Evolución de los parámetros dificultad y discriminación en el ejercicio de examen MIR. Análisis de las convocatorias de 2009 a 2017

Jaime Baladrón, Fernando Sánchez-Lasheras, José M. Romeo-Ladrero, José Curbelo, Paloma Villacampa-Menéndez, Paula Jiménez-Fonseca

Introducción. En España, el procedimiento reglado de acceso a la formación médica especializada tipo MIR exige la superación de un examen. El examen es un ejercicio tipo test de respuesta múltiple. Superado éste, se bareman los méritos académicos previos de los aspirantes. La ponderación de ambos valores permite sumarlos y ordenar a los aspirantes por su puntuación total. Los aspirantes tienen la oportunidad de elegir especialidad y centro de formación en función del número de orden obtenido.

Sujetos y métodos. La base de datos utilizada en el presente trabajo contiene las respuestas de un total de 23.809 candidatos presentados a los exámenes MIR de las convocatorias comprendidas entre 2009 y 2017. La información disponible se analizó utilizando el índice de dificultad, índice de dificultad con corrección de los efectos del azar, índice de discriminación e índice de correlación puntual biserial, y las dificultades y discriminaciones se calcularon usando el modelo de dos parámetros de la teoría de respuesta al ítem.

Resultados. Los resultados obtenidos en la serie temporal analizada permiten afirmar que la dificultad del examen ha disminuido en las tres últimas convocatorias. Además, la discriminación alcanza el valor mínimo en la convocatoria de 2016.

Conclusiones. Dado el incremento en los últimos años del número de médicos españoles recién graduados que se presentan al examen, sería deseable mejorar la calidad psicométrica de las preguntas con el fin de que resulten al menos de la misma calidad como lo eran en los exámenes de convocatorias previas. La inclusión, por parte de la comisión calificadora, de criterios psicométricos de anulación de preguntas ayudaría a conseguir este objetivo.

Palabras clave. Estudiantes. Medicina. Mediciones educativas. Prueba MIR. Psicometría.

Evolution of the difficulty and discrimination of the MIR test in recent years. A study of the calls since 2009 to 2017

Introduction. In Spain, the regulated procedures of specialized medical training impose the passing of the MIR test. After this, the candidates choose the preferred specialty according to the order number obtained. The order number depends on the score of the applicant in the MIR exam and their academic record.

Subjects and methods. The database used in the present research contains the responses of a total of 23,809 candidates submitted to the MIR exams since 2009 to 2017. The available information was analyzed using the difficulty index, difficulty index with correction of random effects, discrimination index, biserial point correlation index, as well as the difficulties and discriminations calculated using a two parameters model of the Item Response Theory.

Results. The results obtained let us to say that the difficulty of the MIR test has decreased in the last three calls. In addition, the discrimination reaches the minimum value of the time series in the 2016 MIR test.

Conclusions. Given the increase in recent years in the number of recently graduated Spanish doctors who take the MIR exam, it would be desirable to try to improve the psychometric quality of the questions in order to be at least as discriminative as the tests of previous calls. The inclusion by the committee of psychometric criteria of cancellation of questions would help achieve this objective.

Key words. Educational measurements. Medicine. MIR exam. Psychometrics. Students.

Director del Curso Intensivo MIR Asturias; Oviedo (J. Baladrón). Departamento de Matemáticas; Facultad de Ciencias; Universidad de Oviedo; Gijón (F. Sánchez-Lasheras). Editor del blog MIRentrelazados; Zaragoza (J.M. Romeo-Ladrero). Servicio de Medicina Interna; Hospital Universitario La Princesa; Madrid (J. Curbelo). Graduada en Medicina; Avilés (P. Villacampa-Menéndez). Servicio de Oncología Médica; Hospital Universitario Central de Asturias; Oviedo, Asturias, España (P. Jiménez-Fonseca).

Correspondencia:
Dr. Fernando Sánchez Lasheras. Departamento de Matemáticas. Facultad de Ciencias. Universidad de Oviedo. Federico García Lorca, 18. E-33007 Gijón (Asturias).

E-mail:
sanchezfernando@uniovi.es

Recibido:
03.01.18.

Aceptado:
16.01.18.

Conflicto de intereses:
No declarado.

Competing interests:
None declared.

Introducción

El ejercicio como médico especialista en España requiere estar en posesión del título oficial de la especialidad correspondiente. Desde la década de los años cincuenta, la obtención del título de especialista ha estado sometida a diferentes regulaciones. La Ley de 20 de julio de 1955 sobre enseñanza, título y ejercicio de las especialidades médicas, y los Reales Decretos 2015/1978, de 15 de junio por el que se regula la obtención del título de médico especialista [1], y el RD 127/1984, de 11 de enero, por el que se regula la formación médica especializada y obtención del título de médico especialista [2] fueron, sucesivamente, los marcos regulatorios de las enseñanzas médicas especializadas.

Ya en el presente siglo, el modelo de acceso a la formación médica especializada ha quedado regulado por la Ley de Ordenación de las Profesiones Sanitarias (LOPS), Ley 44/2003, de 21 de noviembre de 2003 [3] y sus sucesivas regulaciones. La existencia de la prueba MIR se remonta a 1978. Es una prueba nacional de carácter selectivo que se convoca anualmente y que depende del Ministerio de Sanidad, Consumo y Bienestar Social y Ministerio de Educación y Formación Profesional.

En los últimos años, la convocatoria de la prueba se publica oficialmente a finales del mes de septiembre y el examen tiene lugar a finales del mes de enero o primeros de febrero del año siguiente. El examen de la prueba selectiva de acceso evalúa conocimientos y habilidades clínicas y comunicativas de los aspirantes.

Si bien a lo largo de los años la preferencia de los candidatos por las distintas especialidades ha ido variando [4-7], existen especialidades cuya elección ha quedado históricamente reservada a los candidatos situados en el primer cuartil.

Desde la convocatoria 2009, el ejercicio de examen ha consistido en un test de 225 preguntas, las primeras de ellas vinculadas a imágenes diagnósticas. Desde la convocatoria 2015, las preguntas tienen cuatro alternativas de respuesta, cuando las preguntas de las convocatorias anteriores contaban con cinco opciones. Además, el mencionado ejercicio se completa con otras diez preguntas de reserva que se activan en caso de que alguna de las 225 primeras preguntas sea anulada por la comisión calificadora del examen. La duración del examen es de cinco horas y su contenido está relacionado con las diferentes asignaturas que los médicos han cursado a lo largo de la licenciatura o grado de medicina.

El objetivo del presente trabajo es evaluar la evolución de la dificultad y discriminación, medidas por distintas métricas, de los ejercicios de examen de las convocatorias comprendidas entre 2009 y 2017. Desde el punto de vista de los autores, dicha evaluación permitirá explicar los cambios que se han producido en los resultados obtenidos en las tres últimas convocatorias y, fundamentalmente, lo ocurrido en el examen MIR de la convocatoria 2016.

Sujetos y métodos

Base de datos. Muestra poblacional

Las bases de datos que se utilizaron en este estudio corresponden a las respuestas a las preguntas, que los propios examinados de cada una de las convocatorias de la prueba MIR 2009-2017 introdujeron en una aplicación web desarrollada para este propósito por el Curso Intensivo MIR Asturias. Tras la realización del ejercicio de examen, todos los examinados participantes que quisieron pudieron introducir en la aplicación sus respuestas a las 235 preguntas del examen y su baremo académico, con objeto de acceder al servicio de corrección del examen y de estimación del número de orden que obtendrían en su convocatoria. La base de datos obtenida, una vez eliminados los considerados espurios, es la que se ha empleado para este estudio.

El número de ejercicios corregidos disponible en cada convocatoria para su análisis ha sido diferente, con un mínimo de 1.698 en la convocatoria de 2011 y un máximo de 4.190 en la de 2017. La media de ejercicios registrados en la aplicación para las convocatorias analizadas fue de 2.684. El número exacto de examinados que introdujeron los resultados en cada una de las convocatorias se recoge en la figura 1. Como se observa, el porcentaje de médicos de los que se dispone de información relativa a sus respuestas a las preguntas, en las últimas convocatorias, supera el 30% del total de presentados al examen.

A la hora de analizar los resultados para cada una de las convocatorias se han considerado únicamente las 225 preguntas que se han empleado para el cálculo de la puntuación en cada una de ellas. Es decir, no se han tenido en cuenta las preguntas anuladas ni las preguntas de reserva que no fueron activadas por la comisión calificadora de la prueba en la plantilla de respuestas definitivas del examen.

Clasificación de las preguntas

Cada pregunta de cada convocatoria se ha clasificado atendiendo a su área temática o asignatura, y al tipo de pregunta. Respecto al área temática se han

clasificado en: bioestadística y medicina preventiva, microbiología y enfermedades infecciosas, aparato digestivo y cirugía digestiva, neumología y cirugía torácica, cardiología y cirugía cardiaca, nefrología y urología, ginecología y obstetricia, neurología y neurocirugía, endocrinología y cirugía endocrina, hematología, reumatología, pediatría, psiquiatría, cirugía ortopédica y traumatología, dermatología, farmacología clínica, otorrinolaringología, inmunología, oftalmología, anatomía patológica, fisiología, genética, gestión clínica, anatomía, geriatría, oncología, habilidades comunicativas, cuidados paliativos, cirugía vascular, urgencias, cirugía maxilofacial, cirugía plástica y reparadora, medicina legal y bioética, bioquímica y radiología.

En lo que respecta al tipo de pregunta, se han agrupado en cuatro categorías: casos clínicos con imagen, casos clínicos sin imagen, preguntas negativas y preguntas de test. Bajo la categoría de casos clínicos, se agruparon todas aquellas preguntas de la prueba en las que el enunciado contiene numerosos datos descriptivos del cuadro patológico de un paciente o de un problema bioestadístico, diferenciándose entre las que van asociadas a una imagen y las que no. Se denominan preguntas negativas aquellas en donde en su enunciado se pregunta por la opción incorrecta. Es decir, el enunciado incluye palabras tales como 'falso', 'excepto', 'incorrecta', 'no incluye', etc. Finalmente, bajo la denominación genérica de preguntas de test, se incluyen todas las que no se han clasificado dentro de ninguna de las categorías anteriores, siendo habitualmente preguntas directas con enunciados y opciones de respuesta cortas.

Índice de dificultad

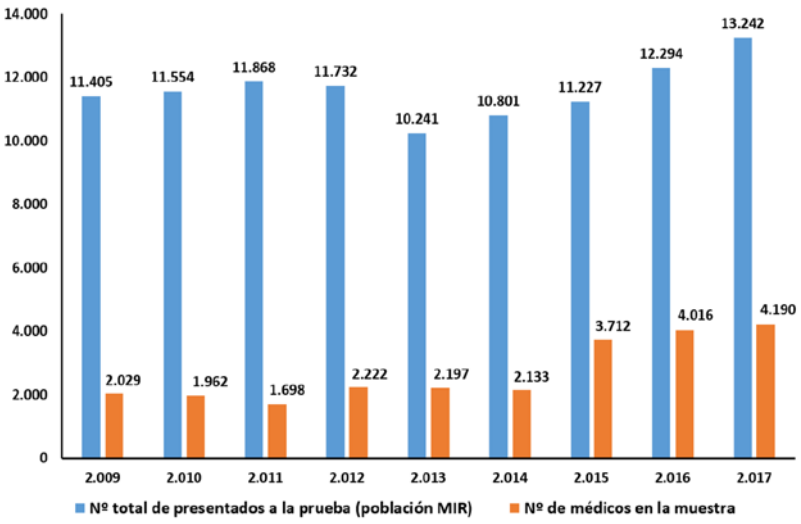
El primer indicador que utilizaremos es el índice de dificultad. Se define como tal el cociente que se obtiene de dividir el número de examinados que aciertan la pregunta (A) entre el número total de examinados que se presentaron al examen (N); su fórmula sería: $D = A / N$.

Tal y como está definido este índice, un valor más alto del índice de dificultad significa que se trata de una pregunta más fácil, con un mayor porcentaje de examinados que la aciertan.

Índice de dificultad con corrección de los efectos del azar

Como se hizo en las publicaciones previas [8,9], con el fin de calcular el índice de dificultad de cada una de las preguntas corrigiendo los posibles efectos de acertar por azar, se aplicó la fórmula:

Figura 1. Número total de presentados a la prueba MIR y número de examinados que introdujeron los resultados de su prueba en la aplicación web de corrección del examen.



$$\mathrm{ID} = \frac{A - \dfrac{E}{K-1}}{N}.$$

En esta ecuación, A representa el número de examinados que responden de forma correcta a la pregunta; E, el número de examinados que la fallan, incluyendo en esta categoría a todos aquellos que dejaron la pregunta sin contestar, y N, el número total de presentados a la prueba. Finalmente, K es el número de opciones que presentaba la pregunta.

Como se ha señalado, hasta la convocatoria 2014, cada pregunta tenía cinco posibles opciones de respuesta, pero a partir de la convocatoria 2015, el número de opciones se redujo a cuatro.

En la bibliografía, existen puntos de corte que permiten clasificar las preguntas en función del valor de su índice de dificultad corregido. Para la consulta de los puntos de corte, véase un trabajo previo de los autores en el que se analiza la prueba MIR de 2015 [8]. Finalmente cabe destacar que, al igual que en el índice de dificultad, un valor más alto del índice de dificultad con corrección de los efectos del azar significa que se trata de una pregunta más fácil.

Índice de discriminación

La discriminación de una pregunta es la capacidad que ésta tiene de distinguir entre los examinados

que obtienen puntuaciones altas y bajas en la prueba. Así, se dice que una pregunta es discriminativa si los que la aciertan son, mayoritariamente, los que obtienen las puntuaciones más altas en el test. En el presente trabajo, se define el índice de discriminación por medio de la fórmula:

$$DS = 2\,\frac{F - D}{N_1 + N_2},$$

siendo F el número de respuestas correctas en el grupo fuerte; D, el número de respuestas correctas en el grupo débil; N_1, el número total de examinados que respondió a la pregunta en el grupo fuerte, y N_2, el número total de examinados que respondió a la pregunta en el grupo débil. Se definen como grupo débil y grupo fuerte los grupos formados, respectivamente, por el 27% de examinados que obtuvieron las peores calificaciones y el 27% de los examinados que obtuvieron las mejores calificaciones en el ejercicio de examen de la prueba MIR.

El índice de discriminación representado según esta fórmula puede tomar valores entre −1 y 1. Cuanto más alto sea el valor del índice, más discriminativa resulta la pregunta. En el caso de este índice, al igual que en el caso del índice de dificultad con corrección de los efectos del azar, existe una clasificación de los valores en diferentes categorías con puntos de corte que, aunque no se utilizan en el presente trabajo, pueden también consultarse en un trabajo previo que analiza el ejercicio de examen de la convocatoria 2015 [8].

Índice de correlación biserial puntual

A través del índice de correlación biserial puntual se mide de forma genérica la fuerza de asociación entre dos variables. En el caso del presente estudio, lo que se mide es la asociación entre el resultado obtenido por el examinado en el ejercicio de examen y si éste acertó o falló la pregunta. El índice de correlación biserial puntual se expresa por medio de la siguiente fórmula:

$$\rho_{bp} = \frac{\mu_p - \mu_q}{\sigma_x} \times \sqrt{\frac{ID}{1 - ID}},$$

siendo μ_p la puntuación media en el test de los examinados que aciertan el ítem; μ_q, la puntuación media en el test de los examinados que fallan el ítem; σ_x, la desviación típica de la puntuación total del test, e ID, el índice de dificultad del ítem. Este índice representa la proporción de examinados que aciertan el ítem ($D = A / N$).

Los resultados de este coeficiente permiten clasificar las preguntas en las siguientes categorías: excelente, si el valor que se obtiene es igual o superior a 0,40; buena, cuando el valor del índice es igual o superior a 0,30 pero inferior 0,40; regular, si el valor del índice es igual o superior a 0,20 pero no llega a 0,30; pobre, si el valor está comprendido entre 0 y 0,20, y pésimo, si el valor es inferior a 0.

De la propia definición del índice se deduce que cuanto mayor sea su valor, mayor será la relación entre la obtención de una puntuación alta en el test y el haber contestado correctamente la pregunta.

Índices de dificultad y discriminación según la teoría de respuesta al ítem

Como ya se explicó en un artículo previo sobre la evaluación psicométrica de la prueba MIR [10], la teoría de respuesta al ítem (TRI) propone una serie de modelos explicativos de la forma en la que los individuos responden a las preguntas. Es decir, a través de la TRI es posible conocer la probabilidad que cada uno de ellos tiene de acertar las distintas preguntas que se le presentan en el ejercicio de examen de la prueba MIR. Así, los modelos matemáticos de la TRI son capaces de calcular los dos parámetros fundamentales de cada pregunta, como son la dificultad y la discriminación.

Para el estudio del ejercicio de examen, y teniendo en cuenta los resultados obtenidos por los autores en estudios previos [10], se ha optado por utilizar el modelo denominado de dos parámetros. En este modelo, de respuesta dicotómica, si el examinando acierta una pregunta, ésta se codifica como 1, mientras que si la pregunta se falla o no se contesta, la respuesta se codifica como 0.

Representando el nivel de conocimiento de cierto examinado por la variable θ, el modelo de dos parámetros que representa la probabilidad de que el i-ésimo examinado responda de manera correcta a la pregunta j-ésima se puede expresar por la ecuación:

$$P(u_j = 1 | \theta_i, a_j, b_j) = \frac{\exp[-1{,}7 \times a_j \times (\theta_i - b_j]}{1 + \exp[-1{,}7 \times a_j \times (\theta_i - b_j]},$$

donde θ_i es el nivel de conocimiento del i-ésimo examinado, a_j es el valor de discriminación de la j-ésima pregunta y b_j es el nivel de dificultad de esa misma pregunta. Así, cuanto mayores sean los valores de a_j y b_j, mayor será la capacidad de discriminación y dificultad de la pregunta. El cálculo de los

valores de dificultad y discriminación para cada una de las preguntas propuestas se realiza a través de la metodología del estimador de máxima verosimilitud, estando disponible su explicación detallada en la bibliografía clásica de TRI [11].

Estudios estadísticos

Se estudió la normalidad de las variables continuas a través del test de Anderson-Darling [12], resultando que ninguna de las variables presentaba normalidad. A continuación, se aplicó el test no paramétrico de Kruskal-Wallis [13] con el fin de encontrar diferencias estadísticamente significativas en las medianas de los grupos considerados. En aquellos casos donde se rechazó la hipótesis nula de igualdad de todas las medianas de los grupos, se aplicó el test *post-hoc* de Dunn [14,15]. En general, el nivel de significación alfa empleado en el presente trabajo ha sido del 5%, considerándose un nivel alfa equivalente del 0,7% en caso del test *post-hoc* de Dunn con el fin de tener en cuenta el número de comparaciones múltiples realizado.

Resultados

Número de preguntas por asignatura

La tabla I muestra la distribución por asignaturas de las preguntas en todos los exámenes de las convocatorias MIR entre 2009 y 2017. La mitad de las asignaturas son responsables de más del 80% de las preguntas del examen MIR. El 81,14% de las preguntas utilizadas para la determinación de la puntuación de los examinados correspondieron a las asignaturas de bioestadística y medicina preventiva, aparato digestivo y cirugía digestiva, neumología y cirugía torácica, microbiología y enfermedades infecciosas, cardiología y cirugía cardiaca, nefrología y urología, obstetricia y ginecología, neurología y neurocirugía, endocrinología y cirugía endocrina, hematología, reumatología, pediatría, psiquiatría, cirugía ortopédica y traumatología, dermatología y farmacología.

Se observa un ligero aumento en el número de asignaturas preguntadas, pasando de un total de 30 asignaturas en el examen de la convocatoria 2009 a 36 en el examen de la convocatoria 2017.

Algunas asignaturas han aumentado sus preguntas en el último examen analizado. Si se compara el número medio de preguntas por asignatura en el examen de las convocatorias 2009 a 2016 con el de la convocatoria 2017, se aprecia que se ha incrementado en dos el número de preguntas de urgencias y genética. En las asignaturas de oncología, cirugía maxilofacial, anatomía, bioética, cardiología y cirugía cardíaca, inmunología y medicina legal, se ha producido un aumento del número de preguntas en la última convocatoria superior a una e inferior a dos, cuando se compara con el promedio de los exámenes de las convocatorias 2009 a 2016. Además, también cabe destacar que en la prueba MIR de 2017 por primera vez aparece una pregunta directamente relacionada con la asignatura de radiología y dos con la de bioquímica.

Como el número de preguntas se mantiene, al haber asignaturas que aumentan su peso en la prueba, otras deben reducir el suyo. Por ello, en el último examen analizado, otras asignaturas han experimentado una disminución de su número de preguntas. Así, se observa una reducción de 4,1 preguntas de reumatología en el examen de la convocatoria 2017 (6 preguntas), comparando éstas con el promedio de las convocatorias 2009 a 2016 (10,1 preguntas). En la asignatura de aparato digestivo y cirugía digestiva, se ha producido una reducción de 3,6 preguntas, y en neumología y cirugía torácica, de 2,5 preguntas. En el caso de hematología, la reducción fue de 2,3 preguntas, y en el de neurología y neurocirugía, de 2,1; no existía ninguna otra asignatura en la que la reducción hubiera sido superior a una pregunta.

Número de preguntas por tipo

La tabla II muestra la distribución de las preguntas en función de su tipo. El promedio de casos clínicos, con y sin imágenes, en el examen de las convocatorias 2009 a 2016 fue de 137,3 frente a 125 en el examen de la convocatoria 2017.

En relación con las preguntas negativas, su número medio en las convocatorias comprendidas entre 2009 y 2016 fue de 29,5, y en la de 2017, de 42.

Por último, el promedio de preguntas de test de las convocatorias comprendidas entre 2009 y 2016 fue de 58,25, y en la convocatoria 2017, de 58.

Índice de dificultad

La tabla III muestra los valores medios y desviaciones estándares de las variables dificultad, índice de dificultad con corrección de los efectos del azar (dificultad corregida), índice de discriminación, coeficiente de correlación punto biserial, e índice de dificultad y discriminación según la TRI.

Cuando se analiza el promedio del índice de dificultad corregida por los efectos del azar de las pre-

Tabla I. Preguntas del ejercicio de examen por asignatura y convocatoria de prueba MIR entre 2009 y 2017. Porcentaje de preguntas por asignatura y porcentaje de preguntas acumulado con las asignaturas ordenadas de mayor a menor importancia.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total	Porcentaje	Acumulado
Bioestadística y medicina preventiva	23	23	25	22	21	21	15	17	17	184	9,09%	9,09%
Aparato digestivo y cirugía digestiva	18	20	20	18	20	20	21	20	16	173	8,54%	17,63%
Neumología y cirugía torácica	19	16	15	16	15	16	14	13	13	137	6,77%	24,40%
Microbiología y enfermedades infecciosas	11	13	15	15	16	14	16	14	15	129	6,37%	30,77%
Cardiología y cirugía cardiaca	13	13	16	15	13	12	13	14	15	124	6,12%	36,89%
Nefrología y urología	14	13	10	13	12	14	12	13	13	114	5,63%	42,52%
Obstetricia y ginecología	14	13	11	10	9	10	11	11	11	100	4,94%	47,46%
Neurología y neurocirugía	12	12	12	11	11	12	10	9	9	98	4,84%	52,30%
Endocrinología y cirugía endocrina	10	11	11	12	10	13	10	10	10	97	4,79%	57,09%
Hematología	10	10	11	9	12	12	10	9	8	91	4,49%	61,58%
Reumatología	11	11	9	8	13	9	10	10	10	89	4,40%	65,98%
Pediatría	8	8	10	13	10	10	11	9	6	87	4,30%	70,27%
Psiquiatría	10	6	9	7	7	8	8	8	8	71	3,51%	73,78%
Cirugía ortopédica y traumatología	7	11	5	7	4	8	6	6	7	61	3,01%	76,79%
Dermatología	6	5	5	4	5	6	3	5	5	44	2,17%	78,96%
Farmacología clínica	4	6	6	5	5	4	4	5	5	44	2,17%	81,14%
Otorrinolaringología	5	3	4	5	4	4	3	3	5	34	1,68%	82,81%
Inmunología	4	3	4	3	4	3	4	4	2	33	1,63%	84,44%
Oftalmología	5	3	4	4	3	3	3	3	4	32	1,58%	86,02%
Anatomía patológica	4	2	0	4	4	4	5	4	4	31	1,53%	87,56%
Fisiología	3	2	3	3	3	2	5	5	3	29	1,43%	88,99%
Genética	1	4	4	2	3	3	2	3	5	27	1,33%	90,32%
Gestión clínica	4	3	2	2	2	2	4	3	3	25	1,23%	91,56%
Anatomía	1	0	2	2	3	3	4	4	4	23	1,14%	92,69%
Geriatría	1	2	2	2	3	2	3	3	4	22	1,09%	93,78%
Oncología	3	0	2	2	3	1	3	4	3	21	1,04%	94,81%
Habilidades comunicativas	0	4	1	0	0	1	6	3	2	16	0,79%	95,60%
Cuidados paliativos	1	2	2	2	2	0	3	2	0	15	0,74%	96,35%
Cirugía vascular	1	1	1	2	1	3	1	1	4	15	0,74%	97,09%
Urgencias	0	1	1	3	3	1	1	1	2	13	0,64%	97,73%
Cirugía maxilofacial	1	2	1	0	1	1	2	2	3	13	0,64%	98,37%
Anestesiología	1	1	0	2	2	1	1	1	1	10	0,49%	98,86%
Cirugía plástica	0	1	0	1	0	1	1	2	2	8	0,40%	99,26%
Medicina legal	0	0	2	0	1	1	0	2	1	7	0,35%	99,60%
Bioética	0	0	0	1	0	0	0	2	2	5	0,25%	99,85%
Bioquímica	0	0	0	0	0	0	0	0	2	2	0,10%	99,95%
Radiología	0	0	0	0	0	0	0	0	1	1	0,05%	100,00%
Total	225	225	225	225	225	225	225	225	225	2.025	100%	

guntas de los diferentes exámenes, se observa una tendencia al descenso progresivo de la dificultad de los exámenes a lo largo de la serie temporal analizada. El examen de la convocatoria 2016 resultó el más fácil de los años estudiados, con un valor de 0,6909 frente a 0,6750 para el examen de la convocatoria 2017 y un promedio de 0,6252 en los exámenes de las convocatorias 2009 a 2015.

Índice de dificultad con corrección de los efectos del azar

Cuando se analiza el índice de dificultad corregida por los efectos del azar, se observa un comportamiento equivalente, siendo el examen de la convocatoria 2016 el más fácil de toda la serie temporal analizada, con un valor de 0,5971, frente a 0,5740 del examen de la convocatoria 2017 y un promedio de 0,5462 en los exámenes de las convocatorias 2009 a 2015.

Índice de discriminación

En lo referente al promedio de la discriminación, el examen de la convocatoria 2017 fue también el menos discriminativo de toda la serie temporal analizada, con una diferencia de 0,0910 con respecto al promedio de los exámenes de las convocatorias anteriores.

Índice de correlación biserial puntual

Lo mismo ocurre en relación a la discriminación medida mediante el coeficiente de correlación biserial puntual, donde el valor promedio obtenido para el examen de la convocatoria 2017 es el más bajo de toda la serie temporal analizada y junto con el de 2016 los únicos que, redondeado a dos decimales, entrarían en la categoría de discriminación 'regular' en lugar de en la categoría 'buena' de los exámenes 2009 a 2015.

La tabla IV muestra la capacidad discriminativa de las preguntas del examen de las convocatorias MIR entre 2009 y 2017. Hay una reducción en el número de preguntas de alta calidad discriminativa en los dos últimos exámenes analizados respecto al promedio de los de convocatorias anteriores. El promedio de preguntas con una discriminación buena o excelente en el examen de las convocatorias 2009 a 2015 fue de 127,86, que supone más de 37 preguntas por encima de las que presentaron de media las pruebas de las convocatorias 2016 y 2017.

Si se realiza la comparación únicamente con el examen de la convocatoria 2015, la reducción en el de la convocatoria 2016 del número de preguntas con discriminación buena o excelente fue de 26, mientras que si se compara con el de 2017, la reducción fue de 29. También es muy destacable el aumento en el examen de 2016 de las preguntas con discriminación regular, que ha pasado de 49 preguntas en la convocatoria 2015 a 81 en dicha convocatoria, un aumento de 35 preguntas, mientras que en la de 2017 se reducen hasta 65. Nótese que el examen de la convocatoria de 2016 es el que mayor número de preguntas tiene en esta categoría de los ocho años analizados.

Los exámenes con mayor número de preguntas con discriminación 'pésima' de las consideradas para el cálculo de las puntuaciones de los examinados fueron los de las convocatorias 2013, 2015 y 2017, los dos primeros con un total de 10 preguntas en esta categoría y el último con un total de 11.

Tabla II. Preguntas en el ejercicio de examen, por tipo y convocatoria, en las pruebas MIR entre 2009 y 2017.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Casos clínicos	92	112	102	109	111	105	115	104	94	944
Casos clínicos con imagen	25	30	29	34	34	34	30	32	31	279
Preguntas negativas	35	22	27	32	29	28	26	37	42	278
Preguntas de test	73	61	67	50	51	58	54	52	58	524
Total	225	225	225	225	225	225	225	225	225	2.025

Índices de dificultad y discriminación según la TRI

En lo referente a los coeficientes de dificultad y discriminación obtenidos por medio de la TRI, también se observa que el examen de la convocatoria 2016 arroja el valor promedio más bajo de toda la serie temporal analizada, coincidiendo este hecho con el resultado de los valores psicométricos descritos anteriormente utilizando parámetros de la teoría clásica de los tests (Tabla III).

Análisis de dificultad y discriminación por número de repeticiones del concepto preguntado

Otra posible forma de clasificar las preguntas del ejercicio de examen de la prueba MIR es a partir del número de veces que el concepto preguntado lo ha sido en los exámenes de convocatorias anteriores de la prueba. Así, se completa el apartado de resultados con un análisis de dificultad y discriminación

Tabla III. Valor medio y desviación estándar para las variables dificultad, índice de dificultad con corrección de los efectos del azar, índice de discriminación, coeficiente de correlación puntual biserial, índice de dificultad e índice de discriminación según la teoría de respuesta al ítem (TRI) de las preguntas de los ejercicios de examen de las pruebas MIR entre 2009 y 2017.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	Promedio
Dificultad	0,6159 (0,2338)	0,5862 (0,2506)	0,6310 (0,2366)	0,6429 (0,2260)	0,6199 (0,2269)	0,6367 (0,2301)	0,6439 (0,2171)	0,6909 (0,2182)	0,6750 (0,2221)	0,6313 (0,2368)
Dificultad corregida	0,5205 (0,2557)	0,5142 (0,2941)	0,5645 (0,2813)	0,5754 (0,272)	0,5454 (0,2736)	0,5656 (0,277)	0,5379 (0,2837)	0,5971 (0,2874)	0,5740 (0,2922)	0,5552 (0,2807)
Discriminación	0,3613 (0,1687)	0,3356 (0,1613)	0,3357 (0,1862)	0,3596 (0,1793)	0,367 (0,1764)	0,3321 (0,1649)	0,3075 (0,1545)	0,2552 (0,1302)	0,2407 (0,1320)	0,3203 (0,1645)
Coeficiente de correlación puntual biserial	0,3 (0,1136)	0,2989 (0,1223)	0,3034 (0,1154)	0,3313 (0,1333)	0,3295 (0,1347)	0,2967 (0,1204)	0,2977 (0,1381)	0,2693 (0,1108)	0,2556 (0,1320)	0,2973 (0,1259)
Dificultad TRI	−0,8637 (7,039)	−0,5353 (7,875)	−0,5182 (4,343)	−0,4371 (7,631)	−0,4636 (8,813)	−0,3985 (9,434)	−0,3191 (6,216)	−1,6325 (3,584)	−0,7262 (8,3274)	−0,7692 (6,6858)
Discriminación TRI	0,7476 (0,3895)	0,7859 (0,4314)	0,7801 (0,4065)	0,8142 (0,4314)	0,8751 (0,4849)	0,7214 (0,375)	0,7639 (0,4761)	0,7104 (0,4183)	0,6701 (0,4035)	0,7617 (0,4269)

Tabla IV. Capacidad discriminativa de las preguntas de los ejercicios de examen de las pruebas MIR entre 2009 y 2017. medidas por el índice de correlación biserial puntual.

	2009	2010	2011	2012	2013	2014	2015	2016	2017	Total
Excelente	56	50	51	73	74	51	56	25	18	454
Buena	61	73	72	76	72	68	62	67	71	622
Regular	63	54	70	41	38	59	49	81	65	520
Pobre	41	39	26	29	31	39	48	45	60	358
Pésimo	4	9	6	6	10	8	10	7	11	71
Todos	225	225	225	225	225	225	225	225	225	2.025

de las preguntas de los ejercicios de examen de las pruebas MIR comprendidas entre las convocatorias 2009 y 2017, teniendo en cuenta el número de veces que el concepto del ítem se ha preguntado en exámenes de convocatorias anteriores de la prueba. Para ello, se revisaron todas las preguntas correspondientes a los exámenes de las convocatorias comprendidas entre 1980 y 2017. Los resultados se muestran en la tabla V. En esta tabla se observa que el promedio de la dificultad de cada convocatoria siempre es menor cuando se trata de conceptos que se han preguntado tres veces o más, en relación con los ítems que presentan conceptos que se han preguntado dos veces, o bien conceptos preguntados únicamente en una convocatoria. Este mismo comportamiento se observa también para el caso del índice de dificultad corregido. En el caso del índice de discriminación, el resultado convocatoria por convocatoria no es concluyente porque a veces el valor medio del índice de discriminación resulta inferior en los ítems con conceptos nunca repetidos que en aquellos en los que el concepto se había preguntado con anterioridad, mientras que en otros años ocurre lo contrario. En lo relativo al coeficiente de correlación puntual biserial, salvo en el caso del ejercicio de examen de la convocatoria 2016, siempre se observa que el valor de la discriminación medida de dicho coeficiente es superior en las preguntas relativas a conceptos que se preguntaron tres veces o más, seguidos por los conceptos preguntados dos veces, respecto a los valores de las preguntas que versan sobre conceptos nunca repetidos. Finalmente, el análisis de la dificultad TRI indica que éste muestra el mismo comportamiento que la dificultad y la dificultad corregida, salvo en el ejercicio de examen de la convocatoria 2016, donde las preguntas relativas a conceptos repetidos dos veces resultan ligeramente más fáciles que las preguntas con más de dos repeticiones. Este mismo fenómeno se observa también en el valor de discriminación calculado según la TRI en las convocatorias de 2016 y 2017.

La figura 2 muestra los valores promedios de las variables dificultad, dificultad corregida, discriminación y coeficiente de correlación puntual biserial para los exámenes MIR entre las convocatorias 2009 y 2017. Se aprecia que tanto la dificultad como

Tabla V. Valor medio para las variables dificultad, índice de dificultad con corrección de los efectos del azar, índice de discriminación, coeficiente de correlación puntual biserial, índice de dificultad según la teoría de respuesta al ítem (TRI) e índice de discriminación según la TRI de las preguntas de los ejercicios de examen de las pruebas MIR entre 2009 y 2017, divididas en función del número de veces que se ha repetido el concepto.

	Tipo de pregunta	Dificultad	Dificultad corregida	Discriminación	Coeficiente de correlación puntual biserial	Dificultad TRI	Discriminación TRI
Año 2009	Preguntado tres o más veces	0,6598	0,5835	0,3774	0,3389	–2,3850	0,9068
	Preguntado dos veces	0,5743	0,4717	0,3573	0,3024	–0,3191	0,7512
	Preguntado por primera vez	0,5305	0,4287	0,3514	0,2733	0,1121	0,6423
Año 2010	Preguntado tres o más veces	2,1600	0,6244	0,3537	0,3569	–1,1720	0,9813
	Preguntado dos veces	1,8852	0,5225	0,3238	0,2866	–0,7867	0,7343
	Preguntado por primera vez	1,4508	0,3526	0,3291	0,2543	0,0702	0,6371
Año 2011	Preguntado tres o más veces	0,7269	0,6591	0,3147	0,3344	–1,4423	0,9180
	Preguntado dos veces	0,6656	0,6016	0,3663	0,3223	–1,2015	0,8866
	Preguntado por primera vez	0,5099	0,3969	0,3480	0,2656	0,6183	0,6170
Año 2012	Preguntado tres o más veces	0,7119	0,6528	0,3762	0,3809	–1,3713	0,9693
	Preguntado dos veces	0,6026	0,5209	0,3852	0,3234	–0,0410	0,7767
	Preguntado por primera vez	0,5228	0,4197	0,3380	0,2728	0,5749	0,6301
Año 2013	Preguntado tres o más veces	0,6839	0,6174	0,3826	0,3754	–0,7130	1,0606
	Preguntado dos veces	0,5934	0,4967	0,3407	0,2998	–1,2118	0,7289
	Preguntado por primera vez	0,5264	0,4166	0,3577	0,2874	0,0343	0,7110
Año 2014	Preguntado tres o más veces	0,6837	0,6209	0,3389	0,3278	–1,7010	0,8399
	Preguntado dos veces	0,6143	0,5332	0,3644	0,3082	–1,3702	0,7460
	Preguntado por primera vez	0,5331	0,4299	0,3074	0,2501	1,4547	0,5526
Año 2015	Preguntado tres o más veces	0,7017	0,6359	0,3327	0,3469	–1,7496	0,9712
	Preguntado dos veces	0,6531	0,5703	0,3522	0,3274	0,0252	0,8281
	Preguntado por primera vez	0,5678	0,4603	0,2683	0,2452	0,8630	0,5782
Año 2016	Preguntado tres o más veces	0,7778	0,7205	0,2623	0,3059	–2,0799	0,8565
	Preguntado dos veces	0,7007	0,6209	0,2703	0,3079	–2,1739	0,9299
	Preguntado por primera vez	0,5942	0,4775	0,2447	0,2283	–1,1139	0,5354
Año 2017	Preguntado tres o más veces	0,7757	0,7090	0,2316	0,2852	–2,3674	0,8197
	Preguntado dos veces	0,7081	0,6214	0,2584	0,2915	–1,8409	0,8221
	Preguntado por primera vez	0,5915	0,4689	0,2439	0,2276	0,6617	0,5342
Promedio	Preguntado tres o más veces	0,8757	0,6471	0,3300	0,3391	–1,6646	0,9248
	Preguntado dos veces	0,7775	0,5510	0,3354	0,3077	–0,9911	0,8004
	Preguntado por primera vez	0,6474	0,4279	0,3098	0,2561	0,3639	0,6042

Figura 2. Valores promedios de las variables dificultad (a), dificultad corregida (b), discriminación (c) y coeficiente de correlación puntual biserial (d) para los exámenes MIR entre 2009 y 2017, clasificando las preguntas en función del número de repeticiones del concepto en el ejercicio de examen de la prueba MIR.

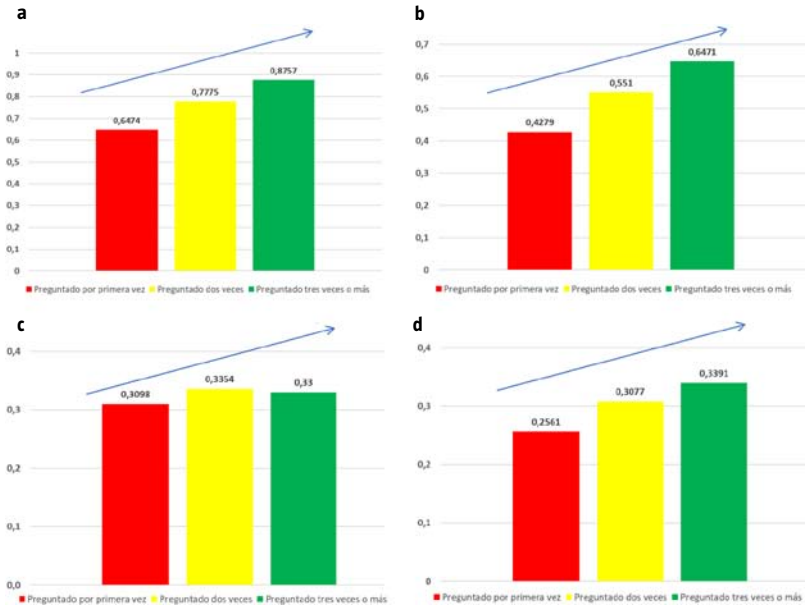
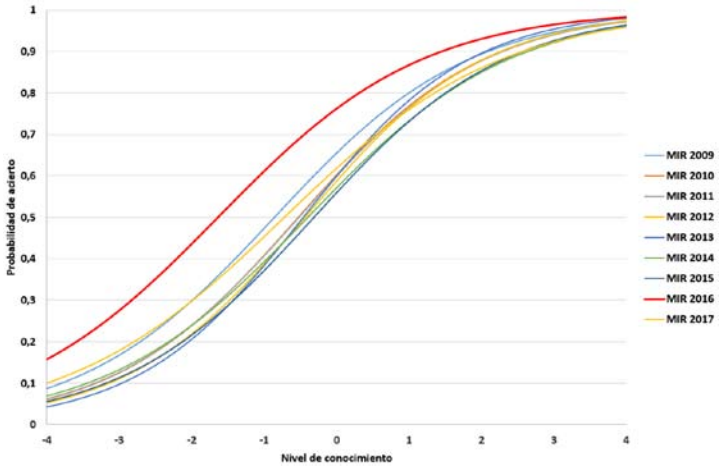


Figura 3. Curvas de probabilidad de los ejercicios de examen de las pruebas MIR entre 2009 y 2017, representadas tomando para cada curva el valor medio de probabilidad y discriminación de todas las preguntas del examen de la prueba MIR representada.



la dificultad corregida incrementan su valor promedio (es decir, las preguntas son más fáciles) en las preguntas relativas a conceptos repetidos frente a la que alcanzan en conceptos nunca preguntados con anterioridad. En lo relativo a la discriminación, las preguntas sobre conceptos ya preguntados con anterioridad resultan más discriminativas y con valores más altos de coeficiente de correlación puntual biserial que las preguntas que versan sobre conceptos no repetidos, en todas las convocatorias.

La figura 3 muestra las curvas de probabilidad de los ejercicios de examen de las pruebas MIR entre 2009 y 2017. Estas curvas se han representado utilizando como coeficientes, en el modelo de dos parámetros, los valores promedios de dificultad y discriminación correspondientes a cada uno de los ejercicios de examen de las convocatorias analizadas. Se observa claramente que en el ejercicio de examen de la convocatoria 2016, para todos los niveles de conocimiento, la probabilidad de acierto de los examinados es superior a la que tuvieron en las convocatorias comprendidas entre 2009 y 2015, así como en la de 2017.

Discusión

Fiabilidad de la prueba MIR

Tanto la fórmula número 21 de Kuder-Richardson como el alfa de Cronbach son parámetros utilizados como indicadores de la fiabilidad de una escala; en la bibliografía se considera que valores de dichos parámetros superiores a 0,8 son suficientes para garantizar la alta fiabilidad de un test.

En un trabajo previo publicado por los autores acerca del ejercicio de examen de la convocatoria 2015 [4], se obtuvo un valor de 0,9459 para la fórmula número 21 de Kuder-Richardson y de 0,9579 para el alfa de Cronbach. En el caso de las convocatorias de 2016 y 2017, el resultado de la fórmula número 21 de Kuder-Richardson fue de 0,9403 y 0,9353, respectivamente, y el del alfa de Cronbach, de 0,9288 y 0,0212, valores similares a los obtenidos en el ejercicio de examen de la convocatoria 2015.

A pesar de esto, es necesario tener en cuenta lo indicado por Muñiz [16] a la hora de hacer una interpretación rigurosa del valor de dichos coeficientes. En primer lugar, se debe considerar que el valor, tanto de la fórmula número 21 de Kuder-Richardson como del alfa de Cronbach, aumenta cuando se incrementa el número de ítems, y en segundo lugar, en aquellos casos en los que se proponen ítems similares, estos índices también aumentan. Sin que sea necesario entrar a analizar la similitud de las preguntas propuestas en el ejercicio de examen de la prueba MIR, el número de ítems que lo componen es ya lo suficientemente elevado como para que los valores de ambos coeficientes estén por encima de 0,8.

Cambios en dificultad y discriminación en el MIR 2016 y 2017

A pesar de que los ejercicios de examen de las convocatorias 2016 y 2017, según estos dos indicadores de fiabilidad, podría considerarse como muy buenos, se ha constatado también que los parámetros relativos a su dificultad y discriminación han cambiado, siendo el examen propuesto en la convocatoria de 2016 más fácil y menos discriminativo que el resto de exámenes de la serie temporal analizada.

En opinión de los autores, el origen probable de la menor dificultad de los exámenes de las dos últimas convocatorias es multifactorial, con componentes atribuibles a cambios en las preguntas del examen y a cambios en la población de examinados. Desde la convocatoria 2015, las preguntas tienen cuatro alternativas de respuesta, mientras que las preguntas de las convocatorias anteriores tenían cinco opciones. A pesar de la reducción en el número de respuestas, la penalización por fallar la pregunta se ha mantenido en el equivalente a perder una pregunta acertada por cada tres preguntas erróneas. En consecuencia, la penalización estadística por contestar preguntas al azar se ha reducido desde 2015, lo que ha contribuido a disminuir el número de preguntas no contestadas por la población de examinados en las dos últimas convocatorias. Esto probablemente haya favorecido un aumento en las puntuaciones de los examinados con menor nivel de conocimientos y peor manejo del 'riesgo' de contestar dudando entre varias opciones de respuesta, lo que se ha reflejado en que, cada año, la nota de corte del examen ha afectado a una menor proporción de examinados. Por otra parte, entre las causas de la aparente menor dificultad del examen, habría que citar el aumento del nivel de conocimiento 'promedio' derivado de los cambios en la población de examinados formada, cada año, por un mayor número y proporción de médicos españoles recién graduados, tal y como se explicará en el apartado siguiente.

Desde nuestro punto de vista, el cambio de los valores medios de dificultad de las preguntas del examen de la convocatoria de 2016 puede ser responsable, en parte, del descenso en su valor promedio de discriminación. En un examen más fácil existe un mayor número de examinados con una probabilidad alta de acertar las preguntas y eso hace que las preguntas resulten menos discriminativas. Existe cierta correlación entre dificultad y discriminación, pero no en todos los casos una pregunta difícil es discriminativa, y al revés. En la convocatoria de 2017, este efecto se ha atenuado por una dificultad media de las preguntas algo superior.

Cambios en la composición de los subconjuntos de médicos presentados a las dos últimas convocatorias de la prueba MIR

¿Son las preguntas de los ejercicios de examen las únicas responsables del cambio en la dificultad y discriminación promedio de los exámenes de las últimas convocatorias MIR? No debe perderse de vista el hecho de que la composición de la población de médicos presentados a la prueba MIR ha cambiado mucho en los últimos años, siendo cada vez un conjunto menos heterogéneo y con mayor proporción de presentados pertenecientes al subconjunto que mejor resultado obtiene en la prueba: los recién graduados de universidades españolas. En opinión de los autores, estos cambios en la composición de los subconjuntos de médicos presentados en las dos últimas convocatorias de la prueba realizadas puede ser responsable, al menos en parte, de la disminución de la dificultad y discriminación del ejercicio de examen.

Estudios previos han demostrado que el resultado en el ejercicio de examen de la prueba de los médicos españoles es mejor que el de los extranjeros [17]. Así, de los presentados al examen en la convocatoria 2017, un 74,44% eran médicos españoles, porcentaje que en la de 2012 fue del 61,21%, y en la de 2009, del 55,82%. Es decir, el porcentaje de médicos españoles que se presentan al MIR ha pasado de poco más de la mitad de la población en la convocatoria 2009 hasta casi el 75% en la convocatoria 2017. Como los resultados promedio en el examen MIR son mejores para el grupo de españoles presentados que para el grupo de extranjeros, este cambio en la composición de la población MIR puede ser uno de los factores responsables de la aparente mayor facilidad de los exámenes de las últimas convocatorias.

A ello se suma que, en las últimas convocatorias, ha aumentado el número y porcentaje de médicos españoles que además son recién graduados, tanto por el aumento del *numerus clausus* de las facultades existentes como por la apertura de nuevas facultades públicas y privadas en nuestro país. Existen estudios [17] e informes del Ministerio de Sanidad [18] que comunican mejores resultados en el examen MIR en el subconjunto de médicos presentados más jóvenes, respecto a los no recién graduados de mayor edad. El análisis de los motivos de esta realidad excede el objetivo y alcance del presente trabajo. Podríamos resumir los cambios demográficos en la frase: 'el MIR es cada año más español y más joven'. Este cambio demográfico influye en la misma dirección y sentido que el anterior-

mente descrito, haciendo que la población MIR sea cada año menos heterogénea en su composición y mejor preparada, lo que a su vez puede ser responsable de parte del descenso de los índices de dificultad y discriminación observados en el análisis psicométrico de los ejercicios de examen MIR de las convocatorias de 2016 y 2017.

Desde el punto de vista de los autores, otro motivo que ha contribuido a la menor capacidad discriminativa de los ejercicios de examen de las convocatorias de 2016 y 2017 ha sido la anulación de un menor número de preguntas respecto a lo que suele ser habitual. En el examen de estas dos últimas convocatorias se anularon sólo tres y cuatro preguntas, respectivamente, frente a las siete en 2014 y ocho en 2013. Un número bajo de preguntas anuladas no sería problemático en sí, salvo porque en la plantilla activa del examen se dejaron siete preguntas sin anular con una discriminación pésima, es decir, aquellas cuya probabilidad de ser acertada no se correlacionaba con el conocimiento médico expresado en el resto de preguntas del examen. Tal y como se presentó en un trabajo anterior [9], una anulación de preguntas que invalide aquellas con una menor calidad psicométrica contribuye a mejorar la capacidad discriminativa de la prueba. El descenso del promedio de la discriminación de las preguntas del examen de las convocatorias de 2016 y 2017 demuestra que la anulación de un menor número de preguntas no significa que la calidad general de las preguntas restantes en la prueba sea mejor que las de años anteriores y, por tanto, que el examen resultante posea una mejor capacidad discriminativa. En opinión de los autores, dejar preguntas no discriminativas sin anular, pudiendo hacerlo por quedar preguntas de reserva disponibles no utilizadas, supone una pérdida de oportunidad de mejorar la calidad de la prueba. En este sentido, parece interesante sugerir la posibilidad de usar una metodología de anulación de preguntas automatizada, similar a la que se usa en Chile en el Examen Único Nacional de Conocimientos de Medicina (EUNACOM) [19]. Con el fin de no ser reiterativos y teniendo en cuenta que ya se explicó en un trabajo anterior [9], solamente recordaremos que antes de la prueba EUNACOM se fijan unos valores de corte para la dificultad y discriminación, y todas las preguntas que no cumplen con dichos criterios se eliminan. A modo de ejemplo, en 2012, se fijó como criterio de corte de discriminación el que las preguntas tuvieran un valor del coeficiente de correlación puntual biserial de al menos 0,15. La imposición de un criterio de anulación como el aquí presentado, junto con otras normas, supone que el número de preguntas válidas del EUNACOM varíe de un año a otro, pero consigue preservar una alta calidad psicométrica de la prueba. De este modo, en el caso chileno no se tienen reparos a la hora de anular preguntas, siempre y cuando dicha anulación contribuya a mejorar la calidad global de la prueba. Bonillo [17], en su análisis de las pruebas MIR de 2005 y 2006, ya puso de manifiesto la conveniencia de recurrir a la anulación de un mayor número de preguntas del examen. Podríamos resumirlo con la frase: 'en el MIR, menos no es más', y menos preguntas anuladas no implica más discriminación o mayor calidad del examen.

Por último, y siendo conscientes del reto que ello supone, insistimos en la necesidad de conseguir un examen MIR de la máxima discriminación posible porque, en general, una disminución en la discriminación hace que se incremente el efecto del azar en las puntuaciones obtenidas. En un examen menos discriminativo, muchas de las puntuaciones de los examinados se concentran en un rango estrecho y pequeñas variaciones en el resultado de un examinado, fruto del azar, se traducen en grandes cambios en el número de orden final que los examinados obtendrán en la prueba y, por tanto, en la posibilidad de acceder o no a la especialidad deseada.

Además, tampoco hemos de olvidar que la tendencia iniciada en los últimos años de disminución del número de médicos extranjeros que se presentan a la prueba, así como el aumento en términos absolutos y proporcionales del número de médicos españoles recién egresados, con unos conocimientos más homogéneos que el de los médicos que se presentaban en 2009-2010, hacen cada año más difícil para los examinadores el conseguir un examen suficientemente discriminativo. Desde nuestro punto de vista constituye todo un reto, y a la vez una necesidad, intentar aumentar la capacidad discriminativa del examen MIR, como mínimo hasta los valores previos al examen de la convocatoria 2016. Con ello se conseguiría que las puntuaciones obtenidas midieran de forma más precisa los conocimientos del candidato y redujeran el efecto del azar en la puntuación de la prueba.

Una pregunta a la que contestan de forma correcta los médicos que menos conocimiento han demostrado en el resto de preguntas, es decir, con discriminación negativa, parece evidente que debería anularse porque probablemente se haya contestado al azar. Cada pregunta anulada por discriminación negativa aumenta la discriminación del examen final, que es el promedio de la discriminación de las preguntas que lo constituyen. En nuestra opinión, sería conveniente utilizar al menos la totali-

dad de las diez preguntas de reserva para sustituir, durante el proceso de anulaciones realizado por la comisión calificadora, a otras tantas de peor calidad psicométrica que se hubieran incluido entre las 225 primeras del examen.

Como limitación fundamental de este trabajo destaca que, a diferencia del estudio de Bonillo [17], en nuestro análisis no hemos podido acceder a la información de los resultados de todos los candidatos presentados a la prueba, sino solamente a una muestra de ellos; si bien en todas las convocatorias está formada por más de 1.600 candidatos (y en las últimas convocatorias MIR supera el 30% de los examinados), presenta un sesgo porque no se accede de forma aleatoria a examinados de todos los niveles de conocimiento, sino únicamente de los que introdujeron los resultados de su examen por decisión propia en la aplicación informática del Curso Intensivo MIR Asturias.

Como reflexión final del presente trabajo, cabe destacar que la confección de un examen de alta calidad se hace cada año más exigente para el examinador, que debe seleccionar o elaborar un conjunto de preguntas con una dificultad y discriminación adaptadas a una población formada por un mayor número y una mayor proporción de médicos recién egresados de facultades de medicina españolas. Si esto no se tiene en cuenta, la prueba podría perder calidad. Una de las herramientas de la que se dispone para mejorar la calidad psicométrica sería la anulación *a posteriori* de algunas de las preguntas propuestas.

Bibliografía

1. Real Decreto 2015/1978, de 15 de julio, por el que se regula la obtención de títulos de especialidades médicas. Boletín Oficial del Estado, n.º 206, de 29 de agosto de 1978. p. 20172-4.
2. Real Decreto 127/1984, de 11 de enero, por el que se regula la formación médica especializada y la obtención del título de médico especialista. Boletín Oficial del Estado, n.º 26, de 31 de enero de 1984. p. 2524-8.
3. Ley 44/2003, de 21 de noviembre, de ordenación de las profesiones sanitarias. Boletín Oficial del Estado, n.º 280, de 22 de noviembre de 2003. p. 41442-58.
4. Murias-Quintana E, Sánchez-Lasheras F, Fernández-Somoano A, Romeo-Ladrero JM, Costilla-García SM, Cadenas-Rodríguez M, et al. Análisis de la elección de la especialidad de radiodiagnóstico en el examen MIR desde el año 2006 hasta 2015. Radiología 2017; 59: 232-46.
5. Curbelo J, Romeo JM, Fernández-Somoano A, Sánchez-Lasheras F, Baladrón J. Endocrinología y nutrición: evolución de la elección de la especialidad en los últimos años. Endocrinología, Diabetes y Nutrición 2017; 64: 329-31.
6. Curbelo J, Fernández-Somoano A, Romeo JM, Villacampa T, Sánchez-Lasheras F, Baladrón J. La elección de la especialidad medicina intensiva: análisis de los últimos 10 años. Medicina Intensiva 2018; 42: 65-8.
7. Curbelo J, Galván-Román JM, Sánchez-Lasheras F, Romeo JM, Fernández-Somoano A, Villacampa T, et al. Aparato digestivo: evolución de la elección de la especialidad en los últimos años. Rev Esp Enf Dig 2017; 109: 614-8.
8. Baladrón J, Curbelo J, Sánchez-Lasheras F, Romeo-Ladrero JM, Villacampa T, Fernández-Somoano A. El examen al examen MIR 2015: aproximación a la validez estructural a través de la teoría clásica de los tests. FEM 2016; 19: 217-26.
9. Baladrón J, Sánchez-Lasheras F, Villacampa T, Romeo-Ladrero JM, Jiménez-Fonseca P, Curbelo J, et al. Propuesta metodológica para la detección de preguntas susceptibles de anulación en la prueba MIR. Aplicación a las convocatorias 2010 a 2015. FEM 2017; 20: 161-75.
10. Baladrón J, Sánchez-Lasheras F, Villacampa T, Romeo-Ladrero JM, Jiménez-Fonseca P, Curbelo J, et al. El examen MIR 2015 desde el punto de vista de la teoría de respuesta al ítem. FEM 2017; 20: 29-38.
11. Lord FM. Applications of item response theory to practical testing problems. Hillside, NJ: Lawrence Erlbaum; 1980.
12. Anderson TW, Darling DA. Asymptotic theory of certain 'goodness-of-fit' criteria based on stochastic processes. Annals of Mathematical Statistics 1952; 23: 193-212.
13. Kruskal WH, Wallis A. Use of ranks in one-criterion variance analysis. J Am Stat Assoc 1952; 47: 583-621.
14. Dunn OJ. Multiple comparisons among means. J Am Stat Assoc 1961; 56: 52-64.
15. Dunn OJ. Multiple comparisons using rank sums. Technometrics 1964; 6: 241-52.
16. Muñiz J. Teoría clásica de los tests. Madrid: Pirámide; 2002.
17. Bonillo A. Pruebas de acceso a la formación sanitaria especializada para médicos y otros profesionales sanitarios en España: examinando el examen y los examinados. Gac Sanit 2012; 26: 231-5.
18. Ministerio de Sanidad, Servicios Sociales e Igualdad. Formación médica especializada. URL: http://sis.msssi.es/fse/. [11.12.2017].
19. Examen Único Nacional de Conocimientos de Medicina. Resultados EUNACOM-ST 2011. URL: http://www.eunacom.cl/resultados/resultados-actuales.html. [29.11.2017].