ÍNDICES DE DIFICULTAD Y DISCRIMINACIÓN
¿SON BUENOS INDICADORES?

Walter Alvarez Villar

Universidad Católica del Uruguay
Avda 8 de Octubre 2738 Montevideo Uruguay CP 11600
walvarez@ucu.edu.uy

EXTENSO

Antes de comenzar a desarrollar el tema vamos a recordar algunas definiciones:
El **índice de dificultad** de un ítem se define como la frecuencia relativa de respuestas incorrectas, es decir, como el cociente entre el número de respuestas incorrectas y el número total de respuestas. Por lo tanto, este índice es un número comprendido entre 0 y 1. Es una manera de medir el grado de dificultad: un índice cercano a 1 indica un ítem de gran dificultad, en tanto uno próximo a 0 señala uno fácil. Por otro lado, el índice de dificultad promedio de los ítems de la prueba sirve para medir la dificultad global de la misma.
El **índice de discriminación** tiene una definición un poco más complicada. En primer lugar se constituyen dos grupos, que se llaman de los sobresalientes y de los deficientes, constituidos respectivamente por quienes han obtenido una puntuación comprendida en el 27% superior o inferior de las registradas en la aplicación de la prueba. [1]El índice de discriminación se define como la diferencia entre la frecuencia relativa de respuestas correctas en el grupo de los sobresalientes y la frecuencia de respuestas correctas en el grupo de los deficientes. Por lo tanto, éste índice es un número comprendido entre –1 y 1. Un ítem cuyo índice de discriminación sea negativo no cumple con la finalidad de distinguir a los que han tenido un buen resultado en la prueba de los que no. Entre tanto, un ítem con un índice de discriminación mayor que 0,4 se considera altamente discriminador.
Se admite, además, que calidad de la prueba está relacionada directamente con la cantidad de ítems que discriminan adecuadamente. De la misma manera que para el índice de dificultad, el promedio de los índices de discriminación es un indicador del poder global que tiene la prueba para distinguir entre buenos y malos desempeños.
La **noción de confiabilidad** de una prueba hace referencia a dos factores: el primero es la condición de que los resultados efectivamente obtenidos en ella no difieran mayormente de los que obtendrían los mismos participantes en una prueba equivalente; el segundo es que estos resultados no dependan del azar.
Dado que es muy difícil establecer que dos pruebas son equivalentes, una forma de analizar la confiabilidad de una prueba es a través del procedimiento de dividirla en dos mitades al azar y comparar estadísticamente los resultados de los participantes en estas dos mitades como si fueran dos pruebas separadas. Este procedimiento da origen al llamado coeficiente de Spearman-Brown. La confiabilidad de la prueba resulta estar directamente relacionado con el valor de este coeficiente.
Un segundo indicador de la confiabilidad lo constituye la desviación estándar de los puntajes obtenidos.
Las teorías de medición brindan el marco teórico en el diseño e implementación de pruebas de múltiple opción. Estas teorías indican la metodología para la asignación de puntajes, y marcan las características de las preguntas o ítems, para que a partir de los resultados obtenidos podamos realizar otros análisis de interés. En este trabajo trataremos de hacer notar las carencias mas visibles de una de las principales teorías que se utilizan en el ámbito de medición educacional: la teoría clásica. En tren de vislumbrar soluciones citaremos la teoría de respuesta al ítem, que presenta algunas ventajas de

[1] La elección de este porcentaje del 27%, que parece arbitraria, tiene que ver con el uso estadístico de este índice como estimador de la probabilidad de que un ítem tenga un índice de dificultad intermedia.

aplicabilidad aunque también tiene sus limitaciones, dado el esfuerzo a realizar para la construcción de un banco de items adecuado.

Una forma de medición que se utiliza con mucha frecuencia en pruebas a gran escala, como las de admisión a la universidad (por ejemplo, en el SAT de Estados Unidos y en la antigua PAA y en la PSU chilenas), es la llamada teoría clásica (TC). En teoría clásica, el indicador de la *habilidad* de un estudiante corresponde al puntaje de éste en la prueba, construido a partir del número de respuestas correctas (o en algunos casos del número de respuestas correctas netas) que obtuvo. Como indicadores de la calidad de la prueba se utilizan los índices de dificultad y de discriminación, anteriormente definidos. Ambos descriptores de los ítems pueden calcularse ya sea en el contexto de una prueba piloto o experimental, o en la mayoría de los casos a posteriori de la prueba definitiva u operacional, pero en el caso de una prueba piloto la validez de los valores obtenidos dependerá de que el comportamiento de la muestra sea exactamente igual al de la población.

Como se puede apreciar, en teoría clásica el grado de habilidad de una persona depende del grupo de ítems (vale decir, de su nivel de dificultad y discriminación) que contiene la prueba. Por ejemplo, si la prueba es fácil, un mismo alumno tendrá un puntaje mayor que si la prueba es difícil.

Con esto resulta difícil o casi imposible hacer comparaciones entre estudiantes que han rendido pruebas diferentes. A su vez, los índices de dificultad y de discriminación de los ítems dependen del grupo de personas que rinden la prueba. Así, un mismo ítem puede ser catalogado como fácil si el grupo que rindió la prueba es excepcionalmente capacitado, pero como difícil si el grupo que rindió la prueba es menos aventajado. Con respecto a la discriminación, un ítem puede aparecer muy discriminatorio en el contexto de un grupo con nivel heterogéneo de habilidades, pero poco discriminatorio si el grupo que rindió la prueba es muy homogéneo (es decir, si todos los estudiantes tienen un nivel de rendimiento similar). Esta dependencia de la habilidad de un estudiante con respecto al grupo de ítems de la prueba, junto con la dependencia de los descriptores de los ítems con respecto a las características del grupo, se le llama *dependencia circular.*

Otra debilidad de la teoría clásica es que supone que la precisión con que se hace la medición es igual para todos los examinados, independientemente de su nivel de habilidad. Este supuesto es bastante discutible. Intuitivamente es claro que una prueba en que, por ejemplo, la mayor parte de sus preguntas son difíciles, va a distinguir más finamente entre dos personas con habilidad superior a la media que entre dos personas menos hábiles. Los que tienen habilidades inferiores obtendrán una estimación menos precisa de su habilidad, ya que son pocas o ninguna las preguntas de la prueba que responderán correctamente, y que, por lo tanto, servirían para distinguirlos. Esta debilidad, junto con la dependencia circular que se genera en el cálculo de la habilidad de los examinados y la dificultad y discriminación de los ítems, ha obligado a buscar métodos que permitan obtener una medida de la habilidad de los examinados independiente de los ítems propuestos. Lo que se debería tener es una caracterización de los ítems independiente de la población a la que se aplican, y al mismo tiempo una medida más fiel de la precisión con que se está midiendo la habilidad. Para respaldar estas afirmaciones, a continuación presentaremos los resultados obtenidos para los mismos ítems propuestos en una prueba de diagnóstico a los ingresantes a los cursos universitarios en diferentes carreras y facultades públicas y privadas de la ciudad de Montevideo.

Estos valores son altamente demostrativos respecto a los diferentes comportamientos observados frente al test.

Es importante hacer notar la relevancia que tiene para nuestro país esta investigación ya que no hay antecedentes de investigaciones conjuntas entre universidades privadas y públicas, ya que no existen intercambios académicos entre las diferentes universidades.

Esta prueba de diagnóstico se hace sólo en algunas facultades y es obligatoria, pero no tiene carácter de acreditación, ni consecuencias respecto a los cursos. En uruguay el ingreso a la universidad es irrestricto para todos los que culminen el bachillerato dcorrespondiente.

Lo que tenemos en la próxima página es un cuadro comparativo con los índices correspondientes a los mismos 10 ítems propuestos en distintas facultades de dos universudades del Uruguay, la UDELAR y la UCU.

ÍNDICE DE DISCRIMINACIÓN

Fac. Med. UDELAR MEDICINA	Fac. Cien. UDELAR BIOL_CIENCIAS	Fac. Cien. UDELAR FIS_MAT_CIENCIAS	FIT UCU INGENIERÍA	FCE UCU EMPRESARIALES
31,08%	39,47%	72,31%	75,00%	58,14%
53,78%	63,82%	56,92%	64,29%	51,16%
32,67%	51,32%	50,77%	82,14%	27,91%
41,43%	51,32%	64,62%	60,71%	30,23%
44,62%	57,89%	36,92%	28,57%	58,14%
57,77%	59,21%	46,15%	28,57%	69,77%
38,25%	15,13%	73,85%	75,00%	48,84%
51,79%	72,37%	36,92%	32,14%	44,19%
39,44%	50,00%	32,31%	85,71%	46,51%
12,35%	15,13%	73,85%	28,57%	9,30%

ÍNDICE DE DIFICULTAD

Fac. Med. UDELAR MEDICINA	Fac. Cien. UDELAR BIOL_CIENCIAS	Fac. Cien. UDELAR FIS_MAT_CIENCIAS	FIT UCU INGENIERÍA	FCE UCU EMPRESARIALES
81,05%	76,95%	47,11%	50,00%	66,88%
59,85%	60,99%	45,04%	33,33%	55,00%
84,82%	81,03%	53,72%	51,96%	78,13%
66,95%	70,04%	44,21%	46,08%	78,75%
45,53%	46,99%	17,77%	21,57%	41,25%
47,69%	50,71%	67,77%	29,41%	45,00%
76,32%	81,38%	49,59%	45,10%	71,25%
27,34%	38,12%	20,25%	13,73%	26,88%
72,98%	73,94%	16,94%	50,98%	80,00%
73,09%	81,38%	41,32%	71,57%	75,63%

A simple vista se nota la gran heterogeneidad de los valores de los indicadores según la facultad donde se aplicó el test.

Las pruebas estadísticas aplicadas aseguran que existen diferencias significativas entre los valores de los índices para un mismo ítem según la facultad donde éste fue propuesto. Además se comprobó que habría homogeneidad entre los valores de los índices entre la orientación Medicina y Ciencias Biológicas y también entre Ingeniería y Físico Matemáticas.

Vamos ahora a ver comenzando en la próxima página algunos de los diez ítems propuestos y observar su comportamiento en los diferentes lugares, a los efectos de tener un panorama más claro de la problemática del uso de la teoría clásica, y además podremos sacra interesantes conclusiones sobre equivalencias y diferencias entre los alumnos de diferentes orientaciones.

1) La expresión algebraica: $Y = (2X - 1)^2$ representa la siguiente relación entre los números naturales Y y X

A) Y es el doble del cuadrado del anterior de X
B) Y es el doble del anterior de X al cuadrado
C) Y es el cuadrado del doble del anterior de X
D) Y es el cuadrado del anterior del doble de X

		Medicina	Biología	Fis-Mat	FIT UCU	FCE UCU
ÍNDICE DE DISCRIMINACIÓN		53,78%	63,82%	56,92%	64,29%	51,16%
ÍNDICE DE DIFICULTAD		59,85%	60,99%	45,04%	33,33%	55,00%
PUNTAJE PROMEDIO		39,83%	39,01%	54,96%	66,67%	45,00%
DESVIO DE LOS PUNTAJES		59,85%	48,82%	49,86%	47,37%	49,91%

2) En una viaje en taxi , cae una ficha cada 0.2 km. La "bajada de bandera" cuesta $18 y cada ficha cuesta $2 ¿Cuál de las siguientes funciones representa el costo del viaje en taxi?(x es la distancia recorrida en km.)
A) f(x)=2+0,2x
B) f(x)=18x+2
C) f(x)=2x+18
D) f(x)=10x+18

		Medicina	Biología	Fis-Mat	FIT UCU	FCE UCU
ÍNDICE DE DISCRIMINACIÓN		53,78%	63,82%	56,92%	64,29%	51,16%
ÍNDICE DE DIFICULTAD		59,85%	60,99%	45,04%	33,33%	55,00%
PUNTAJE PROMEDIO		39,83%	39,01%	54,96%	66,67%	45,00%
DESVIO DE LOS PUNTAJES		59,85%	48,82%	49,86%	47,37%	49,91%

3) La fórmula $R = \frac{8l}{r^4}$, permite el cálculo de la resistencia R de un fluido en un tubo en función de la longitud del tubo L y del radio r. Si dados 2 tubos de igual longitud, el segundo tiene radio igual a la mitad del radio del primero, la resistencia del segundo
A) Se duplica.
B) Se divide por 2.
C) Se multiplica por 16.
D) Se divide entre 16.

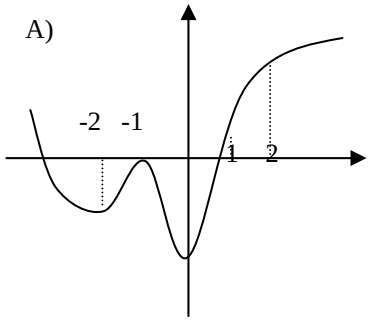
		Medicina	Biología	Fis-Mat	FIT UCU	FCE UCU
ÍNDICE DE DISCRIMINACIÓN		41,43%	51,32%	64,62%	60,71%	30,23%
ÍNDICE DE DIFICULTAD		66,95%	70,04%	44,21%	46,08%	78,75%
PUNTAJE PROMEDIO		32,51%	29,96%	55,79%	53,92%	21,25%
DESVIO DE LOS PUNTAJES		66,95%	45,85%	49,77%	50,09%	41,04%

4) Considere los siguientes conjuntos:
E = {x / x es estudiante de Economía}
I = {x / x es estudiante de Ingeniería}
U = {x / x estudiante universitario}.
Entonces, la afirmación "Cualquier estudiante de Economía o de Ingeniería es estudiante universitario" se expresa como:
A) $E \cap I \subset U$
B) $U \subset E \cup I$
C) $(E \cup I) \cup U$
D) $E \cup I \subset U$

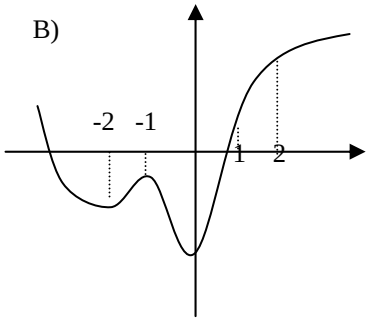
	Medicina	Biología	Fis-Mat	FIT UCU	FCE UCU
ÍNDICE DE DISCRIMINACIÓN		44,62%	57,89%	36,92%	28,57%	58,14%
ÍNDICE DE DIFICULTAD		45,53%	46,99%	17,77%	21,57%	41,25%
PUNTAJE PROMEDIO		53,93%	53,01%	82,23%	78,43%	58,75%
DESVIO DE LOS PUNTAJES		45,53%	49,95%	38,30%	41,33%	49,38%

5) Se sabe que los valores de una función en –2,-1,0,1 y 2 están en la siguiente relación: f(0)<f(-2)<f(-1)=0<f(1)<f(2). Indique cuál de las siguientes gráficas es compatible con los datos.

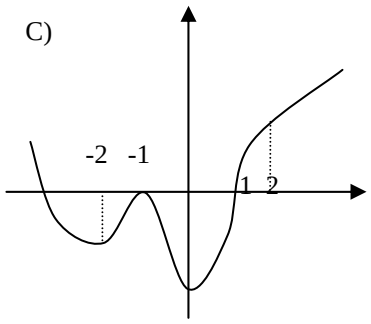
A)



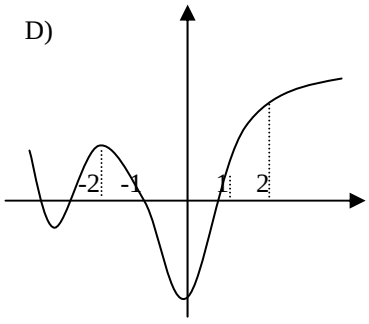
B)



C)



D)



	Medicina	Biología	Fis-Mat	FIT UCU	FCE UCU
ÍNDICE DE DISCRIMINACIÓN		57,77%	59,21%	36,92%	28,57%	69,77%
ÍNDICE DE DIFICULTAD		47,69%	50,71%	20,25%	29,41%	45,00%
PUNTAJE PROMEDIO		51,67%	49,29%	79,75%	70,59%	55,00%
DESVIO DE LOS PUNTAJES		47,69%	50,04%	40,27%	45,79%	49,91%