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

METHODOLOGICAL ARTICLE

Relationship between the Difficulty and Discrimination Indices

Relación entre los Índices de Dificultad y Discriminación

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

Received: 15/12/17 **Revised:** 19/03/18 **Accepted:** 08/05/18 **Published:** 30/06/18

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.

Keywords:
difficulty index, discrimination index, reliability, validity, educational measurement.

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.

CITE AS: Hurtado Mondoñedo, L. L. (2018). Relación entre los índices de dificultad y discriminación. *Revista Digital de Investigación en Docencia Universitaria*, 12(1), 266-292. doi: http://dx.doi.org/10.19083/ridu.12.614

***email:** 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.

The psychometric analysis of the questions comprising a performance test include the calculation of indices to typify them (Mejía, 2005). Two of the main indices are the difficulty index and the discrimination index. The difficulty index of a question, as its name suggests, is given by the numerical expression of the difficulty for test takers in answering the question. The discrimination index of a question divides, distinguishes, differentiates, and classifies test takers with higher and lower performance in the test. The quality of a performance test is largely defined by generating indices within suitable ranges. This begs the questions of how to calculate them, whether the difficulty and discrimination indices are connected, and which criteria determine they are within suitable ranges. This paper shows how these indices are determined and connected, as well as how the discrimination rule influences the valuation of quality of the test.

Objectives

a. To determine the connection between difficulty and discrimination indices of test questions.

b. To show how the discrimination rule influences the quality valuation of a test.

THEORETICAL FRAMEWORK

Answer Pattern.

In a performance test, difficulty and discrimination indices of questions are determined by answer patterns provided by test takers in the test questions. Based on Guttman's frame of reference, people are classified based on their total score, which is their index of skill, and items are arranged based on their total score (Andrich, 2008). The answers of every person to each question are the gross data we started with (Wright & Stone, 1979). Regardless of a

question's number of options, if these are dichotomous, the test takers will only have two answer options: correct or incorrect. We registered the answers to each test question per test taker. We agreed to use 1 to symbolize correct answers and 0 for incorrect answers. The answer pattern is set as a table where the first column registers the identification of each test taker and the following columns the answer code (1 or 0) to each question. The answer pattern allows us to determine the test taker's score and number of correct answers per question. The score is found by counting the number of correct answers of the test taker in each question of the test; that is, by counting the number of codes 1 in the row of each test taker. The number of correct answers per question can be found by counting the number of test takers who answered them correctly; that is, by counting the number of codes 1 in each question column. We can rank, sort and distinguish between test takers with higher and lower performance based on these scores. The tally of correct answers and the distinction between groups are necessary to determine the aforementioned indices. Table 1 shows the ranking and number of correct answers, drawn from the answer pattern of test takers from our example (see Table 1).

Table 1

Example of ranking and number of correct answers from a group of test takers

ANSWER PATTERN AND SCORE RANKING OM		Q1	Q2	Q3	Q4	SCORE
1	Test taker 6	1	1	1	1	4
2	Test taker 3	1	1	1	0	3
3	Test taker 8	1	1	1	0	3
4	Test taker 2	1	0	1	1	3
5	Test taker 1	1	1	0	0	2
6	Test taker 4	1	0	1	0	2
7	Test taker 5	0	0	1	1	2
8	Test taker 7	1	0	0	0	1
Number of correct answers		**7**	**4**	**6**	**3**	

Difficulty Index.

The question Q1 had the highest number of correct answers (7), whereas Q4 had the lowest number of correct answers (3). Since Q1 was answered correctly by more test takers than Q4, the former turned out to be easier for the group. We will say then, that Q4 is "harder" than Q1 for this group of test takers. In order to determine a measurement of how hard the question was for a group, we will compare the number of test takers who answered incorrectly with the optimal number. We will call difficulty index (*IDif*) to the comparison between the number of incorrect answers $N - C$ for the question and the total number of test takers N.

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

If we decompose the fraction: $IDif = \frac{N}{N} - \frac{C}{N}$

The first fraction corresponds to 1 and the second one to the so-called easiness index.

$$IDif = 1 - Ifac$$

This relationship shows that the difficulty and easiness of a question are exclusive and complementary notions regarding the unit. Easy and Difficult are polar adjectives. A question will be easier for a group of test takers if a higher number of people answer it correctly; therefore, the more people answer it incorrectly, the more difficult it will be. In order to analyze the difficulty of a question, we should count the number of incorrect answers instead of correct answers. In specialized literature, the relationship between the number of correct answers and the total of test takers is frequently named as "difficulty index" or "level of difficulty" (Canales, 2005; García-Cueto, 2005; Gronlund, 1999; Tristán, 2001). According to this definition, the higher the index, the greater the number of correct answers and thus the easier the question, which is contrary to the difficulty. From a purely semantic point of view, the term easiness index (García-Cueto, 2005) is more accurate to refer to the relationship between the number of correct answers and the total of test takers, as we had previously considered. The distinction is made in this paper by referring to the numerical expression of how difficult the target group finds it to answer a question correctly when we talk about difficulty index. This is given by:

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

Where C is the number of correct answers per question and N is the number of test takers (see Table 2).

Table 2
Estimated difficulty indices

	Q1	Q2	Q3	Q4
Number of correct answers (C)	7	4	6	3
N	8	8	8	8
IDif = 1 - C/N	0.125	0.5	0.25	0.625

The difficulty index (*IDif*) can only use values within an interval. If every test taker answers a question correctly, we will find that the number of correct answers (C) is equal to the number of test takers (N), in this case $C=N$, and therefore the difficulty index will be $IDif = 1 - \frac{N}{N} = 0$. On the other hand, if none of the test takers answers a question correctly, the number of correct answers will be zero ($C = 0$) and therefore the difficulty index will be $IDif = 1 - \frac{0}{N} = 1$. In general, since $0 \leq C \leq N$, then $0 \leq \frac{C}{N} \leq 1$, implying that $-1 \leq \frac{C}{N} \leq 0$, and therefore $0 \leq 1 - \frac{C}{N} \leq 1$. The value of difficulty index can thus be between 0 and 1, including these ones. The higher the difficulty index is, the more difficult the question is.

$$0 \leq IDif \leq 1$$

Discrimination Index.

According to Bazán (2000), “the discrimination of a question is measured by how it helps to expand the estimated differences between those who got a relatively high total test score and those who got a relatively low score” (p. 6). So, the discrimination index is the numerical expression of how a question divides the test takers with the highest performance from those with the lowest performance. These groups, referred to herein as upper group (UG) and lower group (LG), are determined using the score average as cut-off point, which is 2.5 in our example (see Table 1). The UG will consist of test takers with higher scores than 2.5 and the LG of test takers with lower scores than 2.5. Tables 3 and 4 will present the scores and answer patterns of the individuals from each group.

Table 3
Answer pattern of the upper group

		Q1	Q2	Q3	Q4	Score
1	Test taker 6	1	1	1	1	4
2	Test taker 3	1	1	1	0	3
3	Test taker 8	1	1	1	0	3
4	Test taker 2	1	0	1	1	3
	Number of correct answers	**4**	**3**	**4**	**2**	

Table 4
Answer pattern of the lower group

		Q1	Q2	Q3	Q4	Score
5	Test taker 1	1	1	0	0	2
6	Test taker 4	1	0	1	0	2
7	Test taker 5	0	0	1	1	2
8	Test taker 7	1	0	0	0	1
	Number of correct answers	**3**	**1**	**2**	**1**	

Since it is not possible to directly observe the test takers' real level of knowledge on the topic discussed in this test, it must be inferred. Although test scores are an ordinal measurement, these can traditionally be indicators of the level of knowledge of the person being tested. Occasionally, the ratio between the number of correct answers and test questions is used as an indicator. Similarly, a measurement of a group's level of knowledge to a certain question is determined by the ratio between the number of correct answers and test takers of that group. The more correct answers the group has, the more homogeneous their level of knowledge will be on the topic discussed in the question. The more correct answers the group has, the higher the ratio between the number of correct answers and test takers in the group and thus the more homogeneous the group will be. A question that aims at differentiating test takers with higher performance form those with lower performance would have to compare the ratio between the number of correct answers and test takers per group. But this comparison must be by excess; the greater the difference between the UG and LG's ratio of correct answers and test takers is, the higher the measurement of discrimination will be. The traditional theory of the test states that high discrimination is interpreted as a desirable feature and a key quality indicator of an item (Masters, 1988).

Let us consider a group of test takers N, the UG and LG determined according to the average

would each have $N / 2$ test takers. If the number of correct answers of UG is represented by C_S and the number of correct answers of LG by C_i, the discrimination measurement of UG from LG is determined by the difference $C_S \ / \ (N/2) - C_i \ / \ (N/2)$. This difference is referred to as discrimination index, which will be represented by $IDisc$ and is calculated using the following formula:

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

Table 5
Estimated discrimination indices

	Q1	Q2	Q3	Q4
UG Number of correct answers (Cs)	4	3	4	2
LG Number of correct answers (Ci)	3	1	2	1
N	8	8	8	8
IDisc = (Cs-Ci)/(N/2)	0.25	0.5	0.5	0.25

The discrimination index ($IDisc$) can only select values within an interval. If a question is answered correctly by every test taker from UG and none from LG, we will have that $C_s = N/2$ and $C_i = 0$, so the discrimination index will be $IDisc = \frac{(0 \ \ N/2 - 0)}{(N/2)} = 1$. On the other hand, if none of the test takers from LG answer a question correctly but everyone from LG does, then we will have that $C_s = 0$ and $C_i = N \ / \ 2$, which will result in an index as follows: $IDisc = (0 - N/2) \ / \ (N/2) = -1$. A middle case would be when none of the test takers from both groups answers the question correctly ($C_s = 0$ y $C_i = 0$), leading to the following index: $IDisc = (0 - 0) \ / \ (N/2) = 0$. This way, the discrimination index can be between -1 y 1, including these values.

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

We can say that a question with $IDisc = -1$ is totally discriminatory, while, at the other end, a question with $IDisc = 1$ is erroneously discriminatory. If the goal is to measure the difference between test takers of the UG and those of the LG regarding a higher level of knowledge, a negative $IDisc$ would indicate a higher level of knowledge in the LG than in the UG, an unacceptable situation as it contradicts the sense of the index.

The discrimination index has been calculated and described based on opposite groups. This is an easier way to determine discrimination than other indices. A common way to

calculate this index is by a question-answer biserial correlation. A question is properly discriminatory in a test if it can be used to differentiate, distinguish, and separate individuals with higher and lower scores. When the question is properly discriminatory, the immediate consequence is that a positive correlation will take place between the question and test scores (García-Cueto, 2005). This way, the discrimination index is a statistical index that describes to what extent a question is in line with the other ones that seek to discriminate between people (Andrich, 2008). The type of correlation to be used will depend on the measurement features of the questions and test, such as, for example, if both variables are dichotomous, dichotomized, continuous, or a combination of all of them. Generally, the greater this correlation, the higher the discrimination; though there are some exceptions where a high discrimination is not expected. According to Garret (1966), the benefit of other methods of discrimination index calculation "is judged based on the extent to which they are able to give results that approximate those obtained through biserial correlation" (p. 403).

The problem of optimal difficulty

The questions that are too easy or turn out to be too difficult would lead to asymmetrical distributions regarding their percentage of correct and incorrect answers. An easy question shows a strong negative asymmetry; and the asymmetry is positive when the question is more difficult, being completely asymmetrical if its difficulty index is 0.5 (García-Cueto, 2005). In an experimental study of the distribution of test scores with difficulty values for items, Ebel (1977) notes: "the results of this research confirm the recommendation to use questions of intermediate difficulty when drafting a performance test" (p. 491). Along the same lines, García-Cueto (2005) maintains that "tests will generally get the best results when most of the items are at an intermediate difficulty" (p. 60). Tristán (2001) expresses an opposing opinion: "it is advisable to have a reagent in every range of difficulty and not only reagents focused on difficulty at 50% so that we are able to accurately measure the knowledge of each person" (p. 7). Tristán takes as an example the body temperature measured with a water thermometer, properly calibrated in a range from 0 to 100, where the optimal temperature is not at 50 °C. Similarly, "the goal of including reagents with different levels of difficulty is to have a well-calibrated scale"1; hence the importance of considering the full range of difficulty for test questions. The standard tests –such as entrance tests– allow us to classify test takers, organizing them in a common scale and differentiating between high and low performing groups. In classroom assessments, a teacher does not normally seek to turn their performance test into a well-calibrated scale. It is more likely that their rank changes when applied to different groups. The low number of test takers these tests are usually

1 Tristán (1995, 2001, 2006) explains these ideas when the foundations for the Kalt model were presented.

applied to tends to be the biggest limitation for obtaining accurate scales. Similarly, the limited number of test takers and classroom test questions causes a greater error in the calculation of reliability. On the other hand, the symmetrical distribution of scores should not be a requirement. The students' performance in the test should not necessarily follow a normal distribution. This is more likely when the number of test takers is big and there are a series of differential features. Delgado (2004) refers to the agreement of Bloom, Hastings & Madaus with De Landsheere that "the normal curve is the most appropriate distribution for casual activity, whereas education has a deliberate purpose" (p.165). This paper aims at including both aforementioned opinions. On the one hand, to determine an interval for the difficulty index in a range close to intermediate difficulty and, on the other hand, to distribute them in this interval as a well-calibrated scale. Graphically, both approaches can be represented in what we propose as optimal area.

METHOD

We will begin by analyzing the case of certain UG and LG based on the average. This will allow us to determine an area of admissible values for the indices. Separately, each of the indices can only have values within an interval of values. However, they must also be located within an area in a bidimensional plane when analyzed together as an ordered pair. Our first task will be to mathematically identify this area and therefore determine the relationship between $IDif$ and $IDisc$. Once it has been defined, we will try to find a desirable part of this area for these indices, given a discrimination rule. Next, we will seek to optimize this area so that it has the highest number of ordered pairs of the form of ($IDif$, $IDisc$) associated with the performance test questions. Finally, we will apply these areas to groups of n individuals, where $n < N$.

Difficulty and Discrimination

The existence of a connection between $IDif$ and $IDisc$ is generally mentioned, though nothing else is said other than what can be gathered from the critical values. A question with a difficulty of 0 or 1 has a discrimination of 0; and a question with a difficulty of $0{,}5$ has a discrimination of 1. We will seek to deepen this relationship. From the results presented in previous tables, we can summarize the indices for the four questions of our example in the following table:

Table 6
Estimated difficulty and discrimination indices

	Q1	Q2	Q3	Q4
IDif	0.125	0.500	0.250	0.625
IDisc	0.250	0.500	0.500	0.250

These indices, gathered after implementing the test, allow us to describe the questions and analyze the results. This analysis includes a characterization based on the question's level of difficulty, a characterization based on its level of discrimination, and a quality assessment of the test based on the location of these indices in an interval of admissible values. There is a connection between *IDif* and *IDisc*. The group of possible values for one of them is related to the value of the other. A comparison of the indices shown on table 5 with a particular discrimination or difficulty rule would make us accept some questions and review (or discard) others. Whereas one of the goals of a performance test concerning standards is to organize test takers according to their level of knowledge of the topic assessed on the test, the power of discrimination of a question is more important. A good question must be answered correctly by a higher number of individuals with higher scores in the test than those with lower scores (García-Cueto, 2005).

Area of Admissible Values

This paper was conducted based on the Kalt model2; however, we have based our analysis on the number of correct answers rather than percentages as Kalt shows. On the one hand, this will allow us to simulate extreme and intermediate cases for the answer pattern of test takers with a simple presentation and, on the other hand, mathematically formulate what will refer to as area of admissible values, normed area, and optimal area. We will begin with the case of UG and LG groups defined by the average.

Table 7
Extreme cases for UG and LG determined by the average

	CASE A	CASE B	CASE C	CASE D
Cs	N/2	N/2	0	0
Ci	N/2	0	N/2	0
C = Cs + Ci	N	N/2	N/2	0
Cs - Ci	0	N/2	-N/2	0
N	N	N	N	N
IDif	0	0.5	0.5	1
IDisc	0	1	-1	0

The UG consists of individuals with higher scores than average, and the LG by those with lower scores than the average. Since there is no middle group, individuals from UG and

2 The Kalt model is used for the analysis of computer-based questions by IEIA in Mexico.

LG comprise the total of test takers, which is why the total number of correct answers in each question will be equal to the sum of correct answers in both UG and LG to that question; that is $C=C_s + C_i$.

According to this, we can distinguish four extreme cases:

Case A: All test takers answer correctly: $C_s = N/2,\ C_i = N/2$ and $C = N$.

Case B: Only individuals from UG answer correctly: $C_s = N/2,\ C_i = 0$ and $C = N/2$.

Case C: Only individuals from LG answer correctly: $C_s = 0,\ C_i = N/2$ and $C = N/2$.

Case D: All test takers answer incorrectly: $C_s = 0,\ C_i = 0$ and $C = 0$.

Table 8 shows $IDif$ and $IDisc$ values for the aforementioned cases.

We will assume the case of 80 test takers ($N=80$) to explain this point more clearly. Both UG and LG would each have 40 test takers and thus the number of correct answers C_s or C_i could not be higher than 40. The theoretical assumption is that the UG comprises test takers with higher performance, and the LG those with lower performance. The optimal behavior in a discriminatory question of these groups is that everyone from UG answers it correctly and everyone from LG fails. An erroneous behavior would arise if everyone from UG fails and everyone from LG answers correctly. Among both behaviors –optimal and erroneous–, we can find others by making the behavior of one of the groups fixed and the other variable.

Table 8

Cases AB. Optimal behavior of UG ($C_s = 40$) and variable behavior of LG.

	CASE AB1	CASE AB2	CASE AB3	CASE AB4	CASE AB5
Cs	40	40	40	40	40
Ci	0	10	20	30	40
C = Cs + Ci	40	50	60	70	80
Cs - Ci	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
	Intervals	0. ≤ *IDif* ≤ 0.5	0 ≤ *IDisc* ≤ 1		

Table 9
Cases BD. Optimal behavior of LG ($C_i = 0$) and variable behavior of UG.

	CASE BD1	CASE BD2	CASE BD3	CASE BD4	CASE BD5
Cs	0	10	20	30	40
Ci	0	0	0	0	0
C = Cs + Ci	0	10	20	30	40
Cs - Ci	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
	Intervals	$0.5 \leq IDif \leq 1$	$0 \leq IDisc \leq 1$		

Table 10
Cases DC. Erroneous behavior of UG ($C_s = 0$) and variable behavior of LG.

	CASE DC1	CASE DC2	CASE DC3	CASE DC4	CASE DC5
Cs	0	0	0	0	0
Ci	0	10	20	30	40
C = Cs + Ci	0	10	20	30	40
Cs - Ci	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
	Intervals	$0.5 \leq IDif \leq 1$	$-1 \leq IDisc \leq 0$		

Table 11
Cases CA. Erroneous behavior of LG ($C_i = 40$) and variable behavior of UG.

	CASE CA1	CASE CA2	CASE CA3	CASE CA4	CASE CA5
Cs	0	10	20	30	40
Ci	40	40	40	40	40
C = Cs + Ci	40	50	60	70	80
Cs - Ci	-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
	Intervals	$0 \leq IDif \leq 0.5$	$-1 \leq IDisc \leq 0$		

The cases described above allow us to have a group of possibilities for both $IDif$ and $IDisc$. Table 12 shows the $IDif$ and $IDisc$ in descending order according to difficulty.

Table 12

Cases of $IDif$ and $IDisc$ arranged in order of difficulty

CASE	*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

We will create ordered pairs of the form of $(IDif, IDisc)$ for each case considered and we will put them in a bidimensional plane ($IDif$ **VS** $IDisc$). The horizontal axis considers the values of $IDif$ and the vertical axis the values of $IDisc$. Figure 1 shows the group of points of the form $(IDif, IDisc)$ for each case in table 12. The junction of points would indicate the outline of the

rhombus where extreme possibilities for points with coordinates indicated by ordered pairs of the form (*IDif, IDisc*) would be located. Thus, for example, an extreme possibility would occur in a question where the 40 test takers from UG and only 6 from the LG answer correctly, so the point $(0.575, 0.850)$ associated with this question would be part of the outline of the rhombus. The non-extreme possibilities would be located inside. For example, if 37 test takers from the UG and 9 from the LG answer correctly, we would get the point $(0.575, 0.700)$ that is located below the point $(0.575, 0.850)$ from the outline of the rhombus. This way, both the perimeter and the inside of the rhombus would contain the set of all possibilities of ordered pairs of the form (*IDif, IDisc*) for each question of the performance test.

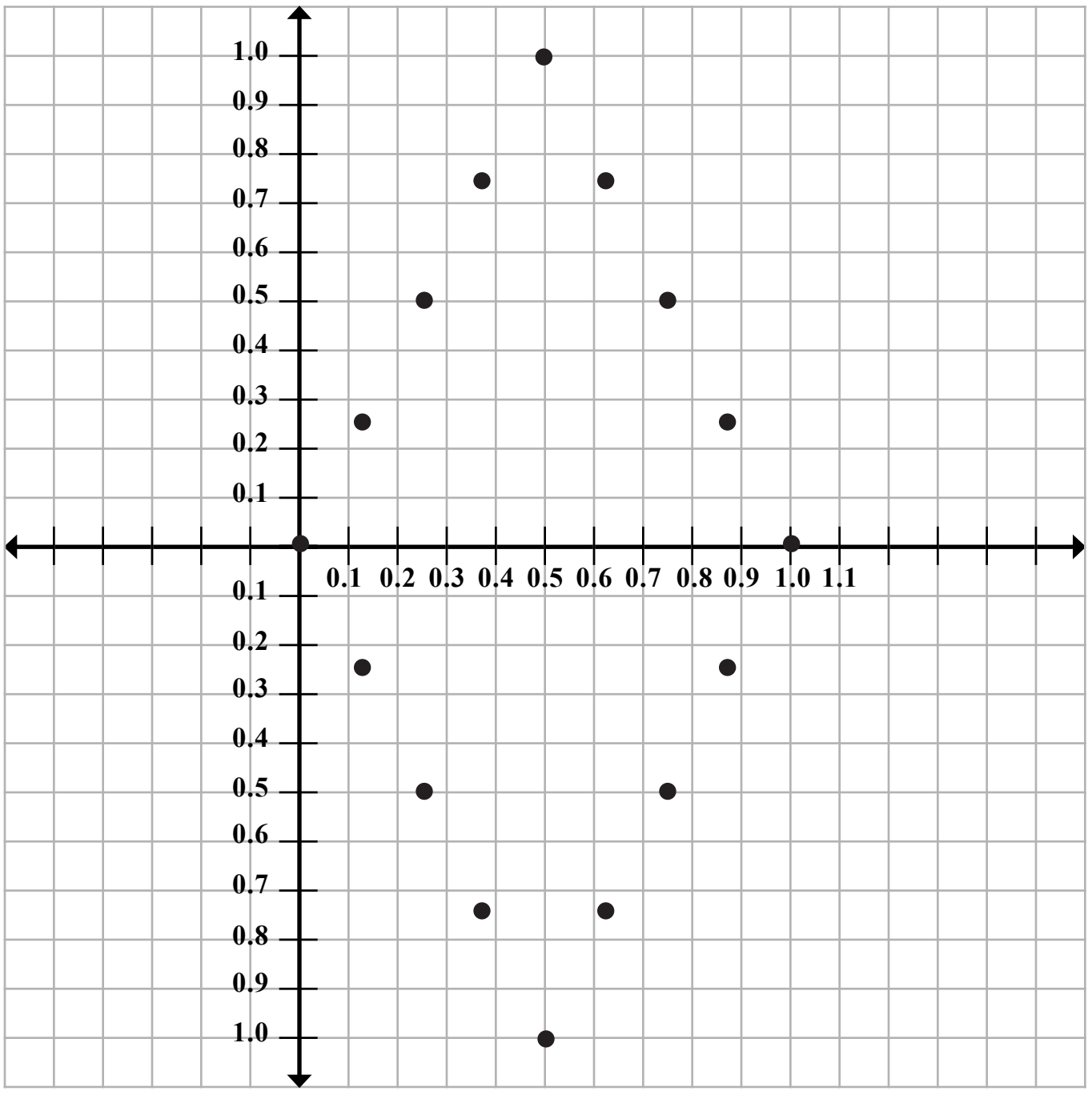

Figure 1. Distribution of points (*IDif,IDisc*) for the cases on table 12

If we refer to the vertices of the rhombus as $P(0.0,0.0)$, $Q(0.5,1.0)$, $R(1.0,0.0)$ and $S(0.5,-1.0)$, and in the form this paper has defined difficulty and discrimination indices, the ordered pair with *IDif* as first component and *IDisc* as second component must belong to the area restricted by the rhombus PQRS, as shown in figure 2. Since a negative *IDisc* is not admissible, we should only consider the cases with non-negative *IDisc* in the aforementioned area of the rhombus; that is, only the cases AB and BD.

This would make us redefine our area as the one bounded by the triangle PQR in the first quadrant where the admissible values of *IDisc* would be located, as shown in figure 3. We can therefore say that the points of the form of (*IDif,IDisc*) associated to each question of a

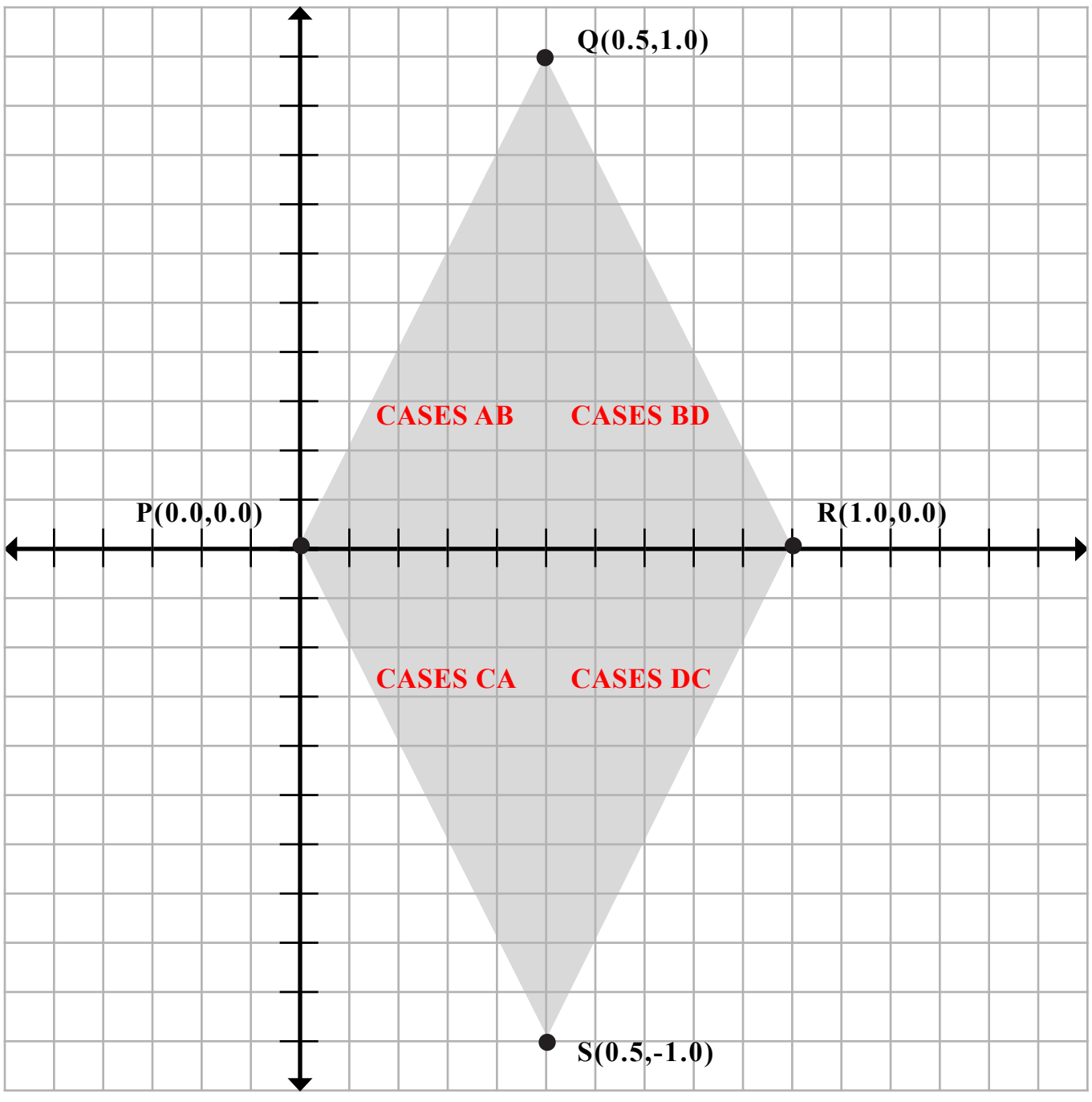

Figure 2. Area created by the group of points (*IDif,IDisc)*

performance test must belong to the triangular area bounded by the points $(0.0,0.0)$, $(0.5,1.0)$ and $(1.0,0.0)$, referred to as area of admissible values.

Mathematical Formulation for the Area of Admissible Values

The triangular area PQR will become our area of interest and it can be mathematically defined through inequalities. When an inequality only has two variables, the solution set is graphically represented by a half-space in the Cartesian plane. The half-space of an inequality of the type $y \leq ax+b$ consists of the respective line and every point below it. If the inequation is of the type $y \geq ax+b$, the half-space includes the line and every point above it. The $IDif$ will be used as the x variable and $IDisc$ as the y variable. We will first look for the linear equations that include the segments PQ and QR, which are two of the limits of our triangular area. A point-gradient form will be used.

The segment PQ is bounded by the points $P(0,0)$ and $Q(0.5,1)$, so its gradient is determined by $\frac{1 - 0}{0.5 - 0} = 2$. Using $P(0,0)$ as crossing point with a gradient of $m=2$, we state a linear equation including PQ:

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

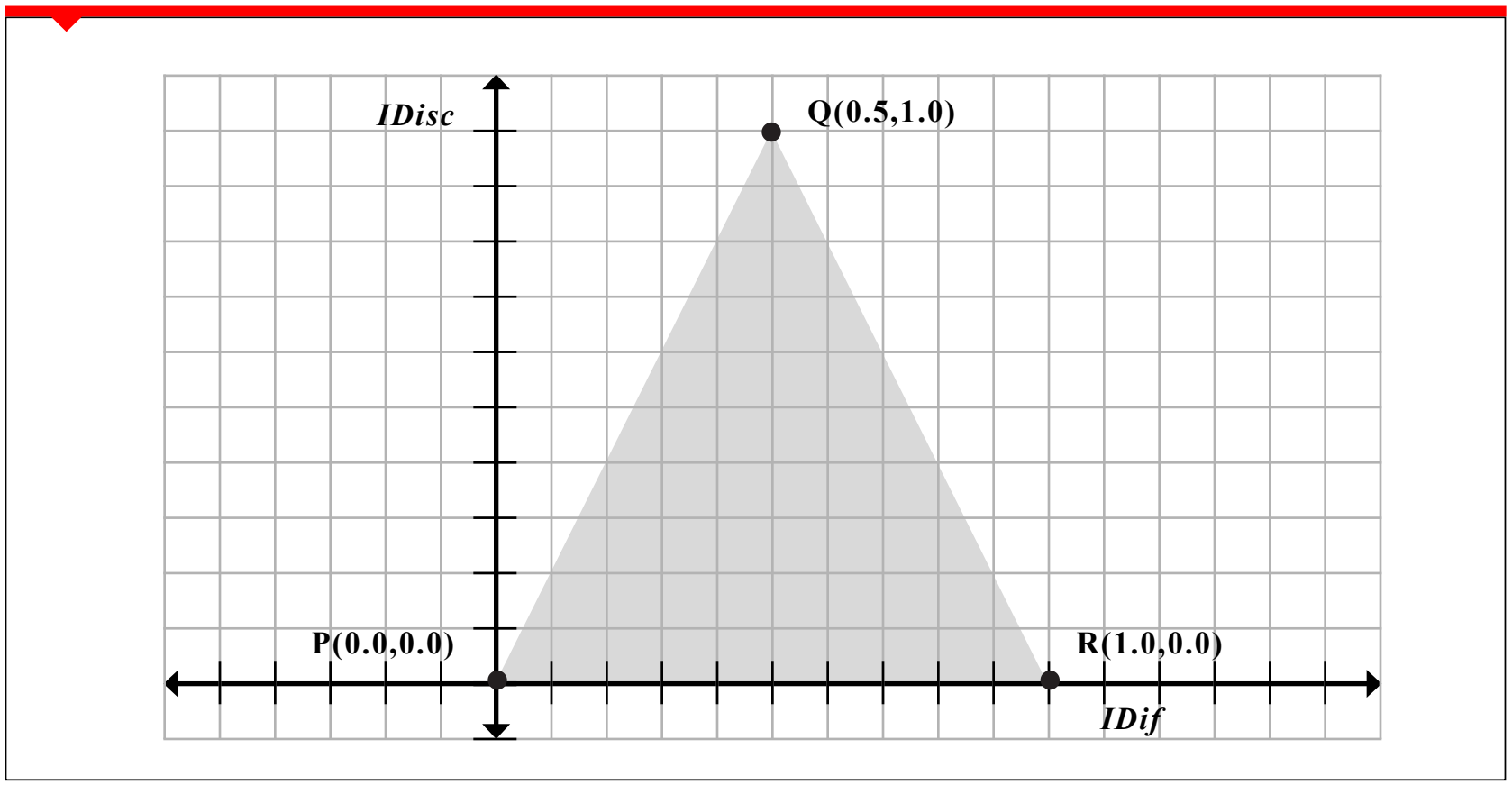

Figure 3. Area of admissible values

The segment QR is bounded by the points $Q(0.5,1)$ and $R(1,0)$, so its gradient is determined by $\frac{0-1}{1-0.5} = -2$. Using $R(1,0)$as crossing point with a gradient of $m=-2$, we estate a linear equation including QR:

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

The equations found belong to two lines that bound the triangular area PQR. The third one includes the segment PR ; that is, the horizontal axis, with the equation $y = 0$ (L_{PR}: y =0). These equations allow us to define the triangular area PQR as the intersection of three half-spaces:

Half-space 1: Consisting of all the points in the line LPQ and below it: $y \leq 2x$.

Half-space 2: Consisting of all the points in the line LQR and above it: $y \leq -2x + 2$.

Half-space 3: Consisting of all the points in the line LPR and above it: $y \geq 0$.

That is,

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

We can also redefine the triangular area as the junction of two right triangles. If we refer to the midpoint of the PR segment as T, then the PQR area is the junction of both triangular areas PTQ and RTQ. These latter two are defined by conveniently restricting the interval of the variable x. We have as follows:

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

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

The PQR area has three edge points. The point P with coordinates x=0, y=0 ; the point Q with coordinates $x = 0.5, y = 1$; and the point R with coordinates x=1, y=0. Or the equivalent form:

Area PQR = Area PTQ ⌣ Area RTQ

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

Since x and y correspond to $IDif$ and $IDisc$ respectively, the area PQR corresponds to the set of possible ordered pairs of the form of $(IDif, IDisc)$ for certain question. We will refer to this zone as AREA OF ADMISSIBLE VALUES (AAC) of indices for every performance test question to be analyzed.

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

Based on the AAV, we can mathematically formulate the relation between $IDif$ and $IDisc$ when the groups have been determined according to the average.

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

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

This means that the interval of potential values for $IDisc$ is determined by the value of $IDif$. So, for example, if we have a question with $IDif = 0.35$, its discrimination index must be between 0 and 2 x 0.35; that is: $0 \le IDisc \le 0.70$. If other question has a difficulty index of 0.75, its discrimination index must be between 0 and $2 - 2$ x 0.75; that is: $0 \le IDisc \le 0.50$.

Normed Area

The triangular area PQR described in the item above shows us the AAV; that is the area of the plane $IDisf$ vs $IDisc$, where the indices of questions are distributed. This is a theoretical area, but not necessarily optimal. We can determine a value for the $IDisc$ above which the discrimination indices of questions should be placed. We will refer to it as *discrimination standard value*, and it will be represented by $IDisc_{norma} = k$, where k is a positive constant lower than 1. On the other hand, we know that the closer to 0.5 the difficulty indices of the questions, the higher the reliability for the test will be and the better the scores will be distributed. There must be an interval of values for the difficulty index that enables a higher reliability without sacrificing the discrimination between groups; that is, there must be a closed interval where desirable difficulty indices are located. If we define the lower end of the interval with the value D and, given the symmetry of the admissible value area, the higher value of the interval will be equivalent to $1 - D$. We will refer to D as a *difficulty normed value* and it will be represented by $IDif_{norma} = D$.

We will outline a horizontal line corresponding to the normed value of discrimination. Every question whose points belong to the AAV and are inside or above the horizontal line would discriminate according to the norm. Figure 4 represents this with the horizontal line $IDisc = k$. Simultaneously, the closed interval $[D, 1-D]$ has been defined, where we can find the normed values for the difficulty index.

All questions whose points belong to the AAV and are between the vertical lines outlined

by the endpoints of the interval will have a desirable difficulty. This way, questions whose points with coordinates $(IDif,\ IDisc)$ are located within the AAV, between the vertical lines $IDif = D$ and $IDif = 1 - D$, and inside or above the horizontal line $IDisc = k$ will belong to the normed area.

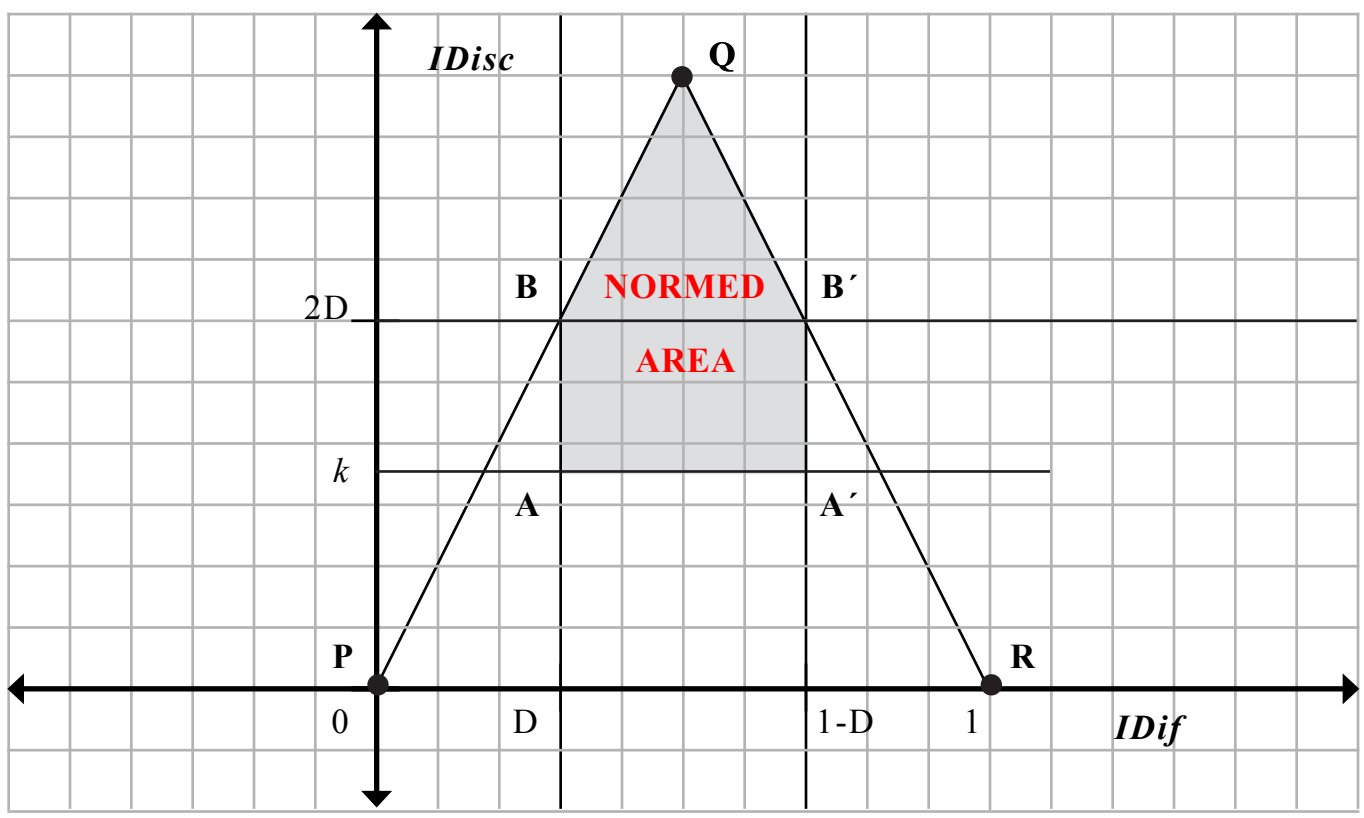

Figure 4. Normed area for the case: GS = GI = N/2

The normed area corresponds to the inside and outline of the polygon $ABQB'A'$. The coordinates of the polygon's vertices are determined by $A(D,k);\ B(D,2D);\ Q(0.5,1.0);\ B'(1-D,2D)$ and $A'(1-D,k)$. The points B and B' are determined by replacing $x=D$ and $x=1-D$ in the linear equations of L_{PQ} and L_{QR}, respectively.

Optimal Area

The more points within the normed area, the better the behavior of the questions and therefore the better constructed the performance test will be. These points must be distributed covering the interval of desirable difficulty and above the discrimination normed value, covering a part of the polygonal area. A greater number of these dispersed points, following both rules –difficulty and discrimination– would create an optimal area determined by the biggest area possible of the polygon $ABQB'A'$. We are facing an optimization problem whose objective function is the area function of the polygon $ABQB'A'$.

With the help of figure 4, we can model the function area of the polygon.

Area of polygon $ABQB\acute{A}'$ = Area of rectangle $ABQB'A'$ + Area of triangle BQB'

Area of polygon $ABQB\acute{A}' = (1-2D)(2D-k) + \frac{1}{2}(1-2D)(1-2D)$

Area of polygon $ABQB'A' = -2D^2 + 2kD - k + 0.5$

When analyzing a test, we compare the indices of questions with some qualifying criterium. If we accept certain discrimination rule, the value of k is constant; therefore, the polygon $ABQB'A'$ area is expressed as a function of the difficulty normed value D. We use S to represent the area of the polygon $ABQB\acute{A}'$, so:

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

In order to maximize this function, we must first find its critical value; that is, the value of D that maximizes the function $S(D)$. For this reason, we found the first derivative of the function and set it equal to zero.

$$S(D) = -4D + 2k$$

$$-4D + 2k = 0$$

$$D = \frac{k}{2}$$

The second derivative of the function $S(D)$ is determined by $S''(D) = -4$. So, for $D = \frac{k}{2}$ we have that $S'' = \left(\frac{k}{2}\right) < 0$. This shows that the function $S(D)$ reaches its maximum value when $D = \frac{k}{2}$. In other words, for a normed difficulty value of $D = \frac{k}{2}$, the part of desirable area is the biggest possible. With $D = \frac{k}{2}$ we can find the coordinates for the polygon vertices, so we have that $A\left(\frac{k}{2}, k\right)$; $B\left(\frac{k}{2}, k\right)$; $Q\,(0.5, 1.0)$; $B'\left(1-\frac{k}{2}, k\right)$ and $A'\left(1-\frac{k}{2}, k\right)$. It should be noted that both points A and B, and A' and B' have the same coordinates for the value of $D = \frac{k}{2}$ which makes the polygonal area optimal. In other words, the polygonal area becomes optimal when it turns into the triangular area BQB'.

This means that a performance test with questions that discriminate equal to or above a discrimination normed value k must have difficulty indices within the interval $\left(\frac{k}{2}, 1-\frac{k}{2}\right)$ so that the greatest number of questions with coordinates ($IDif$, $IDisc$) will be located in the optimal area. As a result, a performance test will have a higher quality as long as it includes a greater number of questions in the optimal area. It should not be surprising that the optimal

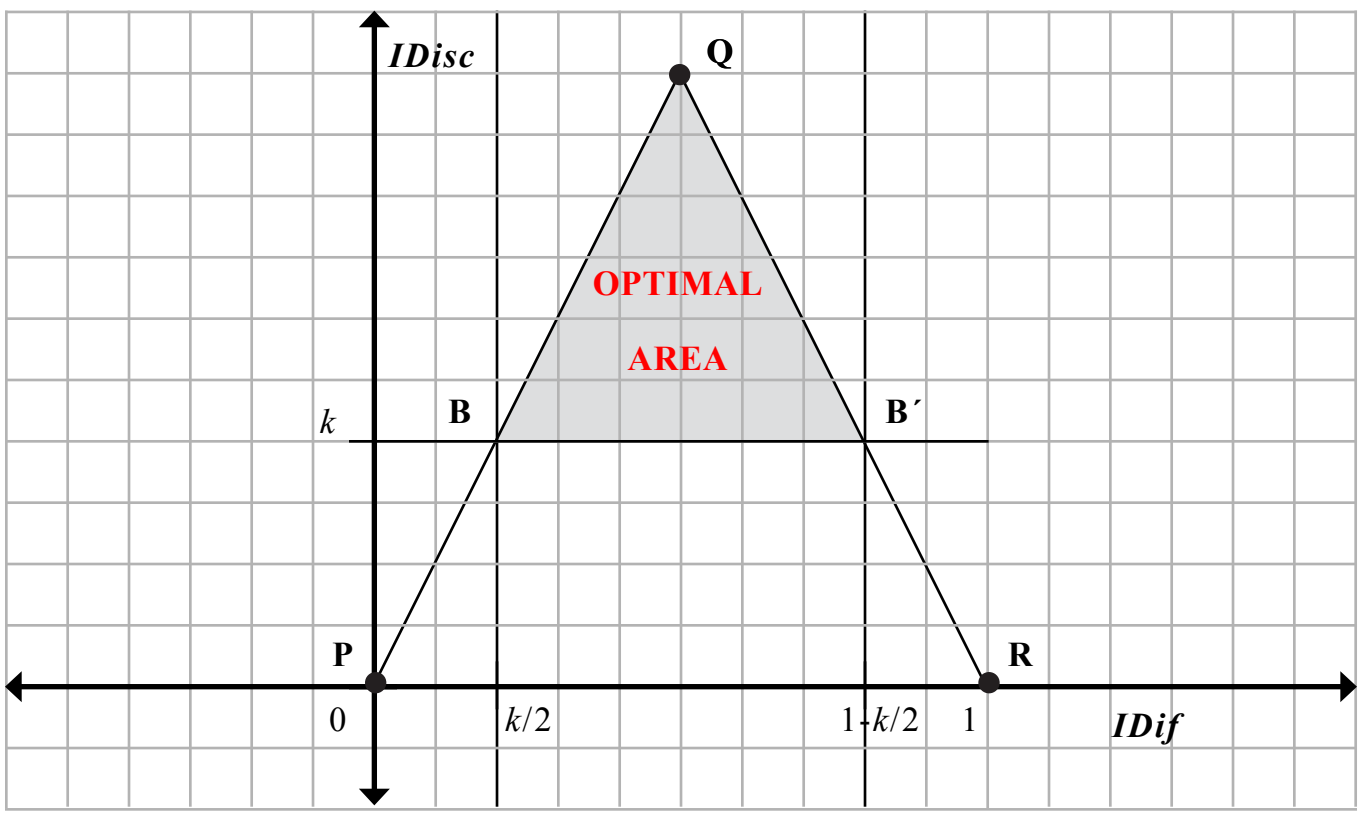

Figure 5. Optimal area for the following case: GS = GI = N/2 and discrimination rule k

area has turned out to be a triangle. When looking for an interval for the difficulty index in a range close to the value of $IDif = 0.5$, we must cover the area with points horizontally distributed and increasingly closer to 0.5. Similarly, when looking for a discrimination above a value of $IDisc = k$, we must cover the area with points vertically distributed approaching 1. The most plausible distribution for this double behavior is the triangular distribution. We can find points horizontally distributed that approach 0.5 and vertically distributed above k that approach 1 in the optimal area.

Indices for groups of “n” individuals (n<N/2)

In order to calculate the $IDif$, we counted the number of incorrect answers of test takers in total, which is separate from the way the upper and lower groups have been divided. On the contrary, the calculation of $IDisc$ depends on the way they are divided. The cut-off point setting for groups is not standardized among assessors. The first part of this paper has considered the score average as cut-off point for UG and LG. Occasionally, when we have large groups of test takers, we use quartiles or deciles. The discriminatory power of a question has proved to be more accurately determined if the groups are based on the upper or lower 27% instead of any other distribution percentage (Garret, 1966). Though this is the optimal percentage, Ebel (1977) states that “they are actually not much better than the 25 or 33% groups” (p. 476). One of

the reasons for not choosing the average is a better division of groups without being affected by the scores of test takers with average performance.

Considering that the UG and LG each consist of n individuals, the discrimination index would be calculated based on

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

Since each group will have n individuals, the difference in correct answers will be divided by n and not by half of the total of test takers $\frac{N}{2}$ as we did for the division according to the average. The calculation of $IDif$ will not be affected by this division since it depends on the total number of correct answers and not its sum per group. It is important to make this distinction since the opposite; that is, using the relation $1-\frac{C_s + C_i}{N}$ to calculate the $IDif$, would give us an interval of $\left[1-\frac{2n}{N}, 1\right]$ different than the theoretical one of $[0,1]$ for the difficulty index. We created table 13 based on the information above, which includes extreme cases. Ruling out the cases III and IV for being opposite to the sense of discrimination index, we have the following critical points:

$$A_1\ (0,0);\ B_1\left(\frac{n}{N}, 1\right);\ C_1\left(\frac{2n}{N}, 0\right);\ D_1\left(1\ \frac{2n}{N},\ 0\right);\ E_1\left(1\ \frac{n}{N}, 1\right);\ F_1\ (1,0)$$

that define a trapezoidal area (AAV) when graphed, as shown in figure 6.

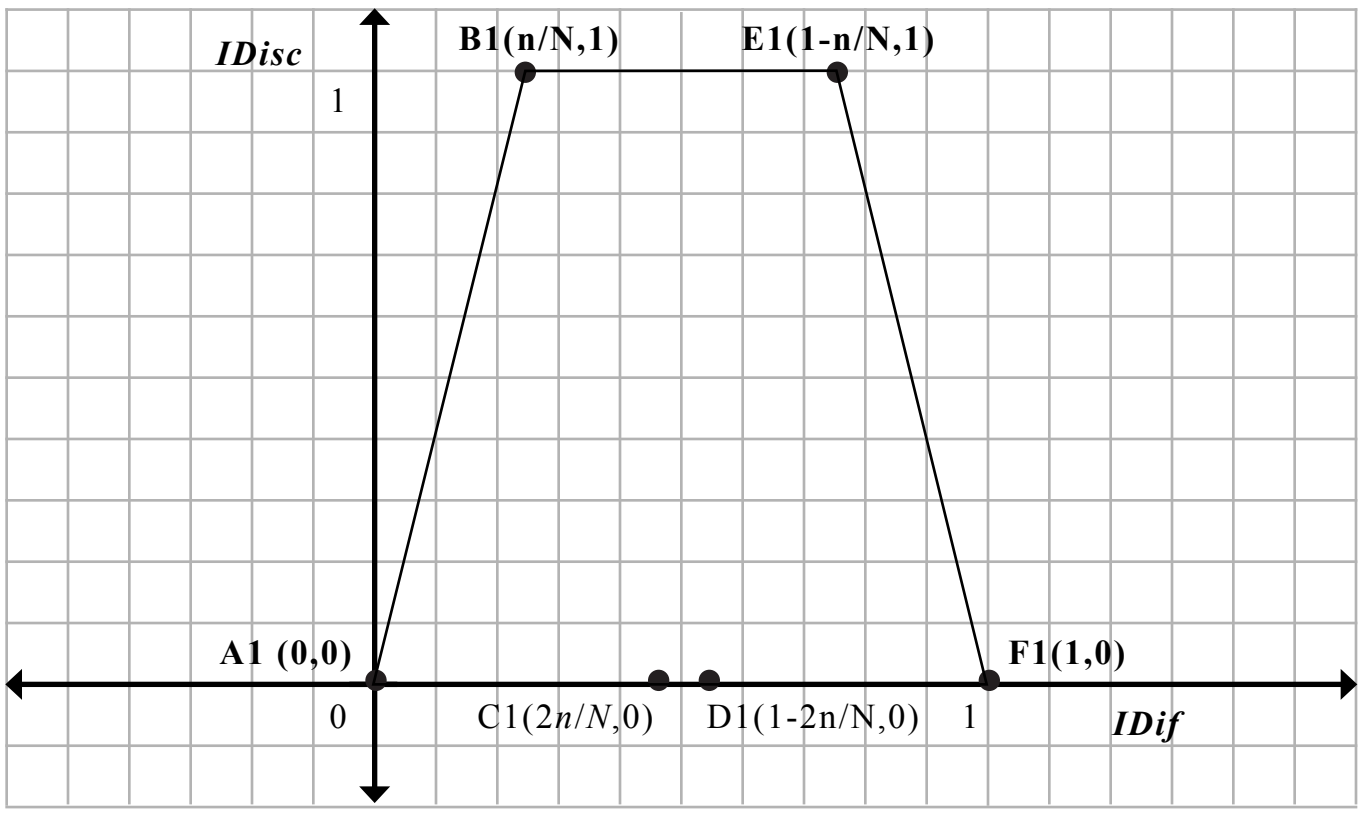

Figure 6. Area of admissible values for the case: GS = GI = n (n < N/2)

Table 13

	Case I	Case II	Case III	Case IV	Case V	Case VI	Case VII	Case VIII
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	2n/N	1
IDisc	1	1	-1	-1	0	0	0	0

Area of Admissible Values for Groups of ¨n¨ Individuals (n<N/2)

Following a similar methodology as for groups created according to the average, we can demonstrate that the AREA OF ADMISSIBLE VALUES (AAV) for groups with n individuals, with $n < \frac{N}{2}$, is determined by:

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

If the performance test has been applied to a big group of N test takers, it is recommended to divide it in three groups: UG, MG (middle group) and LG, where each end group consists of n individuals ($n < \frac{N}{2}$). In this case, we have an area defined by a trapezoid and not a triangle as for the groups with $\frac{N}{2}$ individuals each. Note that the points C_1 and D_1 are not vertices of the trapezoid, but only points included in the segment $A_1 F_1$. Figure 6 shows that the points $B_1\left(\frac{n}{N},1\right)$ and $E_1\left(1\ \frac{n}{N},1\right)$ are located at a distance of $d\,(B_1E_1) = 1 - \frac{2n}{N}$ away from one another.

Taking it to the limit, when n approaches 0 , the distance between both points would be 1.

$$lim\ d\,(B_1E_1) = \ lim\ \left(1 - \frac{2n}{N}\right) = 1$$

This limit (and impossible) case would occur when $n = 0$, turning the trapezoidal area into a rectangular area determined by $[0,1]$. $[0,1]$, which corresponds to the theoretical ranges of indices and would show no distinction between groups. The opposite case would occur when n approaches $\frac{N}{2}$, since now the distance between points would be 0.

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

According the latter, B_1 and E_1 would only meet at one point with coordinates $(0.5,1.0)$.This way, the trapezoidal area $A_1B_1E_1F_1$ would turn into the triangular area of the case previously analyzed; that is, the groups divided according to the average. We can also demonstrate that, in the limit case, the points C_1 and D_1 would meet in the middle point of the segment A_1F_1;

that is, in the point with coordinates $(0.5,0.0)$. The more n moves away below the value of $\frac{N}{2}$, the points B_1 and E_1 would be farther away and therefore the area of admissible values would be bigger.

Normed Area

Based on the AAV, we will outline a horizontal line corresponding to the discrimination normed value $IDisc = k$ and two vertical lines outlined by the edges of the interval for desirable difficulty $[D,1-D]$. Every question with points that belong to the AAV and are inside or above the horizontal line and between the vertical lines will belong to the normed area. Figure 7 shows this polygonal area $P_1Q_1B_1E_1Q_1'P'_1$. Since k is constant, these conditions can demonstrate that the area$S$ of the normed area is function of D and is determined by:

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

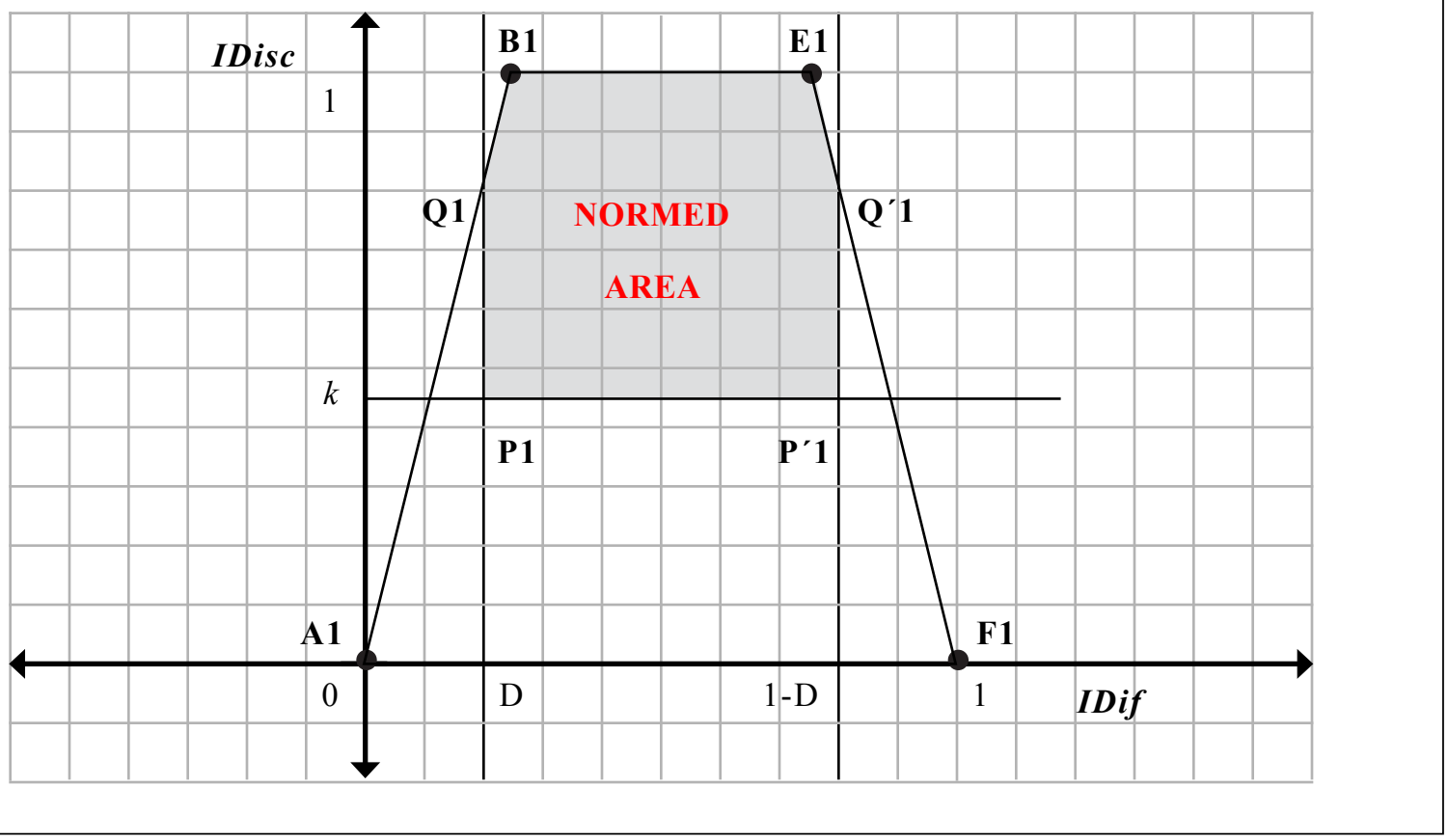

Figure 7. Normed area for the case: GS = GI = n (n < N/2)

The same that maximizes for a value of $D = \frac{n}{\mathrm{N}}\ k.$. The coordinates, both for the points P_1 and Q_1 and P_1' and Q_1' are equal for this value, resulting in a trapezoidal area.

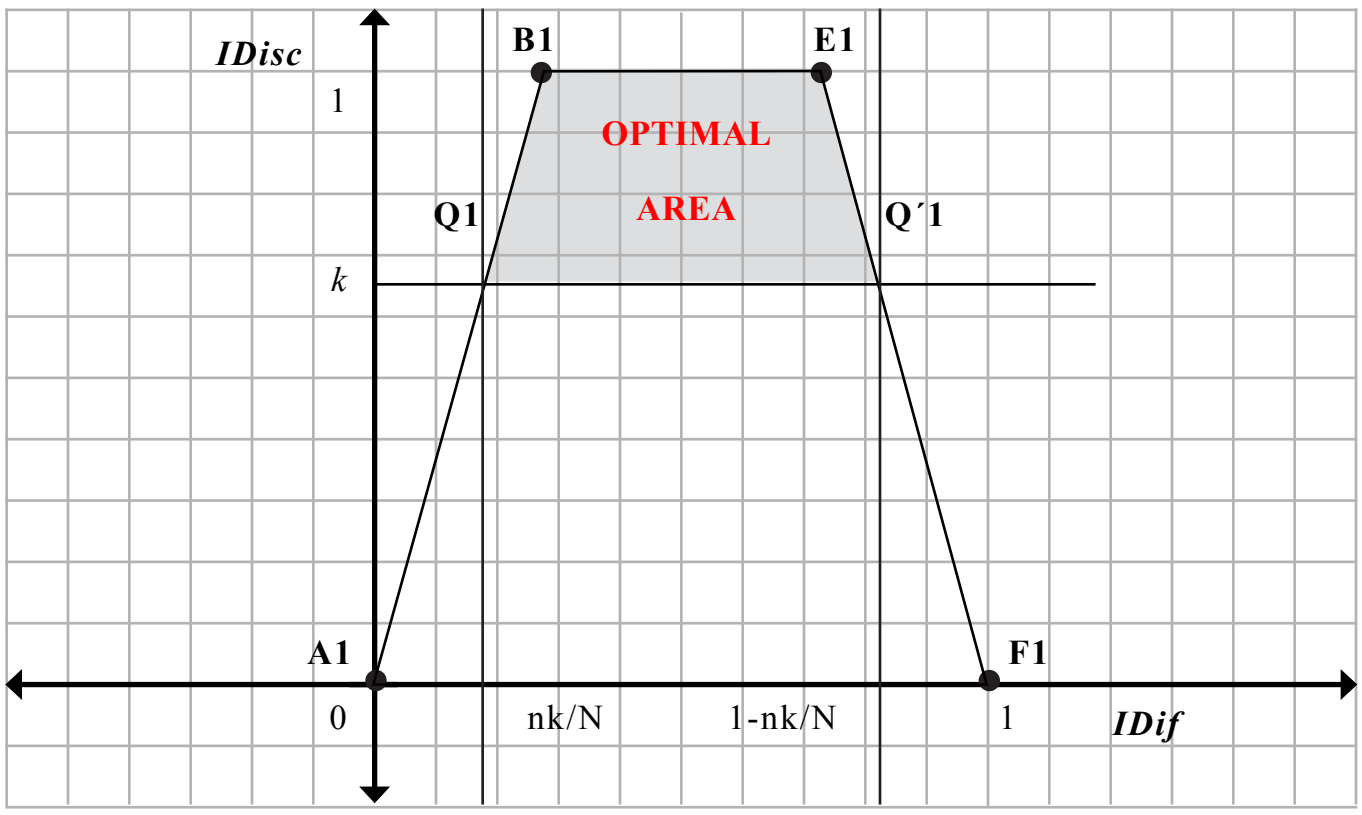

Figure 8. Optimal area for the case: GS = GI = n (n < N/2) and discrimination rule "k"

In sum, if the groups UG and LG consisting of n individuals each are considered in a performance test whose questions discriminate equal to or above a discrimination normed value k, the difficulty indices must belong to the interval $\left[\ \frac{n}{\mathrm{N}}\ k,\ 1 - \frac{n}{\mathrm{N}}\ k\ \right]$ so that the higher number of questions with coordinates of the form of $(IDif,\ IDisc)$ are inside the optimal area. As a result, a performance test will have higher quality if it has a higher number of questions within the optimal area. So, for example, if the UG and LG comprise 27% of test takers with higher and lower scores, respectively; we have that $n = 27\%\ N$, therefore $\frac{n}{\mathrm{N}} = 0.27$.

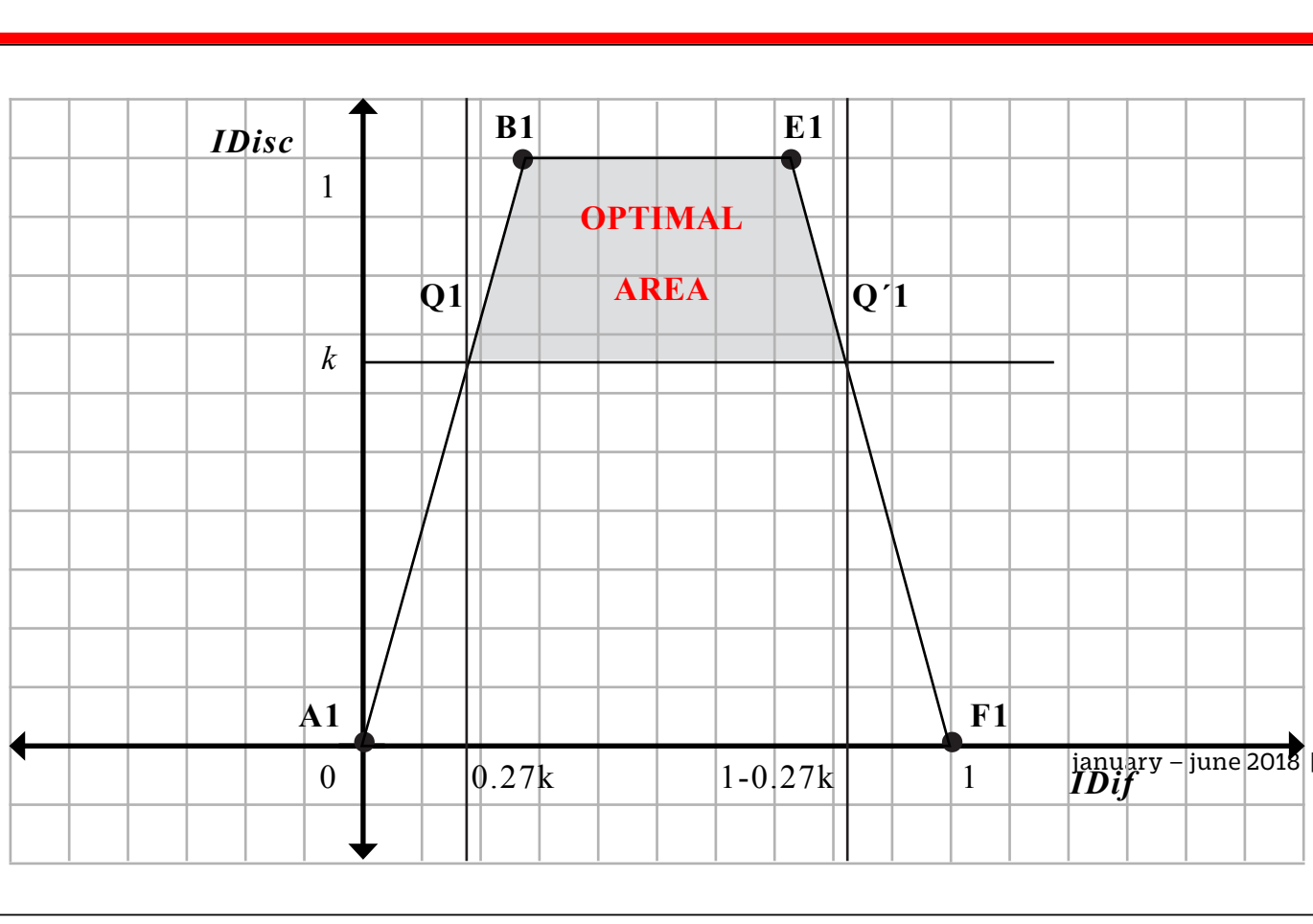

Figure 9. Optimal area for the case: GS = GI = 0.27N and discrimination rule "k"

For a discrimination rule k, we would have the following interval for $IDif$: $[0.27k, 1-\ 0.27k]$. Figure 9 shows the optimal area where the coordinates of the form $(IDif,\ IDisc)$ should be located for each question of a performance test where groups were set by the rule of 27%.

The areas shown in figures 5 and 9 are optimal areas where the indices of the form $(IDif,\ IDisc)$ should be found for each test question. The more points in this area, the greater the number of questions with an optimal behavior in the test, and therefore a higher quality.

FINAL COMMENTS

Based on this paper, and for a performance test, we can conclude as follows:

a. With the difficulty index $(IDif)$ and the discrimination index $(IDisc)$ of a question, we can create ordered pairs of the form $(IDif,\ IDisc)$ for each test question.

b. There is a zone in the bidimensional plane where the pairs $(IDif,\ IDisc)$ of each question are located. This zone is clearly defined and will be called area of admissible values (AAV).

c. The AAV has a mathematical formulation.

d. The $IDif$ and $IDisc$ for each question are not separate values. They are linked through a mathematical relation that stems from the mathematical formulation of AAV.

e. Given a discrimination normed value k and a difficulty normed value D, there will be a normed area within the AAV where the pairs $(IDif,\ IDisc)$ of questions that behave according to the rules would be placed.

f. There is a value D in terms of k that optimizes the normed area. This is referred to as optimal area or area of desirable behavior (ADB).

g. The ADB allows the fulfillment of two desirable conditions when designing a performance test: i) questions that discriminate equal to or above the standard, and ii) questions with difficulty distributed in a range close to intermediate difficulty.

h. The way the upper group (UG) and lower group (LG) are determined influences the mathematical formulation of the AAV, the mathematical relation between indices, and the value D that optimizes the normed area.

i. The ADB is a triangular area in the case of groups determined by the average.

j. The ADB is a trapezoidal area if the groups consist of n individuals of the N test takers, where $n < \frac{N}{2}$.

k. If groups are determined by the average, the optimizing value is $D < \frac{k}{2}$.

l. If both UG and LG groups consist of n individuals out of the N test takers, with $n < \frac{N}{2}$, then the optimizing value is $D = \frac{n}{N}\ k.$.

m. The more pairs of the form of $(IDif,\ IDisc)$ belong to the ADB, the better behavior of questions and the higher quality of performance test.
n. Determining the discrimination standard value affects the area of ADB and thus the number of questions in it.
o. Determining the discrimination standard value influences the quality interpretation of a performance test.

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