Análisis de pruebas de opción múltiple en carreras de la salud de la Universidad Mayor

Elisa Giaconi[a,T,*], María Elisa Bazán[a,b,‡], Manuel Castillo[a,§], Amelia Hurtado[a,Δ], Héctor Rojas[Φ,(†)], Valentina Giaconi[c,ℓ], Ernesto Guiraldes[a,d,◊]

Facultad de Medicina

Resumen

Introducción: La Facultad de Ciencias de la Universidad Mayor utiliza regularmente pruebas de opción múltiple (POM) en la evaluación del aprendizaje en estudiantes de pregrado. Desde hace 10 años estas pruebas son corregidas con sistemas automáticos; esto permitió generar un diagnóstico global de las POM utilizadas.

Objetivo: Analizar el índice de dificultad, el de discriminación, el coeficiente de fiabilidad y los distractores de las POM aplicadas en las distintas carreras de la salud y corregidas de manera automatizada y centralizada en la Facultad de Ciencias de la Universidad Mayor, en el periodo 2013-2017.

Método: Para este estudio cuantitativo, descriptivo, no experimental, transversal y retrospectivo, de 2,640 pruebas corregidas por lector óptico, aplicadas por 7 carreras de la salud, se seleccionaron aleatoriamente 337 por muestreo probabilístico estratificado con selección sistemática. Se estimaron indicadores psicométricos en el marco de la Teoría Clásica de Test y se utilizó el test de ANOVA para la comparación entre carreras.

Resultados: Los índices de dificultad, de discriminación y el coeficiente de fiabilidad presentan promedios de 68%, 0.23 y 0.50 respectivamente. Se observaron diferencias significativas entre carreras en los índices de dificultad y discriminación. En relación a los distractores,

[a] Oficina de Educación en Ciencias de la Salud (OFECS), Facultad de Ciencias, Universidad Mayor, Santiago, Chile.
[b] Escuela de Kinesiología, Facultad de Ciencias, Universidad Mayor, Santiago, Chile.
[c] Instituto de Ciencias de la Educación, Universidad de O'Higgins, Rancagua, Chile.
[d] Escuela de Medicina, Facultad de Ciencias, Universidad Mayor, Santiago, Chile.
[(†)] El autor falleció en abril de 2021.
ORCID ID:
[†] https://orcid.org/0000-0003-0981-1590
[‡] https://orcid.org/0000-0001-8790-3576
[§] https://orcid.org/0000-0002-4614-2317
[Δ] https://orcid.org/0000-0003-0313-5982
[Φ] https://orcid.org/0000-0002-3169-6800
[ℓ] https://orcid.org/0000-0002-5166-5673
[◊] https://orcid.org/0000-0001-5535-8155

Recibido: 14-abril-2021. Aceptado: 8-julio-2021.
* Autor para correspondencia: Elisa Giaconi. Oficina de Educación en Ciencias de la Salud, Universidad Mayor, Camino La Pirámide 5750 Huechuraba, Santiago, Chile. Celular (+56 9) 91580562.
Correo electrónico: elisa.giaconi@umayor.cl

en promedio, 1.51 distractores son funcionales (1.52 para preguntas de cuatro alternativas, y 1.49 para preguntas de cinco alternativas).
Conclusiones: Los resultados develan situaciones críticas en evaluación de los aprendizajes, que demandan profundizar la reflexión de directivos y docentes de las distintas carreras para asegurar la validez y calidad de las POM.

Palabras clave: *Medición educacional; preguntas de exámenes; psicometría; control de calidad; formación médica de pregrado.*

Item analysis of multiple choice question tests in undergraduate health programs of Universidad Mayor

Abstract

Introduction: Multiple-choice question tests (MCTs) are widely used to assess undergraduate students' learning at the Health Sciences Schools (HSS) of our Universidad Mayor. In the last decade, the process of checking MCQs has been accomplished by optical mark recognition. This fact allowed this research study to obtain a diagnostic vision of the quality of this process.
Objective: To analyze the difficulty, discrimination, reliability and distractors of MCTs applied in seven HSS of Universidad Mayor during the 2013-2017 period.
Method: For this quantitative, descriptive, non-experimental, cross-sectional, and retrospective study, of the population of tests under study, i.e. 2640 MCTs, 337 were randomly selected by stratified probabilistic sampling with systematic selection. Psychometric indicators were estimated from the framework of Classical Test Theory. ANOVA tests were used to compare between programs.
Results: For item difficulty, item discrimination, and reliability coefficient, the respective means were: 68%, 0.23 and 0.50. Only for item difficulty and discrimination were there significant differences among the participating HSS. Regarding distractors, on average, 1,51 distractors were found to be functional (1,52 for items with four choices and 1,49 for items with five choices).
Conclusions: These results reveal that there is considerable room for improvement in the application of MCTs in the assessment of students' learning in our HSS. An in-depth reflection between faculty and university authorities should be carried out to ensure the future validity inferences and quality of MCTs.

Keywords: *Educational measurement; examination questions; psychometrics; quality control; undergraduate medical education.*

INTRODUCCIÓN

En las tendencias actuales de educación en ciencias de la salud, los programas adhieren a una formación por competencias y se utiliza una diversidad de instrumentos para evidenciar integradamente los aprendizajes. Las pruebas de opción múltiple (POM) contribuyen eficazmente con la evaluación de conocimientos y su aplicación, permitiendo evidenciar evolución de competencias[1,2]. Así, se requieren instrumentos que cuenten con garantías de validez y fiabilidad, lo cual demanda realizar evaluaciones sistemáticas de sus resultados, como la que se presenta en este trabajo.

Para facilitar la corrección de las POM, muchas universidades emplean procesos automáticos, generando una gran cantidad de datos, que pocas veces son analizados con el propósito de la mejora de estos procesos e instrumentos. Los sistemas automáticos reducen drásticamente el tiempo de corrección y el error humano, y proporcionan información para un análisis de la formulación de los ítems y el instrumento. Al respecto, numerosa literatura menciona y avala recomendaciones pertinentes[3-6].

El análisis de ítems posterior a un test puede usarse para mejorar la calidad de este, considerando los siguientes índices: dificultad, discriminación,

coeficiente de fiabilidad y distractores[7,8]. El índice de dificultad corresponde a la proporción de examinados que respondió correctamente cada pregunta. Además de la complejidad propia de cada pregunta, este puede verse afectado por defectos de construcción, como son planteamientos ambiguos y elementos irrelevantes. Por otra parte, el índice de discriminación de una pregunta refleja su capacidad para diferenciar estudiantes con buen desempeño o con bajo desempeño. Esto debería propiciar una retroalimentación diferenciada y medidas remediales a los estudiantes. Los distractores son opciones no correctas de respuesta[8-10], las que para ser funcionales deben ser verosímiles y no dar claves[5,11-13]. Finalmente, el coeficiente de fiabilidad de un test indica la consistencia (o precisión) del instrumento de medición[14]. Esta información es fundamental también como evidencia adicional de validez, que justifique las interpretaciones y usos de una evaluación[15].

Los estudios consultados que analizan estos índices post-aplicación, abordan generalmente evaluaciones específicas de determinadas asignaturas. Menos frecuentemente se encuentran estudios que aborden globalmente los índices obtenidos a nivel de una facultad o instituto y/o definan un sistema de seguimiento en que se utilicen estos datos como material para mejorar la calidad de las evaluaciones. Algunos ejemplos más recientes de investigaciones de este tipo, que analizan poblaciones grandes o moderadamente amplias de examinados, están a continuación. Abozaid et al[16] analizan en dos etapas tres exámenes finales sumativos en medicina, pediatría y cirugía, estudiantes de medicina, sexto año; Mehta y Mokhasi[17] estudian un examen de anatomía; Mukherje y Lahiri[18] revisan una prueba de medicina comunitaria, segundo año; Taib & Bahri[19] examinan una prueba de pediatría, medicina, cuarto año; Hingorjo y Jaleel[20] analizan un examen de fisiología, primer año Odontología; y Pérez et al[21] revisan una prueba de cátedra de cito-histología. En estos estudios se obtienen diversos resultados, que en su mayoría son considerados para definir mejoras a los instrumentos de evaluación. No siempre, de estudios como los mencionados, surge la proposición de formular un diagnóstico a nivel institucional (macro), el que puede permitir generar planes de mejora de mayor envergadura que los implementados a nivel solamente de asignaturas (micro) e involucrar a actores con distintos roles y con niveles de decisión más resolutivos. Esto último es una condición necesaria para que las evaluaciones promuevan efectivamente el aprendizaje de los estudiantes de manera sistemática y a nivel institucional[22].

OBJETIVO

En función de lo anterior, considerando que no hay suficientes estudios que analicen a nivel institucional macro la calidad de pruebas de opción múltiple y con el propósito de identificar oportunidades de mejora, se plantea como objetivo de nuestro estudio: Analizar índice de dificultad, de discriminación, coeficiente de fiabilidad y distractores de las POM aplicadas en las distintas carreras de la salud y corregidas de manera automatizada y centralizada en la Facultad de Ciencias de la Universidad Mayor en el periodo 2013-2017.

MÉTODO

Este es un estudio cuantitativo, descriptivo, no experimental, transversal y retrospectivo que considera las POM de 7 carreras de la salud, corregidas con un sistema automatizado en el periodo de 2013 a 2017 (5 años). Las carreras: enfermería, kinesiología (fisioterapia), medicina, nutrición, obstetricia, tecnología médica y terapia ocupacional, fueron anonimizadas y codificadas (desde C1 hasta C7).

Para determinar la muestra de POMs en estudio, se utilizó muestreo aleatorio estratificado con selección sistemática; el nivel de significación y error base asumidos fueron de 5%. De un total de 2,640 pruebas, el tamaño de muestra se estableció en n = 337 instrumentos. Respecto al número de estudiantes que respondieron las pruebas que conforman la muestra, el total corresponde a 18,759, lo que implica un promedio de 56 estudiantes por prueba.

Para evaluar los reactivos y las pruebas se tomó como marco la Teoría Clásica de Test, que permite estimar distintos indicadores psicométricos. Los indicadores utilizados para describir los reactivos de las pruebas fueron: dificultad (porcentaje de estudiantes que responde correctamente el ítem), discriminación (correlación punto biserial del puntaje del reactivo con el puntaje de la prueba) y funcionalidad de los distractores (distractor funcional es considerado al que fue respondido por más del 5% de los examinados). Los in-

Tabla 1. Promedio del índice de dificultad y discriminación según carrera

Carrera	Reactivos	Índice dificultad*[1]		Índice discriminación*[2]	
		Media	DE	Media	DE
C1	2,659	68.32	24.44	0.21	0.17
C2	1,713	63.87	23.67	0.24	0.15
C3	3,564	74.17	22.64	0.22	0.20
C4	1,058	66.29	23.54	0.26*	0.16
C5	1,138	68.96	25.11	0.23	0.18
C6	2,119	62.27	24.89	0.23	0.18
C7	1,178	68.54	23.20	0.25	0.19
Total	13,429	68.26	24.20	0.23	0.18

DE: Desviación estándar muestral.
* Existen diferencias significativas entre carreras
Valores recomendados:
Dificultad[1]: 50% a 75%[7].
Discriminación[2]: 0.1 a 0.5[7].

Tabla 2. Estadísticos descriptivos respecto a distractores funcionales y no funcionales

Tipo de reactivo	Reactivos con todos los distractores funcionales	Total distractores no funcionales		Total distractores funcionales	
	Porcentaje	Media	DE	Media	DE
4 alternativas	18.34%	1.48	0.98	1.52	0.98
5 alternativas	4.93%	2.51	1.14	1.49	1.14
Todas	14.11%	1.80	1.14	1.51	1.03

DE: Desviación estándar muestral.

dicadores utilizados para describir las pruebas fueron: coeficiente de fiabilidad (estimada con el coeficiente KR20), porcentaje de ítems con discriminación mayor a 0.1, y porcentaje promedio de respuestas correctas.

Se emplearon registros digitales de correcciones de POM realizadas con lector óptico. Se registró toda la información estadística de la prueba: número de estudiantes, número de preguntas, puntajes mínimos y máximos, coeficiente de fiabilidad, índice de dificultad, índice de discriminación, distribución porcentual de distractores, entre otros. Para la lectura, el procesamiento y análisis de los datos se utilizaron las plataformas Python, R y SAS. El análisis descriptivo contempló la estimación de estadísticos descriptivos (media y desviación estándar) y la realización de diagramas de caja para la visualización de la distribución de indicadores psicométricos. El análisis inferencial contempló la comparación de medias de los indicadores psicométricos en los grupos definidos por la carrera y el correspondiente año, por medio de un test ANOVA de una vía.

Consideraciones éticas

Se tiene la aprobación del Comité Ético Científico de la Universidad Mayor, siendo anonimizadas las carreras analizadas. El estudio no incluyó trabajo con humanos, solo se analizaron bases de datos secundarias innominadas.

RESULTADOS

En relación al número de reactivos presentes en las 337 pruebas, estos suman 13,429 reactivos, lo que implica un promedio de 40 por prueba. La **tabla 1** presenta los valores promedios y la desviación estándar de los índices de dificultad y discriminación de los reactivos según carrera. El índice de dificultad promedio de todas las carreras fue 68.3%. El índice de discriminación promedio de todas las carreras fue 0.23%. En ambas variables se encontraron diferencias significativas de medias entre las carreras.

La **tabla 2** presenta valores promedios y desviación estándar del número de distractores funcionales y no funcionales por reactivo. El análisis de

Tabla 3. Promedio de indicadores a nivel de prueba según carrera

Carrera	Pruebas	Porcentaje promedio de respuestas correctas*		Porcentaje de reactivos con discriminación mayor a 0.1*		Coeficiente fiabilidad*[3]	
		Media	DE	Media	DE	Media	DE
C1	61	70.27	9.37	78.21	10.86	0.48	0.21
C2	55	63.63	7.94	82.89	8.17	0.40	0.19
C3	76	74.38	9.76	73.74	13.65	0.57	0.19
C4	31	66.02	10.34	83.30	11.02	0.54	0.23
C5	27	69.79	12.55	78.43	11.42	0.51	0.23
C6	47	63.32	10.24	77.53	12.70	0.55	0.26
C7	40	68.32	6.11	78.60	9.66	0.45	0.23
Total	337	68.48	10.21	78.40	11.74	0.50	0.22

DE: Desviación estándar muestral.
* Diferencias significativas entre carreras.
La literatura, en general, recomienda un valor para el coeficiente de fiabilidad[3] mayor a 0.7[14], indicador kr20.

distractores, distinguió los reactivos de cuatro alternativas (68.5%) de los de cinco alternativas (31.5%).

El análisis de distractores consideró "distractor funcional" al respondido por más del 5% de los examinados. Se observa que los reactivos de 5 opciones tuvieron en promedio un distractor no funcional adicional, en relación a los de 4 alternativas.

Respecto a los distractores funcionales, los reactivos de 4 y 5 opciones presentaron la misma cantidad de ellos.

Las estadísticas descriptivas de indicadores sobre las pruebas se presentan en la **tabla 3**. El porcentaje promedio de respuestas correctas fue 68.48%. El porcentaje de ítems con discriminación mayor a 0.1, fue en promedio 78.4%. Finalmente, el coeficiente de fiabilidad promedio de todas las carreras fue 0.50. En todos estos indicadores se encontraron diferencias significativas de medias entre carreras. Es interesante notar que la carrera cuyas pruebas en promedio tienen menor coeficiente de fiabilidad (C2) presenta un porcentaje de ítems con discriminación mayor a 0.1 entre los más altos del grupo.

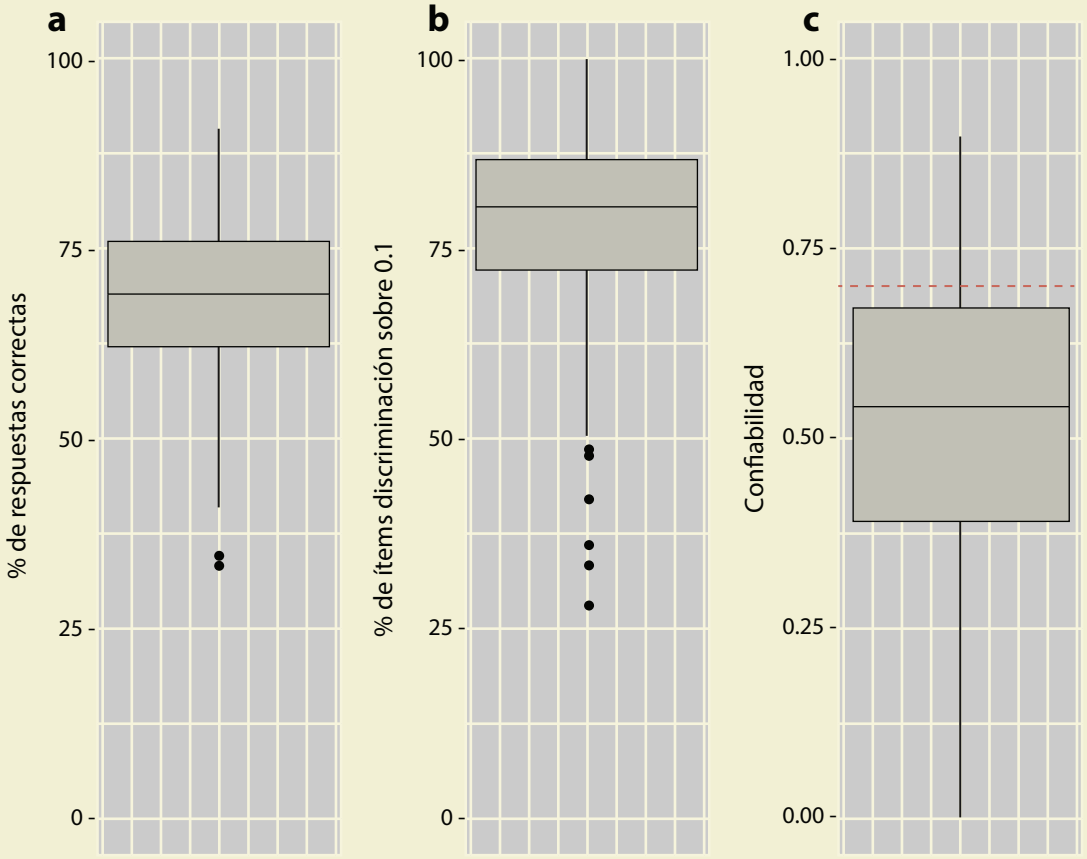
La **figura 1** muestra la distribución de los indicadores antes mencionados. Respecto al porcentaje promedio de respuestas correctas, se observa que la mayoría de las pruebas tiene sobre un 60% de respuestas correctas **(figura 1a)**. En relación al porcentaje de reactivos con discriminación menor a 0.1, se observa que la gran mayoría de las pruebas no logra que todos sus reactivos tengan discriminación aceptable, y que en más de un cuarto de las pruebas menos del 75% de los ítems tienen discriminación aceptable **(figura 1b)**. Finalmente, se observa que tres cuartos de las evaluaciones tienen un coeficiente de fiabilidad menor a 0.7, lo que evidencia una situación crítica **(figura 1c)**.

DISCUSIÓN

En las carreras de la salud, el conocimiento está en la base del desempeño clínico y la solución de problemas, y su evaluación es relevante. Las POM –bien elaboradas y pertinentes– son un método práctico y confiable para evaluar el conocimiento[2,23]. Su uso facilita la evaluación de gran cantidad de estudiantes en poco tiempo, brindando estandarización, objetividad y efectividad[5,9,10]. Las POM de calidad se basan en viñetas con situaciones clínicas realistas, que contribuyen a evaluar la aplicación del conocimiento y el uso de procesos cognitivos superiores, en lugar de la simple memorización de hechos[23]. Sin embargo, su uso sin aseguramiento de criterios de calidad puede generar consecuencias no deseadas en términos de las inferencias de validez de las evaluaciones y el aseguramiento de la adquisición de competencias de los estudiantes.

Respecto a las características psicométricas de las evaluaciones, en este estudio, el índice de dificultad promedio de las preguntas fue de 68.32% y está dentro de la categoría de dificultad media. Dependiendo del autor, se encuentran diferentes interpretaciones; para Tavakol y Dennick[7], y Violato y Violato[8] el rango de dificultad "media" es aquel de 50% a 75%, para Aubin et al[24], basándose en cálculos propios y

Figura 1. Distribución de indicadores en las pruebas analizadas



generados por la población que estudiaron, rotulan como "difícil" una media de 65%, como de "dificultad promedio", 80% y "fácil", 90%. Esto implica que la dificultad promedio observada en nuestro estudio es adecuada y además se observó una amplia variabilidad (DE = 24.2), lo que es deseable. Sin embargo, al no haberse realizado una evaluación del contenido de cada prueba –proceso de gran importancia en evaluación–, esta interpretación descansa solamente en los resultados numéricos globales correspondientes. En las POM es posible encontrar reactivos cuya mayor dificultad es artificial e irrelevante, producto de fallas en la construcción de la pregunta, que pueden comprometer la validez de inferencias del test[4,25-27]. Este tipo de situaciones, obviamente, no fue factible analizarlo en este estudio ni era su propósito.

El índice de discriminación de una pregunta, refleja la capacidad de esta para diferenciar entre los estudiantes con buen desempeño de aquellos con bajo desempeño[7,8,26,28]. Este estudio encontró un índice promedio de 0.23 considerando todas las carreras; este valor es mínimamente aceptable, en tanto se define como valor bajo a moderado al que está entre 0.1 y 0.5[7,14]. Los reactivos con distractores más efectivos (o verosímiles) en general discriminan mejor[10,29]. Se ha demostrado que reactivos confusos, ambiguos, con distractores poco verosímiles o de tipo verdadero/falso discriminan en forma insuficiente[10,29]. Si bien no existe una cifra que defina unánimemente el índice de discriminación óptimo de un ítem, se recomienda no utilizar aquellas preguntas con bajo índice (≤ 0.10 a ≤ 0.15), dependiendo del propósito

del test[8,14,26] y cumpliéndose el prerrequisito de que el número de preguntas de una prueba sea mayor a 40[21,40]. Se preferirá seleccionar preguntas que hayan demostrado índices de discriminación elevados si el test está referido fundamentalmente a normas, es decir, si está destinado a seleccionar los mejores candidatos dentro de una cohorte[14]. Por otra parte, esto no es absolutamente esencial si la prueba está referida a criterios, es decir a evaluar si el conjunto de los examinados maneja apropiadamente la mayoría de los contenidos de aprendizaje en un determinado programa[14]. Los resultados de este estudio indican que la discriminación de las preguntas es un área crítica en que se debe trabajar para mejorarla.

En relación a los distractores, se encontró que, en promedio, había aproximadamente 1.51 distractores funcionales por pregunta. Este resultado coincide con la literatura médica revisada: en promedio, solo 2 distractores resultan eventualmente ser funcionales[5,13]. Además, ello muestra un mejor comportamiento de las preguntas de cuatro alternativas, ya que a pesar de tener un distractor menos, presentaron prácticamente la misma cantidad de distractores funcionales que las de cinco alternativas. Esto reafirma investigaciones recientes sobre el número óptimo de distractores, que sugieren ampliamente que cuatro opciones (una respuesta "correcta" y tres distractores) son tan efectivas como cinco[6,29,30] (una repuesta "correcta" y cuatro distractores). Un buen distractor está destinado a identificar a los examinados con conocimientos parciales o superficiales[5,31]; por ello debe ser verosímil en forma y contenido y no dar claves indirectas que orienten a la respuesta[12,13]. A mayor cantidad de distractores no funcionales en un test, es menor el poder de discriminación de las preguntas[5,16]. Es decir, los resultados de estas aproximan –más que separan– a los estudiantes de mejor rendimiento a los de menor rendimiento, discriminando en contra de los primeros. Las preguntas que apelan solo a la memoria, suelen contener más distractores no funcionales que las que indagan funciones cognitivas más elevadas[30]. En tanto no se optimicen la calidad y el número de distractores en las POM, existirá, por consiguiente, un desperdicio considerable de opciones que son a la larga inútiles. Ello conlleva pérdida de tiempo y de insumos, además de provocar efectos pedagógicos regresivos.

El coeficiente de fiabilidad promedio (representativo de la consistencia interna de cada test), de todas las carreras fue 0.5. valor convencionalmente considerado bajo, ya que la norma habitual es calificar como deseable un coeficiente ≥ 0.7 en el ámbito educativo[14]. En algunos casos, un factor que pudo haber influido en este coeficiente bajo en nuestro estudio fue el nivel de discriminación que se observó en las preguntas correspondientes, ya que ambos indicadores están directamente relacionados. Otro aspecto a considerar es que se encuentra un mejor coeficiente de fiabilidad en las pruebas con mayor número de preguntas (> 50 o 60 o más), las que además permiten –en el análisis *post-hoc*– eliminar preguntas con fallas, sin afectar los parámetros de calidad del test. El promedio de reactivos por prueba fue de 40, es decir una cifra mínimamente garante de un adecuado coeficiente de fiabilidad. Puede postularse, por lo tanto, que en muchas de las pruebas analizadas, el bajo número de reactivos fue un factor que conspiró en contra de su mejor fiabilidad. Tavakol y Dennick[31] han analizado este tema en profundidad y han sugerido cautela en la interpretación literal de este coeficiente. Es importante agregar que una prueba puede poseer un alto coeficiente de fiabilidad, pero haberlo obtenido más que nada por redundancia de ítems, es decir por estar construido por reactivos muy semejantes entre sí y/o derivados de un único constructo.

En función de lo comentado, una sugerencia derivada del estudio es que es un deber de los docentes velar por la estructura general de las preguntas de sus pruebas, así como por una apropiada discriminación de estas. Igualmente, han de promover que en las POM se cuente con un número mínimo de preguntas, que garanticen su fiabilidad, una de las condiciones necesarias para argumentar en favor de la validez de los resultados de dichas pruebas[24].

La principal limitación de nuestro estudio fue no contar con información sobre el uso en cada carrera de las POM (formativo vs sumativo, criterios vs. normas), lo que permitiría obtener conclusiones más certeras sobre las implicancias de los resultados. Queda pendiente como proyección de este estudio profundizar en los resultados para diferenciar estos objetivos en las evaluaciones realizadas en nuestra

facultad, lo que permitirá propuestas de mejora diferentes en unos u otros casos.

En relación con las fortalezas del estudio, su principal aporte es haber contado con una amplia base de datos, que posibilitó derivar conclusiones robustas y pormenorizadas. Otra fortaleza es que permite reflexionar sobre los índices psicométricos de las evaluaciones aplicadas en diversas carreras de la salud y develar situaciones críticas, que demandan una profunda reflexión entre directivos y docentes del área de la salud, para asegurar validez y pertinencia de las POM[22] efectuadas al alero de la institución.

CONCLUSIONES

El panorama observado en este estudio muestra la relevancia de cuidar el desarrollo y calidad de las POM en las carreras de salud.

Los resultados psicométricos muestran que la discriminación de los ítems y la fiabilidad de las evaluaciones fueron áreas críticas que es necesario abordar con mayor profundidad. Hay muchas publicaciones que colaboran a diseñar buenas preguntas de opción múltiple en educación en ciencias de la salud[10-12,25,32,33]. Al elaborar preguntas, la revisión por pares calificados ayuda a mejorar las características psicométricas de dichos reactivos[16]. Es necesario implementar estas recomendaciones y evaluar si se presentan mejoras en los resultados psicométricos, por medio de un sistema de control de calidad de evaluaciones. Este sistema no es factible de ser desarrollado a nivel individual por los docentes, sino que debe estar organizado a nivel institucional, para promover políticas de evaluaciones de calidad[22].

Respecto a los distractores funcionales y no funcionales, en este artículo el análisis de 13,429 preguntas permite afirmar que, en promedio, en las preguntas de 5 alternativas el cuarto distractor no aporta ninguna ventaja. Esta conclusión tiene una gran relevancia ya que generar menos distractores baja la carga de trabajo para los constructores de ítems y de lectura para los examinados.

Entendiendo que la creación de POM de alta calidad requiere experiencia, conocimiento experto y gran cantidad de tiempo de dedicación, es que se postula que al menos aquellas evaluaciones de mayor trascendencia, como exámenes o evaluaciones complejas si debieran contar con el respaldo de expertos como las Unidades de Educación en Ciencias de la Salud[34].

Nuestras conclusiones aportan a las recomendaciones de La Federación Mundial de Educación Médica (WFME) que sugiere desarrollar un sistema de evaluación que permita asegurar a los estudiantes retroalimentación pertinente que identifique sus fortalezas y debilidades, ayude a consolidar su aprendizaje, y garantice el apoyo necesario para la toma de decisiones sobre la progresión y graduación de los estudiantes, por una parte, y mejore el desempeño del personal académico, los cursos y la institución, por otra[35].

CONTRIBUCIÓN INDIVIDUAL

- EG: Conceptualización, diseño, dirección y gestión del proyecto. Revisión bibliográfica y redacción manuscrito.
- MEB: Conceptualización, diseño y gestión del proyecto. Revisión bibliográfica y redacción manuscrito.
- MC: Conceptualización, diseño y gestión del proyecto. Revisión bibliográfica y redacción manuscrito.
- AH: Conceptualización y diseño.
- HR: Análisis estadístico y redacción del manuscrito.
- VG: Análisis estadístico y redacción del manuscrito.
- EG: Revisión bibliográfica y redacción manuscrito.

AGRADECIMIENTOS

Ninguno.

PRESENTACIONES PREVIAS

Congreso Internacional de Educación en Ciencias de la Salud 2019.

FINANCIAMIENTO

Universidad Mayor.

CONFLICTO DE INTERESES

Ninguno.

REFERENCIAS

1. Scallon G. L'évaluation des apprentissages dans une approche par compétences. Bruselas: De Boeck Université; 2004.

2. Epstein RM. Assessment in medical education. N Engl J Med. 2007;356(4):387-96.
3. Paniagua M, Swygert K. Constructing written test questions for the basic and clinical sciences. Philadelphia: National Board of Medical Examiners (US); 2016.
4. Wood T, Cole G, Lee C. Developing multiple choice questions for the RCPSC certification examinations. Ottawa Canada: Royal College of Physicians and Surgeons Canada; 2011.
5. Tarrant M, Ware J, Mohammed AM. An assessment of functioning and non-functioning distractors in multiple-choice questions: a descriptive analysis. BMC Med Educ. 2009;9(1):1-8.
6. Tarrant M, Ware J. A comparison of the psychometric properties of three-and four-option multiple-choice questions in nursing assessments. Nurse Educ Today. 2010;30(6):539-43.
7. Tavakol M, Dennick R. Post-examination analysis of objective tests. Med Teach. 2011;33(6):447-58.
8. Violato EM, Violato C. Multiple choice questions (MCQs) in a nutshell: Theory, practice, and post-exam item analysis. Acad Med. 2019;95(4):659.
9. McCoubrie P. Improving the fairness of multiple-choice questions: a literature review. Med Teach. 2004;26(8):709-12.
10. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Appl Meas Educ. 2002;15(3):309-33.
11. Coughlin PA, Featherstone CR. How to write a high quality multiple choice question (MCQ): A guide for clinicians. Eur J Vasc Endovasc Surg. 2017;54(5):654-8.
12. Brame CJ. Writing good multiple choice test questions. [Internet] Vanderbuilt Univ Cent Teach; 2013 [citado 2021 Marzo 11] Disponible en: https://bit.ly/3hYAObW
13. Jurado-Núñez A, Flores-Hernández F, Delgado-Maldonado L, Sommer-Cervantes H, Martínez-González A, Sánchez-Mendiola M. Distractores en preguntas de opción múltiple para estudiantes de medicina: ¿cuál es su comportamiento en un examen sumativo de altas consecuencias? Investig en Educ Médica. 2013;2(8):202-10.
14. Shultz KS, Whitney DJ, Zickar MJ. Measurement theory in action: Case studies and exercises. 2nd. ed. New York: Routledge; 2013.
15. American Educational Research Association, American Psychological Association, National Council on Measurement in Education. Estándares para Pruebas Educativas y Psicológicas (Original work published 2014). Lieve M, translator. Washington, DC: American Educational Research; 2018.
16. Abozaid H, Park YS, Tekian A. Peer review improves psychometric characteristics of multiple choice questions. Med Teach. 2017;39(sup1):S50-S54.
17. Mehta G, Mokhasi V. Item analysis of multiple choice questions-an assessment of the assessment tool. Int J Heal Sci Res. 2014;4(7):197-202.
18. Mukherjee P, Lahiri SK. Analysis of multiple choice questions (MCQs): Item and test statistics from an assessment in a medical college of Kolkata, West Bengal. IOSR J Dent Med Sci. 2015;1:47-52.
19. Taib F, Yusoff MSB. Difficulty index, discrimination index, sensitivity and specificity of long case and multiple choice questions to predict medical students' examination performance. J Taibah Univ Med Sci. 2014;9(2):110-114.
20. Hingorjo MR, Jaleel F. Analysis of one-best MCQs: the difficulty index, discrimination index and distractor efficiency. JPMA-Journal Pakistan Med Assoc. 2012;62(2):142-7.
21. Pérez Tapia JH, Acuña Aguilar N, Arratia Cuela ER. Nivel de dificultad y poder de discriminación del tercer y quinto examen parcial de la cátedra de cito-histología 2007 de la carrera de medicina de la UMSA. Cuad Hosp Clínicas. 2008;53(2):16-22.
22. Banta TW, Palomba CA. Assessment essentials: Planning, implementing, and improving assessment in higher education. San Francisco, CA: John Wiley & Sons; 2014.
23. Tangianu F, Mazzone A, Berti F, Pinna G, Bortolotti I, Colombo F, et al. Are multiple-choice questions a good tool for the assessment of clinical competence in Internal Medicine? Ital J Med. 2018;12(2):88-96.
24. Aubin A-S, 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):151-6.
25. Pugh D, De Champlain A, Gierl M, Lai H, Touchie C. Using cognitive models to develop quality multiple-choice questions. Med Teach. 2016;38(8):838-43.
26. Dory V, Allan K, Birnbaum L, Lubarsky S, Pickering J, Young M. Ensuring the quality of multiple-choice tests: An algorithm to facilitate decision making for difficult questions. Acad Med. 2019;94(5):740.
27. Meneses J, Barrios M, Bonillo A, Cosculluela A, Lozano LM, Turbany J, Valero S. Psicometría. Barcelona: UOC; 2014.
28. Hasty BN, Lau JN, Tekian A, Miller SE, Shipper ES, Merrell SB, et al. Validity evidence for a knowledge assessment tool for a mastery learning scrub training curriculum. Acad Med. 2020;95(1):129-35.
29. Kilgour JM, Tayyaba S. An investigation into the optimal number of distractors in single-best answer exams. Adv Heal Sci Educ. 2016;21(3):571-85.
30. Testa S, Toscano A, Rosato R. Distractor efficiency in an item pool for a statistics classroom exam: assessing its relation with item cognitive level classified according to Bloom's taxonomy. Front Psychol. 2018;9:1-12.
31. Tavakol, M, Dennick, R. Making sense of Cronbach's alpha. International journal of medical education. 2011;2:53.
32. Haladyna TM, Rodriguez MC, Stevens C. Are multiple-choice items too fat? Appl Meas Educ. 2019;32(4):350-64.
33. McCarty T. How to Build Assessments for Clinical Learners. En: Weiss Roberts L , editor. Roberts Academic Medicine Handbook. 2nd. Ed. Cham: Springer; 2020. p. 83-90.
34. Aguayo-Albasini JL, Atucha N. and García-Estañ J. Las unidades de educación médica en las facultades de Medicina y de Ciencias de la Salud en España, ¿son necesarias? Educación Médica. 2021;22:48-54.
35. World Federation for Medical Education (WFME). Basic medical education WFME global standards for quality improvement, The 2020 Revision. [Internet]. [Consultado 01 Jul 2021]. Disponible en: https://bit.ly/3hZf2EV