Matas Terrón, A. (2003). Estudio diferencial de indicadores de rendimiento en pruebas objetivas. *RELIEVE, v. 9*, n. 2, p. 184-197. http://www.uv.es/RELIEVE/v9n2/RELIEVEv9n2_5.htm

RELIEVE

Revista ELectrónica de Investigación y EValuación Educativa

ESTUDIO DIFERENCIAL DE INDICADORES DE RENDIMIENTO EN PRUEBAS OBJETIVAS

(Differential analysis of indicators of achievement in objective tests)

por

Article record
About authors
HTML format

Antonio Matas Terrón (amatas@us.es)

Ficha del artículo
Sobre los autores
Formato HTML

Abstract

Often objective tests to evaluate student qualifications are used. However, the selection between different evaluation criteria is in the hands of teachers or evaluators preferences. This investigation offers, briefly, a revision of some strategies utilized for objective evaluation. We also present other alternative strategies to the most classical ones. Afterwards, the strategies of evaluation have been analized with data from 114 subjects, who were evaluated with a test. The results suggest there are two kinds of strategies, one where only students' performance are evaluated, and other, where the test and the students are evaluated at the same time. We also show the fail of concordance between the indicators used

Keywords

Achievement assessment, objective tests, bayesian estimation, item analysis.

Resumen

La utilización de pruebas objetivas para la calificación del alumnado es muy habitual. Sin embargo, la elección de un criterio de valoración u otro está, a menudo, en manos de las preferencias del profesor o evaluador. En esta investigación se ofrece brevemente una revisión de algunas de las estrategias utilizadas para ello, presentando alternativas a las más clásicas. Posteriormente, con los datos procedentes de una prueba real administrada a 114 sujetos, se han analizado los resultados generados a partir de las distintas aproximaciones valorativas. Se concluye que existen dos estrategias básicas para afrontar el problema, una de ellas basadas en la valoración exclusivamente de la ejecución de los sujetos ante los ítems, y otra que al mismo tiempo permite una evaluación de la prueba. Se muestra también la falta de concordancia entre los indicadores utilizados.

Descriptores

Evaluación del rendimiento, pruebas objetivas, estimación bayesiana, análisis de items

Introducción

Este trabajo se centra principalmente en indicadores aplicados a la evaluación del rendimiento académico universitario. No obstante, antes de entrar en el mismo, se pretende ofrecer un breve acercamiento al concepto de "rendimiento académico" para establecer un convenio con el lector en el punto de partida. Intentar ofrecer una delimitación de "rendimiento académico" requiere un esfuerzo destacable en virtud de la compleji-

dad y multidimensionalidad de este concepto (Pérez Serrano, 1981).

Generalmente, las investigaciones realizadas se han centrado en analizar las variables que determinan el rendimiento de los escolares. Autores como Kazcynska (1965), o Muñoz Arroyo (1977), relacionaban el rendimiento con la voluntad o capacidad del sujeto hacia la tarea académica. Otros, lo han vinculado al esfuerzo personal (Secadas, 1952), y a la experimentación continuada (Bloom, 1972). En general, los investigadores han entendido el rendimiento como un producto de múltiples factores, entre los que se encuentran la familia, el sistema educativo, el sujeto, su esfuerzo, su motivación, autoestima, etc. (Plata Gutiérrez, 1969; González Fernández, 1975; Forteza, 1975; García y Musitu, 1993; García y Doménech, 1997). En consonancia con estos últimos autores, puede considerarse que el rendimiento académico es el resultado que el alumnado obtiene en su proceso de enseñanza y aprendizaje.

Otro aspecto importante, es tener en considerar la medición de este rendimiento. Centrarse solamente en los procesos de medición pueden hacer derivar en un concepto simplemente operativo del rendimiento, entendiéndolo exclusivamente como el resultado de las mediciones sociales y académicas de interés (Carabaña, 1979). Si bien este trabajo trata sobre la forma de medir el rendimiento académico, esto no agota este ámbito de estudio. La multidimensionalidad intrínseca al rendimiento exige recurrir a una diversidad de variables, objetivos, e instrumentos, generando estrategias distintas de análisis y medición. Normalmente, se han utilizado dos tipos de estrategias de valoración del rendimiento, por un lado las calificaciones escolares (obtenidas a su vez por multitud de medios) y las pruebas objetivas (Álvaro Page, Bueno, Calleja, Cerdán, Echevarría, García, et al., 1990; Jornet y Suárez, 1996). Estas últimas, pueden estar o no estandarizadas.

Estas pruebas objetivas presentan una serie de ventajas destacables, entre las cuales se encuentran el ofrecer resultados que permiten la comparación entre grupos, centros, etc., o la obtención de medidas que suelen ser más fiables y válidas que en otros instrumentos. Igualmente presentan inconvenientes como por ejemplo, valorar una gama limitada de actividades intelectuales.

En definitiva, la utilización de pruebas objetivas para la valoración del rendimiento tiene una indiscutible presencia. Su versatilidad, rapidez de ejecución, y el supuesto grado de objetividad que confieren a la evaluación de conocimientos los hacen instrumentos muy demandados tanto por el profesorado como por parte del alumnado. No obstante, la utilidad de las pruebas objetivas no sólo depende de su formato, sino que está asociado al tratamiento que se realiza sobre la información recopilada.

El elemento central de la prueba objetiva lo constituye el ítem. Un enunciado que actúa como un supuesto al que el sujeto ha de otorgar un juicio (uno sólo) siguiendo ciertas instrucciones formales (Álvaro Page, 1993). Generalmente, la calificación de un test o examen de este tipo, se realiza utilizando los índices de dificultad aplicados sobre los sujetos y no sobre los ítems. Las expresiones pueden ser varias en función de los intereses del evaluador:

- Proporción de aciertos, excluyendo los ítems no alcanzados, cuando por ejemplo, no ha dado tiempo a contestar:

$$C = \frac{A}{N-f} \cdot 10 \quad \text{(Exp. 1)}$$

Donde A es el número de aciertos, N el número de preguntas del test, y f el número de preguntas no alcanzadas.

- Proporción de aciertos excluyendo los ítems omitidos:

$$C = \frac{A}{N-O} \cdot 10 \quad \text{(Exp. 2)}$$

Siendo O el número de ítems omitidos

- Proporción de aciertos penalizando los errores:

$$C = \frac{A - \frac{E}{K-1}}{N} \cdot 10 \quad \text{(Exp. 3)}$$

Donde E es el número de ítems erróneos y K, las alternativas de respuesta de los ítems.

- Proporción de aciertos excluyendo a los ítems no alcanzados y omitidos:

$$C = \frac{A}{N - f - O} \cdot 10 \quad \text{(Exp. 4)}$$

- Proporción de aciertos penalizando los errores y excluyendo los ítems no alcanzados:

$$C = \frac{A - \frac{E}{K-1}}{N - O} \cdot 10 \quad \text{(Exp. 5)}$$

- Proporción de aciertos penalizando los errores, excluyendo a los ítems no alcanzados y omitidos:

$$C = \frac{A - \frac{E}{K-1}}{N - f - O} \cdot 10 \quad \text{(Exp. 6)}$$

Como se ha apuntado anteriormente, estos índices son una adaptación de los indicadores de dificultad. Cualquier adaptación tiene sus problemas, y en este caso no sería menos. Bajo una perspectiva estadística, cualquier índice de dificultad debería estar cercano a una proporción de 0,5. Esta posición maximiza la varianza del ítem y por tanto del test. Una variabilidad menor sugiere homogeneidad entre los sujetos, lo que impide su discriminación. Por el contrario, cuando se utilizan estos índices para calificar a los alumnos, se espera una mínima variabilidad de los sujetos en sus respuestas a los ítems. La homogeneidad sugiere que el alumno se encuentra en un nivel de conocimiento específico, mientras que una alta variabilidad indicaría la existencia de problemas en el diseño del test, las condiciones de aplicación, o en la forma de contestar del alumnado.

En cualquier caso, la utilización de estos indicadores sólo permiten analizar o bien el comportamiento de los ítems respecto al grupo de sujetos, o bien los sujetos sobre la prueba. Para superar estas limitaciones se ha propuesto una alternativa basada en la perspectiva bayesiana (Serrano, 1990; Serrano, 1997; Serrano y Matas, 1999).

Cuando un sujeto afronta la tarea de contestar a un ítem, pueden suceder varias cosas, entre ellas, interesan dos especialmente, que el sujeto conozca la respuesta correcta o bien que no la conozca. De forma similar, cuando se elige una opción, pueden darse dos condiciones básicamente: el sujeto sabe la respuesta y contesta correctamente, o bien el sujeto no sabe la respuesta pero también contesta correctamente. Existe otra posibilidad, donde el sujeto conoce la respuesta aunque responde de forma incorrecta. Esta situación está asociada más a aberraciones del estado psicológico del sujeto, y en principio no se incluye en el supuesto inicial.

Por lo tanto, del conjunto de sujetos que contesta a un ítem, se puede suponer la existencia de un subgrupo que conoce la respuesta y otro que no. De forma similar, existe un grupo de sujetos que contesta correctamente, parte de ellos porque conocen la respuesta y otros porque no la conocen. Inicialmente se asume que quien contesta erróneamente es porque no conoce la respuesta.

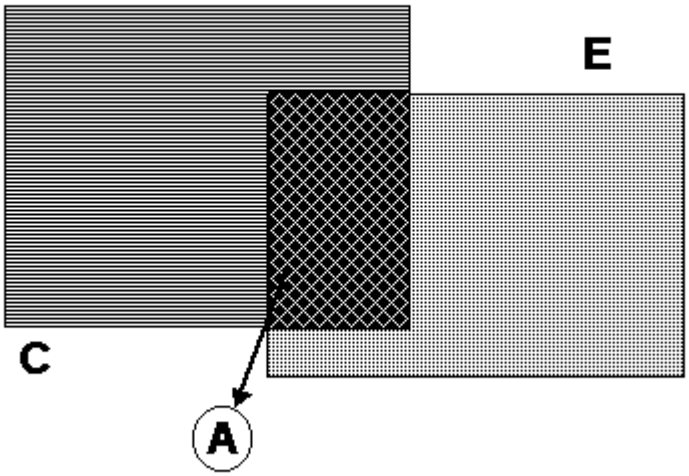

Gráfico 1. Representación de subgrupos

En el gráfico 1 el grupo C está formado por los alumnos que contestan correctamente. El grupo E lo forman los sujetos que no conocen la respuesta. El grupo A es la intersección entre ambos, el grupo de sujetos que no conocen la respuesta pero contestan correctamente. El resultado de extraer del grupo C los correspondientes al grupo A, genera el grupo S de sujetos que contestan correctamente porque conocen la respuesta. A partir de esta situación se podría preguntar cuál es la probabilidad de que un sujeto que haya contestado correctamente al ítem (C), pertenezca al grupo de sujetos que conocen la respuesta (S).

Esta pregunta puede afrontarse desde el teorema de Bayes, de forma que la probabilidad de que un sujeto conteste correctamente porque conoce la respuesta ($P_{(c|s)}$) será:

$$P_{(c|s)} = \frac{P_{(s|c)} \cdot P_{(c)}}{P_{(s)}} \quad \text{(Exp. 7)}$$

Donde $P_{(S|C)}$ es la probabilidad de saber la respuesta y contestar correctamente, $P_{(S)}$ hace referencia a la probabilidad de que el sujeto conozca la respuesta, y $P_{(C)}$ a que conteste correctamente. Desarrollando la expresión 7 para todos los ítems de un test se llega a la siguiente expresión (Serrano, 1997) para calcular la calificación de un sujeto en una prueba:

$$C = \frac{N - \sum_{i=1}^{n} \left(\frac{E_i}{(k-1)A_i} \right) \cdot 10}{I} \quad \text{(Exp. 8)}$$

Así como el siguiente índice de dificultad:

$$ID = 1 - \frac{E}{(k-1)A} \quad \text{(Exp. 9)}$$

Donde A es el número de sujetos que contestan correctamente a un ítem. E es el número de sujetos que contestan erróneamente. N el número total de ítems que el sujeto contesta correctamente y k el número de alternativas de los correspondientes ítems. I es la suma total de todos los índices de dificultad de los ítems que componen el test

La calificación de cada sujeto (Exp., 8) es resultado de sumar los correspondientes índices de dificultad (Exp., 9) de los ítems que ha contestado correctamente.

Siguiendo con esta misma "lógica", puede asumirse que la calificación del alumnado sería estimable a partir de los índices de dificultad, sumando los valores de éstos (independientemente del índice utilizado) correspondientes a los ítems contestados correctamente. Basándose en este procedimiento, en los próximos epígrafes se proponen nuevos indicadores junto con el desarrollado por Serrano (1990, 1997).

En esta investigación se parte de una muestra de sujetos que han atendido una prueba objetiva real. El objetivo principal consiste en analizar cómo influye el procedimiento de valoración en el resultado que obtienen los sujetos (calificaciones) y en el propio análisis de los elementos del test (índices de dificultad). Sobre los datos recopilados por tanto, se estudiaron los índices de dificultad generados a partir de las distintas expresiones anteriores, así como las calificaciones que los sujetos obtendrían en función de la aproximación evaluativa elegida. Se analizaron posteriormente, los resultados derivados de cada expresión extrayendo conclusiones sobre la influencia de los procesos de valoración en la medición del rendimiento del alumnado.

Otro objetivo secundario, que también se pretendía conseguir, consistía en ofrecer un acercamiento a la incorporación de la perspectiva bayesiana en los procesos de evaluación de pruebas objetivas.

Método

Instrumento y Muestra

El estudio ha sido realizado sobre un grupo de 114 alumnas y alumnos de segundo curso de la titulación de Pedagogía. De edades comprendidas entre los 19 y 23 años, con un 93% de genero femenino, y un 7% de hombres.

La prueba administrada consistía en un test de 30 ítems relativo al contenido de la asignatura "Análisis de datos para la investigación educativa". Todos los ítems tenían tres opciones de respuesta, con una sola correcta. En el diseño de esta prueba se evitó la presencia de alternativas negativas del tipo, por ejemplo "La desviación típica no es un estadístico de posición", opciones de multiplicidad encubierta como por ejemplo, "todas las opciones anteriores son correctas", e ítems de juicio creciente similares a "elige la opción más correcta".

Procedimiento y análisis

- Índices de dificultad

Inicialmente se calcularon los índices de dificultad de los ítems considerando los errores exclusivamente, los errores junto con las omisiones, además de aplicar la perspectiva bayesiana de Serrano. Los índices de dificultad de los ítems (tabla 1) se han calculado según las siguientes expresiones (Álvaro Pages, 1993):

$$ID_1 = \frac{A}{N} \quad \text{(Exp. 10)}$$

$$ID_2 = 1 - \frac{E}{(k-1)A} \quad \text{(Exp. 9)}$$

$$ID_3 = \frac{A - \frac{E}{K-1}}{N-O} \quad \text{(Exp. 11)}$$

$$ID_4 = \frac{A - \frac{E}{K-1}}{N} \quad \text{(Exp. 12)}$$

Donde A es el número de ítems correctos, N el número total de ítems, K el número de opciones de respuesta, y E el número de errores. A los datos de la tabla 1 se aplicó un ANOVA de un factor intragrupo con cuatro niveles, correspondientes con cada uno de los cuatro índices. La prueba indicó que no hay diferencias entre los índices (F-ratio= 1,052; p=0,374).

Ítem	ID_1	ID_2	ID_3	ID_4	Varianza	Ítem	ID_1	ID_2	ID_3	ID_4	Varianza
1	0,96	0,98	0,95	0,95	0,03	**16**	0,97	0,99	0,96	0,94	0,03
2	0,94	0,97	0,91	0,9	0,06	**17**	0,68	0,76	0,52	0,46	0,22
3	0,85	0,91	0,77	0,76	0,13	**18**	0,48	0,46	0,22	0,18	0,25
4	0,82	0,89	0,74	0,7	0,15	**19**	1	1	1	0,97	0
5	0,67	0,75	0,5	0,43	0,22	**20**	0,72	0,8	0,58	0,5	0,2
6	1	1	1	1	0	**21**	0,88	0,93	0,82	0,79	0,11
7	1	1	1	0,92	0	**22**	0,69	0,78	0,54	0,51	0,21
8	0,7	0,78	0,54	0,49	0,21	**23**	0,58	0,63	0,36	0,34	0,25
9	0,71	0,79	0,56	0,54	0,21	**24**	0,87	0,92	0,8	0,79	0,12
10	0,84	0,91	0,76	0,68	0,13	**25**	0,8	0,88	0,71	0,6	0,16
11	0,76	0,84	0,63	0,48	0,19	**26**	0,88	0,93	0,82	0,64	0,11
12	0,56	0,61	0,35	0,29	0,25	**27**	0,83	0,9	0,74	0,68	0,14
13	0,09	-4,31	-0,37	-0,3	0,08	**28**	0,19	-1,2	-0,22	-0,16	0,15
14	0,79	0,86	0,68	0,56	0,17	**29**	0,73	0,82	0,6	0,5	0,2
15	0,74	0,83	0,61	0,6	0,19	**30**	0,65	0,73	0,47	0,42	0,23

Tabla 1. Distintos índices de dificultad

Análisis de las calificaciones

a) Análisis de las puntuaciones directas:

Posteriormente se calculó la calificación de cada sujeto en función de las distintas expresiones, según las distintas estrategias resultantes de las expresiones 3, 5 y 8. Además se han añadido dos nuevos indicadores. Para ello se otorgó a cada sujeto el resultado de sumar los valores de dificultad de cada uno de los ítems que contestaron correctamente. Se utilizaron los índices de dificultad con penalización de errores (exp. 12), y el correspondiente a la penalización de errores al mismo tiempo que se tienen en cuenta las omisiones (exp. 11). La expresión resultante es la siguiente:

$$C_j = \frac{\sum_{i=1}^{n} ID \cdot R_i}{I} \cdot 10 \qquad \text{(Exp. 13)}$$

Donde R_i es el valor de la respuesta dada al ítem *i* por el sujeto *j* mientras que ID la expresión del índice de dificultad elegido.

En función de esta expresión se han calculado las calificaciones que obtendrían los sujetos al utilizar los índices de dificultad 11 y 12 (señalados en las tablas como 13_11 y 13_12 respectivamente), junto con las obtenidas por la expresión 3 (calificación penalizando los errores), 5 (penalizando los errores y teniendo en cuenta las omisiones) y 8 (perspectiva bayesiana). Los estadísticos descriptivos básicos para cada una de estas nuevas variables se presentan en la tabla 2.

Calificación	Media	Mediana	Varianza	Min	Max
Exp. 3	5.715	5.83	2.80171	1.83	9.5
Exp. 5	6.36825	6.63	3.20497	2.2	9.5
Exp. 8	7.44272	7.555	1.4801	4.53	9.67
Exp. 13_11	7.95377	8.1	1.29273	4.97	9.98
Exp. 13_12	7.35307	7.49	1.1062	4.6	9.23

Tabla 2. Descriptivos básicos de las calificaciones

El anova de un factor ha mostrado la existencia de diferencias significativas entre las distintas formas de valorar el rendimiento analizadas (tabla 3).

Fuentes	**g.l.**	**Sum. Cuadrados**	**Med. Cuadraticas**	**F**	**Prob**
Const.	1	27663.8	27663.8	13992	≤ 0,0001
Grp.	4	373.36	93.34	47.21	≤ 0,0001
Error	565	1117.09	1.97714		
Total	569	1490.45			

Tabla 3. Anova de un factor *sobre técnicas de calificación*

b) Análisis de las puntuaciones agrupadas:

Por último se analizó cómo afecta al alumnado la utilización de una u otra estrategia. Para ello se categorizaron los sujetos de forma similar a como se hace en una evaluación en las actas oficiales de calificación, es decir, suspenso (de 0 a 4,99), aprobado (de 5 a 6,99), notable (de 7 a 8,99) y sobresaliente (de 9 a 10). Se trataba de comprobar hasta qué punto, un alumno puede pasar de estar en una u otra categoría (suspenso, aprobado, notable, sobresaliente) en función de la forma de calcular su calificación. Para ello se organizaron los datos en una matriz de confusión aplicando el coeficiente Kappa de concordancia. Los cambios quedan reflejados en la siguiente matriz de confusión (tabla 4) donde se resumen los cambios que puede experimentar un alumno/a entre categorías en función de la estrategia elegida.

		Exp 3				Exp 5				Exp 8				Exp 13_11				Exp 13_12			
		1	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**	**1**	**2**	**3**	**4**
	1	31				21	9	1		5	25	1		3	25	3		1	20	10	
Exp	**2**		52				30	22			6	46			7	45			2	50	
3	**3**			30				29	1			22	8			27	3			8	22
	4				1				1				1			1					1
	1					21				5	15	1		3	17	1		1	16	4	
Exp	**2**						39				10	29			11	28			3	36	
5	**3**							52			6	38	8		4	45	3		3	27	22
	4								2			1	1			2				1	1
	1									5				3	2			1	4		
Exp	**2**										31				27	4			18	13	
8	**3**											69			3	66				55	14
	4												9			6	3				9
	1													3				1	2		
Exp	**2**														32				20	12	
13_11	**3**															76				56	20
	4																3				3
	1																	1			
Exp	**2**																		22		
13_12	**3**																			68	
	4																				23

Tabla 4. Matriz de confusión

Los niveles de concordancia entre registros son bastante bajos. Tan sólo la expresión 9 mantiene un acuerdo significativamente alto con la expresión 13_11 (Kappa de 0,7455). El resto de índices ofrecen valores inferiores, así entre la expresión 3 y 5 la concordancia tiene un kappa de 0,5703; el valor Kappa para la expresión 3 y 9 baja hasta un 0,0033; etc., (tabla 5).

	Kappa	P(o)	Se		Kappa	P(o)	Se
Exp 3 X 8	0,0033	0,2983	0,0517	**Exp 8 X 13_11**	0,7455	0,8684	0,0726
Exp 3 X 5	0,5703	0,7105	0,0620	**Exp 8 X 13_12**	0,5231	0,7281	0,0640
Exp 3 X 3_11	0,0199	0,3246	0,0523	**Exp 5 X 13_11**	0,1886	0,5175	0,0645
Exp 3 X 3_12	-0,2069	0,0965	0,0430	**Exp 5 X 13_12**	-0,0953	0,2807	0,0530
Exp 8 X 5	0,1461	0,4696	0,0621	**Exp 13_11 X 13_12**	0,4504	0,7018	0,0640

Tabla 5. Concordancia entre indicadores (P(o): probabilidad observada; Se: error típico)

Aunque estos análisis han mostrado diferencias, la importancia de utilizar una u otra expresión radica en las consecuencias que tiene sobre el alumnado. En este sentido, se ha comprobado la falta de concordancia entre los índices de calificación (tabla 5). Así, 25 de los alumnos aprobados según la expresión 8, pasarían a estar suspensos si se utiliza la expresión 3. De 31 aprobados con la expresión 8, 15 de ellos estarían suspensos con la expresión 5 y 4 estarían en la categoría de notables con la expresión 13_11, etc. Por tanto, utilizar una u otra expresión afecta sustancialmente al alumnado.

Por último, se observa que las expresiones de calificación basadas en índices de dificultad (expresiones 8, 13_11 y 13_12), tienden a resultados homogéneos, disminuyendo el número de puntuaciones extremas y agrupándolas alrededor de las posiciones centrales. Esto es resultado del acercamiento normativo que generan dichas expresiones, frente a los basados en la penalización de los errores del sujeto (expresión 3 y expresión 5).

Conclusiones

Antes de comenzar con las conclusiones propiamente, se quiere hacer hincapié en las vías de análisis mostrados aquí. Es posible afirmar que las expresiones se pueden agrupar en dos categorías o tipos:

-Tipo A: Las basadas en sumar un punto por cada ejecución correcta de los sujetos a los ítems. Donde se penaliza el error del sujeto exclusivamente.
-Tipo B: Las basadas en sumar el valor de cada índice de dificultad de los ítems contestados correctamente para cada sujeto. Donde se penaliza el error del sujeto me-

diado por las características del test respecto al grupo de administración.

En el primer caso, solamente se valora la ejecución de los sujetos ante la tarea. Establecen criterios externos y no garantizan la fiabilidad del proceso. Siempre puede quedar la duda del nivel de conocimiento real del sujeto sobre la materia evaluada respecto a otras pruebas. En el segundo de los casos, se realiza una evaluación conjunta de los ítems con los sujetos. La desventaja radica que cualquier ítem puede resultar poco adecuado (por mala construcción, falta de validez, etc.) de forma que el profesor se puede encontrar ante la posibilidad de tener que desechar parte de la prueba. En estos indicadores, la estrategia es normativa, califican en función de la respuesta de todos los sujetos.

Dentro de estos últimos indicadores, la expresión 8 no permite la existencia de índices de dificultad negativos. Como mal menor, si esto sucede en una prueba para pocos ítems (siempre menos del 5% del total de ítems de la prueba, como recomendación) se puede asignar el valor 1.

Tipo A	Tipo B
- Se basa en la suma de un punto por cada contestación correcta. - Son fáciles de calcular - Son sencillos de argumentar y explicar. - Tienen problemas de fiabilidad. - Se basa en una evaluación criterial - Sólo valoran la ejecución del alumnado ante el ítem.	- Se basa en sumar el valor del índice de dificultad de los ítems contestados correctamente. - Tienen más dificultad de calcular - La explicación es algo más compleja. - Generan indicadores normativos. - Valoran la ejecución del alumnado al mismo tiempo que la bondad del test (por lo tanto, evalúa la labor del evaluador al construir el test).

Tabla 6. Características generales de los procesos de calificación

La razón se encuentra en que para estas expresiones, los valores inferiores a 0 indican una mala construcción, o un ítem no alcanzable por el alumnado. Para ítems como estos, cualquier alumno que conteste correctamente debe obtener, por lo menos, la puntuación máxima que puede asignarse a un ítem (es decir, un 1). Si el número de ítems con índices de dificultad inferior a 0 es elevado, el profesor o autor debería plantearse analizar la situación de examen. En este sentido, la utilización de las expresiones de este tipo debe asumirse como parte de la autoevaluación del docente.

Por otro lado, el desarrollo del proyecto ha permitido cumplir con los objetivos propuestos. Se han mostrado las diferencias que pueden darse en cuanto a la "calificación" que un alumno o alumna recibe, exclusivamente en función de la estrategia utilizada para valorar la prueba. Igualmente se ha mostrado cómo la perspectiva bayesiana tiene mucho que decir en la evaluación del rendimiento académico en pruebas objetivas. En cada uno de los puntos es posible extraer muy diversas consideraciones, no obstante, a continuación se exponen de forma muy resumida, algunas cuestiones para el debate:

a) La técnica utilizada en el análisis de pruebas objetivas influye decisivamente en la medición del rendimiento académico. Si este concepto es el resultado de un proceso, donde intervienen aspectos como la motivación del sujeto, su interés por la material, el contexto de aprendizaje, etc., no debe olvidarse que en la valoración del rendimiento (y debería analizarse hasta qué punto en el propio rendimiento) el proceso de análisis determina sustancialmente el resultado.
b) Este trabajo muestra la variabilidad que puede experimentar un alumno o alumna en función de la técnica de valoración de las pruebas que realiza. En otros términos, los resultados académicos que obtiene el alumnado, dependen (tal vez en exceso) de las estrategias de evaluación, incluso, como es el caso expuesto en estas páginas, cuando se utiliza una prueba de las llamadas "objetivas". Por tanto, será necesario reflexionar sobre en qué consiste el fracaso académico y cuales son los factores realmente decisivos.
c) Si bien los investigadores tienden al estudio de grupos, la educación, como un proceso que afecta a cada individuo en todo su desarrollo vital, y por tanto la investigación en educación, también se interesa por el resultado en cada sujeto. En este caso, se ha mostrado la importancia de realizar el seguimiento de individuo por individuo. La educación no sólo debe afrontar la cuestión de la individualidad en el sentido de la atención diferenciada en el proceso de enseñanza, sino también en la acción valorativa de las tareas académicas. Esto no quiere decir que se defienda una relatividad en los hechos, es decir, una tarea estará bien o mal realizada, y esta situación, como hecho contrastable no es susceptible de duda. Sin embargo, la pregunta debe estar más en la raíz del problema: ¿la tarea expuesta es la adecuada para evaluar el contenido tratado?. Esto encaja directamente con el debate sobre validez de contenido de los instrumentos, y la adaptación de los instrumentos a la realidad de los sujetos evaluados.

La investigación presentada, no obstante, tiene evidentes limitaciones, entre ellas destaca el hecho de estar reducida a datos obtenidos de una muestra relativamente escasa. En este sentido, las conclusiones obtenidas deben considerarse bajo un sentido crítico a falta de una mayor indagación. Igualmente debe comprobarse como puede afectar el corte de las puntuaciones a la hora de agrupar las calificaciones, en la concordancia entre expresiones.

La investigación, por tanto, no está concluida. Existen numerosas líneas de continuidad entre las cuales se perciben como destacables la utilización de datos simulados (por el método Monte Carlo por ejemplo) que permitan un "barrido" de todas las posibilidades y el comportamiento diferencial de los distintos indicadores utilizados aquí. Otra vía de continuidad se concreta en el análisis del potencial de los indicadores tipo B, como recursos para la evaluación, entre ellos se asume a priori un alto grado de utilidad en los procesos de auto-evaluación del profesorado. Por último, es ineludible la necesidad de seguir desarrollando nuevos indicadores que permitan una mejor aproximación a la evaluación del proceso, y del resultado, evitando los problemas de fiabilidad, y sobre todo los de validez que se han apuntado en estas líneas.

En resumen, el ámbito del rendimiento académico es un área problemática que no está agotada sino que, por el contrario, presenta múltiples problemas que son necesarios afrontar. Todo esto está vinculado con los procesos de medición en educación y con la realidad de la evaluación y el diagnóstico en los programas formativos. Son estos, frentes que deben abordarse desde la ciencia, con inmediatez y voluntad de ofrecer soluciones contrastadas.

Referencias

Álvaro Page, M. (1993). *Elementos de psicometría*. Madrid: EUDEMA.

Álvaro Pagés, M., Bueno Monreal, Calleja, J.A., M.J., Cerdán, J., Echevarría, M.J., García, C., Gaviria, J.L., Gómez, C., Jiménez, S.C., López, B., Martín-Javato, L., Mínguez, A.L., Sánchez, A., y Trillo, C. (1991). *Hacia un modelo causal del rendimiento académico*. Madrid: C.I.D.E..

Bloom, B. (1972). *Taxonomía de los objetivos de la educación*. Alcoy: Marfil.

Carabaña, J. (1979). Origen social, inteligencia y rendimiento académico al final de la EGB. En I.N.C.I.E., (Eds.) *Temas de investigación educativa*. Madrid: MEC.

Forteza, J. (1975). Modelo instrumental de las relaciones entre variables motivacionales y rendimiento. *Revista de Psicología General y Aplicada, 132*, 75-91.

García Bacete, F.J., y Doménech, F. (1997). Motivación, aprendizaje y rendimiento escolar. *Revista Electrónica de Motivación y Emoción, 0*(1), 1-16. Consultado en http://reme.uji.es/reme/numero0/indexsp.html (Marzo de 2003)

García, F.J., y Musitu, G. (1993). Rendimiento académico y autoestima en el ciclo superior de EGB. *Revista de Psicología de la Educación, 4*(11), 73-87.

González Fernández, D. (1975). Procesos escolares inexplicables. *Aula Abierta, 11*, 12.

Jornet, J.M., y Suárez, J.M. (1996). Pruebas estandarizadas y evaluación del rendimiento: usos y características. *Revista de Investigación Educativa, 14*(2), 141-163.

Kaczynska, M. (1965). El rendimiento escolar y la inteligencia. Madrid: Espasa.

Muñoz Arroyo, A. (1977). Valoración del rendimiento de centros docentes de EGB. Badajoz: ICE de la Universidad de Extremadura. [VI Plan de Investigación Educativa del MEC].

Pérez Serrano, G. (1981). *Origen social y rendimiento escolar*. Madrid: CIS.

Plata, J. (1969). La comprobación objetiva del rendimiento escolar. Madrid: Magisterio Español.

Secadas, F. (1952). Factores de personalidad y rendimiento escolar. *Revista Española de Pedagogía, 37*.

Serrano, J. (1990). Inferencia bayesiana sobre una proporción. *Revista de Investigación Educativa, 8* (16), 509-516.

Serrano, J. (1997). Valoración de test con ítems de respuestas múltiples. En AIDIPE (Comp.) *Actas del VIII Congreso Nacional de Modelos de Investigación Educativa*. Sevilla: AIDIPE. (584-588).

Serrano, J., y Matas, A. (1999). Valoración del rendimiento en pruebas de opción múltiple. En AIDIPE (Comp.) *Nuevas realidades educativas, nuevas necesidades metodológicas*. Málaga: Centro de ediciones de la Diputación de Málaga. (356-361).

Anexos

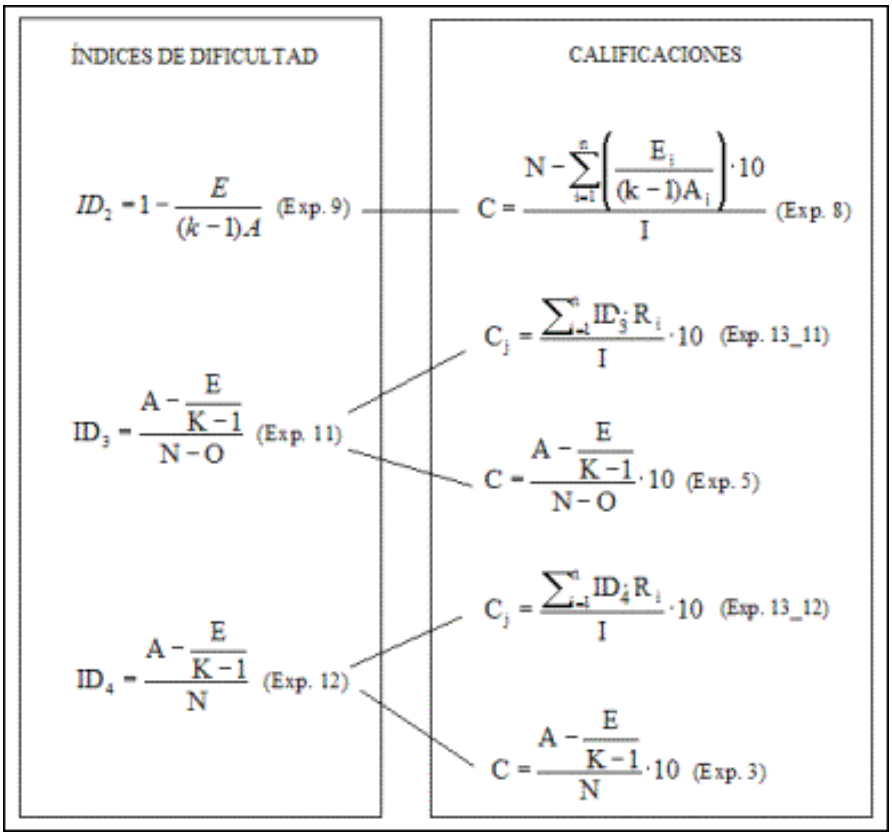

Figura 1. Relaciones entre índices de dificultad y expresiones de calificación

ABOUT THE AUTHORS / SOBRE LOS AUTORES

Antonio Matas Terrón (amatas@us.es). Profesor de Departamento de *Didáctica y Organización Escolar y Métodos de Investigación y Diagnóstico en Educación* de la Facultad de Ciencias de la Educación de la. Universidad de Sevilla (cuya dirección postal es Avenida Camilo José Cela s/n-Sevilla. España). Sus principales líneas de investigación son la evaluación de programas y la medición educativa. Es miembro del grupo HUM-369 del Plan Andaluz de Investigación.

ARTICLE RECORD / FICHA DEL ARTÍCULO

Reference / Referencia	Matas, Antonio (2003). Estudio diferencial de indicadores de rendimiento en pruebas objetivas. *Revista ELectrónica de Investigación y EValuación Educativa, v. 9*, n. 2. http://www.uv.es/RELIEVE/v9n2/RELIEVEv9n2_5.htm. Consultado en (*poner fecha*).
Title / Título	Estudio diferencial de indicadores de rendimiento en pruebas objetivas [*Differential analysis of indicators of achievement in objective tests*]
Authors / Autores	Antonio Matas Terrón
Review / Revista	Revista ELectrónica de Investigación y EValuación Educativa (RELIEVE), *v. 9*, n. 2
ISSN	1134-4032
Publication date / Fecha de publicación	2003 (**Reception Date**: 2003 March 11; **Approval Date:** 2003 Oct. 15 **Publication Date**: 2003 Oct. 17)
Abstract / Resumen	*Often objective tests to evaluate student qualifications are used. However, the selection between different evaluation criteria is in the hands of teachers or evaluators preferences. This investigation offers, briefly, a revision of some strategies utilized for objective evaluation. We also present other alternative strategies to the most classical ones. Afterwards, the strategies of evaluation have been analized with data from 114 subjects, who were evaluated with a test. The results suggest there are two kinds of strategies, one where only students' performance are evaluated, and other, where the test and the students are evaluated at the same time. We also show the fail of concordance between the indicators used* La utilización de pruebas objetivas para la calificación del alumnado es muy habitual. Sin embargo, la elección de un criterio de valoración u otro está a menudo, en manos de las preferencias del profesor o evaluador. En esta investigación se ofrece brevemente, una revisión de algunas de las estrategias utilizadas para ello, presentando alternativas a las más clásicas. Posteriormente, con los datos procedentes de una prueba real administrada a 114 sujetos, se han analizado los resultados generados a partir de las distintas aproximaciones valorativas. Se concluye que existen dos estrategias básicas para afrontar el problema, una de ellas basadas en la valoración exclusivamente de la ejecución de los sujetos ante los ítems, y otra que al mismo tiempo permite una evaluación de la prueba. Se muestra también la falta de concordancia entre los indicadores utilizados.
Keywords Descriptores	Achievement assessment, objective tests, bayesian estimation, item analysis Evaluación del rendimiento, pruebas objetivas, estimación bayesiana, análisis de items
Institution / Institución	Universidad de Sevilla (España)
Publication site / Dirección	http://www.uv.es/RELIEVE
Language / Idioma	Español (Title, abstract and keywords in english)

Revista **EL**ectrónica de **I**nvestigación y **EV**aluación **E**ducativa
(**RELIEVE**)

Electronic Journal of Educational Research, Assessment and Evaluation

[ISSN: 1134-4032]

© Copyright, RELIEVE. Reproduction and distribution of this articles it is authorized if the content is no modified and their origin is indicated (RELIEVE Journal, volume, number and electronic address of the document).
© Copyright, RELIEVE. Se autoriza la reproducción y distribución de este artículo siempre que no se modifique el contenido y se indique su origen (RELIEVE, volumen, número y dirección electrónica del documento).