REVISTA ARGENTINA DE EDUCACIÓN MÉDICA

Artículo original

Di Bernardo, Juan José y Col.; Vol. 9 Nº 2 Septiembre, 54-62, 2020

Calidad psicométrica de los instrumentos de evaluación parcial con preguntas de estímulo y de respuesta

Di Bernardo, Juan José; Andino, Gerardo Marcelo; Urbanek, Luisa Carolina; Cardozo, Samantha; Mariño, Laura Cecilia; Navarro, Viviana de los Angeles

RESUMEN

Introducción: La asignatura Medicina Hombre y Sociedad tiene dos tramos de cursado con cinco evaluaciones sumativas parciales (3 individuales y 2 grupales) al final de cada tramo. Los exámenes individuales comprenden dos integradores de ejes, "Alimentación-Actividad física" y "Sistemas de salud-Ambiente", con preguntas de respuesta abierta (PRA), que se evalúan con rubricas sobre "pruebas patrón"; y un integrador de tramo con preguntas de opción múltiple (POM) que se califican con lector óptico. Para valorar la calidad de estas evaluaciones, se plantearon como objetivos: analizar las propiedades psicométricas y el alcance cognitivo de los instrumentos.

Materiales y métodos: Se analizaron las pruebas de un grupo de cursantes seleccionados aleatoriamente. Como variables psicométricas se calculó la confiabilidad (alfa de Crombach), los índices de dificultad (iP) y discriminación (iD) de cada ítem. Para valorar el alcance cognitivo se identificaron los procesos mentales que demanda responder cada pregunta (según taxonomía de Bloom).

Resultados: Se seleccionaron los resultados de 313 estudiantes, las pruebas abarcaron 50 PRA y 80 POM. La confiabilidad (Alfa: 0,74 vs 0,71) y los iP: 0,49±,08 vs 0,51±,19 fueron similares en PRA y POM. Hubo diferencias en los iD: 0,36±,10 vs 0,28±,18 p<0,000; y fuerte correlación entre los puntajes obtenidos en ambas pruebas (r=0,72 - IC95%:0,66-0,77). El alcance cognitivo de las POM requirió recordar, comprender y aplicar; mientras que las PRA con consignas como: comparar, esquematizar, graficar, integrar, justificar, etc, demandó además niveles de análisis, evaluación y creatividad.

Conclusiones: Ambos instrumentos son confiables y se complementan muy bien para el propósito de una evaluación sumativa parcial, las POM por el numero tienen mayor validez de contenido y las PRA más poder de discriminación y exploración de niveles cognitivos superiores, por lo que son herramientas útiles para evaluar competencias.

Palabras clave: educación médica - examen escrito - teoría clásica de los test - multiple choice - alcance cognitivo - bloom

INTRODUCCIÓN

La asignatura Medicina Hombre y Sociedad (MHS) es la primera materia de la Carrera de Medicina de la Universidad Nacional del Nordeste (UNNE) que los alumnos inscriptos deben regularizar y aprobar para ingresar como estudiantes regulares a dicha carrera.

El programa de MHS está orientado en competencias y enriquecido con contenidos de Medicina Familiar y Comunitaria, paradigma en el que está centrado el perfil de graduación. El cursado está estructurado en cuatro ejes temáticos: Alimentación, Actividad Física, Ambiente y Sistemas de Salud; y un eje de Contenidos Transversales dirigido a la formación humanística, científica y metodológica. Tiene una carga horaria total de 400 horas con dos tramos de cursado que finalizan cada uno en cinco evaluaciones sumativas parciales (dos grupales y tres individuales).[1]

Las instancias grupales abarcan, la evaluación de un informe de los "Trabajos en terreno" que los estudiantes realizan en pequeños grupos (observaciones sistemáticas, aplicación de encuestas, realización de entrevistas en la comunidad o a equipos de salud, análisis de los datos epidemiológicos publicados y otros); y una evaluación de "Contenidos transversales" (epistemología, antropología médica, metodología de estudio, informática e inglés) a través de diferentes actividades y recursos del aula virtual y de un e-portfolio.

Las evaluaciones individuales comprenden dos exámenes integradores de ejes: "Alimentación con Actividad física" (ALAF) y "Sistemas de salud con Ambiente" (SISAM), que consisten cada uno, en cinco ítems con cinco consignas o preguntas de respuesta abierta (PRA) por ítem, y que se evalúan por comparación con "pruebas patrón" (respuestas correctas esperadas) con rúbricas específicas (criterios de evaluación y asignación de puntajes), herramientas que son elaboradas para cada instancia por los docentes que diseñan las pruebas. El tercer instrumento individual es un examen integrador de tramo (EIT), cuestionario con 80 preguntas de estímulo, de opción múltiple (POM) formato Tipo A [2] ricas en contexto [3] que se califican con lector óptico.

Para valorar la calidad de estas evaluaciones y en el marco del Proyecto de Investigación PI 17CI01: "Calidad de los Sistemas de Evaluación de la Carrera de Medicina de la UNNE y su relación con el desarrollo de las Competencias" acreditado por SGCT-UNNE (Resolución N° 966/17-CS), se planteó como objetivo: analizar y comparar las propiedades psicométricas y el alcance cognitivo de los instrumentos de aplicación individual para la evaluación sumativa parcial en la asignatura MHS.

MATERIALES Y MÉTODOS

Este es un estudio de enfoque cuantitativo no experimental, con diseño transversal y alcance exploratorio y descriptivo.

Se analizaron los resultados de las tres pruebas individuales (ALAF, SISAM y EIT) del primer tramo de MHS (ciclo 2019) en un grupo de cursantes seleccionados aleatoriamente (alumnos de comisiones pares con número de orden impar).

Como variables psicométricas, aplicando la Teoría Clásica de los Test, se calculó la Confiabilidad (Fiabilidad) de las pruebas, y los Índices de Dificultad y de Discriminación de cada una de las preguntas o ítems. [4, 5]

La confiabilidad, entendida como la probabilidad del examen de arrojar un resultado similar cuando se aplica nuevamente al mismo grupo de estudiantes, se exploró calculando la consistencia interna de los instrumentos mediante el Coeficiente alfa de Crombach, tomándose como aceptables valores de alfa entre 0,70 y 0,90.

El índice de dificultad (iP) de cada pregunta, indica la proporción de estudiantes que la respondió correctamente y fue calculado con la ecuación correspondiente a POM o PRA de la Figura 1. Los valores se interpretaron según los niveles propuestos por Haldyna [6] en Muy fácil (iP= 0,92-1,00); Medianamente fácil (iP=0,76-0,91); Dificultad intermedia (iP= 0,45-0,75); Medianamente difícil (iP= 0,25-0,44); y Muy difícil (iP= 0,00-0,24).

El Índice de discriminación (iD) indica en qué medida esa pregunta permite diferenciar a los estudiantes evaluados que conocen ese contenido (saben más) de aquellos que no lo conocen (saben menos), y para calcularlo es necesario correlacionar el rendimiento del estudiante en cada pregunta con su rendimiento global en el examen (calificación total). Para ello se ordenaron los estudiantes (según la calificación obtenida) de mayor a menor y se seleccionaron dos grupos: el 27% superior (notas más altas) y el 27% inferior (notas más bajas)[5,7]. De esa forma se pudo calcular la diferencia entre las proporciones de aciertos entre ambos grupos de estudiantes aplicando la ecuación correspondiente a POM o PRA de la Figura 1. Los valores se interpretaron según la clasificación propuesta por Ebel (1965) (5) en Discriminación: Alta (iD ≥ 0,40); Aceptable (iD = 0,30 - 0,39); Baja (iD = 0,20 - 0,29); Mala (iD = 0,00 - 0,19); e Inaceptable (iD < 0,00)

Para valorar el alcance cognitivo de las evaluaciones se estimaron los procesos mentales que los estudiantes debían utilizar para responder cada pregunta, y se categorizaron aplicando la

nomenclatura de la nueva taxonomía de Bloom[8] que incluye en forma creciente seis niveles: recordar, comprender, aplicar, analizar, evaluar y crear.

Figura 1

Ecuaciones aplicadas para calcular los índices

	Índice de Dificultad	Índice de Discriminación
POM	$\frac{\text{N aciertos}}{\text{N respuestas}}$	$\frac{\text{« N aciertos »}\ 27\%\ \text{Sup} - 27\%\ \text{Inf}}{\text{N (27\%)}}$
PRA	$\frac{\text{media de puntos}}{\text{puntaje máximo posible}}$	$\frac{\text{« medias »}\ 27\%\ \text{Sup} - 27\%\ \text{Inf}}{\text{puntaje máximo posible}}$

Tratamiento Estadístico:

Los datos se expresan en valores absolutos, porcentajes y proporciones según corresponda. Como medidas de resumen se calcularon los valores de tendencia central y de dispersión. Para la presentación de los datos resumidos se utilizan tablas de columnas. Para la representación gráfica de las variables se utilizan diagramas de caja (box plots), que permiten mostrar una serie de datos numéricos a través de sus cuartiles identificando claramente la media, la mediana, y los valores que se encuentran fuera de los límites aceptables; además gráficos de puntos para mostrar la relación entre dos variables.

Para los análisis estadísticos se utilizó el software XLSTAT (Addinsoft), se realizaron pruebas de “t” de Student tomando como nivel de significancia valores de $p < 0.05$. Para la correlación de las variables se utilizó el coeficiente de correlación (r) de Pearson y se realizó análisis de regresión calculándose el coeficiente de regresión (r), de determinación (r2) y la curva de ajuste con los intervalos al 95% de confianza, asignándose $p<0.05$ como nivel de significación para el ajuste.

RESULTADOS

Se seleccionaron los resultados de las tres evaluaciones individuales de 313 estudiantes que representan el 23% de los cursantes (n: 1362) que completaron el Primer Tramo de MHS en 2019. Las pruebas integradoras de ejes (ALAF y SISAM) se analizaron en forma conjunta, reuniendo así diez ítems con cinco consignas o preguntas cada uno, que totalizaron 50 PRA. Del examen integrador de tramo (EIT) se obtuvieron 80 POM ricas en contexto con tres opciones.

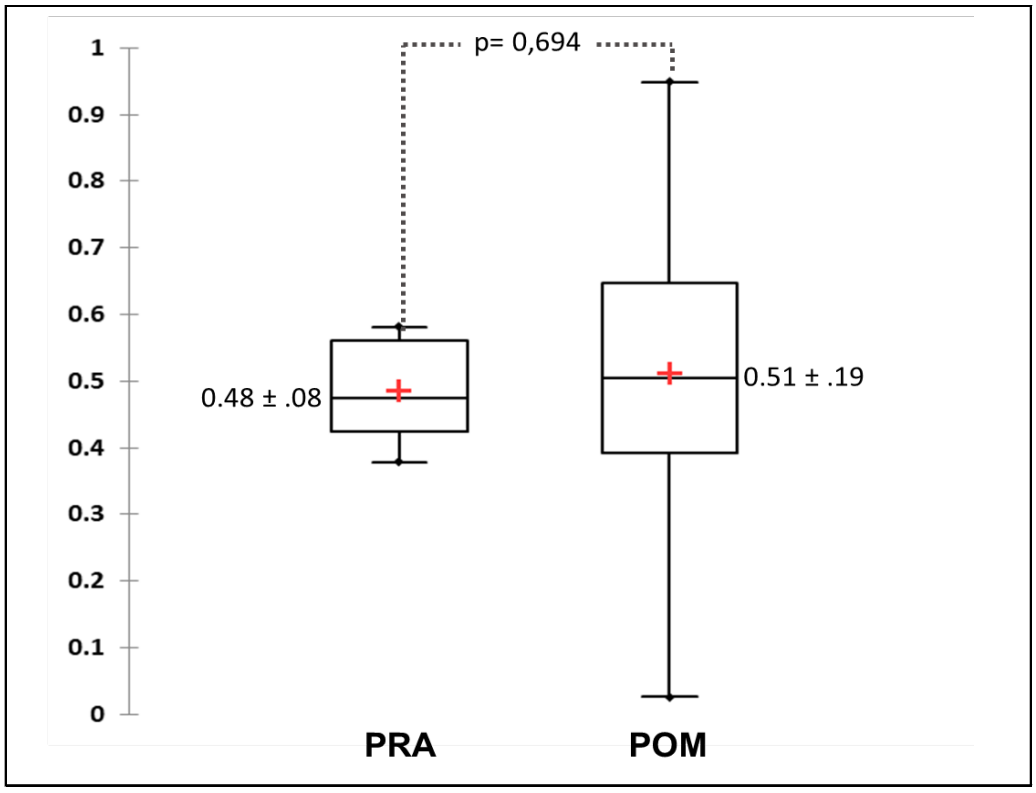
Variables psicométricas

El coeficiente de confiabilidad fue similar en los dos tipos de prueba (PRA y POM), Alfa de Crombach de 0,74 y 0,71 respectivamente y los valores aceptables.

Los índices de Dificultad mostraron también medias muy similares en los dos formatos, en las PRA un iP= 0,49 ± 0,08 y en las POM un iP= 0,51 ± 0,19. (Gráfico 1)

Gráfico 1

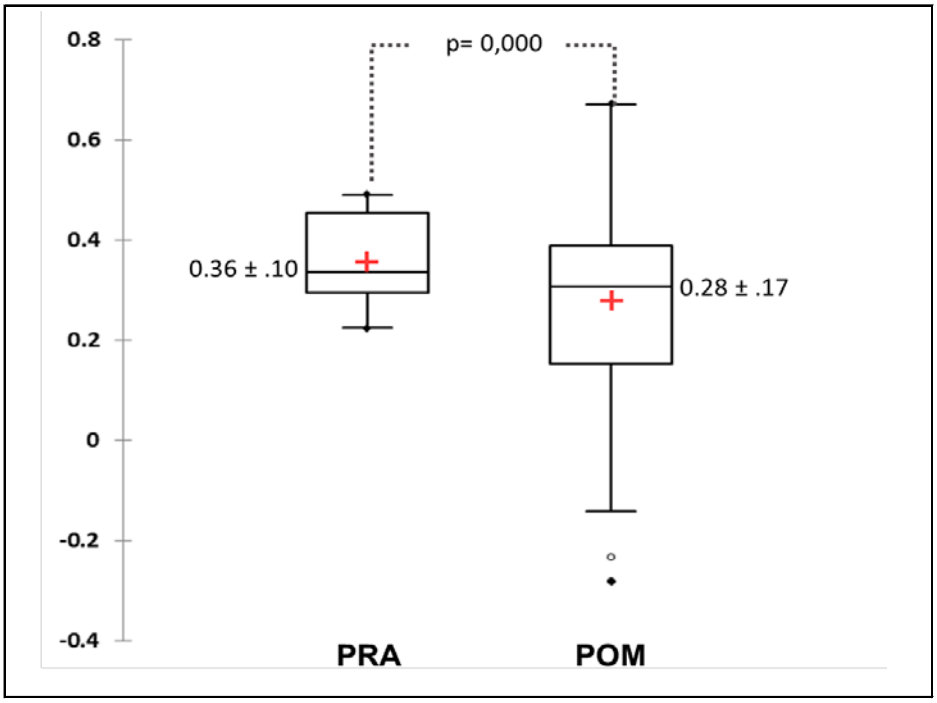
Comparación de los Índices de Dificultad de las preguntas



Sin embargo, en los Índices de Discriminación hubo diferencias significativas entre las PRA y las POM, iD: 0,36 ± 0,10 vs 0,28 ± 0,18 ($p < 0,000$) respectivamente. (Gráfico 2)

Gráfico 2

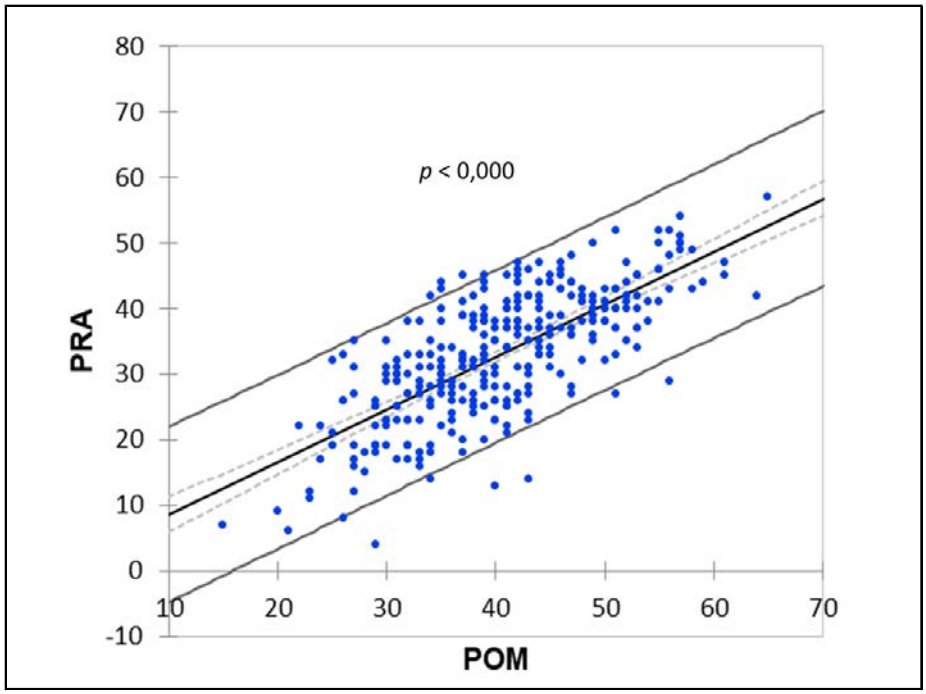
Comparación de los Índices de Discriminación de las preguntas



Además se comprobó una fuerte correlación positiva con r= 0,72 (IC95%: 0,66 - 0,77) entre los puntajes obtenidos en ambas pruebas. (Gráfico 3)

Gráfico 3

Regresión de puntajes de PRA por POM (R^2 = 0.53)



Alcance cognitivo de las preguntas

Analizando las 80 POM del EIT se observó que todas incluyen un breve enunciado relacionado a contenidos de la asignatura que fueron redactados como: viñeta clínica; experiencia (observación) de laboratorio; problema social; situación sanitaria; proceso biológico; mecanismo bioquímico; fenómeno de la naturaleza; teoría científica; suceso histórico; y referencia antropológica. Debajo de cada enunciado se plantea una pregunta con tres opciones de respuesta (solo una es correcta), que lleva a los estudiantes a realizar algunos de estos procesos: Recordar (conceptos, definiciones, propiedades, modelos); Identificar (elementos, compuestos, estructuras, funciones); Interpretar (datos, formulas, tablas, gráficos, mecanismos); Reconocer (características, patrones, factores, categorías); Ordenar (elementos, números, etapas). En consecuencia, el alcance cognitivo de las POM abarca los tres primeros niveles de la Taxonomía de Bloom: Recordar, Comprender y Aplicar. (Tabla 1)

Tabla 1
Alcance cognitivo de cada pregunta según Taxonomía de Bloom

Niveles taxonómicos	POM (n: 80)		PRA (n: 50)	
	n	%	n	%
Recordar	44	55%	6	12%
Comprender	24	30%	15	30%
Aplicar	12	15%	10	20%
Analizar			9	18%
Evaluar			10	20%
Crear			0	0%

Los ítems de los exámenes integradores de ejes utilizaron como disparadores, en el ALAF: etiquetas nutricionales, relato que introduce al metabolismo hidroelectrolítico, listado de nutrientes, dibujo esquemático de la fibra muscular, gráfico de producción de energía en ejercicio; y en el SISAM: familigramas, problema relacionado a determinantes de salud, relato que introduce a la división y ciclo celular, casos de salud en diferentes etapas de la vida, problemas del sistemas de salud. Cada uno de estos ítems comprenden cinco PRA con consignas que llevan a los estudiantes a realizar algunos de estos procesos: interpretar, identificar, ordenar, comparar, seleccionar, combinar, esquematizar, graficar, integrar, valorar, justificar. La utilización de estos procesos dan a las PRA un alcance cognitivo que abarca cinco niveles de la Taxonomía de Bloom: Recordar, Comprender, Aplicar, Analizar y Evaluar. (Tabla 1)

DISCUSIÓN

El sistema de evaluación que aplica la asignatura Medicina Hombre y Sociedad fue reestructurado a partir de 2016 articulando diferentes instrumentos para poder evaluar todo el espectro de las competencias, que es el modelo curricular vigente en la Carrera de Medicina. Los instrumentos de evaluación fueron seleccionados con el propósito de explorar: los dominios declarativos a través de pruebas escritas estructuradas y abiertas (formatos de estímulo y de respuesta); los dominios procedimentales a partir de los informes de los trabajos en terreno, análisis de los datos y diagnósticos de situación; y los dominios orientados a la metacognición mediante un e-portfolio que sirve como instrumento de reflexión y consolidación de los aprendizajes.[1]

Este trabajo de investigación estuvo enfocado solo en una parte del sistema de evaluación, que son los exámenes escritos individuales, y aunque los resultados no permiten valorar en su totalidad dicho sistema, puede servir para mejorar la calidad de estas evaluaciones.

Los dos tipos de pruebas analizadas (POM y PRA) funcionaron como herramientas confiables (Alfa de Crombach > 0,70), lo que es esencial en el contexto que se aplican, pues el propósito de la evaluación es sumativo y los estudiantes deben aprobar ese tramo de cursado para regularizar la asignatura que les posibilita el ingreso a la carrera de medicina. La confiabilidad no es tan importante si se hubiera tratado de evaluaciones formativas o para retroalimentar el aprendizaje.[9]

Tanto las POM como las PRA tuvieron en su mayoría índices Dificultad Intermedia (iP entre 0,45 y 0,75), que es lo recomendado para este tipo de pruebas por los diferentes autores [2, 4, 5, 6]. Las preguntas muy fáciles (iP > 91) o muy difíciles (iP < 24) no brindan mucha información sobre los estudiantes evaluados y podrían indicar que el contenido de las preguntas no corresponde con los conocimientos de los mismos[2]. Una pregunta muy difícil puede indicar que los estudiantes no conocen el contenido, pero también que la pregunta es ambigua, que está mal formulada, que la clave de corrección está equivocada o que hay más de una respuesta correcta4. Para Bonillo[5]: "Si todos los evaluados aciertan una pregunta es como si regaláramos a todos una parte del puntaje; y si todos fallan es como si los penalizáramos".

Los índices de Discriminación fueron aceptables (iD entre 0,30 y 0,39) en la mayoría de las PRA a diferencia de las POM que mayoritariamente mostraron índice bajos o malos. Esto puede explicarse por el formato de la pregunta, pues las PRA requieren "elaborar" la respuesta, mientras que en las POM hay que "elegir" una respuesta, y esto puede hacerse también por azar. En consecuencia, las PRA permitieron diferenciar mejor que las POM, a los estudiantes bien preparados (sabían más) de aquellos poco preparados (sabían menos), que es una de las finalidades de la evaluación sumativa. Sin embargo hay que tener presente que el poder de discriminación depende del grado de dificultad de la pregunta, específicamente de su variancia[5, 7], las preguntas más discriminantes no son por lo general las más fáciles ni las más difíciles, sino las dificultad intermedia[4].

Por otro lado, relacionando los puntajes obtenidos por los cursantes en los dos tipos de pruebas, se observó una fuerte correlación positiva, lo que refleja la concordancia en el rendimiento de los estudiantes expuestos a ambos formatos de preguntas.

El alcance cognitivo de las preguntas de estímulo y de respuesta también fue diferente, las POM solo alcanzaron los tres primeros niveles de la Taxonomía de Bloom, mientras que las PRA tuvieron un alcance más amplio que abarcó cinco niveles (Recordar, Comprender, Aplicar, Analizar y Evaluar). Esto es muy importante pues requerir niveles más altos de habilidades cognitivas conlleva un aprendizaje profundo y a la transferencia de conocimientos y habilidades a una mayor variedad de tareas y contextos[8]. Además, hay evidencia que este tipo de pruebas pueden mejorar el rendimiento posterior

de los estudiantes, permitiéndoles identificar lo que han aprendido y lo que deben aprender, potenciando así la eficiencia de las oportunidades de estudio posteriores.[10]

CONCLUSIONES

Ambos formatos de instrumentos son confiables y se complementan muy bien para el propósito de una evaluación sumativa parcial, las POM por ser más numerosas tienen mayor validez de contenido y las PRA más poder de discriminación y exploración de niveles cognitivos superiores. En consecuencia, la aplicación combinada de preguntas de estímulo y de respuesta, son herramientas muy útiles para la evaluación de competencias.

BIBLIOGRAFIA

1. Di Bernardo, JJ. Navarro, V. Fernández, G. Demuth Mercado, P. Larroza, GO. Medicina, Hombre y Sociedad: adecuando el ingreso a medicina al modelo de competencias y al perfil de graduación. Revista de la Facultad de Medicina de la UNNE, 2017; 37 (3):5-14.
2. Paniagua, MA. Swygert, KA. (Editores). Cómo elaborar preguntas para evaluaciones escritas en las áreas de ciencias básicas y clínicas. Philadelphia. National Board of Medical Examiners (NBME). Cuarta Edición (2016). En https://www.nbme.org/sites/default/files/2020-01/DownloadingtheGoldBook_ES.pdf. Acceso: 31/03/2018
3. Schuwirth, LWT. van der Vleuten, CPM. Different written assessment methods: what can be said about their strengths and weaknesses? Med Educ 2004; 38: 974–979
4. Morales, P. Análisis de ítems en las pruebas objetivas. Madrid: Universidad Pontificia Comillas (2012). En http://educrea.cl/wp-content/uploads/2014/11/19-nov-analisis-de-items-en-las-pruebas-objetivas.pdf. Acceso el 02/10/2014.
5. Bonillo, A. Análisis de los Items. Universitat Oberta de Catalunya. PID: 00198631. Psicometría febrero 2013. En http://openaccess.uoc.edu/webapps/o2/bitstream/10609/69325/7/Psicometr%C3%ADa_M%C3%B3dulo%205_An%C3%A1lisis%20de%20los%20%C3%ADtems.pdf. Acceso: 22/06/2018
6. Haladyna, TM. Downing, SM. Rodriguez, MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education. 2002; 15: 309-334
7. Rodriguez, MC. Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. Educational measurement: issues and practice 2005; 24(2): 3-13.
8. Adams, NE. Bloom's taxonomy of cognitive learning objectives. J Med Libr Assoc. 2015; 103(3):152-153. doi:10.3163/1536-5050.103.3.010
9. Brailovsky, CA. Educación Médica, Evaluación de las competencias. En Aportes para un Cambio Curricular en Argentina 2001. OPS y Facultad de Medicina, UBA; pag.103-122. En http://www.fmv-uba.org.ar/posgrado/proaps/9.pdf. Acceso el 10/09/2008.
10. Marsh, EJ. Roediger, HL. Bjork, RA. Bjork, EL. The memorial consequences of multiple-choice testing. Psychonomic Bulletin & Review 2007; 14 (2):194-199

DATOS DE AUTOR

Título
Calidad psicométrica de los instrumentos de evaluación parcial con preguntas de estímulo y de respuesta

Autores:
Di Bernardo Juan José *
Andino Gerardo Marcelo *

Urbanek Luisa Carolina *
Cardozo Samantha *
Mariño Laura Cecilia *
Navarro Viviana de los Angeles *
*Facultad de Medicina. Universidad Nacional del Nordeste. Corrientes. Argentina

Titulo abreviado: Calidad psicométrica de evaluaciones parciales
Número total de palabras: 2527 (sin bibliografía)
Autor para correspondencia: Juan José Di Bernardo
Correo electrónico de contacto: jjdibernardo@med.unne.edu.ar