Longitud (número de preguntas) y resultado de un examen

R. Sarrias-Ramis, L. Mateu, E. Baillès, J. Pérez

Objetivo. Este artículo presenta dos estudios sobre la relación entre el número de preguntas de un examen y sus resultados. Estos dos estudios se realizaron en respuesta a dos problemas concretos presentados en dos facultades de ciencias de la salud de dos universidades catalanas. **Sujetos y métodos.** El primer estudio, realizado en la Facultad de Medicina de la Universitat Autònoma de Barcelona, compara los resultados reales obtenidos en pruebas de elección múltiple en tres asignaturas con los resultados que hubieran obtenido los estudiantes con la mitad de las preguntas. La dificultad del examen en ambas situaciones fue prácticamente la misma y los resultados académicos también fueron similares. El segundo estudio, llevado a cabo en la Facultad de Ciencias de la Salud y de la Vida de la Universitat Pompeu Fabra de Barcelona, compara los resultados reales obtenidos en pruebas de ensayo de dos asignaturas de los estudios de Biología con los hipotéticos que se hubieran obtenido con la mitad de las preguntas. Los resultados obtenidos globalmente fueron muy parecidos. **Conclusión.** La conclusión general del estudio es que la evaluación del rendimiento académico no dependería fundamentalmente del número de preguntas y que sería más importante buscar la representatividad y relevancia de éstas.

Palabras clave. Evaluación del rendimiento académico. Longitud de los exámenes. Métodos de evaluación.

Length (number of questions) and exams results

Aim. *This paper resents two studies about the relationship between the number of questions that appear in an exam and their results. These studies have been done in response to two concrete problems found in two Spanish universities.* ***Subjects and methods.*** *The first study, done at the Faculty of Medicine of the Autonomous University of Barcelona, compared the real results achieved in three subjects which use MCQ-tests with the theoretical results which the students would have achieved if half of the questions had been used. The difficulty of the exams in both situations was practically the same and the student's academic results were very similar in both situations. The second study, carried out at the Faculty of Health and Life Sciences at the Pompeu Fabra University of Barcelona, compared the real academic results in two subjects which use free response questions with the supposed ones with half of the questions of the exam in two subjects which use many essay questions. The results obtained by the students were practically identical.* ***Conclusion.*** *Our general conclusion is that in the evaluation of academic performance, it is more important how relevant and representative the questions are than the number of them.*

Key words. *Academic assessment. Assessment methods. Length of exams.*

Facultad de Ciencias de la Salud y de la Vida. Universitat Pompeu Fabra. Barcelona, España.

Correspondencia
Dr. Jorge Pérez. Facultad de Ciencias de la Salud y de la Vida. Universitat Pompeu Fabra. Doctor Aiguader, 80. E-08003 Barcelona.

Fax
+34 935 422 802

E-mail
jordi.perez@upf.edu

Introducción

La evaluación del aprendizaje de los estudiantes es esencial en cualquier proceso educativo dado que fundamentalmente nos permite determinar el grado de asunción de los objetivos previstos. Además, la evaluación es esencial porque dirige el tipo de aprendizaje que realizan los alumnos. Éstos van a abordar su forma de aprender en función del sistema de evaluación utilizado [1-5].

Así, dada la trascendencia de la evaluación, es muy importante escoger el método evaluativo más adecuado a los objetivos previstos. Disponemos de numerosos instrumentos de evaluación [1,6,7]

que deberían seleccionarse en función de los objetivos de aprendizaje buscados [4,8]. De todos modos, es muy habitual la utilización de pruebas de elección múltiple (PEM), de preguntas cortas (PC) y de ensayos más extensos [4,9]. En nuestra experiencia docente habitualmente hemos utilizado estos métodos y en algún momento hemos reflexionado sobre la cantidad de preguntas necesarias para que el examen tenga validez.

En el presente trabajo presentamos dos estudios que hacen referencia a la relación entre el número de preguntas y el resultado del examen. Estos dos estudios se realizaron en respuesta a dos problemas concretos planteados en dos centros de ciencias de la salud en dos universidades catalanas.

Estudio 1. Número de preguntas de elección múltiple y resultados de los exámenes

En 1993, la Facultad de Medicina de la Universitat Autónoma de Barcelona (UAB) cambió su currículo organizando los objetivos de aprendizaje en asignaturas amplias que agrupaban contenidos que previamente se habían contemplado en asignaturas independientes. En el primer año de carrera se crearon dos voluminosas asignaturas denominadas 'Morfología, estructura y función del organismo humano' (OH-1 y OH-2), que incluían contenidos de las antiguas asignaturas de Biología, Bioquímica, Fisiología e Histología. Así mismo, en el segundo curso también se crearon dos grandes asignaturas llamadas 'Desarrollo, estructura y función de los sistemas en estado de salud' (AS-1 y AS-2)', que contemplaban objetivos educativos de las antiguas materias de Anatomía, Histología y Fisiología. También en el segundo curso, los estudiantes cursaban la asignatura de 'Psicología' (Psico), que era menos voluminosa.

En la facultad se usaban frecuentemente PEM para evaluar los conocimientos de los estudiantes. Cuando se crearon las nuevas asignaturas se planteó el tema de cuántas preguntas serían necesarias para evaluar las asignaturas amplias. El número de preguntas no podía ser la suma de las que se utilizaban anteriormente en las antiguas asignaturas, ya que el examen sería excesivamente largo. En general, las antiguas asignaturas contemplaban pruebas independientes con un número de preguntas generalmente superior a 60. Por consenso, se determinó que el número de preguntas para las nuevas asignaturas amplias no debería ser muy superior a 100. Ciertos profesores consideraban aceptable la decisión, pero otros dudaban de si el número de preguntas era suficiente para evaluar dichas asignaturas.

Ante dicha duda, los responsables del Gabinete Técnico-Pedagógico de la facultad en aquel momento decidimos hacer un estudio para intentar dar respuesta a las dudas planteadas. Así, estábamos interesados en detectar posibles diferencias en los resultados en las PEM en función del número de preguntas utilizado. Peveíamos que no existirían grandes diferencias en los resultados de los estudiantes en función del número de preguntas.

Sujetos y métodos

El estudio se llevó a cabo en la Facultad de Medicina de la UAB durante el curso 1995-1996. Seleccionamos tres asignaturas con características diferentes: una amplia de primer curso (OH-1), una amplia de segundo curso (AS-1) y una menos voluminosa (Psico). En AS-1 y Psico, las PEM eran de una única respuesta correcta, mientras que la PEM de OH-1 permitía diversas respuestas correctas.

Los participantes fueron los estudiantes de primer y segundo curso, que se examinaron en la primera convocatoria en las asignaturas seleccionadas. El material utilizado fueron las hojas de respuestas para la corrección mecanizada que emplearon los participantes al responder las pruebas.

Una vez que el proceso de evaluación había acabado y los estudiantes habían recibido sus calificaciones, seleccionamos la mitad de las preguntas en cada una de las tres asignaturas. Esta selección se hizo con criterios de representatividad de los contenidos de las antiguas asignaturas, respetando la supuesta validez de contenido de las pruebas. Posteriormente, procedimos a una nueva corrección mecanizada solamente con los ítems seleccionados.

Para cada asignatura determinamos dos puntuaciones para cada alumno: la puntuación real con todas las preguntas (90 en OH-1, 100 en AS-1 y 60 en Psico) y la puntuación supuesta que hubieran obtenido con la mitad de las preguntas (45 en OH-1, 50 en AS-1 y 30 en Psico). También calculamos la consistencia interna (alfa de Cronbach) y el nivel de dificultad en dos asignaturas (AS-1 y Psico). No pudimos determinar estos estadísticos en OH-1, ya que el programa informático no pro-

Tabla I. Índices de dificultad y de fiabilidad de consistencia interna de los exámenes (pruebas de elección múltiple) en ambas situaciones.

		Dificultad		Fiabilidad	
	Ítems	PR	PS	PR	PS
OH-1 [a]	100	–	–	–	–
AS-1	120	0,34	0,34	0,92	0,85
Psico	60	0,32	0,33	0,73	0,60

PR: puntuación real (100% de los ítems); PS: puntuación supuesta (50% de los ítems). [a] Prueba de elección múltiple con más de una respuesta correcta posible (información no proporcionada).

porcionabas dicha información cuando el examen permitía más de una respuesta correcta.

Posteriormente comparamos los valores de fiabilidad y dificultad en ambas situaciones. Finalmente comparamos los porcentajes de estudiantes que superarían o no las pruebas en ambas situaciones. Para el estudio establecimos el criterio de superación habitual en los estudios universitarios españoles, una nota de 5 o superior en una escala de 0 a 10. En la comparación de porcentajes utilizamos la prueba de chi al cuadrado.

Resultados

La tabla I presenta los índices de fiabilidad y de dificultad de los exámenes con todos y con la mitad de los ítems. En AS-1 y en Psico, el nivel de dificultad fue prácticamente idéntico en ambas situaciones y las fiabilidades de consistencia interna fueron ligeramente más bajas en las versiones cortas.

En la tabla II podemos observar el número de estudiantes que pasaría los exámenes con el estándar habitual (nota de 5 o superior). El porcentaje de alumnos con éxito fue muy similar en ambas situaciones. Las pruebas de chi al cuadrado no mostraron diferencias significativas en ninguna de las tres asignaturas.

Un número reducido de estudiantes tendría cambios cualitativos entre una u otra situación. En OH-1 existió mayor éxito en la versión corta, donde nueve estudiantes que superarían el examen largo no lo harían en el corto, pero 26 alumnos que no aprobarían el examen largo sí lo harían en el corto. Por el contrario, en Psico el éxito fue superior en el examen largo, donde 16 estudiantes que superarían el largo no lo harían en el corto y ocho participantes que no superarían el largo sí lo harían en el corto. En AS-1, el número de alumnos con cambios fue el mismo en ambos casos: seis estudiantes que superarían el largo no lo harían en el corto y otros seis que sí lo harían en el corto no lo superarían en el largo.

Discusión

El estudio lo realizamos únicamente con PEM, pero la evaluación definitiva de los estudiantes no se basaba exclusivamente en estas pruebas.

Hemos visto que la dificultad de los exámenes en ambas situaciones, con todas y la mitad de las preguntas, era muy similar. También, como sería de esperar, la fiabilidad de consistencia interna era ligeramente inferior en las versiones cortas, ya que este tipo de fiabilidad aumenta con el número de preguntas. De todos modos, las fiabilidades en las versiones cortas serían aceptables.

Respecto al éxito o fracaso global, en ambas situaciones fue muy similar y en ninguno de los tres casos encontramos diferencias significativas. El porcentaje de estudiantes que superaría la prueba era similar tanto en las versiones cortas como en las largas.

Las variaciones cualitativas de algunos estudiantes serían atribuidas a las variaciones que oscilarían alrededor de la nota 5 que marcaba el criterio de superación. Ante diferentes exámenes es imposible que todos los estudiantes tuvieran exac-

Tabla II. Número y porcentaje de estudiantes que superarían las pruebas de elección múltiple (puntuaciones de 5 o superiores en una escala de 0 a 10).

	Estudiantes	Éxito en PR	Éxito en PS	χ^2
OH-1	330	169 (51,2%)	186 (56,4%)	1,76 (NS)
AS-1	273	210 (76,9%)	210 (76,9%)	0,00 (NS)
Psico	248	214 (86,3%)	206 (83,1%)	0,99 (NS)

PR: puntuación real (100% de los ítems); PS: puntuación supuesta (50% de los ítems); NS: diferencias no significativas.

tamente la misma nota y sería lógico aceptar pequeñas variaciones. Algunas de ellas podrían hacer que notas próximas al 5 unas veces se decantaran por debajo, y otras, por encima. Los resultados de nuestro estudio serían acordes con lo anterior, ya que en una asignatura el éxito era superior con la versión corta (OH-1), en otra con la versión larga (Psico), y en la tercera fue idéntico (AS-1).

A pesar de la ligera reducción de la fiabilidad de consistencia interna, nuestro estudio indicaría que los resultados globales serían muy similares en el examen real y en el examen ficticio, con menos preguntas.

La conclusión del estudio sería que el número de preguntas que se determinó en su momento sería suficiente para evaluar los objetivos docentes de las asignaturas estudiadas. Así, el número de preguntes no afectaría a los resultados globales de los estudiantes. Ante esta conclusión, uno se podría preguntar si el número de preguntas que se utilizaron realmente podría reducirse a la mitad. Nosotros nos inclinamos por la primera opción, ya que la fiabilidad de consistencia interna sería superior. Así, pensamos que alrededor de 100 preguntas en PEM podrían ser suficientes para evaluar asignaturas con contenidos amplios.

Estudio 2. Número de preguntas de ensayo y resultados de los exámenes

La Facultad de Ciencias de la Salud y de la Vida de la Universitat Pompeu Fabra (UPF) de Barcelona inició su actividad con la titulación de Biología, unos estudios con una orientación biosanitaria centrados en la biología humana. Actualmente imparte los nuevos grados de Biología Humana y de Medicina, este último compartido con la UAB. En dicho centro existe un proyecto educativo dirigido por un organismo técnico, la Oficina de Coordinación y Evaluación Académica, donde la evaluación de los aprendizajes está unificada y se realiza de forma colectiva. La evaluación del rendimiento académico se caracteriza por exámenes conjuntos de todos las asignaturas cursadas en las mismas sesiones de evaluación, por la importancia de la evaluación de las habilidades prácticas en la nota final, por el uso de técnicas diferentes de evaluación, por la asunción de criterios externos de superación, por la realización de evaluaciones formativas a la mitad de los cursos y por la posibilidad de compensaciones ante rendimientos que no superarían el estándar en alguna asignatura pero con rendimientos positivos en el resto de materias.

En este centro son frecuentes los exámenes de ensayo, que se utilizan en la inmensa mayoría de asignaturas junto a PEM para evaluar los conocimientos teóricos de los estudiantes. En el inicio de los estudios de Biología, y dentro de la actividad conjunta de evaluación, en algunas asignaturas se plantearon exámenes de ensayo demasiado largos que generaban distorsiones importantes que podrían afectar negativamente el rendimiento de los estudiantes. Los exámenes citados normalmente necesitaban más tiempo para su contestación que el asignado a cada asignatura por la oficina educativa; ello implicaba alargar las sesiones de evaluación, con el consiguiente nerviosismo y malestar en los estudiantes.

Ante dicha situación, la oficina educativa se planteó realizar un estudio con alguna de las asig-

Tabla III. Número y porcentaje de estudiantes que superarían las pruebas de ensayo (puntuaciones de 5 o superiores en una escala de 0 a 10).

	Estudiantes	Éxito en PR	Éxito en PS	χ^2
BGA	62	35 (56,5%)	33 (53,2%)	0,13 (NS)
BC	53	49 (92,5%)	48 (90,6%)	0,12 (NS)

PR: puntuación real (100% de los ítems); PS: puntuación supuesta (50% de los ítems); NS: diferencias no significativas.

naturas cuyos exámenes de ensayo tardaban en contestarse más tiempo del asignado para ver el efecto del número de preguntas de ensayo sobre el resultado global de los estudiantes. Así, pretendíamos comparar los resultados en las pruebas reales con aquellos que hubieran tenido los estudiantes si los exámenes hubieran contado con la mitad de las preguntas.

A partir de los resultados de nuestro primer estudio con PEM, esperábamos que los resultados fueran similares.

Sujetos y métodos

El estudio se realizó durante el curso académico 2000-2001 en la Facultad de Ciencias de la Salud y de la Vida de la UPF con dos asignaturas de la licenciatura de Biología: 'Bioquímica general y aplicada' (BGA), de primer curso, y 'Biología celular' (BC), impartida en el tercer curso de carrera. Los participantes fueron los estudiantes que se examinaron en dichas asignaturas en la primera convocatoria.

El material utilizado fueron los exámenes de ensayo contestados por los alumnos en ambas asignaturas. Todas las preguntas habían sido evaluadas y tenían una puntuación asignada por el profesor responsable. Todas las preguntas tenían el mismo peso sobre la nota del examen, que se evaluó entre 0 y 10.

Cuando todo el proceso de evaluación había concluido, para nuestro estudio en cada asignatura determinamos dos puntuaciones para cada alumno: la obtenida realmente con todas las preguntas (9 preguntas en BGA y 20 en BC) y la puntuación supuesta que hubieran obtenido con la mitad de las preguntas (5 en BGA y 10 en BC). La selección de las preguntas de la situación supuesta se realizó teniendo en cuenta la posible dificultad e idoneidad de las preguntas.

En primer lugar correlacionamos las notas obtenidas en ambas situaciones y posteriormente comparamos el nivel de éxito obtenido en los dos supuestos. En este caso, el nivel de éxito implicaba la superación del estándar habitual, nota de 5 o superior en la escala decimal. Para la comparación de porcentajes utilizamos la prueba de chi al cuadrado.

Resultados

Las correlaciones entre los resultados obtenidos en puntuación real y puntuación supuesta fueron de 0,82 para BGQ y de 0,84 para BC. La tabla III presenta el número de estudiantes que superarían el estándar (notas de 5 o superiores). El porcentaje de estudiantes con éxito fue muy similar en ambas situaciones. La pruebas de chi al cuadrado no mostraron diferencia significativas.

El número de estudiantes con cambios cualitativos entre las dos situaciones fue pequeño. En BGA, ocho estudiantes que pasarían el examen real no lo harían con la versión corta, y seis que no superarían la prueba real sí lo harían con la mitad de las preguntas. En BC, tres alumnos que pasarían la prueba larga no lo harían en la corta, y dos alumnos que no pasarían la prueba real sí lo harían en la más corta.

Discusión

Desde el punto de vista global, en las dos asignaturas estudiadas, los resultados fueron muy similares en ambas situaciones, ya que las correlaciones fueron muy altas y los porcentajes de éxito, muy parecidos.

Como hemos comentado en la discusión del estudio anterior, las variaciones cualitativas entre ambas situaciones serían atribuidas a las pequeñas variaciones que se producirían al hacer pruebas diferentes. Posiblemente, aquellos alumnos con notas cercanas a la nota 5 en algunos casos obtendrían notas inferiores, y en otros, superiores.

El estudio confirmaría nuestras previsiones porque los resultados globales en el rendimiento de los estudiantes en las dos asignaturas con más o menos preguntas fueron muy similares. En el caso que nos ocupa, la reducción del número de preguntas de ensayo haría que los exámenes fueran susceptibles de responderse en el tiempo asignado por la oficina educativa. Ello tendría la ventaja de reducir el estrés de los estudiantes por la percepción de falta de tiempo e, incluso, podría mejorar su rendimiento.

Discusión general

Nuestro objetivo consistía en responder a dos preguntas concretas ante dos situaciones específicas sobre la influencia del número de preguntas en los resultados de los exámenes. Los datos aportados por los dos estudios apuntan a la misma conclusión. Por un lado, a pesar de un ligero decremento en la fiabilidad de consistencia interna, hemos visto que los resultados con PEM globalmente son muy parecidos con más o menos preguntas. En concreto, creemos que un número alrededor de 100 preguntas sería suficiente para evaluar asignaturas con contenidos amplios, ya que también garantizarían una fiabilidad de consistencia interna muy aceptable.

Por otro lado, no es necesario utilizar un gran número de preguntas de ensayo para evaluar conocimientos teóricos, especialmente cuando todas nuestras asignaturas también utilizan PEM. Entenderíamos que si solamente se utilizaran pruebas de ensayo, la disminución de preguntas podría afectar a la validez de contenido, pero dicho problema se reduce si también utilizamos PEM, como es el caso que nos ocupa.

En nuestro trabajo hemos obviado deliberadamente el tema de la validez de las preguntas, tanto en las PEM del estudio de la UAB como en los exámenes de ensayo de la UPF. Nuestros estudios se han limitado a relacionar la longitud de un examen con el resultado obtenido en él.

Los resultados aportados en los dos estudios realizados tendrían repercusiones prácticas. Por un lado, sería aceptable la propuesta que se hizo en su día en la UAB sobre el uso de PEM con un número de preguntas cercano a 100 para evaluar asignaturas con contenidos amplios. Por otro, la reducción del número de preguntas de ensayo en las asignaturas estudiadas facilitaría el proceso de evaluación colectiva que se realiza en el centro estudiado de la UPF.

A partir de nuestros estudios y de nuestra experiencia creemos que es mucho más importante la relevancia y representatividad de las preguntas que el número de ellas. Así, cuando preparamos un examen, el esfuerzo principal deberíamos realizarlo para garantizar la validez de contenido de nuestra prueba considerando los objetivos educativos previstos.

Bibliografía

1. Guilbert JJ. Education handbook for health personnel. 6 ed. Geneva: World Health Organization; 1992.
2. Rolfe I, McPherson J. Formative assessment: how am I doing? Lancet 1995; 345: 837-9.
3. Cohen-Schonatus J. Student assessment and examination rules. Med Teach 1999; 21: 318-21.
4. Wass V, Van der Vleuten CPM, Shatzer J, Jones R. Assessment of clinical competence. Lancet 2001; 357: 945-9.
5. Wong JGWS, Cheung EPP. Ethics assessment in medical students. Med Teach 2003; 25: 5-8.
6. Harden RM. Ten questions to ask when planning a course or curriculum. Med Educ 1986; 20: 356-65.
7. Schuwirth LWT, Van der Vleuten CPM. Evaluación escrita. In Cantillon P, Hutchinson L, Wood D, eds. Aprendizaje y docencia en medicina. Barcelona: Fundación Dr. Antonio Esteve; 2006. p. 47-51.
8. Nolla-Domenjó M. La evaluación en educación médica. Principios básicos. Educ Med 2009; 12: 223-9.
9. Nendaz MR, Tekian A. Assessment in problem-based learning medical schools: a literature review. Teach Learning Med 1999; 11: 232-43.

