TRABAJO ORIGINAL

Proceso de mejoría de pruebas de conocimiento con preguntas de selección múltiple en un curso teórico de pregrado de medicina.

EDUARDO KATTAN T.[*a], GONZALO PÉREZ D.[*a], CATALINA LE ROY O.[****a], MARISOL SIRHAN N.[**,***a], AGUSTÍN GONZÁLEZ C.[*b], TOMÁS RYBERTT L.[*b], LUZ COLLINS V.[**c], NANCY SOLÍS L.[***d], MARGARITA PIZARRO R.[***d], MARCO ARRESE J.[***a] y ARNOLDO RIQUELME P.[**,***e]

RESUMEN

Introducción: La metodología de evaluación ha tenido un crecimiento exponencial en los últimos años. La evaluación escrita con preguntas de selección múltiple (PSM) sigue siendo el instrumento más utilizado para evaluar conocimientos.

Objetivos: Evaluar la calidad de las PSM durante un proceso de implementación de mejorías en la calidad de la construcción de pruebas de conocimiento.

Material y Método: Se analizó la calidad de las PSM utilizadas en las evaluaciones escritas de un curso teórico del Pregrado de Medicina (período 2002-2005) de acuerdo a: validez (contenido, construcción, y predictiva), % de PSM con violaciones de los principios de escritura (criterios de Haladyna) y niveles cognitivos explorados (taxonomía de Bloom). Se utilizó una pauta de construcción de pruebas, con énfasis en la cobertura de los contenidos y el mapeo de los objetivos del curso. En 2006 se crearon PSM nuevas, se evaluó calidad de las PSM y confiabilidad del instrumento según método de Cronbach. Las comparaciones se realizaron con Test Z y corrección Bonferroni posthoc.

Resultados: La cobertura de temas subió progresivamente hasta el 100% (p = 0,001). En el período 2002-2005, un 54% de las PSM presentaba al menos una violación de los principios de escritura, disminuyendo a un 18,8% (p = 0,001) el año 2006. Paralelamente, se observó un aumento significativo de los niveles cognitivos superiores de Bloom (p < 0,001).

Conclusiones: El uso de una pauta para la construcción de una prueba teórica con PSM, así como la revisión sistemática de las PSM de acuerdo a los criterios de Haladyna por educadores médicos con formación en evaluación, permitió ampliar la cobertura de los temas del curso evaluado y mejorar la validez y confiabilidad del instrumento, explorando niveles cognitivos superiores.

Financiamiento: FONDECYT Proyecto N° 1120652 (A.R.).

Palabras clave: Preguntas de elección múltiple, Evaluación de conocimientos, Estudiantes de medicina, Pregrado.

SUMMARY

Improvement process in knowledge tests with multiple-choice questions in a theoretical course of medical undergraduate.

Introduction: Methodology related to assessment has experienced an exponential growth in the last years. The written evaluation with multiple-choice-questions (MCQ) is still the leading instrument used to assess knowledge.

Objectives: To evaluate the quality of MCQ, during the implementation of a quality improvement process of cognitive assessment instruments.

Recibido: el 27/04/14, Aceptado: el 10/06/14.

* Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile.
** Centro de Educación Médica, Pontificia Universidad Católica de Chile, Santiago, Chile.
*** Departamento de Gastroenterología, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile.
**** Departamento de Pediatría, Escuela de Medicina, Pontificia Universidad Católica de Chile, Santiago, Chile.
a Médico.
b Estudiante de Medicina.
c Profesora.
d Bioquímico.
e Médico, Magíster en Educación Médica.

Material and Method: Quality of MCQ used in the period between 2002-2005, for written assessment of a theoric lecture in undergraduate of Medical school was analyzed according to: validity issues (content, constructive and predictive validity), the percentage of MCQ with writing principles violations (Haladyna's criteria) and cognitive levels explored (Bloom's taxonomy). Test construction blueprint was used to ensure the coverage of the course themes and objectives. Through year 2006, a new set of MCQs were created, in which the quality MCQs and reliability (Cronbach's alpha) were evaluated. Comparisons were made with Z test and Bonferroni posthoc correction.

Results: The coverage of themes progressively increased arising 100% (p = 0.001). During years 2002-2005, 54% of MCQs presented at least one kind of violation of writing principles. In the year 2006, a significant improvement was observed decreasing to 18.8% (p = 0.001). According to Bloom's taxonomy assessment, there was a significant increase in the proportion of MCQs exploring higher cognitive levels (p < 0.001).

Conclusions: Using a blueprint for the construction of a cognitive MCQ test, combined with the evaluation of MCQ by medical educators trained in evaluation, demonstrate a significant improvement of the coverage of the themes of a course, increasing the content validity of the MCQ, and exploring higher cognitive levels.

Key words: Multiple-Choice-Questions (MCQ), Knowledge assessment, Medical students, Undergraduate.

INTRODUCCIÓN

Durante el año 1997, se estructuró la reforma curricular de la Escuela de Medicina de la Pontificia Universidad Católica de Chile (EMPUC)[1], en un intento por mejorar las estrategias educacionales descritas en modelo SPICES[2]. A diferencia del modelo anterior, el cual estaba compartimentalizado y sobrecargado de contenidos, se incluyen estrategias educacionales innovadoras[3]. El nuevo modelo se caracterizó por ser centrado en el alumno, sistemático, basado en la comunidad y en la resolución de problemas, integrado, con un núcleo de contenidos central obligatorio y cursos optativos de profundización. Este modelo se construyó a partir de los aspectos destacados en la declaración de Edimburgo[4].

Esta nueva reforma curricular se basó en un enfoque constructivista, en donde el conocimiento se construye a partir de conocimiento previo obtenido, siguiendo un modelo iterativo[5]. Dentro de los cambios implementados en la EMPUC, destacan la reducción del volumen de contenidos impartidos, la promoción de la integración vertical y horizontal, modernización de las metodologías de enseñanza a través de sesiones tutoriales, incorporación de tecnologías de la información, participación activa de los alumnos y educación orientada a resolución de problemas[1].

Existen tres cursos dentro de las fases preclínicas y clínicas del nuevo currículo, según el modelo constructivista, que cumplen la función de integración vertical (entre conocimientos de ciencias básicas y clínicos) y horizontal (entre disciplinas clínicas). Estos cursos integrados abarcan tanto la adquisición de conocimiento, como habilidades clínicas y actitudinales en relación al nivel de formación del alumno. El curso integrado de cuarto año (CICA), es un curso médico-quirúrgico, orientado en el aprendizaje de enfermedades, con un fuerte enfoque en el proceso diagnóstico con integración vertical con las bases fisiopatológicas del curso de tercer año, y los fundamentos terapéuticos (farmacológicos y no farmacológicos) que serán desarrollados con mayor profundidad en quinto año. El CICA se compone de 15 módulos teóricos lectivos, y práctica clínica basada en atención hospitalaria y policlínicos de atención ambulatoria.

El módulo de Gastroenterología es parte de las actividades del primer semestre de cuarto año. Las actividades de aprendizaje incluyen clases lectivas y seminarios orientados a resolver problemas clínicos. Los objetivos y temas del curso están descritos en las Tablas 1 y 2.

La evaluación del CICA se compone de pruebas objetivas para cada módulo teórico basado en preguntas de selección múltiple (PSM), fichas clínicas de pacientes hospitalizados, actividades en policlínicos, un examen clínico objetivo estandarizado (ECOE) y un examen teórico global al final del curso. Posteriormente, se incorporó el uso de Portafolio de casos clínicos reales y virtuales como herramienta de evaluación para promover actitudes vinculadas al profesionalismo[6].

Según Bloom, existen tres dominios mayores de la educación: cognitivo, psicomotor y afectivo. Él describe que el dominio cognitivo está estructurado jerárquicamente. En la base se encuentra el conocimiento, luego la comprensión, aplicación, análisis, síntesis y finalmente la evaluación[7].

Para evaluar este dominio, se han desarrollado diversos sistemas de calificación, incluyendo pruebas orales, pruebas de ensayo, preguntas de verdadero/falso y PSM. Estas últimas son las más utilizadas actualmente y presentan dentro de sus fortalezas alta objetividad, fiabilidad y validez de contenido, son más fáciles de evaluar, versátiles y reutilizables[8].

Los sistemas de calificación pueden ser muy diversos. Es importante, sin embargo, intentar determinar si realmente cumplen sus objetivos. Según Van der Vleuten, diversos factores influyen en la utilidad de una herramienta de calificación, siendo los más importantes la validez y la fiabilidad de ésta[9].

La validez significa que una prueba mida lo que se supone que debería medir. Si el instrumento de evaluación no posee validez, pierde sentido su implementación. La extensión en la cual el contenido ha sido en-

Tabla 1. Objetivos del curso.

1. Identificar e interpretar los síntomas y hallazgos al examen físico principales de pacientes con enfermedades gastrointestinales y hepáticas.
2. Reconocer las características clínicas requeridas para el proceso diagnóstico de enfermedades gastrointestinales y hepáticas prevalentes.
3. Describir los mecanismos fisiopatológicos subyacentes relacionados con enfermedades gastrointestinales y hepáticas prevalentes.
4. Evaluar la utilidad de exámenes de laboratorio (sangre y deposiciones), endoscopía, y exámenes radiológicos en el diagnóstico y terapia de enfermedades gastrointestinales y hepáticas.
5. Diseñar diferentes aproximaciones al manejo de los pacientes, incluyendo terapias de enfermedades gastrointestinales y hepáticas prevalentes, identificando los efectos adversos más frecuentes relacionados a los medicamentos y procedimientos prácticos usados comúnmente en gastroenterología.
6. Describir los principios básicos relacionados con la terapia de enfermedades gastrointestinales y hepáticas poco comunes.
7. Evaluar el pronóstico e impacto de enfermedades gastrointestinales y hepáticas en la calidad de vida de los pacientes, familia y sociedad.

Tabla 2. Temas del curso de gastroenterología integrado de 4to año.

Temas Gastrointestinales:

1. Imágenes radiológicas en gastroenterología.
2. Aproximación al paciente con dolor abdominal agudo.
3. Aproximación al paciente con dolor abdominal crónico.
4. Endoscopía gastrointestinal diagnóstica y terapéutica.
5. Aproximación al paciente con disfagia.
6. Aproximación al paciente con enfermedad por reflujo gastroesofágico.
7. Aproximación al paciente con dispepsia y gastropatía.
8. Aproximación al paciente con cáncer gástrico.
9. *Helicobacter pylori* y aproximación al paciente con enfermedad por úlcera péptica.
10. Aproximación al paciente con sangrado digestivo alto.
11. Aproximación al paciente con pancreatitis aguda.
12. Aproximación al paciente con pancreatitis crónica.
13. Aproximación al paciente con síndrome malabsortivo.
14. Aproximación al paciente con enfermedad celíaca.
15. Aproximación al paciente con constipación crónica.
16. Aproximación al paciente con cáncer colo-rectal.
17. Aproximación al paciente con sangrado digestivo bajo.
18. Aproximación al paciente con enfermedades inflamatorias intestinales (enfermedad de Crohn y colitis ulcerosa).
19. Aproximación al paciente con síndrome intestino irritable (SII).
20. Aproximación al paciente con diarrea aguda.
21. Aproximación al paciente con diarrea crónica.

Temas Hepáticos:

1. Interpretación de exámenes de laboratorio relacionados con enfermedades hepáticas.
2. Metabolismo de la bilirrubina y aproximación al paciente con ictericia.
3. Aproximación al paciente con hepatitis viral aguda.
4. Aproximación al paciente con hepatitis viral crónica.
5. Aproximación al paciente con hepatitis no viral.
6. Aproximación al paciente con enfermedad hepática por alcohol.
7. Aproximación al paciente con cirrosis.
8. Aproximación al paciente con complicaciones relacionadas con cirrosis I (manejo de ascitis y síndrome hepatorrenal).
9. Aproximación al paciente con complicaciones relacionadas con cirrosis II (manejo de encefalopatía hepática y sangrado variceal).
10. Aproximación al paciente con falla hepática aguda.
11. Aproximación al paciente con hepatitis autoinmune, cirrosis biliar primaria y otras enfermedades crónicas hepáticas no virales.
12. Aproximación al paciente con tumor hepático.
13. Aproximación al paciente con hígado graso no alcohólico.
14. Aproximación al paciente con cálculos biliares y enfermedades de la vía biliar.

fatizado durante la instrucción debería estar reflejada en la construcción del instrumento de calificación. Existen distintos tipos de validez; por ejemplo, la *validez de contenido* propone que un instrumento debe representar el dominio de contenidos que está en consideración, por lo tanto, una prueba debería ser representativa de los objetivos del curso. Por otra parte, la *validez de construcción o constructo* corresponde al grado en que el resultado de una prueba puede ser interpretado como un constructo psicológico (construcción teórica utilizada para explicar un comportamiento). La *fiabilidad* es el grado en el cual una prueba mide consistentemente lo que debe medir; en otras palabras, describe la reproducibilidad de los resultados de la prueba e indica cuán confiables pueden ser éstos[9].

El objetivo de este estudio fue evaluar la calidad de las PSM durante un proceso de implementación de mejorías en la calidad de la construcción de pruebas de conocimiento. En cuanto a los objetivos específicos, en primer lugar, fue evaluar si las PSM del curso teórico de Gastroenterología están alineadas con los objetivos y temas del curso, y determinar si son preguntas válidas y confiables. En segundo lugar, fue evaluar la mejoría en la calidad de la prueba de conocimiento teórico, mediante la implementación de una pauta de construcción de pruebas y la evaluación de la mejoría en la calidad de las PSM en base a los niveles cognitivos explorados y el porcentaje de PSM con violaciones de los principios de escritura.

MATERIAL Y MÉTODO

Se analizaron las PSM de la prueba teórica del módulo de Gastroenterología del CICA de la carrera Medicina de la EMPUC, desde los años 2002 a 2005.

Validez y principios de escritura

Para evaluar la validez de contenido, se revisó la cobertura de los objetivos y temas descritos previamente (Tablas 1 y 2). Para la validez de constructo, se calificaron las preguntas como fallidas o normales, dependiendo si presentaban o no una o más violaciones a los principios descritos por Haladyna (Tabla 3)[10]. La validez predictiva se calculó realizando una correlación entre las notas obtenidas por los alumnos año a año en la prueba de Gastroenterología y la prueba final del curso. Se calculó el índice de dificultad de las PSM año a año.

Niveles cognitivos evaluados

Se categorizó cada pregunta utilizada entre el año 2002-2005 según los niveles cognitivos de la clasificación de Bloom[7]. Se calculó la concordancia intra e inter-observador[11].

Intervención para la mejoría de PSM

En una segunda fase de este proyecto se creó una pauta para cubrir todos los objetivos y temas descritos, para así alinear los objetivos curriculares con el contenido que efectivamente se enseña y evalúa. Se creó también una nueva base de preguntas por parte de los profesores encargados de las clases teóricas, las cuales fueron evaluadas y corregidas por dos investigadores expertos en educación médica, explorando niveles cognitivos superiores, y evitando incurrir en violaciones descritas en la Tabla 2. Estas preguntas fueron utilizadas para la prueba teórica del año 2006. Ese año, se calculó la fiabilidad utilizando el método de Cronbach y el resultado se expresó como un coeficiente alfa de Cronbach con rangos de 0 a 1[12].

Análisis Estadístico

Para realizar las comparaciones entre los distintos años de las variables porcentuales, se utilizó Test Z con corrección Bonferroni post-hoc. Se consideró significativo un valor $p < 0,05$. Se utilizó el software IBM SPSS v.21 (Chicago, EE.UU.) para los cálculos.

Tabla 3. Pauta para la escritura de PSM.

Aspectos de contenido:

1. Cada ítem debe reflejar un contenido específico y un comportamiento mental único.
2. Base cada ítem en un contenido importante del aprendizaje, evite el contenido trivial.
3. Use material novedoso para evaluar los niveles más altos de aprendizaje. Parafrasear en una prueba el lenguaje del libro de texto o el lenguaje utilizado durante las clases lectivas evita evaluar solo la memoria.
4. Mantenga el contenido de cada ítem independiente del contenido de otros ítems de la prueba.
5. Evite contenido muy específico o muy general al escribir PSM.
6. Evite ítems basados en opinión.
7. Evite ítems tramposos.
8. Mantenga el vocabulario simple, acorde al grupo de estudiantes evaluados.

Aspectos de Formato:

9. Use el formato que desee dentro de los formatos de PSM (preguntas cruzadas, verdadero/falso, pareo), pero evite el uso de formato de preguntas complejas MT/F (Tipo K).
10. Formatee el ítem vertical en vez de horizontal.

Aspectos de Estilo:

11. Edite y corrija los ítems.
12. Use una correcta ortografía, puntuación, capitalización y sintaxis.
13. Minimice volumen de lectura de cada ítem.

Escritura de Viñetas:

14. Asegúrese que las indicaciones de la viñeta sean muy claras.
15. Incluya la idea central en la viñeta en vez de las opciones.
16. Evite «excesivos adornos» que no aportan al contexto clínico o datos significativos.
17. Redacte la viñeta de forma positiva, no utilice negativos como NO o EXCEPTO. Si utiliza palabras negativas, use las palabras cuidadosamente, y asegúrese que la palabra esté en negrita y capitalizada.

Escritura de Opciones:

18. Desarrolle la mayor cantidad de opciones efectivas que pueda, pero los estudios sugieren que tres opciones son adecuadas.
19. Asegúrese que sólo una opción sea la correcta.
20. Varíe la ubicación de la opción correcta.
21. Ubique las opciones en orden lógico o numérico.
22. Mantenga las opciones independientes, éstas no deben sobreponerse.
23. Mantenga el largo de las opciones semejante.
24. Mantenga las opciones homogéneas en cuanto al contenido y estructura gramatical.
25. Use cuidadosamente «ninguna de las anteriores».
26. Evite utilizar «todas las anteriores».
27. Redacte las opciones positivamente, evite usar negativos como NO.
28. Evite dar pistas sobre la opción correcta, tal como:
    a. Determinantes específicos, como siempre, nunca, completamente o absolutamente.
    b. Asociaciones obvias, opciones idénticas o palabras semejantes a la viñeta.
    c. Inconsistencias gramaticales que guíen al alumno a la opción correcta.
    d. Opción correcta suspicaz.
    e. Pares o tríos de opciones que guíen al alumno a la opción correcta.
    f. Opciones ridículas o absurdas.
29. Haga que todos los distractores sean plausibles.
30. Use los típicos errores de los alumnos para escribir distractores.
31. Use el humor si es compatible con el profesor y el ambiente educacional.

RESULTADOS

Validez y principios de escritura

En cuanto a los objetivos establecidos, no hubo diferencias estadísticamente significativas (Tabla 4). El objetivo número 5 fue el más representado durante los 5 años de estudio. El N° 2 exhibió una representación elevada los años 2002-2003, sin embargo, presentó una tendencia no significativa a la disminución hasta el 2006. Por su parte, el objetivo N°4 mostró una tendencia no significativa a aumentar en el mismo período. El N°6 fue consistentemente el menos representado en todo el período.

Al evaluar los temas cubiertos por la prueba teórica (Tabla 5) se puede ver que el año 2002 cerca de un 77% de los temas fueron cubiertos (tanto de hepatología como de gastrointestinal), sin embargo, fueron aumentando progresivamente hasta el año 2006, donde se cubrieron el 100% de los temas ($p < 0,001$).

Se observó una moderada validez predictiva de la prueba de Gastroenterología en relación con la prueba teórica final del curso integrado, con un coeficiente de correlación con valores en rango de 0,59 (en 2004) a 0,66 (en 2005), con una tendencia no significativa al aumento a través de los años.

El año 2002 cerca del 54% de las preguntas presentaban algún tipo de violación de los principios propuestos en la Tabla 3. Los más frecuentes eran el uso de estructuras complejas (tipo K) y de enunciados planteados en negativo. Hacia el año 2006, éstos disminuyeron a 18,8% ($p < 0,001$). Cabe destacar que el uso de estructuras complejas permaneció siendo el más frecuente, mientras que sólo un 1,6% de las preguntas se plantearon en negativo (Tabla 6). Al evaluar el nivel de dificultad de las preguntas, no se encontró diferencias a través de los años, sin embargo, sí presentan un mayor índice de dificultad las preguntas fallidas que las normales (0,63 ± 0,27 vs 0,80 ± 0,23; $p = 0,039$).

Al evaluar la fiabilidad con el método de Cronbach, se obtuvo para la prueba del 2006, un coeficiente alfa de 0,78. Cabe destacar que el grupo de preguntas fallidas presentó un coeficiente menor que aquellas normales (0,61 vs 0,71).

Niveles cognitivos evaluados

Al evaluar la taxonomía educacional de Bloom (Tabla 7), la mayoría de las PSM se encuentran en las categorías intermedias (3 y 4). Sin embargo, se aprecia una disminución significativa de las categorías más básicas, como el nivel 1 ($p < 0,001$), y un aumento progresivo de aquellas más elevadas, como el nivel 6 ($p < 0,001$) desde el año 2002 al 2006. Los evaluadores presentaron una concordancia intraevaluador de alrededor de 93,5% e interevaluador de 74%.

Tabla 4. Comparación de objetivos cubiertos por las PSM de las pruebas de gastroenterología desde 2002 al 2006.

	2002	2003	2004	2005	2006
Número de items	N = 73	N = 76	N = 74	N = 57	N = 64
Objetivo					
1	4 (5,5%)	2 (2,6%)	6 (8,1%)	3 (5,3%)	7 (11%)
2	18 (24,7%)	13 (17,1%)	12 (16,2%)	9 (15,9%)	8 (12,5%)
3	14 (19,2%)	27 (35,5%)	29 (39,2%)	17 (29,8%)	9 (14,1%)
4	13 (17,1%)	19 (25%)	13 (17,6%)	13 (22,8%)	17 (26,5%)
5	20 (27,4%)	14 (18,4%)	11 (14,7%)	11 (19,3%)	18 (28,1%)
6	1 (1,4%)	1 (1,3%)	1 (1,4%)	2 (3,5%)	0 (0%)
7	3 (4,2%)	0 (0%)	2 (2,8%)	2 (3,5%)	5 (7,8%)

*** Diferencia estadísticamente significativa ($p < 0,001$). Z test, corrección bonferroni posthoc.**

Tabla 5. Comparación de temas cubiertos por las PSM de las pruebas de gastroenterología desde 2002 al 2006.

	2002	2003	2004	2005	2006
Temas gastrointestinales (N = 21)	17 (81%)	17 (81%)	16 (76,2%)	19 (90,5%)	21 (100%)
Temas hepáticos (N = 14)	10 (71,4%)	11 (78,6%)	13 (92,9%)	13 (92,9%)	14 (100%)
Total (N = 35)	27 (77,1%)*	28 (80%)	29 (82,9%)	32 (91,4%)	35 (100%)*

*** Diferencia estadísticamente significativa ($p < 0,001$). Z test corrección bonferroni posthoc.**

Tabla 6. Frecuencia de violaciones a los principios de escritura de PSM en las pruebas de gastroenterología desde 2002 al 2006.

	2002	2003	2004	2005	2006
Número de ítems	N = 73	N = 76	N = 74	N = 57	N = 64
Principio de escritura violado					
# 9 (tipo-K)	23 (31,5%)*	25 (32,9%)	22 (29,7%)	17 (29,8%)	7 (11%)*
# 14 (viñeta poco clara)	2 (2,8%)	1 (1,3%)	2 (2,8%)	1 (1,8%)	0 (0%)
# 17 (NO o EXCEPTO)	8 (11%)	12 (15,8%)	13 (17,6%)*	3 (5,3%)	1 (1,6%)*
# 22 (opciones sobrepuestas)	1 (1,4%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)
# 25 (ninguna de las anteriores)	2 (2,8%)	1 (1,3%)	0 (0%)	0 (0%)	0 (0%)
# 26 (todas las anteriores)	1 (1,4%)	2 (2,6%)	3 (4,1%)	0 (0%)	0 (0%)
# 28a (uso de absolutos)	0 (0%)	1 (1,3%)	0 (0%)	0 (0%)	0 (0%)
#28e (pares o tripletas)	3 (4,2%)	5 (6,6%)	5 (6,8%)	2 (3,5%)	4 (6,2%)
Total	40 (54,8%)	47 (61,8%)*	45 (60,8%)	23 (40,4%)	12 (18,8%)*

*** Diferencia estadísticamente significativa ($p < 0,001$). Z test, corrección bonferroni posthoc.**

Tabla 7. Comparación del nivel cognitivo de Bloom explorado por las PSM de las pruebas de gastroenterología desde 2002 al 2006.

	2002	2003	2004	2005	2006
	N = 73	N = 76	N = 74	N = 57	N = 64
Taxonomía de Bloom					
Nivel 1	14 (19,2%)*	13 (17,1%)	14 (18,9%)	6 (10,5%)	2 (3,2%)*
Nivel 2	6 (8,2%)	5 (6,6%)	5 (6,8%)	4 (7%)	5 (7,8%)
Nivel 3	43 (58,9%)	38 (50%)	40 (54,1%)	31 (54,4%)	23 (35,9%)
Nivel 4	9 (12,3%)	16 (21,1%)	9 (12,2%)	4 (7%)	15 (23,4%)
Nivel 5	0 (0%)	2 (2,6%)	3 (4%)	0 (0%)	0 (0%)
Nivel 6	1 (1,4%)*	2 (2,6%)	3 (4%)	12 (21,1%)	19 (29,7%)*

*** Diferencia estadísticamente significativa ($p < 0,001$). Z test corrección bonferroni posthoc.**

DISCUSIÓN

Dentro de la educación médica, la metodología de evaluación ha tenido un desarrollo exponencial en los últimos años[13]. Se han desarrollado diferentes herramientas para cubrir todos los aspectos de la educación. En la evaluación cognitiva, sin embargo, las PSM aún son el instrumento líder dado sus características descritas previamente[14]. En este estudio, se decidió evaluar las PSM utilizadas en el módulo teórico de Gastroenterología del CICA, aplicar una pauta de creación de la prueba y someter a corrección de las PSM por investigadores expertos en educación médica.

La cobertura de objetivos del módulo de Gastroenterología se puede considerar satisfactoria, con un rango de 85,7 a 100% en los últimos 5 años. Sin embargo, la proporción de ítems cubriendo cada objetivo fue desproporcionada, desde 1,4% al 39,2% de los ítems por objetivo.

Al analizar los objetivos, se puede apreciar que 1 y 2 corresponden a objetivos psicomotores más que cognitivos. Éstos son evaluados principalmente durante la rotación ambulatoria de Gastroenterología y hospitalaria. Asimismo, el objetivo 6, relacionado con terapia, es cubierto principalmente durante el curso integrado de quinto año. Este punto es crítico, ya que es frecuente que se trate de medir aspectos psicomotores con instrumentos confiables como una prueba de conocimiento con PSM, pero que carecen de validez. Por el contrario, varias escuelas de medicina han incluido evaluación de destrezas clínicas con ECOE. Sin embargo, varias estaciones miden conocimiento, constituyendo un desperdicio de recursos, ya que es más barato y eficiente que dichos conocimientos se evalúen con pruebas escritas.

Una de las herramientas descritas para mejorar la cobertura de objetivos, es el desarrollo de pautas para la creación de pruebas[15]. Esto fue realizado el año 2006, no obstante se mantuvo la desproporción descrita. En el período 2007-2013, se ha incorporado una cobertura más equitativa con la implementación de la pauta de contenidos optimizada. El profesor jefe del curso desarrolla las preguntas en conjunto con los docentes encargados de las clases, considerando tanto la cobertura de objetivos y temas así como un adecuado balance de preguntas asignando, de manera intencionada un 25% del total de PSM a ciencias básicas, 50% a diagnóstico y 25% a terapia. El

borrador de la prueba es revisado por una educadora encargada de la calidad de las preguntas de las pruebas en el Centro de Educación Médica (CEM), quien envía los reparos en la construcción de las PSM para su corrección, previo a su aplicación en el curso. Esta intervención redujo aún más el número de preguntas fallidas, ya que aproximadamente 10 a 20% de las preguntas son corregidas una semana antes de ser utilizadas y desde el 2007 no hay preguntas fallidas al momento de ser aplicadas a los estudiantes. Actualmente, los errores más frecuentes son de carácter menor, y principalmente son de las categorías 1, 2, 5, 7, 8, 11, 12, 13,14, 16, 20, 23, 24, y 28 descritas por Haladyna[10].

Al evaluar la cobertura de temas del curso, se puede ver una mejoría progresiva estadísticamente significativa, desde un 77% a un 100%, el año en que se implementó la pauta de creación de la prueba.

Por otra parte, al analizar la utilidad de la prueba, puede verse una validez predictiva moderada, con una tendencia no significativa a mejorar a través de los años. También posee una buena fiabilidad, con factor alfa de Cronbach (consistencia interna) de 0,78. Además, los ítems fallidos demostraron una menor fiabilidad (0,62). Estos ítems, por lo tanto, son más difíciles y menos válidos en cuanto a su construcción, demostrando la importancia de la evaluación continua de la construcción de las pruebas.

Según nuestra experiencia, validamos el proceso de construcción y revisión de PSM de manera centralizada por pocos expertos, en un proceso estable en el tiempo e independiente del encargado del curso o de los docentes que generaron las preguntas. Este modelo se contrapone a la capacitación masiva de todos los docentes involucrados en la confección de PSM, ya que la logística de capacitación de nuevos docentes es difícil de ejecutar, y por otro lado, el contar con planta docente capacitada no implica necesariamente una mejoría significativa en la calidad de las PSM.

Desde el año 2007 a 2010, se consolidó el proceso de generación y evaluación de las PSM y se dividió la evaluación en dos pruebas (tubo digestivo e hígado), lo que hace no comparable esta etapa con el proceso previamente descrito. Desde el 2011 en adelante, se reestructuró el curso hacia un modelo de evaluación *para* el aprendizaje. Éste corresponde a un nuevo paradigma en educación médica, en donde la evaluación está íntimamente integrada al proceso educacional, incorporando elementos e información de diversas fuentes para identificar el perfil propio del alumno, incluyendo sus fortalezas y debilidades, entregando un *feedback* (retroalimentación) efectivo e individualizado, con el objetivo de maximizar su aprendizaje[16].

La frecuencia de violación a los principios de escritura de PSM disminuyó considerablemente a través de los años, en especial el año en que se aplicó la pauta de construcción. Es interesante destacar que las preguntas fallidas presentaron un mayor índice de dificultad que aquellas normales, lo cual puede considerarse como un avance en la representatividad del proceso de evaluación, ya que es muy probable que el grado de dificultad aumentó en las preguntas fallidas por el defecto de redacción más que por un desafío cognitivo superior.

Las PSM son una excelente herramienta para evaluar niveles cognitivos inferiores, sin embargo, la construcción de preguntas que evalúen niveles superiores es difícil y requiere un esfuerzo adicional del diseñador[17]. En este estudio, la capacitación de los diseñadores de la prueba y la implementación de una pauta de construcción, permitió producir un impacto positivo en la calidad de las nuevas PSM utilizadas en la prueba del 2006. Podemos observar una disminución significativa de la categoría 1, así como un aumento significativo de la categoría 6. El desarrollo de PSM basadas en escenarios clínicos es un paso importante en el desarrollo de los niveles cognitivos superiores.

Es importante destacar que la distribución de las PSM según niveles cognitivos debe ser determinado a priori[18], por lo que no es el objetivo final que todas las PSM evalúen el nivel 6 de Bloom, sino más bien, contar con un alto porcentaje de preguntas entre los niveles 4 al 6. Esto se evidencia en el estudio al comparar las categorías de Bloom con los objetivos del curso, ya que, por ejemplo, los objetivos 4 y 5 permiten el desarrollo de interpretación y análisis por parte del alumno[7]. Es importante alinear los objetivos del curso con los niveles cognitivos utilizados para la evaluación. Existen otros sistemas de evaluación que pudieran simplificar el proceso de evaluación de los niveles cognitivos evaluados, como los de Naaera[19] y Buckwalter[20]. Dentro de este estudio, también categorizamos las preguntas según estos sistemas, encontrando resultados semejantes y altos niveles de concordancia intra e interevaluador. Sin embargo, decidimos utilizar sólo el sistema de Bloom dada su relevancia, simpleza y confiabilidad inter e intraevaluador.

CONCLUSIONES

En conclusión, el uso de una pauta para el desarrollo de una prueba teórica de PSM, así como la evaluación de las PSM por expertos capacitados en educación médica, permite mejorar la cobertura de los temas del curso evaluado, mejorar la validez de contenido de las preguntas, y explorar niveles cognitivos superiores.

BIBLIOGRAFÍA

1. Rosso P, Velasco N, Moreno R. Undergraduate curriculum reform at the Pontifical Catholic University Medical School: aims, methodology and advance status. Rev Med Chil 1997; 125(7): 796-807.
2. Sánchez I, Riquelme A, Moreno R, Mena B, et al. Revitalising medical education: the School of Medicine at the Pontificia Universidad Católica de Chile. Clin Teach 2008; 5(1): 57-61.
3. Harden RM, Sowden S, Dunn WR. Some educational strategies in curriculum development: The SPICES model. Med Educ 1984; 18(4): 284-297.
4. World Federation of Medical Education. The changing medical profession: Implications for medical education. World Summit on Medical Education, Edinburgh, August 1993. Med Educ 1993; 27: 291-296.
5. Harden RM, Davis MH. AMEE Guide No. 5: The core curriculum with options or special modules. Med Teach 1995; 17(2): 125-148.
6. Riquelme A, Sirhan M, Delfino A, Méndez B, et al. Diseño e Implementación de un Portafolio de Casos Clínicos en Estudiantes de Medicina de Pregrado. Artículo aceptado para publicación. Ars Médica.
7. Bloom BS, Engelhart MD, Furst EJ, Hill WH, et al. Taxonomy of Educational Objectives: Book 1: The Cognitive Domain. (1956) Londres: Longman Green.
8. Downing SM. Assessment of Knowledge with Written Test Forms, in: Norman GR, Van der Vleuten CPM, Newble DI. (Eds) International Handbook of Research in Medical Education. (2002) Inglaterra: Kluwer Academic Publishers.
9. Van der Vleuten CPM. The Assessment of Professional Competence: Developments, Research and Practical Implications. Adv Health Sci Educ Theory Pract 1996; 1(1): 41-67.
10. Haladyna TM, Downing SM. A taxonomy of multiple-choice item-writing rules. Applied Measurement in Education 1989; 2(1): 37-50.
11. Ebel RL. Estimation of the reliability of ratings. Psychometrika 1951; 16(4): 407-424.
12. Cronbach LJ. Coefficient alpha and the internal structure of tests. Psychometrika 1951; 16(3): 297-334.
13. Hart IR. Trends in clinical assessment, En: Harden RM, Hart IR & Mulholland H. (Eds) International Conference proceedings: Approaches to the Assessment of Clinical Competence (1992), Dundee, Centre for Medical Education.
14. Downing SM. Assessment of Knowledge with Written Test Forms, in: Norman GR, Van der Vleuten CPM & Newble DI. (Eds) International Handbook of Research in Medical Education (2002), Gran Bretaña, Kluwer Academic Publishers.
15. Bridge PD, Musial J, Frank R, Roe T, et al. Measurement practices: methods for developing content-valid student examinations. Medical Teacher 2003; 25(4): 414-421.
16. Schuwirth LW, van der Vleuten CP. Programmatic assessment: From assessment of learning to assessment for learning. Med Teach. 2011; 33(6): 478-485.
17. LaDuca A, Staples WI, Templeton B, Holzman GB. Item modelling procedure for constructing content-equivalent multiple choice questions. Med Educ 1986; 20(1): 53-56.
18. Irwin WG, Bamber JH. The cognitive structure of the modified essay question. Med Educ 1982; 16(6): 326-331.
19. Naeraa N. Objectives for a course of Physiology for Medical Students. (1972) University of Aarhus and British Medical Association. London.
20. Buckwalter JA, Schumacher R, Albright JP, Cooper RR. Use of an educational taxonomy for evaluation of cognitive performance. Journal of Medical Education 1981; 56(2): 115-121.

Correspondencia:
Arnoldo Riquelme, MD, MMedEd.
Marcoleta 367,
Casilla 114-D,
Santiago, Chile.
e-mail: a.riquelme.perez@gmail.com