ARTÍCULO DE INVESTIGACIÓN / INVESTIGATION ARTICLE

Rev. Spirat. 2024;2(3): 9-24
DOI: 10.20453/spirat.v2i1.5248

Análisis de los índices de dificultad y discriminación de las pruebas de opción múltiple: una herramienta para la evaluación formativa

Analysis of the difficulty and discrimination indices of multiple choice tests: a tool for formative evaluation

Liliana Aidee Muñoz, Naolly Casas Tolentino**, Jamine Pozu Franco****
Universidad Peruana Cayetano Heredia (UPCH)
Recibido: 06-2-2024; aceptado: 19-3-2024

Resumen

En un contexto donde existen la formación de competencias y el surgimiento de tecnologías de inteligencia artificial generativas, surge la interrogante sobre la efectividad de la evaluación de los aprendizajes. Entre los instrumentos de evaluación, se encuentran las pruebas de opción múltiple, y el desafío es que puedan ser un complemento para la evaluación formativa; para ello es esencial que su constructo contemple criterios de calidad. *Objetivo:* Determinar los índices de dificultad y discriminación de las pruebas de opción múltiple elaboradas por los docentes de diversas asignaturas del primer año de estudios universitarios. Así también que, con los resultados de su aplicación, brinden información oportuna para el seguimiento y acompañamiento de los estudiantes desde un enfoque formativo de la evaluación. *Materiales y métodos:* El estudio fue de enfoque cuantitativo de corte transversal descriptivo, aplicando la teoría clásica de la medición; se analizaron 140 reactivos distribuidos en 7 pruebas, las cuales se aplicaron a una muestra variada de estudiantes. *Resultado:* El análisis general mostró que las pruebas tuvieron un índice de dificultad (p) ideal (>=45 y <=75) con una media de 0.53 y una frecuencia de 87 (62%) reactivos en el rango de dificultad ideal, y un índice de discriminación (D) excelente (>0,34) con una media de 0.51. Sin embargo, se identificaron 12 reactivos, distribuidos en casi todas las pruebas, que no cumplían con los criterios mínimos de calidad. *Conclusión:* Se determinaron los indicadores de calidad de las pruebas y sus reactivos permitiendo así detectar aquellos que requieren una redefinición y aquellos que ponen de manifiesto áreas de aprendizaje en los estudiantes que necesitan ser reforzadas.

Palabras clave: Pruebas de opción múltiple, índice de dificultad, índice de discriminación.

Abstract

In a context where the formation of skills and the emergence of generative artificial intelligence technologies, the question arises about the effectiveness of learning assessment. Among the instruments of evaluation, there are multiple choice tests, and the challenge is that they can be a complement to

formative assessment, for this it is essential that their construct contemplates quality criteria. *Objective:* Determine the difficulty and discrimination indices of multiple-choice tests prepared by teachers of various subjects in the first year of university studies. Likewise, with the results of their application, they provide timely information for the monitoring and support of students from a formative approach to assessment. *Materials and methods:* The study had a descriptive cross-sectional quantitative approach, applying classical measurement theory; 140 items distributed in 7 tests were analyzed, which were applied to a varied sample of students. *Result:* The general analysis showed that the tests had an ideal difficulty index (p) (>=49 and <=57) with a mean of 0.53 and a frequency of 87 (62%) items in the ideal difficulty range and an excellent discrimination index (D) (>0.34) with a mean of 0.51. However, 12 reagents were identified, distributed in almost all the tests, that did not meet the minimum quality criteria. *Conclusion*: The quality indicators of the tests and their reagents were determined, thus allowing us to detect those that require redefinition and those that reveal learning areas in students that need to be reinforced.

Keywords: Multiple choice tests, difficulty index, discrimination index.

Introducción

La evaluación de los aprendizajes es un componente crucial en la educación superior, ya que además de acreditar los conocimientos adquiridos por los estudiantes, también permite obtener información y evidencias representativas del nivel de desarrollo de los aprendizajes, lo que posibilita la toma de decisiones orientadas a la retroalimentación y la mejora continua de estos (1), siempre enfocada en que el estudiante pueda demostrar un desempeño de manera integral al realizar una actividad o resolver un problema complejo (2).

La formación por competencias propone que la evaluación debe llevarse a cabo mediante diversos instrumentos, todos ellos con estándares de desempeño y criterios específicos que permitan abordar las limitaciones que existen al evaluar las distintas dimensiones de una competencia. Entre estas dimensiones se incluyen la cognitiva, que abarca conceptos, teorías, conocimientos factuales y habilidades cognitivas; la actuacional, que se refiere a habilidades técnicas y procedimentales; y la afectivo-motivacional, relacionada con valores y actitudes (3, 4). Asimismo, es importante establecer una distinción entre el nivel de desarrollo actual de un estudiante, el cual se puede medir mediante una prueba de rendimiento y el nivel de desarrollo potencial o grado de aprendizaje que el estudiante puede lograr con una mejor instrucción y acompañamiento (5).

Los docentes, como parte del sistema de evaluación, suelen aplicar pruebas de opción múltiple (POM) concibiendo generalmente solo su función sumativa y acreditadora; pero debido a sus características y ventajas, estas deberían ser consideradas una evaluación del y para el aprendizaje, en donde, a partir de la intencionalidad de la enseñanza y la evaluación, los estudiantes pueden demostrar su comprensión en una amplia variedad de temas esenciales, recibir la retroalimentación correspondiente y aprender a autorregular sus procesos de aprendizaje (3). En esa línea, este estudio considera el carácter formativo que puede tener la evaluación objetiva, ya que a partir del cual se puede obtener evidencias de aprendizajes para realizar retroalimentaciones con mayor precisión.

La evaluación, sin embargo, enfrenta desafíos propios del contexto neoliberal y demanda procedimientos e instrumentos que permitan evaluar los conocimientos y capacidades de los estudiantes de manera rápida, objetiva y confiable (8, 9). Un ejemplo de este desafío lo encontramos en el creciente uso de herramientas de inteligencia artificial generativa (IA-G) en la educación, como el Chat GPT. Aunque estas herramientas han generado oportunidades y un cambio significativo en las metodo-

logías de enseñanza, también han planteado dificultades y un cierto descontrol en la verificación de los aprendizajes adquiridos al facilitar la elaboración de diversos materiales académicos (6, 7, 20, 23).

Por otro lado, aunque las POM son ampliamente utilizadas por su eficiencia y objetividad y reúnen las ventajas en cuanto a su factibilidad por ser fácilmente administradas y corregidas mediante lectores automáticos, así como en cuanto a su fiabilidad, siempre y cuando estén bien diseñadas (10), no es una práctica habitual realizar análisis postprueba que impliquen la determinación de indicadores de calidad mediante la teoría clásica del test, a pesar de ser una herramienta útil para mejorar el instrumento, e incluso para tomar decisiones sobre la mejora de los aprendizajes de los estudiantes (11).

Como sostienen diferentes autores (9, 10, 12, 13, 17), el análisis mediante la teoría clásica de la medición o de los test, el cual consiste en el cálculo de los índices de dificultad y discriminación, permite caracterizar las preguntas (reactivos) que conforman una prueba de rendimiento académico y representan indicadores de calidad en la medida que se encuentren dentro de los rangos aceptables. Con la aplicación de esta herramienta, el docente puede ser capaz de conocer las áreas temáticas logradas y no logradas por los estudiantes; puede redefinir las preguntas, ajustar la composición de la prueba, así como crear un banco de preguntas (26).

Por lo tanto, el objetivo del presente estudio es determinar los índices de dificultad y discriminación de las pruebas de opción múltiple elaboradas por los docentes de diversas asignaturas del primer año de estudios universitarios, a fin de brindar información sobre estos indicadores de calidad a los docentes para la mejora de los reactivos y el seguimiento y acompañamiento a partir de los resultados obtenidos por los estudiantes acordes con los resultados del aprendizaje.

Materiales y métodos

Es un estudio transversal de tipo descriptivo realizado mediante la teoría clásica de la medición, que comprende métricas fundamentales que se han empleado con el propósito de analizar la validez estructural de las preguntas o reactivos de las pruebas de opción múltiple (14). Esta herramienta parte del análisis psicométrico: proceso que investiga cómo evaluar los constructos del aprendizaje de los estudiantes, posibilita analizar la validez de los instrumentos de evaluación de dichos aprendizajes y propicia el desarrollo de propuestas de mejora de estos (10).

El estudio se realizó con 140 reactivos agrupados en 7 pruebas (P1, P2, P3, P4, P5, P6 y P7) del área de ciencias, las cuales fueron seleccionadas por tener las POM como el instrumento prioritario de evaluación del curso; cada prueba fue aplicada a una muestra relativa de 191, 123, 260, 245, 146, 170 y 196 estudiantes, respectivamente. Las pruebas fueron elaboradas por los docentes responsables de cada asignatura; cada reactivo de dichas pruebas constaba de una raíz y 5 opciones con solo una respuesta correcta y 4 distractores. La respuesta correcta recibió un punto, la respuesta incorrecta recibió cero puntos y los reactivos que no tuvieron respuesta se consideraron como incorrectos. Cada reactivo fue analizado aplicando los índices de dificultad (p) y discriminación (D) de la teoría clásica de la medición.

Los datos se obtuvieron a partir de las hojas de fichas ópticas de cada una de las pruebas y luego se transfirieron a matrices en Microsoft Excel para generar los patrones de respuesta que incluyen el número de aciertos y errores en cada pregunta (12). Como se observa en las tablas 1 y 2, el puntaje total (calificación) de cada estudiante examinado se calculó contando el número de aciertos en las filas y el número de aciertos de cada reactivo se calculó contando los aciertos en las columnas. El 27% de los estudiantes con las calificaciones más altas se consideró como el grupo superior de estudiantes de alto rendimiento (A) y el 27% de los estudiantes con las calificaciones más bajas se consideró como el grupo inferior de estudiantes de bajo rendimiento (B).

El patrón de respuestas se puede construir ordenando las respuestas de tipo dicotómico: correcto o incorrecto, de dos maneras: mediante códigos (1 y 0), como se observa en la tabla 1, o mediante alternativas (A, B, C), como se observa en la tabla 2; para este último método, es necesario especificar la alternativa correcta.

Tabla 1. Patrón de respuestas con códigos (1 y 0) y el ranking de las puntuaciones

N	Estudiantes	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Puntaje total
1	Estudiante 1	1	1	1	1	1	1	1	1	1	1	10
2	Estudiante 2	1	1	1	1	0	1	1	1	1	0	8
3	Estudiante 3	1	0	1	1	0	1	1	0	0	0	6
4	Estudiante 4	0	0	1	1	0	0	1	0	0	0	3
5	Estudiante 5	0	0	0	0	0	0	1	0	0	0	1
	# de aciertos	**3**	**2**	**4**	**4**	**1**	**3**	**5**	**2**	**2**	**1**	

Nota: Adaptado de Hurtado Mondoñedo (12)

Tabla 2. Patrón de respuestas con alternativas y el ranking de las puntuaciones

N	Estudiantes	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10	Puntaje total
	Respuestas correctas	**A**	**B**	**B**	**A**	**C**	**E**	**A**	**E**	**B**	**B**	**10**
1	Estudiante 1	A	B	B	A	C	E	A	E	B	B	10
2	Estudiante 2	A	B	B	A	A	E	A	E	B	C	8
3	Estudiante 3	A	A	B	A	A	E	A	C	C	E	6
4	Estudiante 4	B	A	B	A	E	C	A	B	A	E	3
5	Estudiante 5	C	E	C	B	B	E	C	C	A	B	1
	# de aciertos	**3**	**2**	**4**	**4**	**1**	**3**	**5**	**2**	**2**	**1**	

El cálculo del índice de dificultad y el índice de discriminación se realizó de la siguiente manera:

Índice de dificultad (p)

Es una expresión numérica de la dificultad que representa para los examinados contestar una pregunta. Por lo tanto, entre menor sea la proporción de estudiantes examinados que respondieron correctamente un reactivo, mayor será su dificultad; mientras que, entre mayor sea la proporción de estudiantes examinados que respondieron correctamente un reactivo, menor será su dificultad (9, 15, 10).

De acuerdo con Kumar et al. (9), para calcular la dificultad (p) de un reactivo (i), primero es preciso conocer el grupo de alto rendimiento y al grupo de bajo rendimiento; en este estudio también se consideró el 27% de estudiantes con las calificaciones más altas y más bajas, respectivamente, con la

finalidad de generar grupos superiores e inferiores indudablemente diferentes; tras conocer ambos grupos, se suma el número de examinados que respondieron correctamente en el grupo de alto rendimiento (A) y el número de examinados que respondieron correctamente en el grupo de bajo rendimiento (B) y, finalmente, se divide entre el número total de estudiantes en los dos grupos (N), incluidos los que no respondieron. Como se muestra en la siguiente fórmula:

$$p_i = \frac{A_i + B_i}{N}$$

Los reactivos se pueden clasificar según los intervalos de dificultad. Existen diferentes formas para clasificarlos. En este estudio se consideraron los criterios de la tabla 3 a partir del estudio realizado por Argudín (16) y considerando la heterogeneidad de los estudiantes del primer año de estudios universitarios.

Tabla 3. Clasificación de reactivos según los intervalos de dificultad (p)

Intervalos de (p)	Interpretación
Menor de 0.24	Extremadamente difícil
Entre 0.25 y 0.44	Difícil
Entre 0.45 y 0.75	Ideal (mejores preguntas)
Entre 0.76 y 0.91	Fácil
Mayor de 0.91	Demasiado fácil

Nota: Adaptado de Argudín (16)

Índice de discriminación (D)

El índice de discriminación de una pregunta distingue, diferencia y selecciona entre los examinados de mayor y menor rendimiento en la prueba. Con este índice se espera que el examinado que logró una puntuación alta en toda la prueba tenga mayor probabilidad de responder correctamente al reactivo, mientras que el que tuvo una baja puntuación en toda la prueba deberá tener pocas probabilidades de responder dicho reactivo (17).

De acuerdo al estudio de Date (17), el índice de discriminación se calculó aplicando la siguiente fórmula:

$$D_i = \frac{2 \times (A_i - B_i)}{N}$$

Si todos los examinados del grupo de alto rendimiento (A) contestan correctamente al reactivo (i) y todos los examinados del grupo de bajo rendimiento (B) contestan incorrectamente al mismo reactivo, el índice de discriminación (D) será igual a 1, valor máximo de este indicador. Por lo tanto, entre más alto sea el índice de discriminación, el reactivo diferenciará mejor a los examinados con altas y bajas calificaciones. La clasificación de los reactivos según su índice de discriminación se realizó considerando los criterios establecidos en Date (17).

Tabla 4. Clasificación de reactivos según los intervalos del índice de discriminación (D)

Intervalos de (p)	Interpretación
Menor de 0.00	Pésimo (muy mala pregunta)
Menor de 0.20	Deficiente
Entre 0.21 y 0.24	Aceptable
Entre 0.24 y 0.34	Bueno
Mayor de 0.35	Excelente

Nota: Adaptado de Date (17)

Relación de los índices de dificultad y discriminación

La comparación de los índices de dificultad y discriminación obtenidos con determinadas normas (criterios óptimos de calidad) de dificultad o discriminación nos permitiría aceptar, revisar y descartar reactivos. En ese sentido, los reactivos con índices de dificultad menores de 0.20 y mayores de 0.80 con valores D menores de 0.20 y los reactivos con valores D menores de 0.0 se seleccionan como reactivos por mejorar o descartar (18). En general, se espera que los reactivos de alta calidad sean contestados por la mitad de los estudiantes examinados, siempre y cuando quienes los acierten tengan mayor dominio de la habilidad.

Resultados

En el análisis general de los 140 reactivos distribuidos en 7 pruebas, se obtuvo un índice de dificultad dentro del rango "ideal" (>=45 y <=75) con una media de 0.53 y un índice de discriminación "excelente" (>0,34) con una media de 0.42. Como se detalla en la tabla 5.

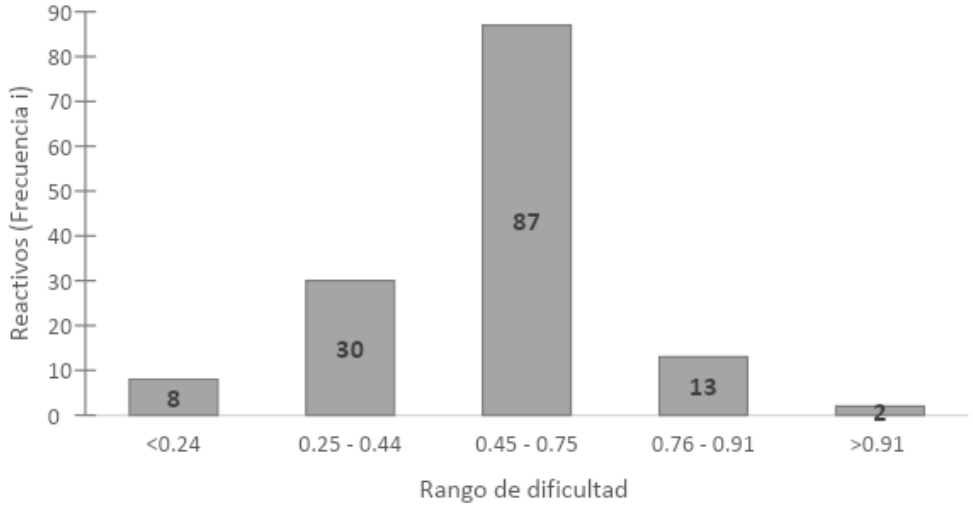
Tabla 5. Resultados descriptivos de las pruebas de opción múltiple

Pruebas (P)	N° estudiantes examinados	N° reactivos	Índice de dificultad (media ± SD)	Índice de discriminación (media ± SD)
P1	191	20	0.49 ± 0.26	0.35 ± 0.18
P2	123	20	0.52 ± 0.12	0.42 ± 0.15
P3	260	20	0.56 ± 0.15	0.48 ± 0.15
P4	245	20	0.49 ± 0.19	0.46 ± 0.23
P5	146	20	0.57 ± 0.19	0.43 ± 0.16
P6	170	20	0.56 ± 0.21	0.36 ± 0.15
P7	196	20	0.54 ± 0.11	0.48 ± 0.19
Total	**1331**	**140**	**0.53**	**0.42**

Nota: SD: Desviación estándar

Para analizar el comportamiento del índice de dificultad de los 140 reactivos se presenta la distribución de las frecuencias absolutas de los valores p en la figura 1, según los rangos descritos en la tabla 1. En esta figura se puede observar que la mayoría de los reactivos se agrupa en una frecuencia de entre 0.45 y 0.75 con una media de 0.58; además, se puede observar que hay un mayor número de reactivos "difíciles" (0.44-0.44) que "fáciles" (0.76-0.91), así como un ligero mayor número de reactivos "extremadamente difíciles" (<0.24) que reactivos "extremadamente fáciles" (>0.91).

Figura 1. Frecuencia de los reactivos según su nivel de dificultad



Clasificando los reactivos según su nivel de dificultad podríamos agruparlos de la siguiente manera: 87 (62%) reactivos tuvieron un índice de dificultad ideal ("mejores preguntas"); 8 (6%) reactivos, un índice de dificultad "extremadamente difícil"; 30 (21%) reactivos tuvieron un índice de dificultad "difícil"; 13 (9%) reactivos, un índice de dificultad "fácil"; y 2 (1%) reactivos, un índice de dificultad "extremadamente fácil".

En cuanto a la distribución de la frecuencia absoluta de los reactivos según su poder de discriminación, se observó que la mayoría de los reactivos se agrupan en una frecuencia de >0.35 con una media de 0.51. Y podríamos clasificarlos de la siguiente manera: 98 (70%) reactivos tuvieron un poder de discriminación "excelente"; 24 (17%) tuvieron un poder de discriminación "bueno"; 5 (4%), un poder de discriminación aceptable; 12 (9%), un poder de discriminación "deficiente"; y 1 (1%) reactivo, un poder de discriminación "pésimo".

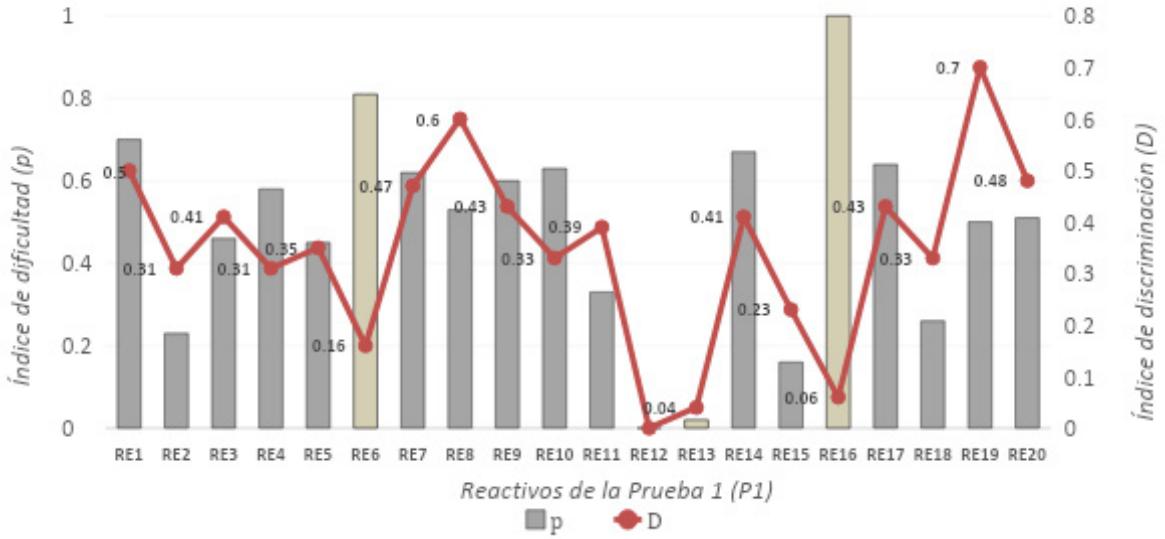
De estos reactivos, aquellos que tuvieron un índice de discriminación menor de 0.20 fueron contrastados con los índices de dificultad para la toma de decisiones sobre la mejora o descarte según corresponda.

En la clasificación de reactivos por prueba, de acuerdo con los índices de dificultad (p) y discriminación (D), se obtuvo que las pruebas 1, 4 y 6 tuvieron casi toda la gama de dificultades, mientras que las pruebas 2 y 7 tuvieron un rango de dificultad de medio a difícil (tablas 1 y 2 en el anexo).

Finalmente, se identificaron los reactivos que necesitaban alguna modificación aplicando los siguientes criterios óptimos de calidad: los reactivos con índices de dificultad menores de 0.20 y

mayores de 0.80 con valores D menores de 0.20, o los reactivos con valores D menores de 0.0 independientemente del valor de dificultad, como se observa en la figura 2, en la que se representa la relación de los índices p y D de la prueba 1 (P1). Según esta gráfica, los reactivos que necesitan ser revisados son el RE6, RE12, RE13 y RE16.

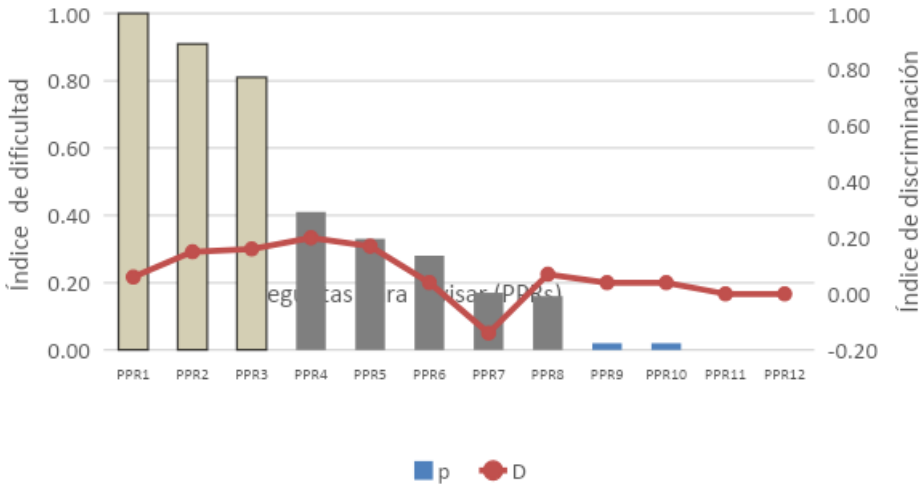
Figura 2. Relación del índice de dificultad (p) y discriminación (D) de la prueba 1



Utilizando estos parámetros en las 7 pruebas, se pudo observar que 12 reactivos no cumplían con las mínimas normas de calidad y se consideraron como reactivos o preguntas para revisar (PPR). Estos reactivos están distribuidos de la siguiente manera: 4 PPR en la prueba 1; 1 en la prueba 2; 2 en la prueba 4; 1 en la prueba 5; 3 en la prueba 6; 1 reactivo en la prueba 7; y cero PPR en la prueba 3. Por lo tanto, los reactivos y las pruebas evaluadas tuvieron una calidad aceptable para ser aplicadas en las asignaturas del presente estudio.

Como se observa en la figura 3, de los 12 reactivos identificados con un poder de discriminación deficiente (D<0.20), se observa que 9 (75%) corresponden a estudiantes de bajo rendimiento, lo cual indica que son reactivos con poco poder de discriminación entre los estudiantes con puntuaciones bajas y se encuentran en las pruebas 1, 2, 4, 6 y 7. Mientras que 3 (25%) corresponden a estudiantes de alto rendimiento, lo cual indica que son reactivos con poco poder de discriminación entre los estudiantes con puntuaciones altas, y estas se encuentran en las pruebas 1 y 5.

Figura 3. Reactivos con bajo poder de discriminación



Estos resultados fueron compartidos con los docentes responsables del diseño de cada prueba con la finalidad de que cuenten con una evidencia cuantitativa de las características de los reactivos. Esta información les permitió identificar qué aspectos temáticos les resultan más complejos a los estudiantes y, a partir de ello, brindar una retroalimentación o reforzamiento oportuno a los estudiantes; además, permite mejorar el instrumento o realizar alguna reestructuración orientando la intencionalidad de las preguntas hacia la comprobación de los aprendizajes que se deseen evaluar.

Discusión

En educación superior, las pruebas de opción múltiple (POM) son uno de los formatos de evaluación escrita más populares, puesto que permiten evaluar una gama amplia de contenidos y objetivos educativos, así como evaluar eficientemente grandes números de examinados, de manera que pueden someterse a un análisis de ítems posterior; lo que le proporciona mayor validez que otros métodos de prueba (26).

Un buen instrumento de evaluación puede proporcionar información valiosa sobre el progreso de los estudiantes; esto permite ajustar la metodología y la estrategia del docente mediador para satisfacer las necesidades de aprendizaje particular de cada estudiante. Asimismo, contribuye a concientizar a los estudiantes sobre su progreso fomentando así su autonomía en el logro de su propio aprendizaje (3, 4, 19).

Para que la evaluación sea efectiva, es esencial que las preguntas de las POM discriminen adecuadamente entre los estudiantes que poseen el conocimiento necesario (un buen rendimiento) y aquellos que no han logrado el nivel esperado (21). Por ende, la determinación de los índices de dificultad y discriminación, así como su relación, deriva en ventajas técnicas (aproximación a la validez estructural) y formativas que se presentan como una herramienta valiosa para la planificación y la mejora de la composición de las pruebas (8, 11, 26).

De acuerdo con Hurtado Mondoñedo (12), esta herramienta, con un enfoque más preciso, contribuye a una evaluación más certera de la habilidad de los estudiantes examinados, siempre y cuando

quienes acierten las preguntas tengan un mayor dominio del conocimiento o habilidad que se pretende valorar. Salazar (22) y Mansour (26) respaldan esta idea destacando que esta herramienta permite optimizar el trabajo metodológico de los docentes en la elaboración de las POM mediante la identificación de las preguntas que no fueron elaboradas correctamente, preguntas poco comprensibles o distractores inefectivos, de manera que puedan ser rediseñadas e incluso eliminadas.

En ese sentido, el aporte formativo del análisis de las POM no solo se presenta en la detección de áreas de aprendizaje no logradas por los estudiantes, a partir de la identificación de las preguntas que resultaron más complejas (valores p <0.24). Esta información presenta las bases para implementar estrategias de reforzamiento y retroalimentación a ese grupo determinado de estudiantes examinados que no respondieron dichos reactivos (24, 25). Por ende, en este estudio se destaca y apoya la integración de las POM como instrumentos oportunos para el proceso de evaluación formativa, sobre todo en un contexto universitario caracterizado por tener un gran número de estudiantes.

Los resultados de este estudio permitieron identificar dichos aspectos. Las pruebas analizadas globalmente tuvieron un índice de dificultad promedio (p 0.49 y 0.56) y de discriminación (D>0.34); mientras que entre el 50 y 85% de los reactivos, individualmente, tuvieron un rango de dificultad media (p 0.45 y 0.75). Al respecto, Argudín et al. (16) recomiendan que la dificultad media de un reactivo sea entre 50 y 60%, mientras que en el estudio de Kumar et al. (9) sostienen que un reactivo ideal será aquel que tenga una dificultad media de entre 30 y 70% y una discriminación mayor de 0.25.

Respecto al poder de discriminación, en este estudio se encontró que la mayoría de reactivos con poco poder de discriminación (75%) estaban en los estudiantes de bajo rendimiento. Según Iñarrairaegui (11), mientras mayor es el índice de dificultad de un reactivo, el poder de discriminación disminuye gradualmente; por ello, sugiere que las preguntas se diseñen de acuerdo al nivel académico de cada grupo de estudiantes.

En cuanto a la distribución de reactivos según su nivel de dificultad, las pruebas P1, P3, P4 y P5 tuvieron una frecuencia con más del 50% de reactivos de dificultad media (p 0.45 y 0.75) y una proporción equilibrada de reactivos “extremadamente fácil” y “extremadamente difícil”, además de reactivos “fácil” y “difícil”; las pruebas P2, P6 y P7 tuvieron una mayor frecuencia de reactivos de dificultad media a difícil. Al respecto, Escudero et al. (18) sostienen que incluir reactivos de toda la gama de dificultades en una prueba es una estrategia acertada, y propuso las siguientes proporciones: 5% de reactivos fáciles, 20% de medianamente fáciles, 50% de dificultad media, 20% de medianamente difíciles y 5% de difíciles; al obtener estas proporciones se logra una escala bien graduada, pertinente para medir con precisión el dominio de los estudiantes con características etéreas.

Estos resultados coinciden con lo propuesto por Vega et al. (10), quienes plantean la importancia de la asimetría en la distribución de rendimientos durante el periodo de aprendizaje; ellos sugieren una distribución asimétrica positiva (mayor proporción de preguntas de dificultad media a fácil) para motivar y estimular un mayor esfuerzo, además de identificar a los estudiantes con alto rendimiento. En cambio, al final del proceso de enseñanza y aprendizaje, sugieren identificar a aquellos estudiantes con un desempeño deficiente mediante una distribución asimétrica negativa (mayor proporción de preguntas de dificultad media a alta) para asegurar el paso al siguiente nivel de estudiantes con el rendimiento esperado y brindar soporte y acompañamiento a aquellos que no lograron los resultados de aprendizaje.

En este estudio se sugiere disponer de reactivos en toda la gama de dificultades con cierta asimetría positiva o negativa, dependiendo del momento en el que se aplique el instrumento, y no solamente de reactivos centrados en el 50% de dificultad, puesto que, en el proceso de enseñanza y aprendizaje de las asignaturas del primer año de estudios universitarios, el rendimiento de los estudiantes no

sigue necesariamente una patrón normal, dadas las características heterogéneas en términos de los aprendizajes previos.

En conclusión, la incorporación de las POM, cuando son diseñadas y analizadas con criterios de calidad y con fines formativos, emerge como una herramienta valiosa con el potencial de proporcionar evidencias sobre las necesidades de aprendizaje de los estudiantes. En ese sentido, es importante promover en la práctica de los docentes su aplicación, siempre y cuando la intencionalidad de estos instrumentos sea con un rol formativo y no solamente para cumplir un rol sumativo o acreditador que no oriente el proceso de evaluación o no aporte en la mejora del aprendizaje de los estudiantes.

Esta modalidad de análisis se puede utilizar en diferentes asignaturas que utilicen las POM como instrumento y puede realizarse periódicamente con la finalidad de proporcionar información que permita extraer conclusiones académicas, técnicas en la elaboración de un banco de preguntas y otras ventajas que aún nos son insospechadas. Finalmente, se confirma la necesidad de integrar más indicadores de calidad, porcentajes de desacierto, realizar estudios enfocados en los distractores y contemplar un análisis de consistencia en el caso de que se esté construyendo un banco de preguntas.

Referencias

(1) Hamodi C, López Pastor VM, López Pastor AT. Medios, técnicas e instrumentos de evaluación formativa y compartida del aprendizaje en educación superior. Perfiles educativos. 2015; 37(147):146-161.

(2) Vaillant D. Análisis y reflexiones para pensar el desarrollo profesional docente continuo. Educar. 2014; 55-66. Disponible en: https://www.redalyc.org/articulo.oa?id=342132562004.

(3) Cañadas L. Evaluación formativa en el contexto universitario: oportunidades y propuestas de actuación. Rev Digit Invest Docencia Univ. 2020; 14(2).

(4) Tobón S. Evaluación socioformativa. Estrategias e instrumentos. Mount Dora, USA: Kresearch. 2017; 98 p. Disponible en: doi: dx.doi.org/10.24944/isbn.978-1-945721-26-7.

(5) Sandoval Oviedo N. La evaluación de los aprendizajes desde un enfoque cognitivo. Itinerario Educativo. 2009; 23(54):97-108.

(6) Romo-Pérez V, García Soidán JL, Özdemir AS, Leiros-Rodríguez RL. ChatGPT ha llegado ¿Y ahora qué hacemos? La creatividad, nuestro último refugio. Revista de Investigación en Educación. 2023; 21(3):320-334.

(7) Gallent-Torres C, Zapata-González A, Ortego-Hernando JL. El impacto de la inteligencia artificial generativa en educación superior: una mirada desde la ética y la integridad académica. RELIEVE. Revista Electrónica de Investigación y Evaluación Educativa. 2023; 29(2):1-21.

(8) Laforcada Ríos C. Grado de dificultad y poder discriminativo de preguntas de elección múltiple en materias de pregrado de la carrera de medicina. Cuad Hosp Clin. 2018; 59(1):62-68. Disponible en: http://www.scielo.org.bo/scielo.php?script=sci_arttext&pid=S1652-67762018000100008.

(9) Kumar D, Jaipurkar R, Shekhar A, Sikri G, Srinivas V. Item analysis of multiple-choice questions: A quality assurance test for an assessment tool. Med J Armed Forces India. 2021; 77(Suppl 1):S85-S89. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7873707/pdf/main.pdf.

(10) Chávez Vega R, Rodríguez Méndez A. Aplicación de la teoría clásica de test a la evaluación de preguntas de opción múltiple. Educ Med Super. 2022; 36(1). Disponible en: http://scielo.sld.cu/pdf/ems/v36n1/1561-2902-ems-36-01-e2228.pdf.

(11) Iñarrairaegui M, Fernández-Ros N, Lucena F, et al. Evaluation of the quality of multiple-choice questions according to the students' academic level. BMC Med Educ. 2022; 22(779). Disponible en: https://doi.org/10.1186/s12909-022-03844-3.

(12) Hurtado Mondoñedo LL. Relación entre los índices de dificultad y discriminación. Rev. Digit Invest Docencia Univ. 2018; 12(1):273-300. Disponible en: http://www.scielo.org.pe/scielo.php?script=sci_arttext&pid=S2223-25162018000100016&nrm=iso.

(13) Reynolds CR, Altmann RA, Allen DN. Análisis de elementos: métodos para adaptar los elementos correctos a la prueba correcta. En: Dominar las pruebas psicológicas modernas. 2021. Springer, Cham. https://doi.org/10.1007/978-3-030-59455-8_7.

(14) Cvetkovic-Vega A, Maguiña JL, Soto A, Lama-Valdivia J, Correa López LE. Estudios transversales. Rev Fac Med Hum. 2021; 21(1):179-185. Disponible en: http://www.scielo.org.pe/pdf/rfmh/v21n1/2308-0531-rfmh-21-01-179.pdf.

(15) Baladrón J, et al. El examen al examen MIR 2015: aproximación a la validez estructural a través de la teoría clásica de los tests. FEM. 2016; 19(4):217-226.

(16) Argudín Somonte E, Díaz Rojas P, Leyva Sánchez E. Índice de Dificultad del examen de Morfofisiología Humana I. Educ Med Super. 2011; 25(2):97-106. Disponible en: http://scielo.sld.cu/pdf/ems/v25n2/ems07211.pdf.

(17) Date AP, Borkar AS, Butwaik RT, Siddiqui RA, Shende TR, Dashputra. Item analysis as tool to validate multiple choice question bank in pharmacology. Int J Basic Clin Pharmacol. 2019; 8:1999-2023. Available at: https://www.ijbcp.com/index.php/ijbcp/article/view/3324/2573.

(18) Backhoff Escudero E, Larrazolo Reyna N, Rosas Morales M. Nivel de dificultad y poder de discriminación del Examen de Habilidades y Conocimientos Básicos (EXHCOBA). REDIE. 2000; 2(1).

(19) Casas Tolentino, NJ. Evaluación formativa: concepciones y práctica asumida por los docentes de una Facultad de Arte y Diseño de una Universidad Particular, Lima [tesis de maestría]. Lima: Universidad Peruana Cayetano Heredia. 2023.

(20) Moya BA, Eaton SE. Examining Recommendations for Artificial Intelligence Use with Integrity from a Scholarship of Teaching and Learning Lens. RELIEVE. 2023; 29(2). Disponible en: https://revistaseug.ugr.es/index.php/RELIEVE/article/view/29295/26519.

(21) Giaconi E, Bazán ME, Castillo M, Hurtado A, Rojas H, Giaconi V, et al. Análisis de pruebas de opción múltiple en carreras de la salud de la Universidad Mayor. Investigación educ médica. 2021; 10(40):61-69. Disponible en: https://www.scielo.org.mx/pdf/iem/v10n40/2007-5057-iem-10-40-61.pdf

(22) Salazar Blanco OF, Vélez CM, Zuleta Tobón JJ. Evaluación de conocimientos con exámenes de selección múltiple: ¿tres o cuatro opciones de respuesta? Experiencia con el examen de admisión a posgrados médico-quirúrgicos en la Universidad de Antioquia. Iatreia. 2015; 28(3):300-311. Disponible en: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-07932015000300008&lang=es.

(23) Mindzak M. What happens when a machine can write as well as an academic? University Affairs. 2020 [cited February 01]. Available at: https://universityaffairs.ca/opinion/in-my-opinion/what-happens-when-a-machine-can-write-as-well-as-an-academic/.

(24) Hernández T, Magid MS, Polydorides AD. Assessment Question Characteristics Predict Medical Student Performance in General Pathology. Arch Pathol Lab Med. 2021; 145(10):1280-1288. doi: 10.5858/arpa.2020-0624-OA.

(25) Aubin AS, Young M, Eva K, St-Onge C. Examinee Cohort Size and Item Analysis Guidelines for Health Professions Education Programs: A Monte Carlo Simulation Study. Acad Med. 2020; 95(1):151-156. doi: 10.1097/ACM.0000000000002888.

(26) Mansour TM, Tarhouny SA, Azab M. Evaluación de la calidad de las preguntas sumativas de opción múltiple (MCQ) para todos los cursos del programa MBCh antes y después de la campaña de concientización del personal, Facultad de Medicina, Universidad de Taibah, KSA, 2022; 11(6). https://typeset.io/papers/assessment-of-multiple-choice-questions-test-item-quality-dtkw20l7.

Anexo 1

Tabla 1. Clasificación de preguntas de acuerdo con el índice de dificultad (p)

PRUEBA 1				PRUEBA 2		PRUEBA 3		PRUEBA 4		PRUEBA 5		PRUEBA 6		PRUEBA 7		PRUEBAS
*p**	Interpretación	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	RE (%)	p (media ± SD)	Media ± SD
<24	**Extremadamente difícil**	4 (20)	0,10 ± 0,11	0 (0)	0,0 ± 0	0 (0)	0,0 ± 0,14	2 (10)	0,09 ± 0,09	0 (0)	0	2 (10)	0,09 ± 0,10	0 (0)	0	**0,04 ± 0,05**
0, 25– 0, 44	**Difícil**	2 (10)	0,30 ± 0,05	7 (35)	0,39 ± 0,05	5 (25)	0,39 ± 0,05	4 (20)	0,37 ± 0,04	6 (30)	0,36 ± 0,07	3 (15)	0,37 ± 0,06	3 (15)	0,33 ± 0,06	**0,36 ± 0,03**
0,45 – 0,75	**Ideal (Mejores preguntas)**	12 (60)	0,57 ± 0,08	13 (65)	0,59 ± 0,09	12 (60)	0,57 ± 0,07	12 (60)	0,54 ± 0,07	10 (50)	0,59 ± 0,08	11 (55)	0,61 ± 0,05	17 (85)	0,57 ± 0,08	**0,58 ± 0,02**
0,76 – 0,91	**Fácil**	1 (5)	0,81 ± 0	0 (0)	0	3 (15)	0,81 ± 0,04	2 (10)	0,80 ± 0,02	3 (15)	0,82 ± 0,05	4 (20)	0,81 ± 0,05	0 (0)	0	**0,58 ± 0,40**
>91	**Extremadamente fácil**	1 (5)	1,00 ± 0	0 (0)	0	0 (0)	0 ± 0	0 (0)	0	1 (5)	0,91 ± 0	0 (0)	0	0 (0)	0	**0,27 ± 0,47**

Nota. p* Índice de dificultad. SD: Desviación estándar

Tabla 2. Clasificación de preguntas de acuerdo con el índice de dificultad (D)

Interpretación		PRUEBA 1		PRUEBA 2		PRUEBA 3		PRUEBA 4		PRUEBA 5		PRUEBA 6		PRUEBA 7		PRUEBAS
D	**Interpretación**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Reactivo (%)**	**D (media ± SD)**	**Media ± SD**
<0	**Muy mala pregunta**	0	0	0	0	0	0	1 (5)	-0,14	0	0	0	0	0	0	-0,02 ± 0,05
<=0,2	**Deficiente. No discrimina**	4 (20)	0,06 ± 0,10	0 (0)	± 0	0	0	1 (5)	0,0	1 (5)	0,15 ± 0	3 (15)	0,09 ± 0,07	2 (10)	0,08 ± 0,05	0,07 ± 0,06
>0,2 a <=0,24	**Aceptable**	1 (5)	0,23 ± 0,00	2 (10)	0,21 ± 0	0	0	0	0	1 (5)	0,23 ± 0	1 (5)	0,22 ± 0	0	0	0,13 ± 0,12
>0,24 a <=0,34	**Bueno. Discrimina bien**	4 (20)	0,32 ± 0,01	1 (5)	0,30 ± 0	4 (2)	0,29 ± 0,02	3 (15)	0,31 ± 0,03	4 (20)	0,29 ± 0,03	5 (25)	0,31 ± 0,03	3 (15)	0,31 ± 0,01	0,30 ± 0,01
>0,34	**Excelente**	11 (55)	0,46 ± 0,10	16 (80)	0,48 ± 0,11	16 (8)	0,53 ± 0,12	15 (75)	0,55 ± 0,13	14 (70)	0,51 ± 0,12	11 (55)	0,47 ± 0,06	15 (75)	0,56 ± 0,11	0,51 ± 0,04

Nota. D*: Índice de discriminación. SD: Desviación estándar

*** Liliana Aidee Muñoz**

Doctora en Educación y Magíster en Docencia Universitaria por la Universidad Nacional Federico Villarreal. Tiene un diplomado en Políticas Docentes por el IIPE-Unesco-Argentina; es experta en Gestión Curricular por el Centro de Investigación en Formación y Evaluación-México; ha llevado cursos en Gestión y Liderazgo Universitario por ANUIS; Formación de Gestores Universitarios por la Universidad Peruana Cayetano Heredia (UPCH). Es docente principal y directora de la Unidad de Formación Básica Integral de la UPCH. Ha sido vicedecana y directora de posgrado de la Facultad de Educación de la UPCH. Ha recibido la Orden Cayetano Heredia en la clase de Comendador y acreedora de la medalla de honor y miembro honorario del Colegio de Profesores del Perú. Ha sido considerada una de las personas más influyentes en la educación básica por el Grupo Educación al Futuro en el año 2022 y 2024. Ha sido reconocida con el Premio Excelencia Cayetano 2022.

Correo: liliana.munoz@upch.pe

ORCID: orcid.org/0000-0002-9791-7370

**** Naolly Casas Tolentino**

Ocupa el cargo de jefa de Gestión Docente en la Unidad de Formación Básica Integral (UFBI) de la Universidad Peruana Cayetano Heredia (UPCH) desde el 2023. Es Magíster en Educación con mención en Docencia e Investigación en la Educación Superior por la Universidad Peruana Cayetano Heredia (UPCH). Es Biotecnóloga de la Universidad Nacional del Santa (UNS). Cuenta con un diplomado en Gestión de la calidad de laboratorios de ensayos físico químicos en la Pontificia Universidad Católica del Perú (PUCP). Ha participado en proyectos de investigación relacionados a biología molecular en la identificación de polimorfismos (SSRs) y micropropagación de especies de plantas nativas del Perú en el Instituto de Biotecnología de la Universidad Nacional Agraria la Molina (UNALM). Actualmente, participa en un proyecto de investigación sobre evaluación de los aprendizajes y el perfil del docente universitario.

Correo: naolly.casas@upch.pe

ORCID: orcid.org/0000-0002-5509-2227

***** Jamine Pozu Franco**

Profesora asociada de la Universidad Peruana Cayetano Heredia (UPCH). Es comunicadora social y tiene especializaciones en educación a distancia por la Universidad Nacional de Educación a Distancia (UNED) y en habilidades docentes por el Tecnológico de Monterrey. También es graduada del Programa Interamericano de Formación en Gestión de Ambientes de Innovación de la Organización Universitaria Interamericana. Ha desempeñado actividades de gestión universitaria, como coordinadora del Centro Editorial, jefa de la Unidad de Educación a Distancia y del Observatorio e Incubadora de Programas Académicos, jefa del Departamento Académico de Educación, secretaria académica de la Escuela de Posgrado, secretaria académica y coordinadora académica en la Unidad Básica de Formación Integral (primer año de la universidad). Como experiencia profesional ha trabajado para organismos nacionales e internacionales como el Ministerio de Educación, Ministerio de Salud, Indecopi, OPS, Intervida y Harvard Business Publishing Education, entre otros. Su línea de investigación abarca áreas como didáctica, tecnologías e innovaciones en educación superior.

Correo: jamine.pozu.f@upch.pe

ORCID: orcid.org/0000-0003-0892-178X