Comparación de resultados obtenidos en exámenes de elección múltiple con y sin penalización por errores

Alfonso M. ALBANESE, Alicia B. MERLO, Adriana V. INGRATTA, Marta G. GÓMEZ, Eduardo F. ALBANESE

Objetivo. Comparar los resultados obtenidos en exámenes de elección múltiple con y sin penalización por errores.

Sujetos y métodos. Se evaluaron 1.059 alumnos de anatomía en exámenes por múltiple opción de 180 preguntas en una doble hoja de respuestas simultáneas, una sin penalización y otra con penalización para respuestas erróneas.

Resultados y conclusión. La comparación entre resultados de ambas hojas pone en evidencia las respuestas en convencimiento y en incertidumbre-azar que el mismo evaluado reconoce en una suerte de autoevaluación.

Palabras clave. Convencimiento. Elección múltiple. Evaluación. Incertidumbre-azar. Respuestas.

Comparison of results obtained in multiple choice exams with and without penalty for errors

Aim. To compare the results obtained in multiple choice exams with and without penalty for errors.

Subjects and methods. 1059 students of anatomy were evaluated in exams by multiple choice of 180 questions in double sheets of simultaneous answers, one without and another with a penalty for erroneous answers.

Results and conclusion. The comparison between the results of both sheets reveals the answers in conviction and in uncertainty-chance that the same evaluated recognizes in a sort of self-evaluation.

Key words. Answers. Conviction. Evaluation. Multiple choice. Uncertainty-random.

Cátedra de Anatomía Normal, de Superficie y por Imágenes e Instituto de Investigaciones. Facultad de Medicina. Universidad del Salvador. Buenos Aires, Argentina.

Correspondencia:
Dra. Alicia Beatriz Merlo. Cátedra de Anatomía Normal, de Superficie y por Imágenes e Instituto de Investigaciones. Facultad de Medicina. Universidad del Salvador. Avda. Córdoba, 1601. CP 1020. Buenos Aires, Argentina.

E-mail:
aliciabeatriz.merlo@gmail.com

Recibido:
23.04.19.

Aceptado:
30.05.19.

Conflicto de intereses:
No declarado.

Competing interests:
None declared.

Introducción

El examen por múltiple opción sin penalización de errores es muy usado en ciencias de la salud, tanto en el grado como en el posgrado y en concursos de oposición [1]. Es igualitario, con un mismo texto para todos los evaluados, puede abarcar gran amplitud de temas, es de rápida corrección y sus constancias escritas son fácilmente comparables [2], pero no permite saber si una respuesta correcta o errónea se realizó con pleno convencimiento, en incertidumbre o por azar.

En una evaluación sin penalizaciones, el evaluado habitualmente responde todas las preguntas, independientemente del grado de su conocimiento, pues no tiene nada que perder por sus errores. En una evaluación con penalizaciones habitualmente no se responden todas las preguntas y, por tanto, se pierde información de la modalidad de actuar del examinado cuando no hay riesgo de pena. Las causas por la que se abstiene a responder pueden relacionarse con factores que involucran el conocimiento o también con la personalidad del evaluado y su actitud ante el riesgo [3].

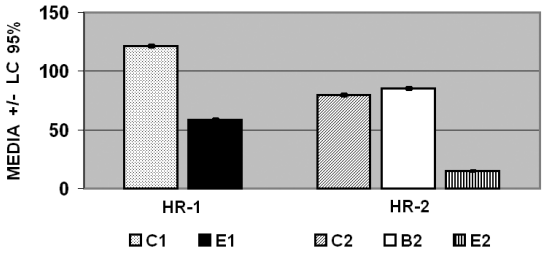
En el método de evaluación presentado, el evaluado debe responder en dos hojas de respuestas una misma batería de preguntas: una hoja sin penalización de las respuestas erróneas y otra con penalización. La diferencia entre el número de respuestas correctas así como de erróneas en ambas hojas evidencia el convencimiento o incertidumbre con que se respondieron las preguntas. El objetivo era comparar los resultados obtenidos en exámenes de elección múltiple con y sin penalización por errores.

Sujetos y métodos

Fueron evaluados por exámenes de múltiple opción 1.059 alumnos de la asignatura de anatomía del primer año de la carrera de medicina de la Universidad del Salvador a través de una batería de 180 preguntas (60 preguntas de cada uno de los tres exámenes formativos que cubrían la totalidad de los temas de la materia), elaboradas de acuerdo con los requerimientos de Tenbrink [4] y del National Board of Medical Examiners [5]. Cada pregunta constaba de una opción correcta y cuatro distractores.

ISSN: 2014-9832
ISSN (ed. digital): 2014-9840

Figura. Exámenes por múltiple opción: media ± límites de confianza al 95% (LC 95%) en 180 preguntas en hojas de respuesta sin penalizar (HR-1) y con penalización (HR-2). C: correctas; E: erróneas; B: en blanco.



La totalidad de los temas objeto de las preguntas se habían desarrollado profundamente en clase. Las preguntas fueron elaboradas cuidadosamente por los mismos profesores que dictaron la materia a fin de evitar ambigüedades en la interpretación de las opciones. La corrección de cada examen fue realizada por docentes de la materia y no por lectura óptica.

Las preguntas se contestaron en dos hojas de respuesta. En una, sin penalización, denominada HR-1, las puntuaciones asignadas a cada tipo de respuesta fueron: respuesta correcta (C1), 1 punto, y respuesta errónea (E1), 0 puntos. En otra hoja de respuesta, con penalización, denominada HR-2, las puntuaciones fueron: respuesta correcta (C2), 1 punto; respuesta errónea (E2), −1 punto, y pregunta no respondida o en blanco (B2), 0 puntos. Para el total de las evaluaciones, el nivel de dificultad media fue de 0,34 ± 0,01, y el índice de discriminación, de 0,43 ± 0,02.

Para cada alumno, en cada hoja de respuesta, se calculó el número de respuestas de cada tipo en las 180 preguntas y su porcentaje (180 preguntas = 100%) y el porcentaje del total de respuestas en blanco en la HR-2 que provenían de respuestas correctas y de respuestas erróneas de la HR-1. Teniendo en consideración que el número de respuestas de las pruebas de un número elevado de estudiantes tiende a seguir una curva en forma de campana de Gauss [6], se justifica el cálculo aplicado que indicamos a continuación.

Para dichos valores se determinaron las medias y se obtuvo por análisis de varianza la significación estadística entre el número de cada tipo de respuesta en ambas hojas de respuesta y entre los porcentajes entre respuestas correctas y erróneas en HR-1 que fueron en blanco en HR-2. Se calcularon los límites de confianza al 95% (LC 95%) de las medias que aportan información a la probabilidad de la significación estadística sobre la relevancia de los resultados [7].

Resultados

Los valores medios de cada tipo de respuesta que se muestran en la figura difieren significativamente ($p < 0{,}001$) entre sí de acuerdo al resultado del ANOVA y de los LC 95% de las medias que no se superponen entre sí. Como puede observarse en la figura, la media del número de respuestas correctas en HR-2 es aproximadamente dos tercios del correspondiente a las respuestas correctas en HR-1, lo cual muestra claramente el efecto de la penalización al indicar que, en HR-1, alrededor de un tercio de preguntas se respondieron correctamente en incertidumbre o por azar.

Las respuestas en incertidumbre-azar se distribuían de forma aleatoria en el grupo estudiado.

La figura muestra también que, en HR-2, la penalización de los errores produce abstención en un importante número de respuestas (respuestas en blanco), que en promedio es ligeramente mayor aunque significativo al número de las correspondientes respuestas correctas. Como consecuencia de la abstención, la media de respuestas erróneas en HR-2 es significativamente más baja que en HR-1.

Respecto a la procedencia de las respuestas en blanco en HR-2, en el grupo de exámenes estudiado, el 49,57 ± 0,73% de las respuestas en blanco provienen de respuestas correctas en HR-1 no respondidas en HR-2, y el 50,43 ± 0,73%, de las respuestas erróneas en HR-1 no respondidas en HR-2. Estos porcentajes no difieren significativamente ($p > 0{,}05$) entre sí de acuerdo al resultado del ANOVA. Los LC 95% de las respectivas medias se superponen entre sí.

A fin de generalizar los resultados para un número diferente de preguntas al utilizado en este estudio, la tabla muestra los valores (media ± LC 95%) de cada tipo de respuesta expresados como porcentajes del total (180 = 100% de respuestas).

Discusión

Los resultados muestran que el uso de dos hojas de respuesta simultáneas (una sin y otra con penalización de errores) para responder una misma batería de preguntas de opción múltiple arroja información complementaria que no se manifiesta en cada una por separado. La penalización disminuye en la HR-2

el número de respuestas correctas y de erróneas con relación a las de la HR-1. Indica que las repuestas en HR-2, tanto correctas como erróneas, se hacen con convencimiento de que se responde lo correcto, manteniendo prudencia ante la incertidumbre o el azar, lo que se puede calcular por diferencia con el número de las respuestas correspondientes en HR-1.

En el grupo de exámenes estudiado, el número de respuestas en incertidumbre o azar en HR-1, tanto correctas como erróneas, que pasan a integrar el 100% de respuestas en blanco en HR-2 lo hacen en proporción similar, lo que muestra que la incertidumbre al responder en HR-1 se manifiesta en las respuestas que se creen ser correctas o que se marcan por azar independientemente que resulten correctas o erróneas.

De acuerdo con Jurado-Núñez y Leenen [8], 'lo que se espera de cualquier método de evaluación educativa es que proporcione una medición lo más certera posible del rasgo latente que se pretende medir'. Incertidumbre y azar son factores que pueden obstaculizar la obtención de ese objetivo.

Los investigadores en el tema han propuesto fórmulas de corrección, como penalizar en fracciones de unidad las respuestas erróneas de acuerdo con el número de distractores utilizado [9] o bien penalizar de acuerdo con un distractor equivocado [10].

Estas estrategias son elaboradas por los investigadores en el tema sin participación de los examinados. En nuestro método de la doble hoja de respuesta, es el mismo examinado quien pone en evidencia, a través de sus abstenciones a responder en HR-2, la seguridad, dudas y azar con que respondió en HR-1. El método propuesto, a la vez, evidencia la gravedad del desconocimiento a través de las respuestas erróneas en HR-2. Como es el mismo evaluado quien reconoce la incertidumbre o azar con que respondió en HR-1, brinda al evaluador una suerte de autoevaluación y obliga al evaluado a tomar conciencia de su desempeño para desarrollar actitudes responsables.

En conclusión, las dobles hojas de respuesta simultáneas del método de exámenes por múltiple opción, en más de un millar de evaluados, proporciona datos complementarios de interés para la evaluación del examen y de la actitud del examinado frente al desafío y permite disponer de un patrón de comparación para detectar desviaciones de casos individuales a ser tratados por personal especializado para su orientación.

Tabla. Media ± límites de confianza al 95% (LC 95%) del porcentaje de respuestas individuales en evaluaciones sin y con penalización (100% de cada hoja de respuestas = 180 respuestas).

	C1	E1	C2	B2	E2
Media	67,46%	32,54%	44,34%	47,42%	8,24%
LC (95%)	0,73%	0,73%	0,75%	0,69%	0,36%

C: correctas; E: erróneas; B: en blanco.

Bibliografía

1. Reid W A, Duvall F, Evans P. Relationship between assessment results and approaches to learning and studying in year two medical students. Med Educ 2007; 41: 754-62.
2. Bauer D, Holzer M, Kopp V, Fischer MR, Pick N. Multiple choice-exams: a comparison of scoring algorithms. Adv Health Sci Educ Theory Pract 2011; 16: 211-21.
3. Bar-Hillel M, Budescu D, Attali Y. Scoring and keying multiple choice tests: a case study in irrationality. Mind & Society 2005; 4: 3-12.
4. Tenbrink TD. Evaluación. Guía práctica para profesores. 4 ed. Madrid: Narcea; 1987.
5. National Board of Medical Examiners. Cómo elaborar preguntas para evaluaciones escritas en el área de ciencias básicas y clínicas. 3 ed. Philadelphia: NBME; 2006. URL: https://www.nbme.org/PDF/IWG-Sp/IWG-Spanish2006.pdf.
6. Dawson-Saunders B, Trapp RG. Bioestadística médica. México DF: El Manual Moderno; 1991.
7. Molina-Arias M. El significado de los intervalos de confianza. Pediatría Atención Primaria 2013; 15: 91-4.
8. Jurado-Núñez A, Leenen I. Reflexiones sobre adivinar en preguntas de opción múltiple y cómo afecta el resultado del examen. Investigación en Educación Médica 2016; 5: 55-66.
9. Lesage E, Valcke M, Sabbe E. Scoring methods for multiple-choice assessment in higher education –is it still a matter of number right scoring or negative marking? Studies in Educational Evaluation 2013; 39: 188-93.
10. Suh Y, Bolt DM. Nested logit models for multiple-choice item response data. Psychometrika 2010; 75: 454-73.