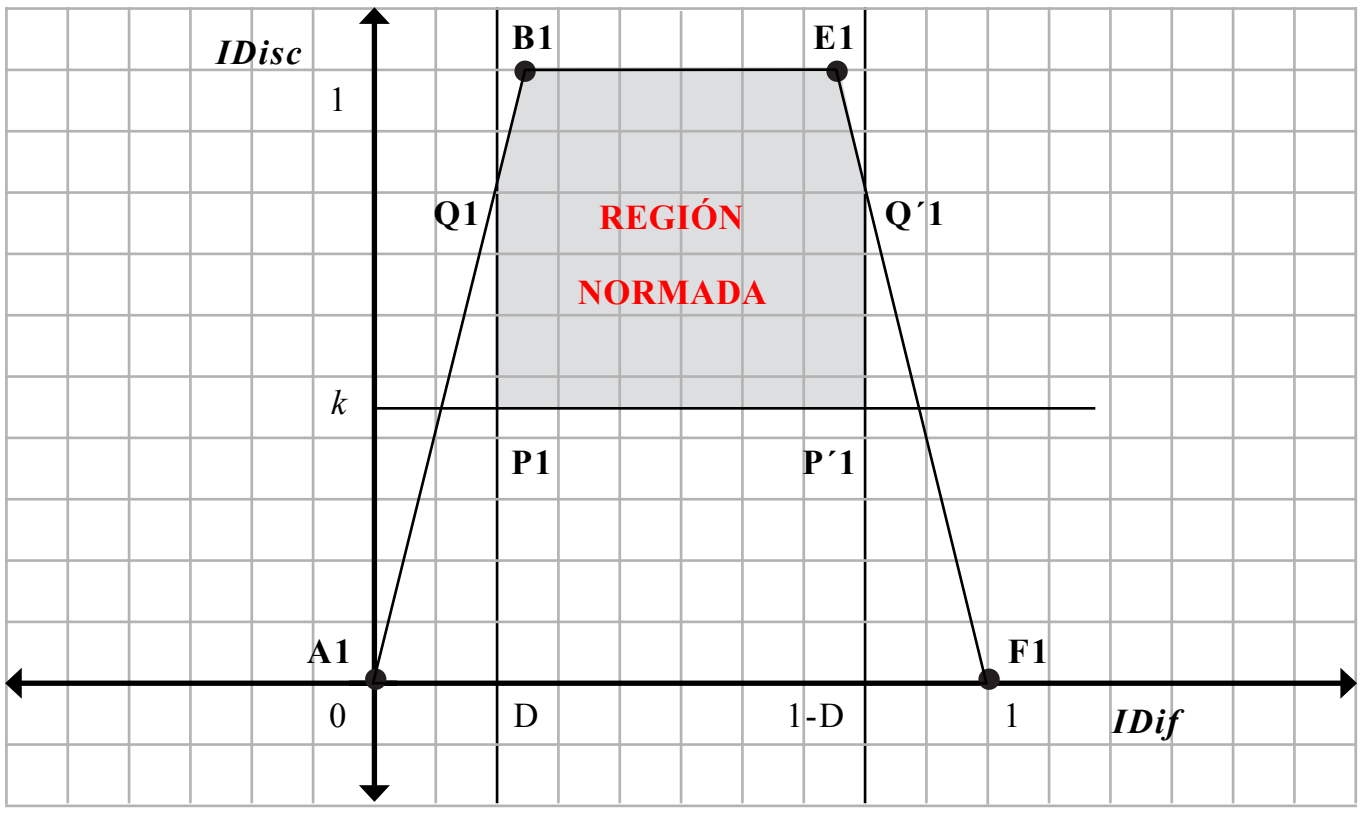

Figura 7. Región normada para el caso: GS = GI = n (n < N/2)

La misma que se maximiza para un valor de $D = \frac{n}{\mathrm{N}}\ k.$ Para este valor las coordenadas, tanto de los puntos P_1 y Q_1 como de los puntos P_1' y Q_1' son iguales, resultando una región trapezoidal.

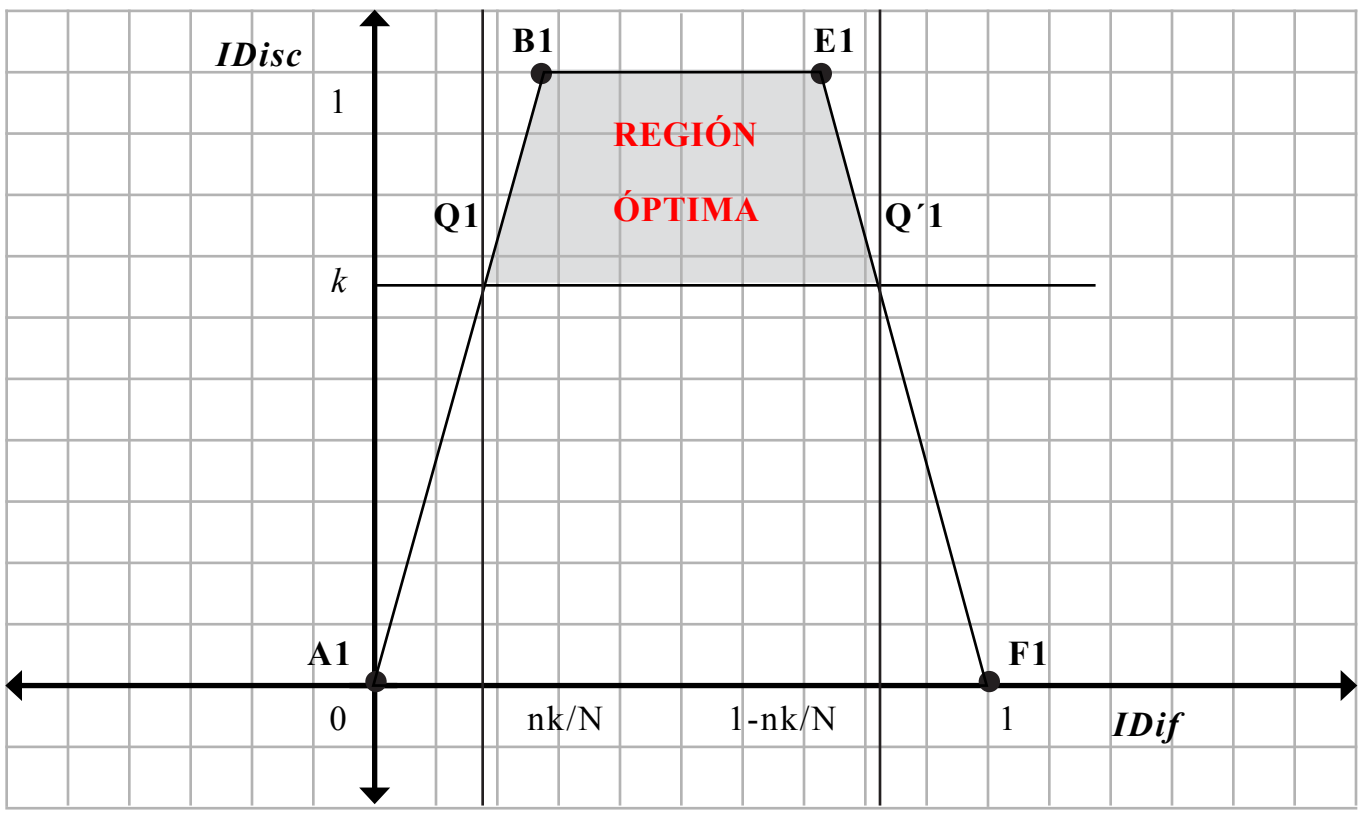

Figura 8. Región óptima para el caso: GS = GI = n (n < N/2) y norma de discriminación "k"

Resumiendo, si se consideran los grupos GS y GI formados por n sujetos cada uno, en una prueba de rendimiento cuyas preguntas discriminen igual o por encima de un valor norma de discriminación k, los índices de dificultad deben pertenecer al intervalo $\left[\frac{n}{\mathrm{N}} k,\ 1 - \frac{n}{\mathrm{N}} k \right]$. para que el mayor número de preguntas cuyas coordenadas son de la forma $(IDif,\ IDisc)$ se encuentren dentro de la región óptima. Como consecuencia, una prueba de rendimiento será de mayor calidad en la medida que contenga mayor número de preguntas en la región óptima. Así por ejemplo, si los GS y GI están formados por el 27% de los examinados, aquellos con puntajes más altos y más bajos, respectivamente, tenemos que $n = 27\%\ N$, con lo cual $\frac{n}{\mathrm{N}} = 0.27$. Para una norma de discriminación k, tendríamos el siguiente intervalo para el $IDif$: $[0.27k, 1-0.27k]$. La figura 9 muestra la región óptima donde se deberían ubicar las coordenadas de la forma $(IDif,\ IDisc)$ para cada una de las preguntas de una prueba de rendimiento donde los grupos se establecieron con la regla del 27%.

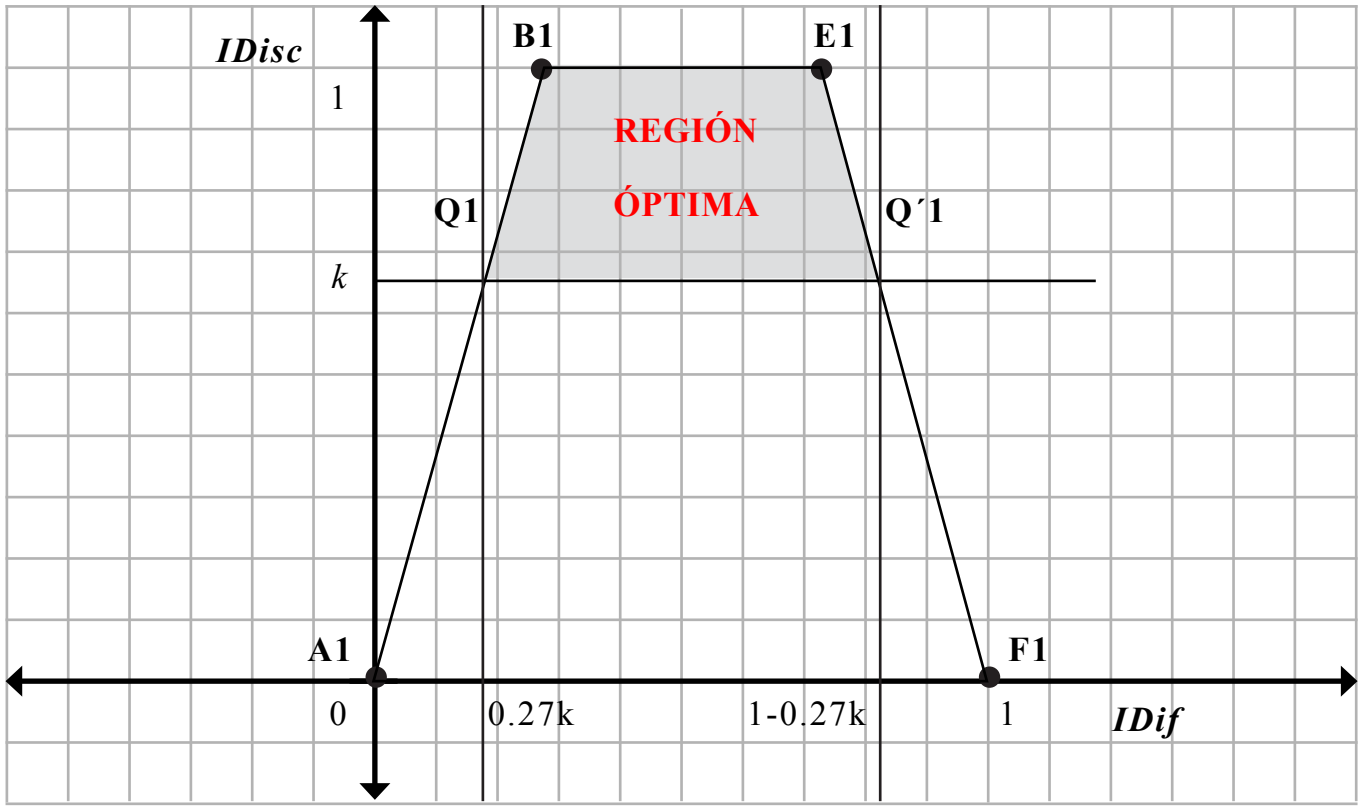

Figura 9. Región óptima para el caso: GS = GI = 0.27N y norma de discriminación "k"

Las regiones presentadas en las figuras 5 y 9 son regiones deseables óptimas donde se deberían encontrar los índices de la forma $(IDif,\ IDisc)$ para cada una de las preguntas de la prueba. Cuántos más puntos se ubiquen en dicha región, mayor número de preguntas tendrán un comportamiento deseable en la prueba y por tanto mayor calidad de esta.

COMENTARIOS FINALES

A partir de lo desarrollado en este trabajo, y para una prueba de rendimiento, podemos concluir:

a. Dado el índice de dificultad $(IDif)$ y el índice de discriminación $(IDisc)$ de una pregunta podemos formar pares ordenados de la forma $(IDif,\ IDisc)$ para cada una de las preguntas de la prueba.
b. Existe una zona en el plano bidimensional donde se ubican los pares $(IDif,\ IDisc)$ de cada una de las preguntas. Esta zona está perfectamente definida y será llamada región de valores admisibles (RVA).
c. La RVA tiene una formulación matemática.
d. Los valores del $IDif$ e $IDisc$ para cada pregunta no son independientes, ellos están ligados por medio de una relación matemática que se deriva de la formulación matemática de la RVA.
e. Dado un valor norma de discriminación k y un valor norma de dificultad D, existirá una región normada dentro de la RVA donde se ubicarían los pares $(IDif,\ IDisc)$ de las preguntas que se comportan según las normas.
f. Existe un valor D en términos de k que optimiza la región normada. Llamamos a esta región óptima o región de comportamiento deseable (RCD).
g. La RCD permite cumplir dos condiciones deseables al momento de diseñar una prueba de rendimiento: i) preguntas que discriminen igual o por encima de la norma y ii) preguntas con dificultad distribuida en un entorno cercano a la dificultad media.
h. La forma como se determinen los grupos superior (GS) e inferior (GI) influye en la formulación matemática de la RVA, la relación matemática entre los índices y el valor D que optimiza la región normada.
i. La RCD es una región triangular en el caso de grupos determinados a partir de la mediana.
j. La RCD es una región trapezoidal si los grupos son formados por n sujetos de los N examinados, donde $n < \frac{N}{2}$.

k. Si los grupos son determinados a partir de la mediana el valor que optimiza es $D = \frac{k}{2}$.

l. Si los grupos GS y GI son formados por n sujetos de los N examinados, con $n < \frac{N}{2}$, el valor que optimiza es $D = \frac{n}{N}\ k.$

m. Cuanto más pares de la forma $(IDif, IDisc)$ pertenezcan a la RCD, mejor comportamiento de las preguntas y mayor calidad de la prueba de rendimiento.

n. La determinación del valor norma de discriminación afecta el área de la RCD y por tanto el número de preguntas en ella.

o. La determinación del valor norma de discriminación influye en la interpretación de la calidad de la prueba de rendimiento.

Referencias

Andrich, D. (2008). Administering, Analysing and Improving Tests. En D. Andrich, & I. Marais (Eds.), *Introduction to Rasch Measurement of Modern Test Theory* (*Reader Semestre 2*). Crawley: UWA.

Bazán, J. (2000). *Evaluación psicométrica de las preguntas y pruebas CRECER 96*. Lima: Unidad de Medición de la Calidad, MINEDU. Recuperado de https://goo.gl/wae1Da

Canales, I. (2005). *Evaluación Educacional*. Lima: UNMSM.

Delgado, K. (2004*). Evaluación y Calidad de la Educación*. Lima: Derrama Magisterial.

Ebel, R. (1977). *Fundamentos de la Medición educacional*. Buenos Aires: Editorial Guadalupe.

García-Cueto, E. (2005). Análisis de los ítems. Enfoque clásico. En J. Muñiz, A. M. Fidalgo, E. García-Cueto, R. Martínez & R. Moreno (Eds.), *Análisis de los ítems. Cuadernos de Estadística N° 30* (pp. 53-79). Madrid: Editorial La Muralla.

Garret, H. (1966). *Estadística en Psicología y Educación*. Buenos Aires: Paidós.

Gronlund, N. (1999). *Elaboración de tests de aprovechamiento*. México: Trillas.

Masters, G. (1988). Item discrimination: When more is worse. *Journal of Educational Measurement*, *25*(1), 15-29. doi: https://doi.org/10.1111/j.1745-3984.1988.tb00288.x

Mejía, E. (2005). *Técnicas e Instrumentos de Investigación*. Lima: UNMSM.

Tristán, A. (1995). *Modelo de análisis de reactivos por computadora*. Primer Foro Nacional de Evaluación. Ceneval, Colima, México, pp.45-68. Tomado de Internet el [25-04-2008] en http://www.ieesa-kalt.com/articulo1_ka.html

Tristán, A. (2001). *Análisis de Rasch para todos*. México: Ceneval.

Tristán, A. (2006). *Fundamentos de la Evaluación del aprendizaje*. México: IEIA.

Wright, B. & Stone, M. (1979). Best *Test Design: Rasch Measurement*. Chicago: Mesa Press.

RIDU / Revista Digital de Investigación en Docencia Universitaria / ISNN 2223-2516