La difícil objetividad de las pruebas de ensayo en la evaluación del rendimiento académico

The hard objectivity of the essay exams in academic achievement assessment

Rafael Torrubia[1] y Jorge Pérez[2]

[1] Facultad de Medicina. Universidad Autónoma de Barcelona (Barcelona)
[2] Facultad de Ciencias de la Salud y de la Vida. Universidad Pompeu Fabra (Barcelona)

Introducción: El objetivo del presente escrito es informar de dos experiencias empíricas donde se pone de manifiesto la poca objetividad de las pruebas de ensayo y de las posibilidades de reducir dicha deficiencia.
Material y Métodos: Los participantes en el estudio (92 profesores de diversas universidades asistentes a un taller de formación y 460 estudiantes de segundo de Medicina) puntuaron en tres situaciones diferentes la respuesta a una pregunta de ensayo sobre el tema de la evaluación del rendimiento académico de los estudiantes. En la primera situación cada evaluador usó sus propios criterios, en la segunda todos los evaluadores tuvieron unos criterios comunes y en la tercera se precisaban las puntuaciones concretas para dichos criterios.
Resultados: Se evidenció la gran disparidad en las puntuaciones. Pero a medida que los criterios de evaluación fueron más precisos las puntuaciones fueron menos dispersas. Cuando la precisión en los criterios no fue máxima, los profesores asignaron mejores puntuaciones al examen que los estudiantes. A pesar de la reducción de la dispersión de las calificaciones cuando la precisión fue máxima, todavía existió una discrepancia considerable entre evaluadores.
Conclusiones: Los resultados del estudio avalan la necesidad de usar criterios de corrección lo más precisos posibles a la hora de evaluar cualquier prueba de ensayo así como la oportunidad de usar pruebas más objetivas.

Palabras clave: Evaluación del rendimiento académico de los estudiantes, Exámenes de ensayo, Objetividad.

Introduction: We explain two empirical experiences about the low objectivity of the essay exams and the possibilities of increasing this objectivity.
Method: Participants in the study (92 university lecturers and 460 medical students) graded an essay exam on student assessment in three different situations: without criteria, with common criteria and with precise common criteria.
Results: The more precise the criteria, the more concordance there was among evaluators. In the two first situations, without precise criteria, lecturers assigned higher scores than students. In spite of precise common criteria, in the third situation there also appeared some discrepancies between evaluators.
Conclusions: The results show the necessity to use precise criteria to grade essay exams and to use objective testing methods to assess the students' academic achievement.

Key words: Student assessment, Essay exams, Objectivity.

Correspondencia:
Jorge Pérez
Facultat de Ciències de la Salut i de la Vida.
Universitat Pompeu Fabra.
c/ Dr. Aiguader, 80, 08003 Barcelona.
e-mail: jperez@imim.es

INTRODUCCIÓN

La evaluación del rendimiento académico de los estudiantes es un paso fundamental de cualquier proceso educativo ya que por un lado nos permite determinar el grado de asunción de los objetivos propuestos[1,2] y, por otro, dirige los aprendizajes de los alumnos [2-5]. No hay ninguna duda de que los estudiantes abordan sus aprendizajes en función del tipo de evaluación a la que serán sometidos haciendo evidente el dicho "dime cómo evalúas y te diré como aprenden tus alumnos".

Así, dada la importancia de la evaluación, es fundamental seleccionar los métodos e instrumentos más adecuados para medir nuestros objetivos educativos. Por suerte, disponemos de un abanico muy amplio de posibilidades técnicas diferentes que se adecuarían selectivamente a los diferentes tipos de objetivos a evaluar tanto de conocimientos como de habilidades como de actitudes o valores [1,6]

En estudios de ciencias de la salud, además de las pruebas de elección múltiple, los exámenes escritos, de ensayo más o menos largo, son muy utilizados para evaluar los conocimientos de los estudiantes. Tal como se ha descrito por expertos, uno de los problemas que tienen estas pruebas hace referencia a su falta de objetividad [1, 7].

Nuestro objetivo consiste en informar de dos experiencias empíricas donde se pone de manifiesto la baja objetividad de las pruebas de ensayo y de las posibilidades de reducir dicha deficiencia. No pretendemos hacer un estudio exhaustivo sobre el problema de la falta de objetividad de las pruebas de ensayo o sobre su pertinencia donde ya existe suficiente evidencia.[8-11]

LAS EXPERIENCIAS

Con profesores universitarios

Desde 1993 los autores de este trabajo venimos impartiendo un taller de formación para profesores sobre la evaluación del rendimiento académico de los estudiantes. Éste, generalmente, ha sido impartido en tres sesiones de tres horas en días diferentes. Entre otros, el taller tenía dos claros objetivos. Por un lado hemos pretendido explicar las posibilidades de las pruebas objetivas y por otro hemos intentado poner de manifiesto la poca objetividad de las pruebas de ensayo y, a su vez, explicitar las posibilidades que tenemos para reducir su subjetividad en caso de utilizarlas.

Para ello, en la primera sesión y antes de recibir ninguna información sobre el tema, los profesores asistentes contestaban de forma anónima una pregunta de ensayo sobre "Las pruebas de ensayo en el proceso educativo: ventajas e inconvenientes". La prueba era contestada durante 10-15 minutos y su extensión no debía ser superior a un folio. De entre todas las respuestas, seleccionábamos tres de ellas a partir de nuestra discreción intentando que tuvieran características diferentes (más o menos extensas, más o menos concretas, etc.).

El una segunda sesión los asistentes al taller evaluaban cuatro respuestas a la pregunta: las tres seleccionadas de los participantes y una cuarta respuesta-señuelo, seleccionada con anterioridad por nosotros, que serviría para la realización del presente estudio ya que fue utilizada en diferentes talleres y fue evaluada por los profesores asistentes a los mismos.

En la tercera sesión se comentaban los resultados y se daba información para aumentar la objetividad de las pruebas de ensayo.

Con estudiantes de Medicina.

Coincidiendo en el tiempo, en la asignatura de Psicología Médica del currículum de Medicina de la Universidad Autónoma de Barcelona también realizábamos una práctica sobre la evaluación del rendimiento académico y que habíamos desarrollado nosotros mismos[12]. Dicha práctica estaba relacionada con el tema de la medida del comportamiento (psicometría) utilizando el ejemplo de los exámenes para aumentar la motivación de los alumnos. En la práctica se incidía en las ventajas e inconvenientes de sus exámenes más habituales, los de elección múltiple y los de ensayo. Entre otras cosas también evaluaban, en las mismas condiciones que los profesores, ciertos exámenes de ensayo. Durante dos cursos, nuestros estudiantes evaluaron la respuesta-señuelo utilizada en este estudio.

MATERIAL Y MÉTODO

Participantes

El estudio fue realizado utilizando dos poblaciones diferentes: profesores y estudiantes universitarios. Respecto al primer grupo, participaron 92 profesores universitarios asistentes a seis talleres realizados en lugares y fechas diferentes; 2 en la Universidad Autónoma de Barcelona, 1998 y 1999; 1 en el Centro de Estudios Jurídicos y Formación Especializada de la "Generalitat" de Cataluña, 1999; 1 en la Universidad Jaume I de Castellón, 1999; 1 en la

Universidad Pompeu Fabra de Barcelona, 2000 y 1 en la Universidad del País Vasco, 2002). También participaron 460 estudiantes de segundo de Medicina de la Universidad Autónoma de Barcelona matriculados en la asignatura de Psicología Básica durante los cursos 1996-97 y 1997-98.

Material

Para la realización de nuestro estudio utilizamos una respuesta dada a la pregunta "Las pruebas de ensayo largo: ventajas e inconvenientes" por un profesor participante en un taller realizado anteriormente (Anexo 1), siendo la respuesta-señuelo que siempre fue objeto de evaluación por los participantes en el estudio.

Procedimiento

Los participantes, profesores asistentes a los talleres y alumnos asistentes a la práctica, debían evaluar la pregunta-señuelo asignando una puntuación entre 0 y 10 en tres situaciones diferentes.

En una primera situación (S-1) y después de leer una monografía sobre el tema[12], cada participante debía asignar una puntuación a partir de sus propios conocimientos y criterios. En una segunda situación (S-2) se calificaba el examen considerando unos criterios concretos que fueron entregados por escrito por los autores (Anexo 2). Finalmente (S-3), se realizaba una tercera evaluación a partir de unas puntuaciones específicas asociadas a cada uno de los criterios citados anteriormente. Dicha información también era presentada por escrito (Anexo 2).

Para la realización de la experiencia, agrupábamos las calificaciones decimales en cuatro intervalos de resultados: deficientes (DEF) para las puntuaciones inferiores a 5; aceptables (ACE) para las puntuaciones entre 5 y 6,9; notables (NOT) para las puntuaciones entre 7 y 7,9; y eficientes (EFI) para las puntuaciones de 8 o superiores.

RESULTADOS

Las tablas 1 y 2 presentan el número de profesores y de estudiantes que asignaron calificaciones dentro de cada intervalo en cada una de las tres situaciones de evaluación. También se presentan las calificaciones decimales más altas y más bajas que fueron atribuidas al examen en las diferentes evaluaciones.

En ambos casos se puede observar una reducción de la dispersión de calificaciones ante cada evaluación. En S-1 se necesitan tres intervalos para agrupar las notas del 75 % de los evaluadores, en S-2 dos intervalos y en S-3 uno solo. También se puede observar que en las dos poblaciones y en las tres evaluaciones se asignaron notas con rangos muy altos.

Tabla 1. Número de profesores y porcentaje entre paréntesis que asignaron calificaciones dentro de cada intervalo en cada una de las tres situaciones de evaluación.

	DEF	ACE	NOT	EFI	Rango
S-1	2 (2%)	27 (29%)	27 (29%)	36 (40%)	4–9
S-2	23 (25%)	44 (48%)	21 (23%)	4 (4%)	2,3–9
S-3	70 (76%)	20 (22%)	2 (2%)	0	1,2–7,6

S–1= sin criterios comunes; S–2= con criterios comunes; S–3= con precisión en los criterios comunes.

Tabla 2. Número de estudiantes de medicina y porcentaje entre paréntesis que asignaron calificaciones dentro de cada intervalo en cada una de las tres situaciones de evaluación.

	DEF	ACE	NOT	EFI	Rango
S-1	55 (12%)	209 (46%)	116 (25%)	80 (17%)	2–9
S-2	206 (45%)	183 (40%)	47 (10%)	24 (5%)	1–9
S-3	343 (75%)	97 (21%)	15 (3%)	5 (1%)	0,9–8

S–1= sin criterios comunes; S–2= con criterios comunes; S–3= con precisión en los criterios comunes.

Los datos presentan algunas diferencias entre las dos poblaciones de sujetos, los profesores, respecto a los estudiantes, en la primera (S-1) y en la segunda (S-2) evaluaciones asignaron menos notas inferiores a cinco y en la última (S-3) no pusieron ninguna nota igual o superior a ocho. En los dos primeros casos las diferencias fueron significativas (S-1, chi2 = 7,92, gl = 1, p < 0,01; S-2, chi2 = 12,36, gl = 1, p < 0,001) y en el tercero no (chi2 = 1,01, gl = 1, n.s.).

DISCUSIÓN

El objetivo de las experiencias comentadas era doble, por un lado pretendíamos hacer patente la difícil objetividad de las pruebas de ensayo pero también queríamos evidenciar, en caso de utilizarlas, la posibilidad que tenemos para reducir al máximo la subjetividad.

Los resultados del estudio avalan de forma clara las dos previsiones. Por un lado queda patente la poca concordancia entre evaluadores cuando se corrigen las pruebas de ensayo y, por otro, vemos como a medida que precisamos los criterios de evaluación la subjetividad disminuye. Tanto con profesores como con estudiantes, la dispersión máxima de notas se produce en la primera situación, ésta se reduce en la segunda evaluación, cuando se establecen ciertos criterios de evaluación, y la dispersión es mucho menor si se establecen puntuaciones concretas para cada uno de dichos criterios.

De los resultados de nuestro estudio se desprende un hecho que no estaba expresamente buscado. Las diferencias encontradas entre las dos poblaciones de evaluadores podrían ser atribuidas a un posible efecto "halo", otro de los grandes inconvenientes que presentan este tipo de pruebas. En la primera evaluación sin criterios (S-1), vemos que los profesores asignaron menos notas inferiores a cinco que los estudiantes (2 % vs 12 %). Seguramente este hecho se debería a que en el caso de los profesores, éstos evaluaron el examen creyendo que era el de un compañero que compartía el taller. Por el contrario, en el caso de los estudiantes se evaluaba un examen de una persona anónima. Este efecto aún se mantuvo en la segunda evaluación cuando ya se incluían criterios concretos (25 % vs 45 %) pero desapareció en la tercera evaluación cuando la precisión para puntuar era máxima (76 % vs 75 %).

En el estudio hemos presentado los resultados acumulados de la evaluación de un examen (respuesta-señuelo) pero quisiéramos destacar que siempre se produjo el mismo tipo de comportamiento en los participantes que evaluaron el caso en las diferentes sesiones. Es más, también hemos observado el mismo patrón de resultados en la inmensa mayoría de los otros exámenes que han sido evaluados (dos o tres más por taller) durante los más de diez talleres de formación impartidos desde 1993.

Nuestro estudio tiene algunas limitaciones que podían haber influido en los resultados. Por un lado los sujetos participantes, tanto profesores como estudiantes, no eran expertos en la materia y, por otro lado, los criterios de evaluación en las situaciones 2 y 3 fueron fijados por los autores del estudio y no acordados por los propios evaluadores.

Las discrepancias entre evaluadores probablemente habrían sido menores si éstos hubieran sido expertos y si hubieran determinado ellos mismos los criterios de evaluación, ya que se puede prestar más atención a los criterios con mayor peso en la puntuación.

También aceptamos como discutible la agrupación de las calificaciones decimales en categorías o los rangos utilizados para determinarlas. Posiblemente en un estudio exhaustivo sobre el tema se tendrían que conocer las puntuaciones directas para poder observar mejor las variaciones intra-examinadores pero recordamos que el objetivo de la experiencia era mucho más humilde. Respecto a los rangos utilizados son muy semejantes a los utilizados en nuestro país para determinar las calificaciones cualitativas.

De todas formas, creemos que los resultados de nuestro estudio permiten confirmar la subjetividad de la pruebas de ensayo, hecho ya puesto en evidencia con correctores con mucha experiencia[11]. Recordemos los altos rangos de puntuaciones existentes en todos los casos y que, incluso existió un 25 % de evaluadores discrepantes cuando la precisión fue máxima. Si bien lo anterior es cierto, los datos también ponen de manifiesto que es posible disminuir la subjetividad, es sorprendente la semejanza en las evaluaciones de profesores y alumnos en la tercera evaluación. Así, el corolario más relevante de nuestro estudio sería la necesidad de usar criterios de corrección lo más precisos posibles a la hora de evaluar cualquier prueba de ensayo.

Finalmente, y dada la dificultad de objetivar las pruebas de ensayo a pesar de la precisión, creemos necesario, siempre que los objetivos a evaluar lo permitan, utilizar pruebas más objetivas. Como han señalado otros autores [2,13] en la educación médica es recomendable utilizar pruebas diversas para evaluar el rendimiento académico de los estudiantes.

BIBLIOGRAFÍA

1. Guilbert JJ. Education handbook for health personnel. 6th ed. Geneva. : World Health Organization, 1992.
2. Wass V, Van der Vluten C, Shatzer J. Jones R. Assessment of clinical competence. Lancet 2001; 357: 945-9.
3. Cohen-Schotanus J. Student assessment and examination rules. Med Teach 1999; 21: 318-21.
4. Rolfe I, McPherson J. Formative assessment: how am I doing?. Lancet 1995; 345: 837-9.
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 Edu 1986; 20: 356-65.
7. Cox KR. ¿Qué tipo de examen escrito debe utilizarse?. En: Cox KR, Ewan CE. La docencia en Medicina. Barcelona: Doyma, 1990 (pp 162- 65).
8. Breland HM. The direct assessment of writing skills: A measurement review. College Board Report nº 83-6. New York: College Entrance Examination Board, 1983.
9. Ebel RL, Frisbie DA. Essentials of educational measurement (4th ed.). Englewood Cliffs: Prentice-Hall, 1986.
10. Millman J, Greene J. The specification and development of tests of achievement and ability. En: Linn RL (Ed.) Educational Measurement (3rd ed.). New York: MacMillan, 1989 (pp 447-74).
11. GRE Board. Writting proeficiency: How is it assesssed. GRE Board Newsletter 1992; 8:3-4.
12. Pérez J, Torrubia R. Proves de rendiment acadèmic. Monografies pràctiques núm. 3. Bellaterra: Unitat de Psicologia Médica. Universitat Autònoma de Barcelona.
13. Nendaz MR, Tekian A. Assessment in Problem-based learning Medical Schools: A literature review. Teach Learn Med 1999; 11: 232-43.

ANEXO 1

Examen señuelo que debían evaluar los participantes.

Pregunta:

Las pruebas de ensayo largo: ventajas e inconvenientes

Respuesta:

Cuando realiza una prueba de ensayo, un alumno ha de demostrar que puede argumentar una respuesta con cierta coherencia y que su razonamiento ha de ser lógico. Las pruebas de ensayo, además de servir para evaluar la comprensión del alumno sobre un tema concreto, indican cual es la capacidad de expresión de este alumno. Es fundamental en asignaturas de lengua, ya que en estos casos es casi tan importante lo que se dice como la manera de decirlo. Sin embargo, las pruebas de ensayo tienen algunos inconvenientes: sólo permiten preguntar sobre una pequeña parte del temario, dejan poca creatividad al alumno (a no ser que, a través de la pregunta del examen, obligues a relacionar los conocimientos) y a menudo dan pié a divagar sobre los temas, aunque no se sepa de qué se habla. Para un profesor, los exámenes de ensayo extenso son los más difíciles de evaluar: el esfuerzo por entender lo que quiere transmitir el alumno es importante, y es necesario tener en cuenta que cada examen estará redactado de una manera diferente, lo cual hará más lenta la corrección.

ANEXO 2

Criterios para la corrección de la prueba de ensayo.

	Puntuación
Referencia al proceso educativo	1
Definición de las pruebas de ensayo	1
Pruebas en función de los objetivos educativos	2
Preparación rápida	0,6
Posibilidad de expresar conocimientos ordenando ideas propias	0,6
Utilidad para medir procesos complejos: síntesis y evaluación	0,6
Poco eficaz para medir conocimientos	0,6
Falta de validez de contenido	0,6
Corrección larga	0,6
No objetividad	0,6
Poco fiable	0,6
Influencias externas del alumno (letra, presentación, etc.) del profesor (emociones, cansancio, efecto halo, etc.)	0,6
Conocimiento demorado de resultados	0,6

Nota: En la situación dos (S-2) los sujetos recibían los criterios sin puntuación asociada. En la situación tres (S-3) los sujetos recibían los criterios con la puntuación asociada.