El examen MIR 2015 desde el punto de vista de la teoría de respuesta al ítem

Jaime Baladrón, Fernando Sánchez-Lasheras, Tomás Villacampa, José M. Romeo-Ladrero, Paula Jiménez-Fonseca, José Curbelo, Ana Fernández-Somoano

Introducción. En España, el acceso a la formación médica especializada, imprescindible para ejercer como médico especialista, se realiza a través de la prueba MIR. Superada esta prueba, los aspirantes pueden acceder a la formación en distintas especialidades ofertadas por numerosos hospitales a lo largo de todo el país.

Sujetos y métodos. Para este trabajo se han utilizado las respuestas al examen MIR 2015 de un conjunto de 3.712 aspirantes.

Resultados. Se calcularon los índices de dificultad y discriminación de todas las preguntas del examen. Las preguntas se analizaron según los valores de dichos índices y se agruparon por asignaturas, bloques y tipos de pregunta. Las preguntas con una mayor dificultad media fueron las pertenecientes a las asignaturas de fisiología, farmacología, geriatría, traumatología, neurocirugía y cuidados paliativos. Las asignaturas cuyas preguntas mostraron valores menores de dificultad media fueron anatomía patológica, anestesiología, cirugía plástica, habilidades comunicativas, genética y enfermedades infecciosas.

Conclusiones. En general, los valores de dificultad y discriminación de las preguntas de la prueba MIR resultan adecuados. La prueba discrimina mejor a los alumnos que demuestran conocimientos más bajos, y el valor óptimo de discriminación se encuentra en torno al percentil 25 de la muestra analizada (con una puntuación equivalente al percentil 41 de todos los médicos presentados al examen MIR 2015). Finalmente, se propone el uso de las metodologías propias de la teoría de respuesta al ítem con el fin de evaluar las preguntas de la prueba candidatas a ser anuladas.

Palabras clave. Estadísticas. Estudiantes de medicina. Mediciones educativas. Prueba MIR. Psicometría.

MIR 2015 exam from the point of view of the item response theory

Introduction. In Spain, in order to gain access to specialised medical training it is mandatory to take the MIR exam. After passing said exam, the candidates can access training in different hospitals all around the country.

Subjects and methods. This research was made using a database of the answers of 3,712 candidates who took the 2015 MIR exam.

Results. The difficulty and discrimination index of all the questions in the exam were calculated. All the questions were analysed, taking into account the values of those parameters and classified by subject, block and kind of question. On average, those questions that were found to be most difficult correspond to the following subjects: physiology, pharmacology, geriatrics, traumatology, neurosurgery and palliative care. The subjects with the least average difficulty were anatomical pathology, anaesthesiology, plastic surgery, communication skills, genetics and infectious diseases.

Conclusions. Overall, the discrimination and difficulty values of the questions in the MIR exam are sufficient. The exam is more discriminatory for those students with the lowest discrimination levels, with percentile 25 having the highest levels. Finally, we propose that item response theory be employed as a support tool in order to decide which exam questions would be nullified.

Key words. Educational measurements. Medicine students. MIR exam. Psychometrics. Statistics.

Director del Curso Intensivo MIR Asturias; Clínica Baladrón de Cirugía Maxilofacial; Oviedo (J. Balandrón). Departamento de Construcción e Ingeniería de Fabricación; Universidad de Oviedo; Gijón (F. Sánchez-Lasheras). Director del Curso Atención Primaria Asturias; Clínica Oftalmológica Villacampa; Avilés (T. Villacampa). Editor del blog MIRentrelazados; Zaragoza (J.M. Romeo-Ladrero). Servicio de Oncología; Hospital Universitario Central de Asturias; Oviedo (P. Jiménez-Fonseca). Servicio de Medicina Interna; Hospital Universitario La Princesa; Madrid (J. Curbelo). IUOPA-Área de Medicina Preventiva y Salud Pública; Departamento de Medicina; Universidad de Oviedo; Oviedo (A. Fernández-Somoano). CIBER de Epidemiología y Salud Pública-CIBERESP; Instituto de Salud Carlos III; Madrid, España (A. Ferneández-Somoano).

Correspondencia:
Dr. Fernando Sánchez Lasheras. Departamento de Construcción e Ingeniería de Fabricación. Universidad de Oviedo. Pedro Puig Adam, s/n. Sede Departamental Oeste. Módulo 5, 1.ª planta. E-33203 Gijón (Asturias).

E-mail:
sanchezfernando@uniovi.es

Recibido:
10.11.16.

Aceptado:
16.11.16.

Conflicto de intereses:
No declarado.

Competing interests:
None declared.

Introducción

El examen MIR lo convocan anualmente, desde 1978, los Ministerios de Sanidad y Educación, y se realiza en el mismo día y hora en toda España. La convocatoria se publica en el Boletín Oficial del Estado unos meses antes de su realización. En los últimos años, la convocatoria se ha publicado en el mes de septiembre y los exámenes se han celebrado a finales de enero o principios de febrero del año inmediatamente posterior. En el caso de la convocatoria de 2015, ésta se realizó por medio de la Orden

de 10 de septiembre de 2015, publicada en el Boletín Oficial del Estado del 18 de septiembre de 2015.

El examen MIR se compone de 225 preguntas de test más 10 preguntas de reserva, de respuesta múltiple, que versan sobre cualquier campo de la medicina, y deben contestarse en un máximo de cinco horas. Cada pregunta acertada suma tres puntos y cada pregunta fallada resta un punto. La nota obtenida en el examen (el 90% de la nota final), junto con la valoración del baremo o expediente académico (el 10% de ella), permite clasificar en orden decreciente de puntuación total a todos los presentados. La nota de corte se fija cada año, y en las últimas convocatorias supone un 35% de la nota de los 10 mejores exámenes de ese año. Los que obtengan puntuaciones que superen la nota de corte, nota mínima exigida para el acceso a una plaza de formación sanitaria especializada, estarán en disposición de escoger la especialidad y el hospital donde realizarán la formación MIR. El MIR es un examen que busca ordenar a los aspirantes en una lista, del primero al último, según su puntuación de examen y baremo académico, para permitir una elección ordenada de las plazas ofertadas anualmente para la formación sanitaria especializada en España.

El número de aspirantes ha oscilado entre 8.000 y 25.000, según las diferentes convocatorias. En el MIR 2015 se ofertaron 6.097 plazas y fueron admitidos al examen 12.427 médicos, de los que finalmente se presentaron 11.227. La nota de corte para dicha convocatoria fue el equivalente a 65,67 preguntas acertadas netas (las preguntas netas son el resultante de restar, a las preguntas válidas, un tercio de las preguntas erróneas). 1.939 médicos (el 17,27% de los presentados) fueron eliminados por no haber obtenido una puntuación de examen superior a dicha nota, en tanto que 9.287 obtuvieron un número de orden en las listas de resultados provisionales del Ministerio y eran potenciales electores de las plazas convocadas. Del grupo de los eliminados, 1.269 eran médicos extranjeros (el 39,33% de los 3.226 médicos extranjeros presentados ese año al MIR) y 670 médicos españoles (el 8,37% de los 8.000 médicos españoles presentados al MIR).

En el MIR 2015 se adjudicaron 6.095 plazas, y quedaron desiertas dos plazas de centros privados que exigían conformidad previa. La última plaza se escogió con el número de orden 7.759 (médico no afectado por el cupo de extranjeros ni perteneciente al turno de discapacidad), y con el 4.547 (médico sí afectado por el cupo de extranjeros). Así, a diferencia de otras convocatorias, en las que todos los presentados que obtuvieron número de orden pudieron escoger plaza, en el MIR 2015 3.193 electores se quedaron sin ella. Estos 3.193 electores sin plaza (a pesar de superar la nota de corte) se pueden dividir en tres subgrupos: 1.268 incomparecencias a los actos de elección de plaza, 1.310 médicos que no pudieron elegir por tener un número de orden superior al agotamiento de la última plaza en el 7.759 y 615 médicos afectados por el cupo de extranjeros que no pudieron elegir por tener un número de orden peor al de agotamiento de dicho cupo en el 4.547. Tampoco pudieron escoger los 1.939 médicos eliminados por la nota de corte antes citados. En resumen, de los 11.227 médicos presentados, 6.095 obtuvieron plaza (54,28%) y 5.132 no la obtuvieron (45,71%), por uno u otro motivo de los enumerados anteriormente [1].

En el MIR 2015, el turno especial de discapacitados partía con una reserva de 427 plazas (el 7% del total de plazas ofertadas). De los 55 admitidos por ese turno que obtuvieron número de orden, 43 de ellos eligieron plaza durante los actos de asignación y 12 no comparecieron al acto de elección. Las plazas desiertas de dicho turno se incorporaron automáticamente al turno general.

La convocatoria de MIR 2015 supuso un cambio en la estructura habitual del examen: se modificó el diseño de las preguntas de respuesta múltiple, que pasó de las cinco opciones de respuesta de las convocatorias 1980-2014 a las cuatro opciones del MIR 2015. Al ser éste el primer MIR de estas características, hemos considerado de interés realizar un estudio de su validez estructural, estudiando sus preguntas desde el prisma de la teoría clásica de los test, ya realizado en un artículo anterior [2], y desde el punto de vista de la teoría de respuesta al ítem, el cual se presenta aquí.

La teoría clásica de los test y la teoría de respuesta al ítem constituyen los dos enfoques principales de la psicometría. Los autores del presente trabajo se proponen el análisis del examen MIR de la última convocatoria (2015), realizado el 6 de febrero de 2016, desde el punto de vista de la teoría de respuesta al ítem. Con este análisis, se completa el estudio comenzado en el artículo anterior [2] y en el que se analizó la validez estructural haciendo especial énfasis en los aspectos medibles desde el punto de vista de la teoría clásica de los test.

La teoría de respuesta al ítem tiene sus fundamentos en los trabajos de Guttman [3], Lord [4] y Rasch [5]. En la actualidad se han desarrollado un gran número de modelos psicométricos que tienen en común la relación matemática de las características latentes (no observables) de los ítems en una prueba y de las personas que las contestan, con el fin de obtener modelos de las probabilidades de

acierto de cada sujeto en cada uno de los ítems en función de su nivel de conocimiento [6].

Dado el escaso número de estudios que analizan los datos de los instrumentos de evaluación del conocimiento médico desde el punto de vista de la teoría de respuesta al ítem [6], los autores de este trabajo consideramos que el análisis que aquí se presenta puede resultar de interés para todos los colectivos implicados en la prueba MIR, así como para los investigadores y evaluadores en el campo de la educación médica.

Sujetos y métodos

Base de datos

Al igual que en el artículo anteriormente publicado por los autores acerca del MIR 2015, la base de datos utilizada en este estudio corresponde a las respuestas a las preguntas del examen que fueron introducidas por los propios examinados del MIR 2015 en una aplicación *ad hoc* creada por Curso Intensivo MIR Asturias. La finalidad de dicha aplicación era que todos los médicos que se presentaron a la convocatoria de 2015 del examen MIR, tras introducir sus respuestas a las preguntas del examen, pudieran conocer, de manera aproximada, el número de orden que obtendrían en la prueba, teniendo en cuenta estimaciones sobre el grado de dificultad de la prueba de ese año, el número de presentados, sus respuestas y su baremo académico. No todos los médicos que se examinaron introdujeron su puntuación en la mencionada base de datos, pero una vez filtrada la información y eliminados los resultados duplicados y los considerados espurios, se obtuvo la información correspondiente a las respuestas de un total de 3.712 examinados.

Modelos de la teoría de respuesta al ítem

La característica fundamental de la teoría de respuesta al ítem es que intenta prever la forma en la que los individuos contestan a las preguntas en función de su nivel de conocimiento. Es decir, la teoría de respuesta al ítem propone una serie de formulaciones sistemáticas que permiten conocer la probabilidad que tiene un individuo de acertar cada una de las preguntas de un test en función de su nivel de conocimiento.

Con el fin de obtener dicha probabilidad, los modelos matemáticos que propone la teoría de respuesta al ítem son capaces de calcular una serie de parámetros propios de cada pregunta, como la dificultad y la discriminación. Por tanto, los modelos propuestos por la teoría de respuesta al ítem asumen que existe una relación funcional entre los valores de la variable que es medida por las preguntas y la probabilidad de obtener una respuesta correcta. La función que representa esta probabilidad se denomina curva característica de los ítems.

En el caso de los exámenes de respuesta múltiple como el MIR, resulta de utilidad el uso de los modelos denominados de respuesta dicotómica [7]. En este tipo de modelos, si la respuesta elegida es correcta, ésta se codifica como 1, y, si es incorrecta, como 0, con independencia de cuál sea la opción elegida. En función del número de parámetros, los modelos dicotómicos se clasifican en modelos de uno, dos o tres parámetros. La selección de un modelo u otro debe realizarse teniendo en cuenta tanto el buen ajuste del modelo a los datos [8] como el número de parámetros que se pretendan analizar desde el punto de vista teórico.

Así, si se representa el nivel de conocimiento por la variable θ, la función de respuesta al ítem que representa la probabilidad de que un examinado con un nivel de conocimiento θ_i responda de forma correcta al *j*-ésimo ítem se puede expresar por la siguiente ecuación [8-10]:

$$P(u_j = 1|\theta_i, a_j, b_j, c_j) = c_j + \frac{(1 - c_j) \cdot \exp[-1{,}7 \cdot a_j \cdot (\theta_i - b_j)]}{1 + \exp[-1{,}7 \cdot a_j \cdot (\theta_i - b_j)]} \quad \text{(ecuación 1)},$$

donde θ_i es el nivel de conocimiento del *i*-ésimo sujeto; a_j, el valor de discriminación de la *j*-ésima pregunta (nótese que el coeficiente de discriminación se relaciona con el valor de la pendiente de la curva en el punto de inflexión); b_j, el nivel de dificultad de la pregunta *j*-ésima, y c_j, la probabilidad de que un sujeto con un nivel de conocimiento muy bajo acierte la respuesta correcta por azar. Si se considera que, teóricamente, es imposible acertar de forma aleatoria la respuesta correcta, este coeficiente tomará el valor de 0.

En otras palabras, el modelo de la ecuación 1, $P(u_j = 1|\theta_i, a_j, b_j, c_j)$, representa la probabilidad de que un sujeto con un nivel de conocimiento θ_i responda de forma correcta a la pregunta *j*-ésima. Tanto la discriminación como la dificultad de cada pregunta determinan cómo de probable es que cada individuo sea capaz de acertar la respuesta, lo que define una curva de probabilidad cuya fórmula es la de la ecuación 1. Este modelo se denomina modelo logístico de tres parámetros. De acuerdo con el modelo propuesto, la probabilidad de obtener una res-

puesta correcta depende por una parte de los parámetros de cada uno de los ítems y por otra del nivel de conocimiento del sujeto. Nótese que, según el principio de independencia local [8], la probabilidad que tiene un examinando de acertar una pregunta depende únicamente de su nivel de conocimiento y de los parámetros de dicho ítem, con independencia de los del resto de preguntas que constituyan el test. Esto, que se conoce como asunción de independencia local [9], se expresa por medio de la siguiente fórmula:

$$P(u_j = 1|\theta) = P(u_j = 1|\theta, u_k, u_l ...)\ (j = k, l ...)$$

(ecuación 2).

La ecuación 1, que representa el modelo logístico de tres parámetros, se puede simplificar de forma que represente tanto el modelo de dos parámetros como el de uno. Así, la diferencia del modelo de tres parámetros con el modelo de dos parámetros es que este último no tiene en cuenta el coeficiente de adivinación (lo supone 0) y basa la probabilidad de acierto únicamente en la dificultad del ítem y en la capacidad de discriminación. Su ecuación es la siguiente, donde todas las variables intervinientes en dicho modelo tienen el mismo significado que en la ecuación 1:

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

(ecuación 3).

A continuación, la ecuación 4 representa el modelo de un parámetro, el cual no considera la existencia de diferencias en la discriminación de los ítems y sólo tiene en cuenta la dificultad de cada pregunta. Nuevamente, tanto θ_i como b_j tienen el mismo significado que en la ecuación 1:

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

(ecuación 4).

Además, para cada ítem se define otra función, denominada función de información, y cuya ecuación es la siguiente [9]:

$$I\{\theta, u_j\} = \frac{\left[\frac{\delta P_j(\theta)}{\delta \theta}\right]^2}{P_j(\theta) \cdot [1 - P_j(\theta)]}$$

(ecuación 5),

donde $P_j(\theta) = P(u_j = 1|\theta_i, a_j, b_j, c_j)$ representa la función de respuesta al ítem. La función de información es la inversa de la precisión con la que el parámetro puede ser estimado. Así, la cantidad de información dada por cada uno de los ítems varía con el nivel de conocimiento θ. Además, la función de información también se puede definir para un test o examen completo, y ésta constituye la suma de las funciones de información para cada uno de los ítems:

$$I\{\theta\} = \sum_{j=1}^{n} I\{\theta, u_j\}$$

(ecuación 6).

El problema de la estimación de los parámetros en la teoría de respuesta al ítem se puede resolver a través de la metodología denominada estimador de máxima verosimilitud, y ésa es la aproximación que se ha adoptado en el presente trabajo. Dicho procedimiento permite la determinación de los coeficientes para todas las preguntas en los modelos logísticos con independencia del número de parámetros. No se profundiza en dicha metodología, cuyo desarrollo se puede consultar en la bibliografía [11], dado que se considera más allá del alcance necesario en este artículo.

Si bien sobre la base de datos del presente estudio se aplicaron los tres modelos de la teoría de respuesta al ítem vistos en este apartado, en la sección de resultados se presentan los correspondientes al modelo con un mejor ajuste. La bondad de ajuste de los modelos utilizados se evaluó a través del criterio de información de Akaike (CIA).

El CIA [12] surge en el marco de la teoría de la información [13] con el fin de proporcionar un criterio objetivo que ayude a los investigadores en la aplicación práctica de los principios teóricos de simplicidad y parsimonia a la hora de construir modelos matemáticos [14,15]. Así, debe tenerse en cuenta que, aunque ningún modelo matemático se puede considerar como absolutamente verdadero, el que mejor se ajuste a los datos y presente el mejor equilibrio entre su complejidad y el ajuste proporcionado debería ser el preferido. En otras palabras, se considera que el modelo que mejor se ajusta a los datos es el que minimiza la pérdida de información. Así, el CIA [16-18] proporciona un método cuantitativo con el fin de determinar qué modelo de entre un conjunto de éstos es el más parsimonioso. Desde el punto de vista práctico, cuando a un mismo conjunto de datos se le aplican diferentes modelos, el que tiene un valor menor de CIA será el que presente un mejor ajuste.

La ecuación del CIA se expresa como:

$$\text{CIA} = -2 \cdot \ln(L) + 2\, N_{\text{param}} \qquad \text{(ecuación 7)},$$

donde N_{param} es el número de parámetros que se debe estimar en cada modelo –en el caso del presente artículo, uno, dos o tres, en función de si el modelo considerado es el de un parámetro, el de dos o el de tres, respectivamente–, y L, el estimador de máxima verosimilitud.

Nótese que, como se comentó anteriormente, el estimador de máxima verosimilitud es un método [18] para la estimación de los parámetros de un modelo estadístico. En los casos en los que se aplica a un conjunto determinado de datos y a un modelo estadístico, nos proporciona una estimación de los parámetros del modelo. En general, se puede decir que el estimador de máxima verosimilitud selecciona los valores de parámetros del modelo que maximizan la coherencia de los datos con el modelo que se propone, minimizándose, por tanto, el error.

Resultados

Al igual que en convocatorias anteriores, el examen MIR de la convocatoria 2015 constó de un total de 235 preguntas, de las cuales las 10 últimas eran de reserva y se utilizarían sólo en el caso de que alguna de las 225 primeras fuera anulada por la comisión calificadora.

Las preguntas propuestas pertenecían a 33 asignaturas diferentes del grado de medicina. La asignatura con mayor número de preguntas en este examen fue aparato digestivo. Dentro de dicha asignatura se incluyen gastroenterología, hepatología y cirugía digestiva, y el total de preguntas pertenecientes a ella fue de 21.

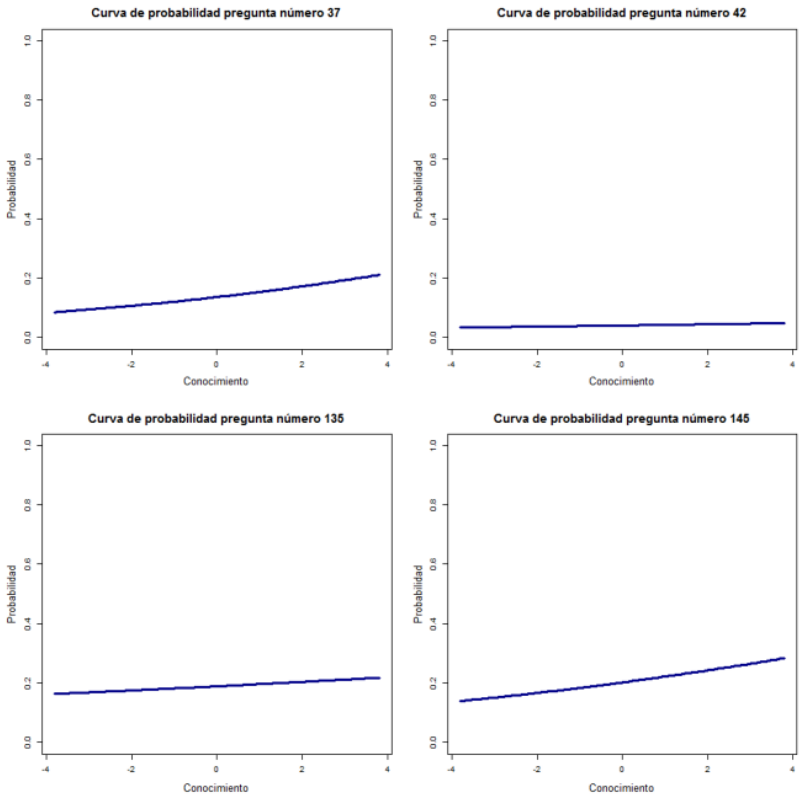
Las diferentes asignaturas que constituyen el examen MIR se pueden dividir en bloques. Así, las nueve especialidades médicas y sus correspondientes especialidades quirúrgicas corresponden al bloque de aparatos y forman el 50,64% del total del examen. El 10,64% de las preguntas integra el bloque de asignaturas básicas. De dicho bloque se han excluido las preguntas de microbiología, dado que se clasificaron dentro de enfermedades infecciosas, y las de bioestadística, que se clasificaron dentro de medicina preventiva. El 38,72% de preguntas restantes se encuadró como correspondiente a otras asignaturas. En la tabla I se clasifica cada asignatura dentro del bloque al que pertenece.

En el examen de la convocatoria analizada se produjeron cuatro anulaciones, las cuales corres-

Tabla I. Valores de dificultad (media y desviación estándar) y discriminación (media y desviación estándar) de las preguntas del examen MIR de 2015 agrupados por asignaturas.

Bloque	Asignatura	N.º de preguntas	Dificultad	Discriminación
Aparatos	Aparato digestivo	21	–0,594 (2,259)	0,695 (0,3467)
	Enfermedades Infecciosas	17	–1,473 (2,135)	0,729 (0,413)
	Neumología	14	–0,725 (3,044)	0,899 (0,451)
	Cardiología	13	–0,48 (1,942)	0,626 (0,243)
	Nefrología	12	0,048 (2,505)	0,753 (0,514)
	Neurología	11	1,294 (9,869)	0,698 (0,398)
	Hematología	10	0,239 (3,562)	1,004 (0,514)
	Reumatología	10	–0,487 (1,392)	0,899 (0,609)
	Endocrinología	10	0,164 (2,652)	0,621 (0,379)
Básicas	Anatomía patológica	5	–8,488 (17,415)	0,913 (0,639)
	Inmunología	4	–0,314 (1,526)	0,909 (0,357)
	Anatomía	4	–0,929 (1,909)	0,511 (0,105)
	Farmacología	4	2,808 (6,994)	0,490 (0,249)
	Genética	2	–1,696 (0,262)	1,704 (0,458)
Otras	Medicina preventiva	15	–1,047 (1,213)	0,989 (0,399)
	Pediatría	13	–0,181 (2,815)	0,838 (0,486)
	Ginecología y obstetricia	11	–0,802 (1,974)	0,891 (0,349)
	Psiquiatría	8	–1,029 (1,388)	1,135 (0,442)
	Traumatología	7	2,0357 (5,541)	0,694 (0,475)
	Habilidades comunicativas	6	–2,511 (0,656)	0,699 (0,137)
	Fisiología	5	14,077 (27,011)	0,486 (0,409)
	Oftalmología	4	–0,594 (1,875)	0,808 (0,389)
	Gestión clínica	4	–0,319 (1,859)	0,831 (0,399)
	Otorrinolaringología	3	–0,894 (1,184)	1,104 (0,156)
	Dermatología	3	–1,006 (0,651)	1,053 (0,265)
	Oncología	3	–0,537 (2,262)	0,821 (0,507)
	Cuidados paliativos	3	0,424 (0,925)	0,322 (0,079)
	Geriatría	3	2,373 (3,966)	0,2617 (0,199)
	Cirugía maxilofacial	2	0,265 (0,839)	0,5085 (0,232)
	Urgencias	1	0,144	0,921
	Cirugía plástica	1	–2,589	0,874
	Anestesiología	1	–2,936	0,838
	Cirugía vascular	1	–0,200	0,564
Total		231	–0,228 (5,859)	0,7914 (0,431)

Figura 1. Curvas de probabilidad de las cuatro preguntas más difíciles del examen.



pondieron a una pregunta de farmacología (pregunta n.º 36), una de cardiología (pregunta n.º 61), una de bioética (pregunta n.º 189) y una de medicina preventiva (pregunta n.º 205). En los análisis realizados no se han tenido en cuenta estas preguntas, pero sí el resto de preguntas constitutivas del examen, incluidas las de reserva, es decir, un total de 231 ítems.

De la aplicación de los modelos de uno, dos y tres parámetros se obtuvo que el modelo cuyo CIA demostraba una mejor adaptación a los datos era el de dos parámetros. Por tanto, serán los resultados correspondientes a dicho modelo los que se presenten en esta sección. Esto supone que los parámetros analizados en cada pregunta han sido su dificultad y discriminación.

Análisis por preguntas

Si se analizan los valores de dificultad de cada una de las pregunta de la prueba MIR, se observa que dichos valores se encuentran comprendidos entre un mínimo de −39,591 y un máximo de 61,876. El rango intercuartílico de los valores de dificultad está entre −1,947 y 0,085. El valor de mediana fue de −1,033 y la media de −0,228, con una desviación estándar de 5,859. Dada la extensión que tendría la tabla que contuviera los valores de dificultad y discriminación de las 231 preguntas del examen, se ha optado por no presentarlos en este artículo, sino que se realizará un análisis posterior por asignaturas, así como por bloques y tipos de preguntas.

En lo relativo a los valores de discriminación de las preguntas de esta prueba, el valor mínimo fue de −0,148, con un máximo de 2,231. El rango intercuartílico estuvo comprendido entre 0,472 y 1,073, con una mediana de 0,748 y un valor de media de 0,791, y una desviación estándar de 0,431. Todas las preguntas salvo dos (el 99,13% del total) presentaron coeficientes de discriminación positivos. Nótese que los valores de discriminación negativos deben considerase anómalos, tal y como se explica más adelante en el presente apartado.

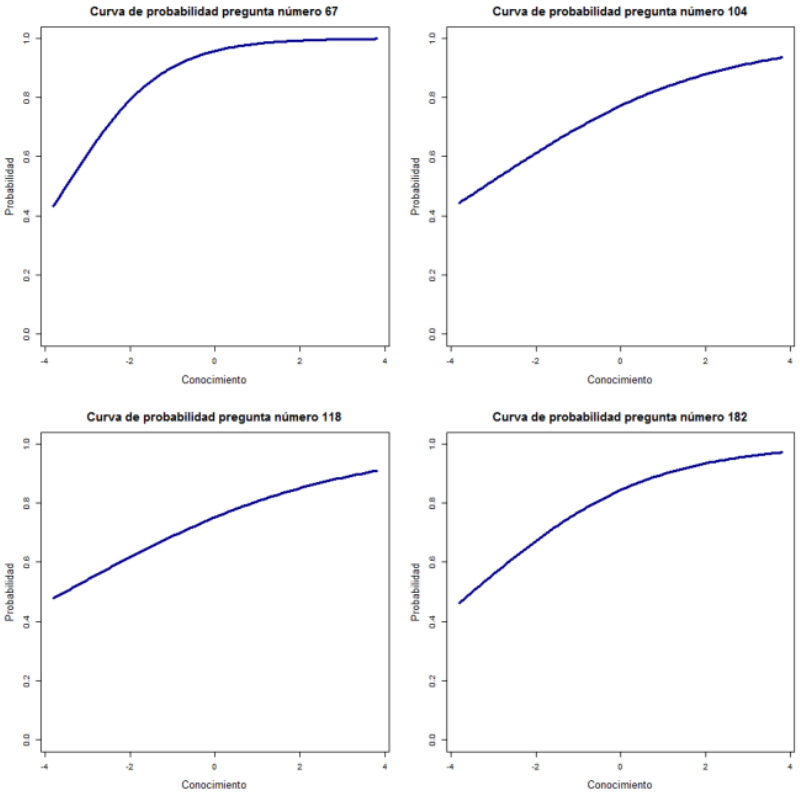
Dado que el enfoque de la teoría de respuesta al ítem se centra en la propuesta de un modelo probabilístico independiente para cada una de las preguntas del examen que permita calcular la probabilidad de acierto de cada individuo en función de su nivel de conocimiento, la presentación de resultados de este artículo se centra en el análisis individual de las preguntas. Así, la figura 1 muestra las curvas de probabilidad correspondientes a las preguntas con un mayor valor del coeficiente de dificultad de todo el examen. Dichas preguntas corresponden a las asignaturas de farmacología (n.º 37), fisiología (n.º 42), neurología (n.º 135) y traumatología (n.º 145). Tal y como se puede observar en todas ellas, la probabilidad de acierto es muy baja para todos los niveles de conocimiento, aunque se incrementa ligeramente según aumenta el nivel de conocimiento de los médicos evaluados, dado que el valor de discriminación de las cuatro preguntas, aunque muy pequeño, es en todos los casos superior a 0. Así, en ninguna de las cuatro preguntas seleccionadas ni tan siquiera los médicos con un mayor nivel de conocimientos son capaces de superar la barrera del 40% de probabilidad de acierto, y destaca entre las demás la pregunta n.º 42, en la que los alumnos más preparados no superan un 20% de probabilidades de acertarla. A través de la página web del Ministerio de Sanidad, Servicios Sociales e Igualdad se dispone de acceso completo al texto de todas las preguntas junto con sus opciones de respuesta [1]. La numeración utilizada para identificar las preguntas en el presente trabajo coincide con la

versión 0 de examen, disponible en la página web del Ministerio.

De forma similar a la de la figura anterior, la figura 2 muestra las curvas de probabilidad de las preguntas más fáciles del examen. En este caso se trata de preguntas correspondientes a las asignaturas de digestivo (n.º 62), enfermedades infecciosas (n.º 104), nefrología (n.º 118) y habilidades comunicativas (n.º 182). En todas estas preguntas se observa que la probabilidad de acierto de los médicos con los niveles más bajos de conocimientos de la muestra analizada se encuentra alrededor del 40%, que se incrementa hasta valores cercanos al 100% para los alumnos con los niveles más altos de conocimiento. Nótese cómo en el caso de las preguntas n.º 67 y 182, los médicos con valores de conocimiento intermedios en la muestra, representados por 0, alcanzan una probabilidad de acierto de estas preguntas superior al 80%, mientras que, en el caso de las preguntas n.º 104 y 118, esta probabilidad supera el 60%.

En relación con los valores de discriminación, en la parte superior de la figura 3 se presentan las dos únicas preguntas con valores de discriminaciones negativos que no fueron anuladas por la comisión calificadora. Se trata de las preguntas n.º 17 y 31, correspondientes a las asignaturas de enfermedades infecciosas y anatomía patológica. El que una pregunta presente una discriminación negativa supone que la probabilidad de responder correctamente a la pregunta por parte de un individuo disminuya a medida que aumenta su nivel de conocimiento. Dicho comportamiento es atípico y, por tanto, creemos que estas preguntas deberían ser anuladas o revisadas (con el fin de asegurarse de si están correctamente formuladas), dado que, si su objetivo es medir el constructo conocimiento médico, no debería ocurrir que, a mayor nivel de conocimiento, los examinados muestren menor probabilidad de acertar los ítems que los evalúan. La parte inferior de la figura 3 presenta las curvas de probabilidad correspondientes a las dos preguntas con mayor valor de discriminación en la prueba MIR. Al tratarse de valores positivos de discriminación, en este caso, a mayor nivel de conocimiento, mayores probabilidades de que el sujeto responda correctamente a la pregunta que se le ha formulado. Las dos preguntas a las que nos referimos son la 44 y la 141, y corresponden respectivamente a las asignaturas de genética y reumatología. Dado el alto grado de discriminación que presentan ambas preguntas, se observa que ambas resultan altamente discriminativas para ciertos niveles de conocimiento, entendiendo por este concepto que existe un intervalo de conocimiento de los examinados entre cuyos extremos se

Figura 2. Curvas de probabilidad de las cuatro preguntas más fáciles del examen.



produce un fuerte incremento de la probabilidad de acertar la pregunta. Así, en el caso de la pregunta n.º 44, se observa cómo los sujetos con un nivel de conocimiento de −2 presentan unas probabilidades de responder la pregunta inferiores al 30%, mientras que los alumnos de nivel de conocimiento 0 incrementan las probabilidades de acierto hasta algo más del 90%. En el caso de la pregunta n.º 141 ocurre algo similar, aunque en este caso se pasa de unas probabilidades de acierto de menos del 20% para un nivel de conocimiento de −3 a una probabilidad de más del 90% para los alumnos con un nivel de conocimiento de 0. Nótese también cómo, en este caso, a partir de dicho nivel de conocimiento la probabilidad de acierto apenas se incrementa, y la curva de probabilidad permanece prácticamente plana.

Análisis por asignaturas

Del análisis de los resultados por asignaturas (Tabla I) se observa que las asignaturas con las preguntas cu-

Figura 3. Curvas de probabilidad de las dos preguntas menos discriminativas del examen (arriba) y de las dos preguntas más discriminativas (abajo).

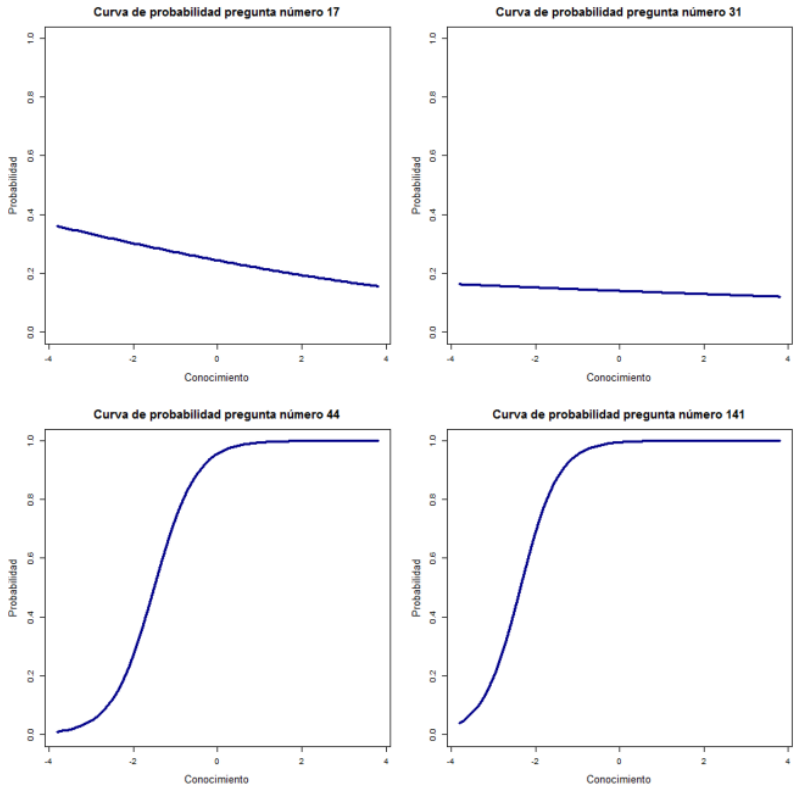
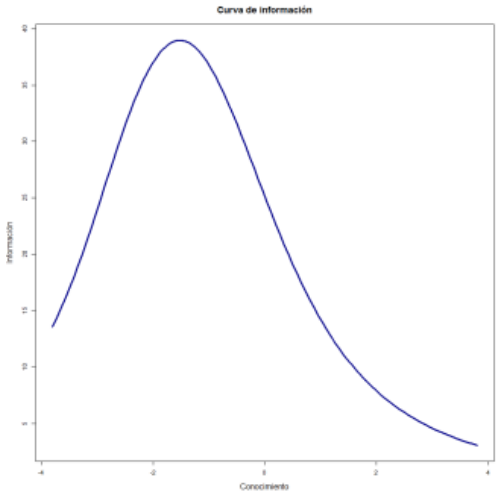


Figura 4. Curva de información correspondiente al examen en su conjunto.



yos valores medios de dificultad fueron más elevados son: fisiología, farmacología, geriatría, traumatología, neurología y cuidados paliativos. Igualmente, las asignaturas cuyas preguntas presentaron una menor dificultad media fueron anatomía patológica, anestesiología, cirugía plástica, habilidades comunicativas, genética y enfermedades infecciosas. En lo relativo a la discriminación, ninguna asignatura presentó promedio negativo de discriminación, y las asignaturas con valores medios más discriminativos fueron genética, psiquiatría, otorrinolaringología, dermatología y hematología.

Análisis por bloques de asignaturas y tipos de preguntas

Si se analiza la dificultad por bloques (Tabla II), el bloque de preguntas que resultan de mayor dificultad media es el de las asignaturas básicas, con un valor de 1,284, mientras que los valores medios de las categorías de aparatos y otras se encuentran muy próximos entre sí (–0,338 y –0,489, respectivamente). Las diferencias de las medias de los bloques de asignaturas en lo relativo a los valores de discriminación son mínimas.

En la tabla II se recogen también los valores medios y desviaciones estándar de las preguntas de la prueba MIR agrupadas según el tipo de preguntas. Así, se observa que la menor dificultad media corresponde a los casos clínicos (media de –0,804), seguidos por las preguntas negativas (0,182) y finalmente los test de preguntas directas (1,168). En relación con la variación de las dificultades, los resultados obtenidos nos permiten afirmar que la mayor variabilidad se encuentra en las preguntas de test (8,99), mientras que la menor la presentan las preguntas negativas (2,891). Los coeficientes de discriminación medios de los tres tipos de preguntas presentan valores muy similares, comprendidos entre los 0,727, con una desviación estándar de 0,468 de las preguntas negativas, y los 0,838 de las preguntas de test, con una desviación estándar de 0,405.

Análisis del examen en su conjunto

La figura 4 muestra la curva de información para el examen en su conjunto. Como ya se comentó, esta curva permite conocer para qué nivel de conocimientos la prueba MIR resulta más discriminativa. Así, dicha curva presenta un valor máximo de información de 38,972, y este máximo se alcanza para los alumnos con un nivel de conocimiento de –1,881.

Por tanto, con la información disponible, el examen MIR a los que mejor discrimina es a los individuos que presentan un nivel de conocimientos por debajo de la media de la muestra analizada. En concreto, a los que se encuentran aproximadamente alrededor del percentil 25 de la muestra analizada (con una puntuación equivalente al percentil 41 de las puntuaciones de examen de todos los médicos presentados al examen MIR 2015), mientras que los niveles más bajos de discriminación se encuentran para los individuos con los mayores niveles de conocimiento. Nótese que, en el caso del percentil 25, existe una gran diferencia entre el valor de los médicos de la muestra (104,42 preguntas netas) frente al valor obtenido por el total de aspirantes presentados al examen MIR 2015 (79,3 preguntas netas). Estas diferencias varían según aumentan los valores de puntuación. Así, en el caso de la mediana de la muestra, el valor de netas es de 127,67 y, en el conjunto de médicos presentados, de 115,67. Así, la mediana de netas de la muestra equivale al percentil 64 de todos los presentados al MIR, mientras que, para el percentil 75 en la muestra, el valor es de 145 y, en el conjunto de la población, de 139,33 preguntas netas.

Tabla II. Valores de dificultad (media y desviación estándar) y discriminación (media y desviación estándar) de las preguntas del examen MIR de 2015 agrupadas tanto por bloque como por tipo de preguntas

		N.º de preguntas (n = 231)	Dificultad	Discriminación
Bloques	Aparatos	118	–0,338 (3,746)	0,759 (0,429)
	Básicas	24	1,284 (15,672)	0,752 (0,511)
	Otras	89	–0,489 (2,482)	0,844 (0,408)
Tipos de preguntas	Caso clínico	150	–0,804 (4,659)	0,786 (0,434)
	Negativa	27	0,182 (2,891)	0,727 (0,468)
	Test	54	1,168 (8,990)	0,838 (0,405)
Total			–0,228 (5,859)	0,791 (0,4309)

Discusión

En el presente artículo se ha realizado el primer análisis conocido de un examen MIR desde el punto de vista de la teoría de respuesta al ítem. Si bien el Ministerio de Sanidad publicó estudios sobre la validez estructural de los exámenes MIR de las convocatorias de 1988 a 1992 [19,20] y además proporcionó los datos necesarios para el análisis de los exámenes de las diferentes profesiones sanitarias de las convocatorias de 2005 y 2006, trabajo realizado por Bonillo [21], los autores no conocen la existencia de ningún otro estudio como el que se presenta en este artículo.

En relación con las limitaciones del trabajo, cabe destacar que, a diferencia de los relacionados en el párrafo anterior, no se analizan los resultados de todos los examinandos, sino de una muestra de 3.712, que supone alrededor de un tercio del total de 11.227 médicos presentados a la prueba el 6 de febrero de 2016. Tal y como ocurría en el estudio publicado anteriormente sobre la misma base de datos, hemos de tener en cuenta que la información de la que disponemos presenta un cierto sesgo, dado que los médicos que obtuvieron en la prueba las puntuaciones más bajas estuvieron menos predispuestos a introducir sus respuestas en la base de datos de la aplicación. La existencia de este sesgo se manifiesta en la mediana de preguntas netas de los médicos de la muestra (128,67 preguntas netas), más alta que la de todos los médicos presentados al examen MIR 2015 (115,67 preguntas netas). Este hecho podría suponer que el coeficiente de dificultad resultante en las preguntas sea ligeramente inferior al que presentarían estas mismas preguntas si se hubiera analizado la población completa.

Por tanto, desde el punto de vista de los autores, los resultados obtenidos reflejarían más fielmente la realidad si se hubiera dispuesto de las respuestas al examen de todos los médicos que se presentaron a la prueba. A pesar de esto, se considera que la aproximación obtenida con la muestra disponible es suficiente.

También nos gustaría señalar que si la comisión calificadora de la prueba dispusiera de la información psicométrica correspondiente a ésta, sobre todo de las curvas de probabilidad correspondientes a cada una de las preguntas, tal y como se presentan en este artículo, así como de los valores de dificultad y discriminación de cada una de las preguntas, su labor de anulación de preguntas resultaría más fácil. Esto es así dado que podrían detectar las preguntas con comportamientos atípicos a través de sus gráficas. Por ejemplo, se observaría la existencia de algunas preguntas con coeficientes de discriminación negativos, como las n.º 17 y 31, y de otras cuya dificultad es muy elevada y la probabilidad de acierto es prácticamente la misma con independencia del nivel de conocimiento de los médicos evaluados, como la pregunta n.º 42. Desde nuestro punto de vista, estas preguntas precisan un análisis

minucioso por parte de expertos con el fin de determinar tanto si su formulación es correcta como si alguna de las respuestas consideradas como incorrectas tendría una formulación que permitiera que dicha respuesta fuera también correcta. No debemos olvidar que la función del examen MIR es ordenar, para lo que se requiere separar (discriminar) entre los distintos niveles de conocimiento de los médicos evaluados en la prueba. Debería considerarse la anulación de las preguntas que no contribuyesen al fin de discriminación del examen, al no separar a los médicos con mayor nivel de conocimiento de los evaluados con menores niveles de éste.

Tal y como puso de manifiesto la curva de información de la prueba MIR en su conjunto, donde menores niveles de discriminación presenta dicha prueba es en su cabeza, lo que pone de manifiesto cómo es de determinante el azar a la hora de que un individuo ocupe una posición u otra dentro del grupo de aspirantes a ocupar los primeros números de orden. Así, que el modelo que mejor ajuste las preguntas del examen sea el modelo de dos parámetros pone de manifiesto la baja probabilidad de acertar por azar las preguntas del examen.

Finalmente, y como otra posible aplicación de los modelos de la teoría de respuesta al ítem, nos gustaría señalar que, disponiendo de las respuestas de un individuo a un subconjunto de las preguntas del examen, y conocidos los parámetros de dificultad y discriminación de las preguntas restantes, sería posible predecir el resultado que el sujeto obtendría en el total de la prueba. Este principio es el que emplean los tests adaptativos computarizados que proporcionan exámenes personalizados. Además, teniendo esto en cuenta, en la actualidad se están realizando investigaciones [22] acerca del rendimiento futuro de estudiantes a partir de sus resultados en las asignaturas previamente cursadas, otro campo muy prometedor para futuros estudios.

Bibliografía

1. Ministerio de Sanidad, Servicios Sociales e Igualdad. Formación sanitaria especializada. URL: http://sis.msssi.es/fse/Default.aspx?MenuId=QE-00. [03.11.2016].
2. 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.
3. Guttman L. A basis for scaling qualitative Data. American Sociological Review 1944; 9: 139-50.
4. Lord F. A theory of test scores (Psychometric Monographs no. 7). Richmond, VA: Psychometric Corporation; 1952.
5. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press; 1980.
6. Leenen I. Virtudes y limitaciones de la teoría de respuesta al ítem para la evaluación educativa en las ciencias médicas. Investigación en Educación Médica 2014; 3: 40-55.
7. Álvarez E, Arcos A, González S, Muñoz JF, Rueda M. Estimating population proportions in the presence of missing data. Journal of Computational and Applied Mathematics 2013; 237: 470-6.
8. Embretson SE, Reise SP. Item response theory for psychologists. Hillside, NJ: Erlbaum; 2000.
9. Lord FM. Applications of item response theory to practical testing problems. Hillside, NJ: Erlbaum; 1980.
10. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. In Lord FM, Novick MR, eds. Statistical theories of mental test scores. Reading, MA: Addison-Wesley; 1968. p. 397-472.
11. Ordóñez-Galán C, Sánchez-Lasheras F, De Cos-Juez FJ, Bernardo-Sánchez AB. Missing data imputation of questionnaires by means of genetic algorithms with different fitness functions. Journal of Computational and Applied Mathematics 2017; 311: 704-17.
12. Akaike H. A new look at the statistical model identification. IEEE Transactions on Automatic Control 1974; 19: 716-23.
13. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretical approach. 2 ed. New York: Springer-Verlag; 2002.
14. Sober E. Instrumentalism, parsimony, and the Akaike framework. Philos Sci 2002; 69: S112-23.
15. Álvarez-Menéndez L, De Cos-Juez FJ, Sánchez-Lasheras F, Álvarez-Riesgo JA. Artificial neural networks applied to cancer detection in a breast screening programme. Math Comput Model 2010; 52: 983-91.
16. García-Nieto PJ, Alonso-Fernández JR, Sánchez-Lasheras F, De Cos-Juez FJ, Díaz-Muñiz V. A new improved study of cyanotoxins presence from experimental cyanobacteria concentrations in the Trasona reservoir (Northern Spain) using the MARS technique. Sci Total Environ 2012; 430: 88-92.
17. Hald A. On the history of maximum likelihood in relation to inverse probability and least squares. Stat Sci 1999; 14: 214-22.
18. Vrieze SI. Model selection and psychological theory: a discussion of the differences between the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC). Psychol Methods 2012; 17: 228-43.
19. Pruebas selectivas para el acceso a plazas de formación de médicos especialistas (1982-1992). Madrid: Ministerio de Sanidad y Consumo; 1993.
20. Pruebas selectivas para el acceso a plazas de formación de médicos especialistas. Validez estructural, diseño y capacidades exploradas (1988-1992). Madrid: Ministerio de Sanidad y Consumo; 1993.
21. 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.
22. Crespo-Turrado C, Casteleiro-Roca JL, Sánchez-Lasheras F, López-Vázquez JA, De Cos-Juez FJ, Calvo-Rolle JL, et al. Student performance prediction applying missing data imputation in electrical engineering studies degree. In Martínez-Álvarez F, Troncoso A, Quintián H, Corchado E, eds. Hybrid Artificial Intelligence Systems. 11th International Conference, HAIS 2016. Switzerland: Springer International Publishing; 2016. p. 126-35.