Gac Sanit. 2012;**26(3)**:231–235

Original

Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados

Albert Bonillo

Departamento de Psicobiología y Metodología de Ciencias de la Salud, Universitat Autònoma de Barcelona, Bellaterra, Barcelona, España

INFORMACIÓN DEL ARTÍCULO

Historia del artículo:
Recibido el 21 de junio de 2011
Aceptado el 14 de septiembre de 2011
On-line el 22 de diciembre de 2011

Palabras clave:
Internado y residencia
Criterios de admisión escolar
Médicos hospitalarios
Psicometría

R E S U M E N

Objetivos: Estudiar las pruebas de acceso a la Formación Sanitaria Especializada de las convocatorias 2005 y 2006. Se pretende evaluar la calidad de los exámenes y explorar las variables de los aspirantes que permiten predecir la puntuación final.
Métodos: El Ministerio de Sanidad y Consumo proporcionó las respuestas de los 23.136 aspirantes de ambas convocatorias, así como variables demográficas y el valor baremado de su expediente académico.
Resultados: Se realiza un análisis de ítems a partir de las respuestas de los aspirantes para así evaluar la fiabilidad de las pruebas. Además, se calculan modelos de regresión lineal para estudiar qué variables permiten predecir la puntuación final de un aspirante.
Conclusiones: Las pruebas de acceso a la Formación Sanitaria Especializada tienen una excelente calidad psicométrica. Serían optimizables reduciendo el número de alternativas y eliminando algunos ítems más a posteriori. Por último, los alumnos españoles son los que mejor nota media ajustada logran.

Access tests to specialized health training for doctors and other healthcare professionals in Spain: examining the exam and the examined candidates

Keywords:
Internship and residency
School admission criteria
Hospitalists
Psychometrics

A B S T R A C T

Objectives: This study examines the accessing tests to Specialised Health Training for 2005 and 2006 calls. It aims to assess the quality of exams and explore candidates' variables that allow predicting the final score.
Methods: The Ministry of Health provided the 23,136 candidates' answer from both calls, plus demographic variables and the normalised value of their student record.
Results: Item's analysis is made from candidates' answers for evaluating the test reliability. In addition, it's been calculated linear regression models for studying which variables allow predicting a candidate's final score.
Conclusions: the accessing tests to Specialised Health Training have excellent psychometric quality. It would be improved by reducing the number of choices and eliminating some items more afterwards. Finally, Spanish students are the ones who achieved best adjusted mean score.

Introducción

La nota final que da acceso a la formación sanitaria especializada se obtiene a partir del expediente académico y del resultado de una prueba específica de cada programa. Ésta consiste en un examen tipo test con cinco opciones de respuesta[1]. El aspirante puede escoger especialidad y centro en función del orden de su nota final. La plaza a la cual se aspira ha dado lugar al nombre «popular» de la prueba: MIR para medicina, FIR para farmacia y EIR para enfermería, entre otros.

La formación sanitaria especializada no ha recibido demasiada atención por parte de la literatura, y las pruebas que rigen su acceso aún menos, con la salvedad del MIR[2]. Se han realizado propuestas de reforma del acceso[3], así como análisis de la validez del formato de respuesta múltiple respecto a otros también habituales en la docencia médica[4]. En otros países, exámenes análogos son estudiados y analizados profusamente; a modo de ejemplo, una búsqueda no exhaustiva en Medline arroja 270 referencias sobre el USMLE (*United States Medical Licensing Examination*).

Existen varias directrices con recomendaciones para redactar óptimamente preguntas de tipo test[5,6], y se ha estudiado profundamente el efecto que tiene el no respetarlas, también en pruebas médicas[7–9]. Las guías indican que un aspecto importante a estudiar son las propiedades psicométricas de un examen, tanto más si éste se repite a lo largo del tiempo y su resultado tiene efectos sociales importantes. En una prueba de rendimiento, como es un examen, esto implica analizar la dificultad y la discriminación de cada una de las alternativas[10]. Además, es de interés estudiar cómo se relacionan las características de los individuos con su rendimiento. Este análisis permite no sólo conocer a los participantes sino corregir posibles sesgos de las pruebas.

En ninguna de las búsquedas realizadas en bases bibliográficas especializadas (Medline, Scielo, Cochrane Plus y Medes) hemos

Correo electrónico: albert.bonillo@uab.cat

0213-9111/$ – see front matter
doi:10.1016/j.gaceta.2011.09.020

Tabla 1
Exámenes de acceso a la formación sanitaria especializada (España 2005-2006): descripción de variables por programa y año de convocatoria

	EIR-05 N = 4.653	FIR-05 N = 829	FIR-06 N = 849	MIR-05 N = 7.908	MIR-06 N = 8.897
Edad, años, media (DT)	27,1 (5,0)	26,7 (3,9)	26,0 (4,1)	29,2 (6,2)	28,5 (6,4)
Años fin estudios, media (DT)	4,8 (4,0)	2,8 (2,9)	3,0 (3,0)	3,5 (4,8)	3,7 (5,0)
Expediente, media (DT)	1,76 (,41)	1,79 (,55)	1,75 (,52)	1,89 (,54)	1,88 (,55)
Examen, media (DT)	85,3 (35,3)	331,6 (125,1)	348,8 (130,0)	297,0 (136,4)	306,9 (142,0)

DT: desviación típica.

encontrado un análisis de las propiedades de las pruebas de acceso a la formación sanitaria especializada, y tampoco ningún trabajo que aporte datos sobre qué características de los aspirantes son predictoras de un buen resultado en las pruebas.

Este trabajo pretende cubrir este vacío, cuantificando dichos aspectos. Considerando la larga historia de las pruebas de acceso a la formación sanitaria especializada, esperamos encontrar que sus propiedades psicométricas sean adecuadas. En cuanto a las características de los participantes, no tenemos ninguna hipótesis, ya que no hay estudios previos sobre ellas.

Método

Tras efectuar una petición, la Subdirección General de Ordenación Profesional del Ministerio de Sanidad, que es la encargada de gestionar las pruebas de formación sanitaria especializada, aceptó proporcionarnos datos relativos a las convocatorias de los años 2005 y 2006, que se realizaron en enero de 2006 y de 2007, respectivamente.

Las matrices contenían, para las convocatorias de MIR, FIR y EIR, datos individuales de cada aspirante presentado, obviamente, sin identificador personal alguno. De la convocatoria de EIR se trabajó sólo con la de 2005 porque los datos de la de 2006 estaban dañados y no podían recuperarse.

Las variables registradas de cada individuo pueden agruparse en tres bloques: por un lado, demográficos (sexo, edad y nacionalidad), por otro, relativas a su currículo académico (nota baremada de su licenciatura o diplomatura [rango teórico entre 1 y 4] y año en que la finalizó); y por último las relativas a la prueba que completó, es decir, la respuesta dada a cada ítem/pregunta del examen y su nota final. La tabla 1 muestra la descripción de estas variables para los 23.136 aspirantes estudiados.

Las pruebas de MIR y de FIR tienen 250 preguntas más 10 de reserva, que se usan para, por orden, sustituir a las que son eliminadas por su mal rendimiento o ambigüedad tras un proceso de análisis interno y de impugnación externa. El examen de EIR sigue el mismo esquema, con la salvedad de que usa 100 preguntas y 10 de reserva. Cada pregunta acertada suma tres puntos, cada una fallada resta uno y se obvian las preguntas no contestadas. El rango teórico de puntuaciones es entre −250 y 750 para MIR y FIR, y de −100 a 300 para EIR.

Para nuestro análisis hemos eliminado las preguntas que fueron anuladas y aquellos aspirantes que no contestaron ninguna. Dejaron su examen completamente en blanco, en las convocatorias de 2005 y de 2006, respectivamente, 1085 (12,1%) y 1359 (14,1%) en el MIR, 119 (12,5%) y 126 (12,9%) en el FIR y 1246 (21%) en el EIR.

En la primera parte de los resultados se presenta el análisis de ítems de cada convocatoria, que se basa en dos índices. Por un lado, la dificultad de una pregunta se entiende como la proporción de personas que la aciertan de entre los que la han contestado. En ítems de elección múltiple (esto es, no binarios) es habitual corregir esta proporción restándole la de errores (dividida entre el número de opciones de respuesta menos uno) para así sustraer del acierto aquella parte que se explica por puro azar[11]. Paradójicamente, un valor alto del índice indica baja dificultad. Por otro lado, se entiende como discriminación de un ítem su capacidad para distinguir entre los participantes con un alto frente a un bajo rendimiento[12]. Puede calcularse mediante la correlación entre la puntuación del ítem y la total de la prueba, o bien utilizando el índice D. Éste, que ha sido el utilizado en el presente trabajo, se calcula mediante la diferencia de proporciones de acertantes entre el grupo de personas con un rendimiento en la prueba superior al 73% del resto frente al grupo que ha obtenido un rendimiento inferior al 23%. Ebel[13] propuso puntos de corte para evaluar el índice D, siendo menor de 0,2 inaceptable, menor de 0,3 revisable, menor de 0,4 aceptable pero mejorable, y mayor de 0,4 óptimo. Además, es deseable que la D sea muy semejante entre distractores de un mismo ítem, ya que sería indicador de que son iguales en atractivo[14].

En la segunda parte del apartado de resultados se utilizan modelos de regresión lineal para explicar la puntuación total de la prueba a partir de las variables demográficas (sexo, edad, nacionalidad) y relativas a la licenciatura/diplomatura (años desde su finalización y puntuación del expediente baremada). Puesto que no se trata de datos muestrales sino poblacionales, no tiene sentido realizar pruebas de significación ni mostrar los errores estándar. Se utiliza esta técnica multivariada para valorar la contribución propia de cada predictor, ajustando su efecto por el resto, y evitar la confusión inherente a todo estudio no experimental[15]. Para poder comparar variables con distintas métricas, o la misma entre diferentes convocatorias, se muestran los coeficientes b y beta.

Se han incluido en las ecuaciones todas las variables predictoras, tanto demográficas como relativas al pasado académico del candidato, que estaban registradas. La variable «nacionalidad» sólo se ha incluido en las convocatorias de MIR por ser éstas las que presentan proporciones no anecdóticas de extranjeros presentados.

Para calcular los índices del análisis de ítems se han creado algoritmos *ad hoc*. Todos los cálculos se han efectuado con SPSS v.19.

Resultados

Análisis de ítems

La figura 1 muestra las curvas de dificultad (ordenando por ésta los ítems en abscisas) de cada una las pruebas. En este tipo de gráfico, una pendiente de 45° indicaría un escalamiento óptimo de los ítems, ya que habría preguntas para todos los grados de dificultad posibles.

En primer lugar, se constata que hay dificultades negativas y que sólo pueden explicarse por aplicar la corrección del azar a ítems muy difíciles. En segundo lugar, cabe tener en cuenta que la prueba de enfermería cuenta con 100 ítems (frente a los 250 del resto de las convocatorias), cosa que explica la evidente diferencia en las pendientes. En tercer lugar, las curvas de las pruebas de farmacia son superiores a las de medicina, lo que indica que son pruebas más fáciles. En cuarto lugar, las curvas son muy semejantes intraprogramas, es decir, la dificultad de las pruebas es muy semejante año tras año.

La figura 2 muestra, mediante un diagrama de cajas, la discriminación de la opción correcta y de cada uno de los distractores, ordenados éstos de mayor a menor, de las cinco convocatorias analizadas (esto es, [250 ítems × 2 programas × 2 años + 100 ítem de

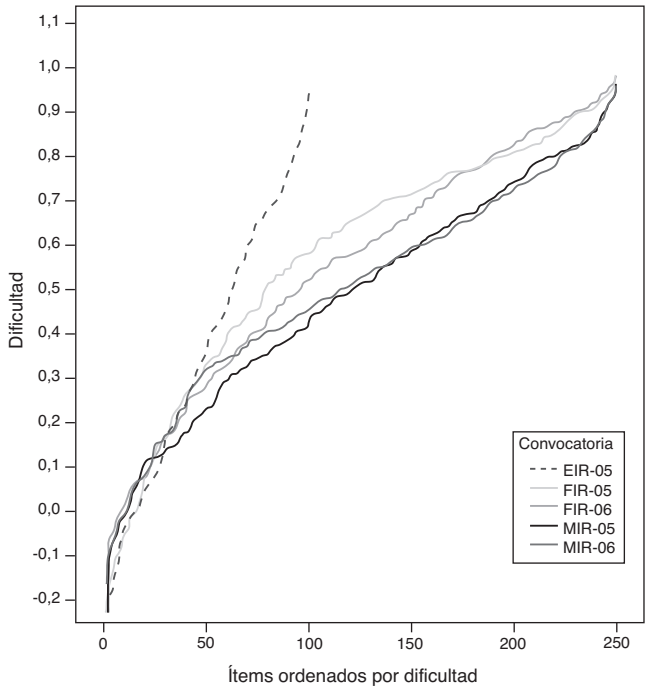

Figura 1. Curvas de dificultad de los ítems por especialidad y año de convocatoria.

EIR] × 5 alternativas = 5500 valores). Así, y para cada ítem, el distractor 1 es el más discriminativo y el 4 el menos. Como es habitual en este tipo de gráficos, las cajas muestran la mediana (en trazo grueso) y los cuartos (en los límites de las cajas). Las patillas (*whiskers*) muestran los valores mínimos y máximos no alejados, que se muestran con puntos, ni extremos, que lo hacen con asteriscos.

Se observa que las discriminaciones de las alternativas correctas son semejantes entre especialidades y convocatorias. La mediana se sitúa entre 0,26 y 0,34, y la anchura de las cajas es acorde a la variabilidad esperable con un número tan alto de ítems. Destacan de las demás las discriminaciones relativas a la prueba de EIR, que son más bajas y menos dispersas. En el análisis de los distractores se observa que hay un escalado entre ellos, pero que se reduce cuantas más alternativas se contemplan; es decir, la diferencia entre la tercera y la cuarta alternativa es mucho menor que entre la primera y la segunda. También se observa que las alternativas tres y cuatro (recordemos que están ordenadas por su discriminación y que no deben identificarse con alternativas de respuesta D y E, por ejemplo) tienen discriminaciones muy bajas o casi nulas. Si observamos que el límite superior de las cajas de la última alternativa es superior a 0, podemos decir que más del 25% de los ítems tienen una alternativa de respuesta con discriminación positiva (es decir, más escogida por el grupo con rendimiento alto). Además, la última alternativa presenta muchos valores extremos y alejados, esto es, ítems en los cuales, por su alta discriminación positiva, sería discutible si la opción dada como correcta verdaderamente lo es, o es la única que lo es.

Al igual que ocurría con la discriminación de la alternativa correcta, la convocatoria de EIR tiene discriminaciones menores en todas sus alternativas que el resto de las convocatorias, que son muy semejantes entre sí.

Análisis de datos

La tabla 2 muestra las regresiones lineales calculadas para cada una de las convocatorias.

El valor de R^2 es especialmente alto en medicina (45% a 48%) y menor, pero destacable, en el resto de las especialidades (16% a 19% en farmacia, 13% en enfermería).

Si observamos los coeficientes beta de cada uno de los modelos, vemos que el orden de su importancia es muy semejante entre especialidades. Excepto en el caso de enfermería, el coeficiente más importante es, sorprendentemente, la edad, y en segundo lugar

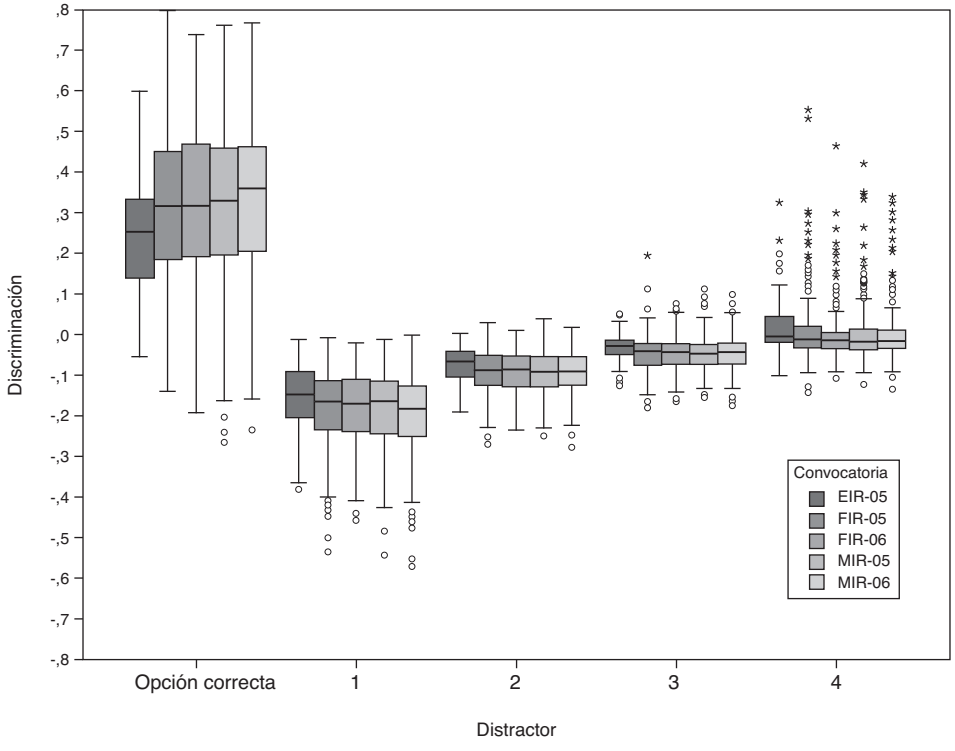

Figura 2. Diagrama de cajas de la discriminación de las alternativas de respuesta.

Tabla 2
Exámenes de acceso a la Formación Sanitaria Especializada (España 2005-2006): modelos de regresión lineal múltiple para explicar la puntuación del examen de acceso a partir de las variables predictoras

	EIR-05 R^2=13%		FIR-05 R^2=16%		FIR-06 R^2=19%		MIR-05 R^2=45%		MIR-06 R^2=48%	
	b	β	b	β	b	β	b	β	b	β
Constante	70,8		635,1		657,2		483,9		533,5	
Sexo[a]	1,2	0,01	−16,5	−0,05	−17,7	−0,05	−18,2	−0,06	−19,7	−0,07
Expediente	24,4	0,29	43,7	0,19	48,3	0,20	83,3	0,33	79,7	0,30
Edad	−1,6	−0,23	−15,3	−0,48	−15,3	−0,49	−11,1	−0,51	−12,3	−0,56
Años fin estudios	2,8	−0,20	10,5	0,24	6,2	0,14	2,2	0,08	2,9	0,10
Europeos[b]			NA				−82,9	−0,11	−94,2	−0,13
Americanos[b]							−85,4	−0,19	−91,7	−0,23
Otros[b]							−108,4	−0,09	−94,0	−,09

NA: no aplicable.
[a] Masculino=0, femenino=1.
[b] Comparación frente a españoles.

el expediente. El coeficiente beta indica que, por cada desviación típica (DT) de edad (entre 4 y 6 años; ver tabla 1) la puntuación del examen disminuye en torno a media DT (-0,48 a -0,56). Examinando el coeficiente b, vemos que cada año que cumple el candidato le «cuesta» 15,3 puntos en el examen FIR y 11,1 a 12,3 en el MIR. En el caso del expediente, cada punto de éste explica entre 43,7 y 83,3 puntos del resultado final del examen. Para enfermería, los factores expediente y edad tienen un peso muy semejante (0,23 a 0,29).

En cuanto al sexo, todos los coeficientes (de nuevo excepto en enfermería) son negativos, lo que indica que, en términos ajustados, los hombres obtienen una mejor calificación que las mujeres. Ahora bien, el valor de beta es muy pequeño (-0,05 a -0,07).

Los coeficientes de la variable «nacionalidad», calculados sólo para el MIR, muestran que los españoles son los que obtienen una nota más alta en el examen, ya que todos los coeficientes son negativos. Los coeficientes b indican que, en términos medios y ajustados, tener una nacionalidad diferente de la española representa obtener entre 83 y 108 puntos menos.

Discusión

A partir del análisis de las dificultades, los resultados muestran que algunas preguntas más podrían haber sido eliminadas del examen, al tener dificultades (corregidas) negativas. Es cierto que se trata de muy pocos ítems (4 o 5 por convocatoria) y que la afectación de no hacerlo en validez de la puntuación total es nula, pero habiendo aún preguntas de reserva no hay ningún motivo para mantenerlas.

Las pruebas presentan un buen escalado, que cubre de forma suficiente todas las zonas de dificultad. Ahora bien, y con la salvedad de EIR, un objetivo a lograr sería que las pendientes fueran algo más cercanas a los 45° que definen lo óptimo. Que las pendientes de la dificultad crezcan antes de llegar a los 50 ítems indica un ligero déficit de preguntas difíciles, que podría ser corregido. Un examen es un instrumento de medida y debe hacer su tarea para todo el continuo del rendimiento. Un déficit de preguntas difíciles produce una peor medida de los mejores aspirantes, que son justamente los que mejor se pretende ordenar para que su oportunidad de escoger las plazas más apetecibles sea lo más justa posible. Por otro lado, la semejanza entre las curvas de un mismo programa aporta evidencia sobre la equivalencia de las pruebas año tras año, garantizando así la equidad (al menos la temporal) en el acceso, que es un objetivo primario de las pruebas[1].

Del análisis de las discriminaciones pueden extraerse varias conclusiones. La más importante de ellas es recomendar reducir las opciones de respuesta a tres, que es el valor aceptado como óptimo en la literatura[16]. En el gráfico se observa que los dos distractores menos discriminantes de cada pregunta aportan muy poco, pues sus alturas son semejantes, y tienen muchos puntos alejados por la parte superior, lo que indica una mala discriminación. Redactar más de dos distractores verosímiles es muy difícil y rara vez aportan algo sustantivo[17], más allá de alargar innecesariamente la prueba.

Los valores de R^2 de las regresiones lineales muestran que se puede predecir la nota en el examen con precisión suficiente a partir de las características del aspirante. Destaca la semejanza entre modelos del mismo programa, lo que indica que la relación entre el perfil del aspirante y su rendimiento es estable. Que esto no se diera no indicaría necesariamente algún problema en las pruebas, pero sí habría que estudiar por qué ocurre.

En las convocatorias de FIR y MIR llama la atención el peso negativo del sexo, esto es, que tras ajustar por el resto de las características, los hombres obtienen una mejor calificación que las mujeres. Aunque el peso del coeficiente beta es pequeño, el b indica una diferencia media de 20 puntos. Este dato invierte totalmente el ofrecido por las medias sin ajustar (10 puntos más para las mujeres, en MIR, y medias prácticamente iguales en FIR). Esta diferencia se explica por el efecto que tiene evitar la confusión ajustando un modelo multivariado. En el caso de los candidatos a FIR, el ajuste se explica por la variable «edad», ya que en términos medios los chicos que se presentan son mayores que las chicas. Al retirar el efecto de la edad, la igualdad entre sexos que sugería la media univariada se muestra como falsa. En el caso de la convocatoria MIR, la principal explicación del ajuste viene dada por la variable «nacionalidad»: buena parte de los extranjeros presentados son hombres. Sobre esta variable, en primer lugar, es evidente que tener un pasaporte u otro no puede incidir sobre el rendimiento de una prueba. Creemos que la nacionalidad es un indicador indirecto (porque seguro que algunos aspirantes españoles se licenciaron en otro país, y viceversa) del país en que el aspirante ha realizado sus estudios. En segundo lugar, el mejor rendimiento lo muestran los estudiantes españoles: en términos medios, entre 83 y 108 puntos, esto es, entre un cuarto y un tercio de las puntuaciones medias. Este dato se explica por dos efectos: por un lado, es un indicio de la calidad de nuestras facultades de medicina, y por otro muestra la focalización de los médicos españoles en la prueba MIR[18]. Conviene recordar la modificación de las condiciones de acceso al examen MIR en la convocatoria de enero de 2011, que establece un cupo máximo de plazas para los no comunitarios de un 10% del total, ampliable a un 15% en caso de haber plazas no cubiertas. Es probable que este cambio haya corregido el sesgo que hemos detectado.

Este estudio es el primero que analiza, con datos reales de las propias pruebas, tanto las preguntas como los aspirantes a la formación sanitaria especializada. En cuanto a las pruebas, los resultados confirman lo esperado, esto es, que son muy adecuadas en términos psicométricos, aunque algunos aspectos son mejorables.

Debe quedar patente que del estudio de la dificultad y de la discriminación de las preguntas de los exámenes no puede extraerse conclusión alguna sobre la bondad global del acceso a la formación

sanitaria especializada. La evaluación del modelo debe hacerse a partir de qué habilidades y competencias queremos que posea el profesional sanitario especializado, y no desde la evaluación de las propiedades matemáticas de las preguntas que ahora se les realizan. Ahora bien, creemos necesario estudiar el examen que, aún a día de hoy, regula el acceso. Coincidimos con Gual y Pardell[19] en que debería valorarse el nivel de competencia global del aspirante y no sólo sus conocimientos teóricos (como ahora ocurre), intentando medir el «currículo oculto», que aúna habilidades transversales, aspectos éticos y de profesionalidad (entre otros), que son de muy difícil cuantificación con preguntas tipo test.

En cuanto a los aspirantes, se aportan evidencias que permitirían predecir su rendimiento en las pruebas y aplicar políticas que hagan posible asegurar aún más, si cabe, su equidad. Finalmente, por la importancia de estas pruebas, creemos que deberían replicarse análisis análogos a éstos y ser publicados año tras año, para así poder mejorarlas.

¿Qué se sabe sobre el tema?

Las propiedades psicométricas de las pruebas de acceso a la formación clínica especializada han sido estudiadas en otros países, pero no en el nuestro.

¿Qué aporta el estudio realizado a la literatura?

Este estudio es pionero es sus objetivos y en sus métodos, en tanto en cuanto no conocemos, en nuestro país, ninguno semejante. Las principales conclusiones que aporta son dos: la excelente salud de las pruebas de acceso a la formación sanitaria especializada (aunque, como todo, son mejorables) y la posibilidad de predecir la puntuación de una persona en ellas a partir de sus variables.

Contribuciones de autoría

A. Bonillo realizó los trámites para conseguir los datos, redactó el artículo y lo revisó.

Financiación

Ninguna.

Conflictos de intereses

Ninguno.

Agradecimientos

A Esther Laso Esteban, entre otras cosas por su exhaustiva revisión.

Bibliografía

1. Programas de Formación Sanitaria Especializada. Ministerio de Sanidad y Consumo. (Consultado el 8/6/11.) Disponible en: http://sis.msc.es
2. Vazquez G, Murillo F, Gómez J, et al. El examen MIR, su cambio como una opción estratégica. Educ Med. 2008;11:203–6.
3. Torres M, Cardellach F, Bundó M, et al. Sistema formativo MIR: propuesta de cambios para la adecuación a las necesidades del modelo sanitario. Med Clin (Barc). 2008;131:777–82.
4. Gómez-Sáez JM, Pujol-Farriols R, Martínez-Carretero JM, et al. El proyecto COMBELL. Un análisis de la competencia clínica médica. Med Clin (Barc). 1995;105:649–51.
5. Haladyna TM, Downing SM, Rodríguez M. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education. 2002;15:309–33.
6. Moreno R, Martínez R, Muñiz J. Directrices para la construcción de ítems de elección múltiple. Psicothema. 2004;16:490–7.
7. Downing SM. The effects of violating standard item writing principles on test and students: the consequences of using flawed test items on achievement examinations in medical education. Advances in Health Sciences Education. 2005;10:133–43.
8. Stagnaro-Green A, Downing SM. Use of flawed multiple-choice items by the New England Journal of Medicine for continuing medical education. Med Teach. 2006;28:566–8.
9. Beullens J, Van Damme B, Jaspaert H, et al. Are extended-matching multiple-choice items appropriate for a final test in medical education? Med Teach. 2002;24:390–5.
10. Muñiz J. Teoría clásica de los test. 5ª ed. Madrid: Pirámide; 1998, 387 p.
11. Page A. Elementos de psicometría. Madrid: Eudema; 1993, 128 p.
12. Guilbert JJ. Educational handbook for health Personnel. 5th ed. Geneva: World Health Organization; 1987, 362 p.
13. Ebel RL. The reliability of an index of item discrimination. Educational and Psychological Measurement. 1951;11:403–8.
14. Murphy KR, Davidshofer CO. Psychological testing: principles and applications. 6th ed. Englewood Cliffs, NJ: Prentice Hall; 2004, 624 p.
15. Rothman KJ, Greenland S. Modern epidemiology. Philadelphia: Lippincott Williams & Wilkins; 1986, 380 p.
16. Rodríguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. Educational Measurement: Issues and Practice. 2005;24:3–13.
17. Budescu DV, Nevo B. Optimal number of options: an investigation of the assumption of proportionality. Journal of Educational Measurement. 1985;22:183–96.
18. Alonso MI. La gestión del proceso de convocatoria de plazas de formación médica especializada. Gac Sanit. 2003;17:289–95.
19. Gual A, Pardell H, coordinadores. El médico del futuro. Barcelona: Fundación Educación Médica (FEM); 2009. 100 p.