Propuesta metodológica para la detección de preguntas susceptibles de anulación en la prueba MIR. Aplicación a las convocatorias 2010 a 2015

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

Introducción. Anualmente, el Ministerio de Sanidad, Servicios Sociales e Igualdad de España convoca el examen de acceso a la formación sanitaria especializada. La comisión calificadora de esta prueba es la encargada de aprobar el cuaderno de preguntas de examen e invalidar las preguntas que considere improcedentes. El presente trabajo propone una metodología de ayuda a la detección de preguntas susceptibles de anulación basándose en métricas de la teoría clásica de los test y de la teoría de respuesta al ítem.

Sujetos y métodos. Para la realización del presente trabajo se usaron las respuestas a las preguntas del examen MIR de muestras de un total de 13.984 examinados de las convocatorias comprendidas entre 2010 y 2015.

Resultados. Se obtuvieron 26 preguntas que, sin haber sido anuladas, serían susceptibles de serlo en función de los valores obtenidos de sus coeficientes de correlación biserial puntual y de discriminación en el modelo logístico de dos parámetros.

Conclusiones. Existe una serie de preguntas que no son anuladas y cuya pobre calidad psicométrica empeora ligeramente la calidad global de la prueba MIR. Dado que esta prueba se realiza con el fin de ordenar a los médicos que desean acceder a una plaza de formación especializada en función de sus conocimientos, desde el punto de vista de los autores, sería recomendable que la comisión calificadora anulase todas las preguntas que presentan una mala calidad psicométrica, aun a costa de que el número final de preguntas válidas fuera ligeramente inferior a 225.

Palabras clave. Estadísticas y datos numéricos. Estándares. Estudiantes de medicina. Mediciones educativas. Métodos. Prueba MIR. Psicometría.

A methodological proposal for questions that would be withdrawn from the MIR test. Applied to the exams from 2010 to 2015

Introduction. Every year, the Spanish Ministry of Health, Social Services and Equality convenes the examination for access to specialized medical health training. The Qualification Commission of the exam is responsible for approving the test questionnaire and withdrawing the questions they consider to be inappropriate. The present work proposes a methodology to help detect questions that might be withdrawn based on the metrics of the Classical Test Theory and the Item Response Theory.

Subjects and methods. For the accomplishment of the present work, the answers given by a total of 13,984 examinees to the questions of the MIR exam have been analyzed, using the exams between 2010 and 2015.

Results. The authors propose the removal of 26 questions which, whilst not having been withdrawn by the Qualification Commission, would be susceptible to withdrawal based on the values of their point-biserial correlation coefficients and on discrimination in the two-parameter logistical model.

Conclusions. There are a number of questions that have not been withdrawn whose poor psychometric quality slightly reduces the overall quality of the MIR test. Given that this test is done in order to classify physicians who want to gain a specialized training place, from the authors' point of view it would be advisable for the Qualification Commission to withdraw all questions that present poor psychometric quality, even at the cost of making the final number of valid questions slightly fewer than 225.

Key words. Educational measurements. Medical students. Methods. MIR exam. Psychometrics. Standards. Statistics and numerical data.

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

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

E-mail:
sanchezfernando@uniovi.es

Recibido:
03.01.17.

Aceptado:
10.01.17.

Conflicto de intereses:
No declarado.

Competing interests:
None declared.

Introducción

Examen MIR

El examen de médico interno residente (MIR) para acceso de médicos a la formación sanitaria especializada en España es convocado anualmente por el Ministerio de Sanidad, Servicios Sociales e Igualdad, previo informe del Ministerio de Educación, Cultura y Deporte y de la Comisión de Recursos Humanos del Sistema Nacional de Salud, e incluye en él la oferta de plazas de formación [1]. La gestión de la prueba está encomendada a la Dirección General de Ordenación Profesional del Ministerio de Sanidad, Servicios Sociales e Igualdad [2]. Este examen evalúa los conocimientos médicos de los facultativos presentados a la prueba mediante un ejercicio compuesto por 225 preguntas de test de opción múltiple más 10 preguntas de reserva, que deben contestarse en un máximo de cinco horas.

En la convocatoria de la prueba selectiva MIR de 2015, el número de opciones de respuesta para cada una de las preguntas se redujo de cinco a cuatro, y, sin embargo, se mantuvo la misma penalización que en las convocatorias anteriores por fallar la pregunta. Cada tres preguntas falladas se resta el equivalente a una pregunta acertada. Como consecuencia de reducir una opción de respuesta por pregunta y mantener el valor de las preguntas acertadas y falladas, y a su vez mantener el tiempo de duración del examen, ha disminuido el riesgo de contestar al azar y ha aumentado el tiempo medio disponible para el análisis de cada una de las respuestas posibles por pregunta.

La nota obtenida en el examen, el 90% de la nota final, junto con la valoración de los méritos académicos, el 10% de ella, permite clasificar a todos los examinados por su puntuación total en la prueba en orden decreciente. Los médicos que obtengan puntuaciones que igualen o superen la nota de corte, nota mínima exigida para el acceso a una plaza de formación sanitaria especializada, serán convocados a los actos de elección de plaza, donde elegirán, con su número de orden, la especialidad y el hospital o unidad docente donde realizarán la formación como MIR. De los 11.227 médicos presentados al examen MIR 2015, 6.095 obtuvieron plaza (54,29%) y 5.132 no la obtuvieron (45,71%) [3].

Comisiones calificadoras de la prueba MIR

La base VIII de las aprobadas por la Orden SSI/1892/2015, de 10 de septiembre de 2015, publicada en el Boletín Oficial del Estado el día 18 del mismo mes [2], por la que se convocaron las pruebas selectivas de 2015 para el acceso en el año 2016 a plazas de formación sanitaria especializada, dispone que, una vez aprobadas las relaciones definitivas de admitidos, la persona titular de la Dirección General de Ordenación Profesional del Ministerio de Sanidad, Servicios Sociales e Igualdad dicta resolución que se publicará en el Boletín Oficial del Estado, anunciando la fecha de realización de los ejercicios y nombrando las comisiones calificadoras, de acuerdo con el artículo 34 del Real Decreto 639/2014, de 25 de julio [1].

La comisión calificadora de la prueba selectiva es el órgano encargado de aprobar el cuadernillo de preguntas del examen, invalidar las que considere improcedentes, resolver las reclamaciones que se produzcan contra ellas y aprobar la plantilla definitiva de respuestas correctas. Para ello puede requerir el asesoramiento de expertos o personas debidamente cualificados.

La composición de la comisión calificadora del MIR 2015 se publicó mediante la Resolución de 4 de enero de 2016, de la Dirección General de Ordenación Profesional. En la convocatoria MIR 2015, por primera vez, y en cumplimiento de lo previsto en el artículo 34 del Real Decreto 639/2014, la presidencia la nombró el Ministerio de Sanidad, Servicios Sociales e Igualdad, y la vicepresidencia, el Ministerio de Educación, y también, por primera vez, las facultades de Medicina tuvieron dos vocales en la comisión calificadora de la prueba. Con anterioridad disponían de un solo vocal.

Las comisiones nombradas se reúnen todos los años en el Ministerio de Sanidad, Servicios Sociales e Igualdad, el mismo día del examen MIR, desde antes de su inicio hasta después de su terminación, para ejercer las funciones que se les encomienda en el artículo 34 del Real Decreto 639/2014, de 25 de julio. La comisión calificadora debate las preguntas del examen y puede anular alguna, y aprueba en esa misma sesión la plantilla de respuestas provisionales que el Ministerio de Sanidad, Servicios Sociales e Igualdad viene publicando, aproximadamente, una semana después de realizado el examen. Si en esa primera reunión decide anular alguna de las 235 preguntas, su solución figurará en blanco en la plantilla de respuestas provisionales.

Una vez publicada la plantilla de respuestas provisionales, los examinados disponen de un plazo de tres días para presentar las reclamaciones a éstas, solicitando por vía telemática la anulación o cambio de respuesta de las preguntas que cada uno estime. Dichas reclamaciones deben estar argumentadas y apoyadas en bibliografía.

Unas semanas después, la comisión celebra una nueva sesión, en los locales del Ministerio de Sanidad, Servicios Sociales e Igualdad, para examinar y resolver las reclamaciones que, en su caso, hubieran podido presentarse a las preguntas y respuestas aprobadas provisionalmente. En dicha reunión pueden decidir anular nuevas preguntas o cambiar algunas de las respuestas inicialmente aprobadas, fijando así la plantilla de respuestas definitivas, que serán las consideradas válidas para el cálculo de las puntuaciones de examen de los médicos presentados a éste. Las preguntas anuladas aparecen en blanco en la plantilla definitiva de respuestas correctas. Una vez anuladas dichas preguntas, automáticamente se activarán para su sustitución, y por riguroso orden, las preguntas de reserva que permitan sumar 225 preguntas 'activas', y se considerarán como tales las primeras 225 preguntas no anuladas.

En general, las razones que pueden motivar la anulación de una pregunta se pueden clasificar en tres categorías fundamentales: defectos en el contenido médico que se pretende evaluar (respuesta errónea o presencia de más de una opción válida de respuesta), defectos técnicos en la redacción de las preguntas que llevasen al examinado a interpretar erróneamente la pregunta o respuesta (incluyendo defectos tipográficos en la elaboración del cuadernillo de preguntas) y defectos en la capacidad discriminativa de la pregunta medida por técnicas psicométricas. Es decir, preguntas cuya probabilidad de acertar no tiene relación con el nivel del conocimiento médico de los examinados, o que aciertan con mayor frecuencia los médicos con un nivel de conocimiento más bajo que los que tienen unos conocimientos superiores.

Los autores del presente trabajo ya se interesaron por el estudio de las características psicométricas de la prueba MIR 2015 tanto desde el punto de vista de la teoría clásica de los test [4] como desde el punto de vista de la teoría de respuesta al ítem [5]. En este tercer trabajo se propone una metodología de ayuda a la detección de preguntas susceptibles de anulación por tener mala o nula discriminación, basándose en dos métricas provenientes de la teoría clásica de los test y de la teoría de respuesta al ítem, respectivamente, que ayudan a mejorar la validez de contenido del test. La metodología propuesta resulta igualmente válida para todos los exámenes MIR, con independencia del número de opciones de sus preguntas. En el presente trabajo, a modo de ejemplo, dicha metodología se ha aplicado a los exámenes MIR comprendidos entre las convocatorias de 2010 y 2015, respectivamente.

Sujetos y métodos

Base de datos

Para elaborar este artículo se han utilizado las respuestas a las preguntas del examen MIR que fueron introducidas por 13.984 examinados de las convocatorias MIR entre 2010 y 2015. Esta información se recogió en una aplicación *ad hoc* creada por el Curso Intensivo MIR de Asturias. La finalidad de dicha aplicación era que todos los médicos que se presentaron a las mencionadas convocatorias, tras introducir sus respuestas a las preguntas del examen, pudieran conocer con antelación, y de manera aproximada, el número de orden que obtendrían en la prueba. De esta forma, se dispone de una muestra de las contestaciones a las preguntas de la prueba MIR de un conjunto de examinados correspondiente a cada uno de los años considerados. Así, para 2010 se tienen las respuestas de 1.962 médicos examinados; para 2011, de 1.695; para 2013, de 2.216; para 2013, de 2.192; para 2014, de 2.207; y para 2015, de 3.712 médicos presentados al examen MIR.

Plantillas de respuestas para la corrección de exámenes

El estudio métrico de las 30 preguntas anuladas desde la aprobación de las plantillas provisionales a las definitivas se ha realizado considerando como válida la respuesta publicada para ellas en la plantilla provisional de respuestas, ya que, una vez anuladas, la solución desaparece de la plantilla de respuestas definitivas. Para el estudio de las 1.372 preguntas restantes se ha considerado como válida la respuesta aprobada en la plantilla de respuestas definitivas. Las preguntas en las que la comisión calificadora decidió cambiar la respuesta desde la plantilla provisional a la definitiva no se han incluido en la categoría de preguntas anuladas analizadas en el presente trabajo y, por lo tanto, su estudio psicométrico se ha realizado únicamente con la solución aprobada para ellas por la comisión calificadora en la plantilla de respuestas definitivas.

Índice de dificultad corregido

Es posible calcular el índice de dificultad teniendo en cuenta la necesaria corrección debida a los efectos del azar [6]. Se trata de la siguiente fórmula:

$$ID = \frac{A - \frac{E}{K-1}}{N} \qquad \text{[ecuación 1]}$$

En esta fórmula se designa como *A* el número de sujetos que aciertan el ítem, *E* representa el número de sujetos que lo fallan sobre el total de *N* sujetos que realizan el examen y *K* designa el número de opciones que tienen las preguntas. En el presente trabajo, al igual que en los realizados con anterioridad [4,5], se considerará que los sujetos que no han contestado a un ítem lo han fallado.

Los valores que se obtienen del índice de dificultad con la corrección del azar son menores que los de dificultad no corregida para cada una de las preguntas analizadas. Con el fin de clasificar las preguntas en función de su nivel de dificultad, se utilizará la siguiente escala [4]:

- *Muy fáciles:* las preguntas cuyo índice de dificultad corregida sea superior a 0,80.
- *Fáciles:* las preguntas cuyo índice de dificultad corregida se encuentre entre 0,66 y 0,80.
- *Óptimas:* las preguntas cuyo índice de dificultad corregida se encuentre entre 0,33 y 0,66.
- *Aceptables:* las preguntas cuyo índice de dificultad corregida se encuentre entre 0 y 0,33.
- *Muy difíciles:* las preguntas cuyo índice de dificultad sea inferior a 0.

Además de haberse utilizado en la publicación de los autores mencionada, esta variable se aplica comúnmente en la teoría clásica de los tests [7], y también se ha empleado en otras publicaciones anteriores que analizaron las pruebas selectivas MIR [8,9].

Índice de correlación biserial puntual

El índice de correlación biserial puntual es una métrica ampliamente utilizada en el marco de la teoría clásica de los tests. Su objetivo es medir la validez discriminativa de las preguntas. Se trata de una extensión de la fórmula de correlación de Pearson para el caso en el que una de las variables sea dicotómica y la otra cuantitativa [4]. Se expresa por medio de la siguiente fórmula:

$$P_{bp} = \frac{\mu_p - \mu_q}{\sigma_x} \cdot \sqrt{\frac{ID}{1 - ID}} \quad \text{[ecuación 2]}$$

siendo μ_p la puntuación media en el test de los sujetos que aciertan la pregunta; μ_q, la puntuación media en el test de los sujetos que fallan la pregunta; σ_x, la desviación típica de la puntuación total del test; e *ID*, el índice de dificultad calculado como la proporción de sujetos que aciertan la pregunta.

A mayor valor del índice de correlación biserial puntual, mayor será la relación entre obtener una puntuación alta en el test y el hecho de haber contestado correctamente a la pregunta.

En función de los resultados del coeficiente de correlación biserial puntual, las preguntas se pueden clasificar en las siguientes categorías [10]:

- *Excelente:* si el valor obtenido del coeficiente de correlación biserial puntual es mayor que 0,39.
- *Buena:* si el valor obtenido del coeficiente de correlación biserial puntual es mayor o igual que 0,30 y menor que 0,39.
- *Regular:* si el valor obtenido del coeficiente de correlación biserial puntual es mayor o igual que 0,20 y es menor que 0,30.
- *Pobre:* si el valor obtenido del coeficiente de correlación biserial puntual es mayor o igual que 0 y menor que 0,20.
- *Pésima:* en los casos en los que el coeficiente de correlación biserial puntual es negativo.

Modelo de dos parámetros de la teoría de respuesta al ítem

La teoría de respuesta al ítem es capaz de predecir cómo los individuos contestarían a las preguntas en función de su nivel de conocimientos. Para ello, la teoría de respuesta al ítem propone unos modelos probabilísticos que estiman con qué probabilidad un individuo será capaz de responder de manera correcta a cierta pregunta.

A partir de los resultados obtenidos en un trabajo previo [5], se determinó que el modelo con el que mejor se podía establecer la relación existente entre la probabilidad de acertar la pregunta por parte de los examinados en la prueba MIR y su nivel de conocimientos era a través del modelo conocido como logístico de dos parámetros (2PL). Dicho modelo se representa por la siguiente ecuación:

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

[ecuación 3]

siendo θ_i el nivel de conocimiento del *i*-ésimo sujeto; a_j, el valor de discriminación de la *j*-ésima pregunta; y b_j, el nivel de dificultad de la pregunta *j*-ésima.

La elección del modelo 2PL se realizó en función de los resultados obtenidos en la aplicación del criterio de información de Akaike (CIA) a los modelos considerados. Esta forma de proceder es habitual para la determinación de la bondad de ajuste [11, 12]. De acuerdo con el modelo propuesto, la probabilidad de obtener una respuesta correcta depende,

por una parte, de los parámetros de cada uno de los ítems (dificultad y discriminación), y, por otra, del nivel de conocimiento del sujeto.

Además, para cada ítem se dispone de otra función denominada función de información [13,14], y cuya ecuación es la siguiente:

$$I\{\theta, u_j\} = \frac{\left[\frac{\delta P_j(\theta)}{\delta \theta}\right]^2}{P_j(\theta) \cdot [1 - P_j(\theta)]} \quad \text{[ecuación 3]}$$

donde $P_j(\theta) = P(u_j = 1 | \theta_i, a_j, b_j)$ representa la función de respuesta al ítem. La función de información de un test así definida para un determinado nivel de conocimiento θ es la inversa de la varianza de sus errores de medida para ese valor. La función de información es un indicador de la precisión del test. Cuanto mayor sea el valor de esta función, menor será el error típico de medida, luego mayor será la información que las estimaciones aportan sobre el parámetro θ. Si la función de información se calcula para todos los niveles de θ, se obtiene la curva de información del test.

Una de las aplicaciones prácticas de la función de información es que permite comparar la eficiencia de dos tests a la hora de medir el nivel de conocimiento (θ) en sus distintos niveles o valores. Se denomina eficiencia relativa de dos tests para un determinado valor de θ al cociente entre la función de información de ambos test para dicho valor de θ.

Los autores proponen la siguiente escala para la clasificación de las preguntas en función de los resultados del coeficiente de discriminación del modelo 2PL, utilizando las mismas categorías empleadas para la clasificación de la discriminación según los valores de correlación biserial puntual.

- *Excelente:* si el valor del coeficiente de discriminación es mayor que 1.
- *Buena:* si el valor del coeficiente de discriminación es mayor o igual que 0,70 y menor o igual que 1.
- *Regular:* si el valor del coeficiente de discriminación es mayor a 0,40 y menor que 0,70.
- *Pobre:* si el valor del coeficiente de discriminación es mayor o igual que 0 y menor que 0,40.
- *Pésima:* si el valor del coeficiente de discriminación es negativo.

Criterios psicométricos propuestos para la anulación de preguntas

A través de la anulación de preguntas se puede mejorar la validez del contenido de un test, dado que así se consigue que éste represente de forma más fiel los contenidos que se pretenden evaluar con él; en el caso del examen MIR, el conocimiento de un médico posgraduado. Además, de esta forma también se conseguirá aumentar la validez aparente [15], es decir, la validez que la prueba tiene a ojos de los que se examinan.

Nótese que la validez aparente resulta muy importante a la hora de conseguir una buena actitud y motivación de los aspirantes que se enfrentan a la prueba MIR, pues, si éstos no considerasen que las preguntas propuestas sirven para la evaluación de sus conocimientos médicos, podrían cuestionar la utilidad real de la prueba como elemento clasificador de cara al acceso de los aspirantes a la formación sanitaria especializada en España.

En el presente trabajo, se proponen como criterios psicométricos que hagan a una pregunta candidata a su anulación los dos siguientes:

- Que su coeficiente de discriminación calculado para el modelo 2PL de la teoría de respuesta al ítem sea negativo.
- Que el valor de su coeficiente de correlación biserial puntual sea negativo.

El cumplimiento de cualquiera de los dos criterios será suficiente para su propuesta como posible pregunta a anular por la comisión calificadora de la prueba. Así, las preguntas que presentan valores negativos de sus parámetros de discriminación para el modelo 2PL se caracterizan porque el valor de la probabilidad de contestar correctamente a estas preguntas disminuye a medida que aumenta el nivel de conocimiento de los médicos examinados, lo cual es contrario a la lógica.

Por tanto, se considera que estas preguntas no son adecuadas para la medición del conocimiento, en el caso que nos ocupa el conocimiento médico. El índice de correlación biserial puntual sirve también para el estudio de la validez discriminativa de las preguntas, pues, cuanto mayor sea este valor, mayor será la relación entre haber obtenido una puntuación alta en el test y haber acertado la pregunta en cuestión.

Con el fin de verificar si los exámenes MIR mejoran anulando las preguntas que cumplan alguno de los dos criterios expuestos con anterioridad, se calculará la modificación de los valores promedio del coeficiente de discriminación, el coeficiente de correlación biserial puntual y el incremento del área bajo la curva de información, lo cual puede interpretarse como la mejora de la eficiencia relativa de discriminación de la prueba MIR considerada para todo el rango de niveles de conocimiento.

Resultados

Preguntas anuladas por las comisiones calificadoras

El número de preguntas anuladas en la plantilla de respuestas definitivas de los exámenes MIR entre las convocatorias de 2010 y 2015 es de 38, de un total de 1.410 preguntas incluidas en dichos exámenes, lo que supone la anulación del 2,7% del total de las preguntas. El número de preguntas anuladas en cada convocatoria varía entre las cuatro preguntas anuladas en 2015 y las ocho de las convocatorias de 2011 y 2013.

De las 38 preguntas anuladas en el período de estudio, ocho lo fueron antes de la publicación de las plantillas de respuesta provisionales (21,05% de las 38 anuladas) y 30 con la publicación de las plantillas de respuesta definitivas (78,95% de las preguntas anuladas). Aproximadamente, de cada cinco preguntas anuladas, una se anula en la plantilla provisional y cuatro en la plantilla definitiva. En lo relativo a las ocho preguntas anuladas antes de la publicación de la plantilla provisional, en la convocatoria de 2012 no se anuló ninguna, en las de 2010 y 2011 se anuló una (pregunta número 160 en 2010 y 210 en 2011), y en las convocatorias de 2013, 2014 y 2015 se anularon dos (preguntas números 121 y 232, 123 y 207, y 61 y 89, respectivamente). Todos los números empleados para identificar las preguntas se refieren a la versión 0 del examen. Nótese que, de estas preguntas, en la plantilla provisional publicada por la comisión calificadora no figura ninguna respuesta como correcta, por lo que no se puede realizar ningún estudio métrico de ellas que permita evaluar su dificultad ni su discriminación. El análisis psicométrico ha incluido el resto de preguntas de los exámenes MIR 2010 a 2015.

Las preguntas anuladas, de las que se dispone de una solución en la plantilla provisional de respuestas de la prueba, se recogen en la tabla I, junto sus características de dificultad y discriminación. El perfil promedio de las 30 preguntas anuladas por la comisión calificadora en el período estudiado es el de una pregunta difícil (promedio de índice de dificultad con corrección del azar de 0,1252), con una discriminación pobre (promedio de correlación biserial puntual de 0,1512 y de coeficiente de discriminación 2PL de 0,3586).

De las 30 preguntas anuladas por las comisiones calificadoras, siete de ellas cumplían los criterios de anulabilidad propuestos en este trabajo (23,3% de las anuladas). De las siete preguntas que fueron anuladas por la comisión calificadora y que también serían anuladas a través del sistema propuesto, una corresponde al MIR de 2010; cuatro, al de 2011; y dos, al de 2013.

Preguntas psicométricamente anulables, pero no anuladas

En el presente trabajo se proponen como anulables todas las preguntas con valores negativos de su coeficiente de discriminación en el modelo 2PL o del valor de su correlación biserial puntual. Con el criterio de anulación propuesto por los autores, 33 de las 1.410 preguntas de los exámenes MIR 2010-2015 no serían discriminativas (el 2,34% del total de preguntas). El número de preguntas que, siguiendo los criterios expuestos, hubieran sido anulables, oscila entre las cuatro de los exámenes de 2010, 2012 y 2014, y las nueve del examen de 2013. De las 33 preguntas psicométricamente anulables con los criterios propuestos, siete fueron anuladas por la comisión calificadora.

La información de la tabla I se completa con la mostrada en la tabla II, que agrupa las 26 preguntas no anuladas cuya correlación biserial puntual es inferior a 0 o poseen un coeficiente de discriminación negativo calculado para el modelo 2PL de la teoría de respuesta al ítem. Nótese que todos los cálculos que se presentan en estos resultados se corresponden con los obtenidos para la base de datos disponible para cada uno de los años según la descripción realizada en el apartado de 'Sujetos y métodos'.

El perfil promedio de las 26 preguntas anulables psicométricamente, pero no anuladas por la comisión calificadora en el período estudiado, es el de una pregunta difícil (promedio de índice de dificultad con corrección del azar de 0,0206), con una discriminación pésima (promedio de correlación biserial puntual de −0,0388 y de coeficiente de discriminación 2PL de −0,1196). Estas 26 preguntas constituyen únicamente el 1,84% del total de preguntas de los exámenes considerados.

En la figura 1 se representan, a modo de ejemplo, las curvas de probabilidad de algunas de las preguntas candidatas a ser anuladas. Se trata, en concreto, de las cuatro preguntas del examen MIR de 2015 que se deberían anular según los criterios propuestos en el presente trabajo. Tal y como se observa, estas cuatro preguntas se caracterizan por tener una curva de probabilidad decreciente. Es decir, a mayor conocimiento del examinado, menor probabilidad de acierto de la pregunta, y se diferencian entre ellas en su pendiente. Nótese cómo la curva de probabilidad de la pregunta 83 es prácticamente plana. Así, en el caso de la prueba de 2015, las cuatro preguntas propuestas para anulación por

Tabla I. Relación completa de preguntas anuladas después de la publicación de la plantilla provisional en el examen MIR en las convocatorias comprendidas entre 2010 y 2015. Anulable: preguntas que serían anulables según el criterio propuesto en el presente trabajo. Índice de dificultad corregido y clasificación de preguntas en sus categorías. Correlación biserial puntual y clasificación de preguntas en sus categorías. Coeficiente de discriminación modelo 2PL de cada una de las preguntas y clasificación de éstas en sus categorías.

Convocatoria	N.º de pregunta	Anulada	Anulable	Índ. dif. corr.	Categorías	Corr. bis. puntual	Categorías	Coef. disc.	Categorías
2010	142	Sí	Sí	–0,1654	Muy difícil	–0,0767	Pésima	–0,2732	Pésima
2010	143	Sí	No	–0,0246	Muy difícil	0,1014	Pobre	0,2774	Pobre
2010	148	Sí	No	0,0845	Difícil	0,1175	Pobre	0,2814	Pobre
2010	209	Sí	No	0,0378	Difícil	0,1709	Pobre	0,4747	Regular
2010	221	Sí	No	–0,1469	Muy difícil	0,0735	Pobre	0,5650	Regular
2011	22	Sí	No	0,4324	Óptimo	0,2615	Regular	0,4875	Regular
2011	60	Sí	No	0,1985	Difícil	0,2781	Regular	0,5839	Regular
2011	102	Sí	No	0,3460	Óptimo	0,1322	Pobre	0,2128	Pobre
2011	109	Sí	Sí	–0,1276	Muy difícil	–0,0495	Pésima	–0,2325	Pésima
2011	121	Sí	Sí	–0,1302	Muy difícil	–0,0185	Pésima	–0,1335	Pésima
2011	133	Sí	Sí	0,1035	Difícil	–0,019	Pésima	–0,0675	Pésima
2011	136	Sí	Sí	–0,0779	Muy difícil	–0,0151	Pésima	–0,0538	Pésima
2012	149	Sí	No	0,0821	Difícil	0,1321	Pobre	0,2910	Pobre
2012	166	Sí	No	0,2657	Difícil	0,1292	Pobre	0,2386	Pobre
2012	199	Sí	No	–0,0264	Muy difícil	0,1998	Pobre	0,6528	Regular
2012	211	Sí	No	0,0469	Difícil	0,2770	Regular	0,7887	Buena
2012	214	Sí	No	0,0366	Difícil	0,2541	Regular	0,7667	Buena
2013	92	Sí	No	0,5836	Óptimo	0,3642	Buena	0,7809	Buena
2013	129	Sí	Sí	–0,0575	Muy difícil	–0,0519	Pésima	–0,1368	Pésima
2013	153	Sí	No	0,1575	Difícil	0,0486	Pobre	0,0539	Pobre
2013	182	Sí	Sí	0,2621	Difícil	–0,0725	Pésima	–0,1423	Pésima
2013	183	Sí	No	0,3134	Difícil	0,1740	Pobre	0,3187	Pobre
2013	230	Sí	No	0,1302	Difícil	0,2727	Regular	0,6712	Regular
2014	128	Sí	No	0,5111	Óptimo	0,3177	Buena	0,6200	Regular
2014	138	Sí	No	–0,0747	Muy difícil	0,1594	Pobre	0,4931	Regular
2014	200	Sí	No	0,3526	Óptimo	0,4045	Excelente	0,9091	Buena
2014	212	Sí	No	0,5556	Óptimo	0,3498	Buena	0,6857	Regular
2014	233	Sí	No	0,1709	Difícil	0,3144	Buena	0,8063	Buena
2015	36	Sí	No	–0,2136	Muy difícil	0,0427	Pobre	0,2198	Pobre
2015	205	Sí	No	0,1312	Difícil	0,2652	Regular	0,6182	Regular

Coef. disc.: coeficiente de discriminación; Corr. bis. puntual: correlación biserial puntual; Índ. dif. corr.: índice de dificultad corregido.

Tabla II. Preguntas psicométricamente anulables, pero no anuladas por la comisión calificadora, y correspondientes a las convocatorias MIR comprendidas entre 2010 y 2015. Índice de dificultad corregido y clasificación de preguntas en sus categorías. Correlación biserial puntual y clasificación de preguntas en sus categorías. Coeficiente de discriminación modelo 2PL de cada una de las preguntas y clasificación de éstas en sus categorías.

Convocatoria	N.º de pregunta	Anulada	Anulable	Índ. dif. corr.	Categorías	Corr. bis. puntual	Categorías	Coef. disc.	Categorías
2010	56	No	Sí	0,0558	Difícil	–0,0424	Pésima	–0,1026	Pésima
2010	158	No	Sí	–0,0129	Muy difícil	–0,0251	Pésima	–0,0654	Pésima
2010	193	No	Sí	–0,0157	Muy difícil	0,0105	Pobre	–0,0029	Pésima
2011	35	No	Sí	0,1149	Difícil	–0,0977	Pésima	–0,2152	Pésima
2011	101	No	Sí	0,1217	Difícil	–0,0256	Pésima	–0,0469	Pésima
2011	209	No	Sí	0,1758	Difícil	–0,0482	Pésima	–0,1165	Pésima
2012	2	No	Sí	–0,0954	Muy difícil	–0,2122	Pésima	–0,4716	Pésima
2012	83	No	Sí	0,1167	Difícil	–0,0265	Pésima	–0,0817	Pésima
2012	222	No	Sí	–0,1889	Muy difícil	–0,0221	Pésima	–0,1159	Pésima
2012	231	No	Sí	0,0613	Difícil	–0,0468	Pésima	–0,085	Pésima
2013	2	No	Sí	0,3628	Óptimo	–0,0112	Pésima	–0,0604	Pésima
2013	65	No	Sí	0,2265	Difícil	–0,0166	Pésima	–0,0461	Pésima
2013	125	No	Sí	–0,1489	Muy difícil	–0,0695	Pésima	–0,2489	Pésima
2013	155	No	Sí	0,2908	Difícil	0,0082	Pobre	–0,0036	Pésima
2013	174	No	Sí	–0,1270	Muy difícil	–0,0759	Pésima	–0,2757	Pésima
2013	199	No	Sí	0,1673	Difícil	–0,1331	Pésima	–0,2613	Pésima
2013	202	No	Sí	–0,0937	Muy difícil	–0,0665	Pésima	–0,2042	Pésima
2014	24	No	Sí	–0,1193	Muy difícil	–0,0409	Pésima	–0,1416	Pésima
2014	42	No	Sí	–0,2199	Muy difícil	–0,0092	Pésima	–0,1312	Pésima
2014	129	No	Sí	0,0437	Difícil	0,0208	Pobre	–0,003	Pésima
2014	165	No	Sí	0,0282	Difícil	–0,0220	Pésima	–0,0794	Pésima
2015	17	No	Sí	0,0027	Difícil	–0,1088	Pésima	–0,2360	Pésima
2015	31	No	Sí	–0,1429	Muy difícil	–0,0154	Pésima	–0,0779	Pésima
2015	42	No	Sí	–0,2410	Muy difícil	0,0527	Pobre	–0,0108	Pésima
2015	135	No	Sí	–0,0693	Muy difícil	0,0027	Pobre	–0,0219	Pésima
2015	175	No	Sí	0,2422	Difícil	0,0118	Pobre	–0,0032	Pésima

Coef. disc.: coeficiente de discriminación; Corr. bis. puntual: correlación biserial puntual; Índ. dif. corr.: índice de dificultad corregido.

esta metodología presentan valores negativos en sus coeficientes de discriminación.

Mejora de la discriminación del examen MIR con la aplicación de los criterios de anulación propuestos

A continuación se muestran los resultados de la mejora de la calidad psicométrica del conjunto de las preguntas del examen como consecuencia de las anulaciones de las preguntas por las comisiones calificadoras, y cómo la aplicación de los criterios psicométricos propuestos conseguiría una mejora aún mayor de la capacidad discriminativa de las preguntas restantes en las pruebas.

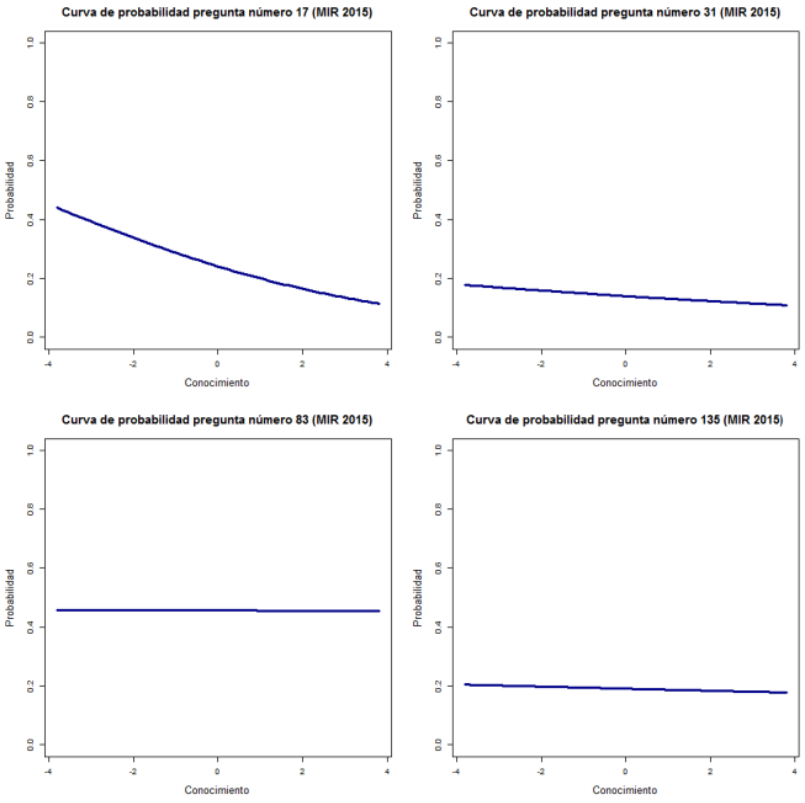
Como consecuencia de la anulación de las preguntas por las comisiones calificadoras, el promedio de correlación biserial puntual de las preguntas de los exámenes MIR objeto del estudio pasó de 0,3038 a 0,308, lo que supone una mejora de la discriminación del 1,38%. Si, además, se anulasen las preguntas con discriminación negativa (medida por los criterios propuestos), los valores promedios de correlación biserial puntual pasarían a 0,3151, con una mejora adicional de la discriminación del 2,31%.

Como consecuencia de la anulación de las preguntas por las comisiones calificadoras, el promedio de coeficiente de discriminación en el modelo 2PL de las preguntas de los exámenes MIR objeto de este estudio pasó de 0,7732 a 0,7853, lo que supone una mejora de la discriminación del 1,56%. Si, además, se anulasen las preguntas con discriminación negativa (medida por los criterios propuestos), los valores promedios de coeficiente de discriminación en el modelo 2PL pasarían a 0,802, lo que supone una mejora adicional de la discriminación del 2,13%.

Otra forma de comprobar la mejora en la discriminación de la prueba con las anulaciones es a través de la modificación de la función de información del examen en su conjunto. La función de información de un test sirve como indicador de su precisión.

La función de información de la prueba MIR de 2010 calculada utilizando las 225 preguntas que fueron consideradas como válidas por la comisión calificadora después de las anulaciones se representa por medio de la curva azul de la figura 2a. En dicha figura, la curva roja representa la función de información que se obtendría si, además de las preguntas eliminadas por la comisión calificadora, se eliminan las otras tres preguntas propuestas por la metodología que se presenta en este artículo (Tabla II). Si se realiza esta eliminación, se produce un incremento del área bajo la curva del 0,8%, lo que se puede asimilar a un incremento de la eficiencia relativa del test en esa misma cantidad.

Figura 1. Curvas de probabilidad de las preguntas números 17 (Enfermedades infecciosas), 31 (Anatomía patológica), 83 (Endocrinología) y 135 (Neumología) de la prueba MIR de 2015.



Si se procede de la misma forma para el resto de exámenes MIR objeto de estudio, en el año 2011 se logra un incremento del área bajo la curva de un 0,74%, aunque en este caso hemos de tener en cuenta que no se dispone de un número suficiente de preguntas que se puedan añadir, por lo que sólo se eliminan las dos primeras preguntas propuestas según la tabla II. En el año 2012 se incrementa el área bajo la curva en un 1,76%; en esta ocasión, y debido a la falta de preguntas adicionales, únicamente se han eliminado las dos primeras preguntas según la propuesta de la tabla II. De la misma forma, en el año 2013, el incremento del área es del 0,98%. En esta ocasión se pueden eliminar las tres primeras preguntas de las cuatro propuestas, dado que no se dispone de suficientes preguntas de reserva. En el año 2014 es posible eliminar tres de las cinco preguntas propuestas y se mejora un 1,14%. Finalmente, en el año 2015, eliminando todas las preguntas que se consideran anulables según el criterio auto-

Figura 2. Curvas de información, comparación de áreas bajo la curva. En azul: curva de información obtenida con las preguntas anuladas por la comisión calificadora. En rojo: curva de información obtenida si, además de las preguntas anuladas por la comisión calificadora, se anulan también las propuestas con la metodología que se expone en el presente artículo. a) Prueba MIR de 2010; b) Prueba MIR de 2015.

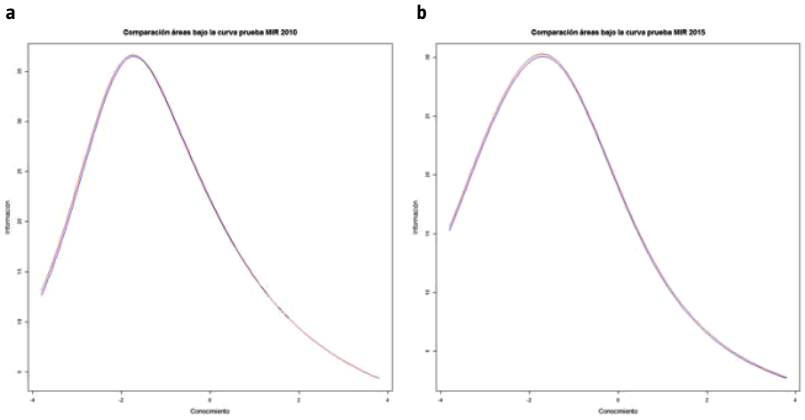
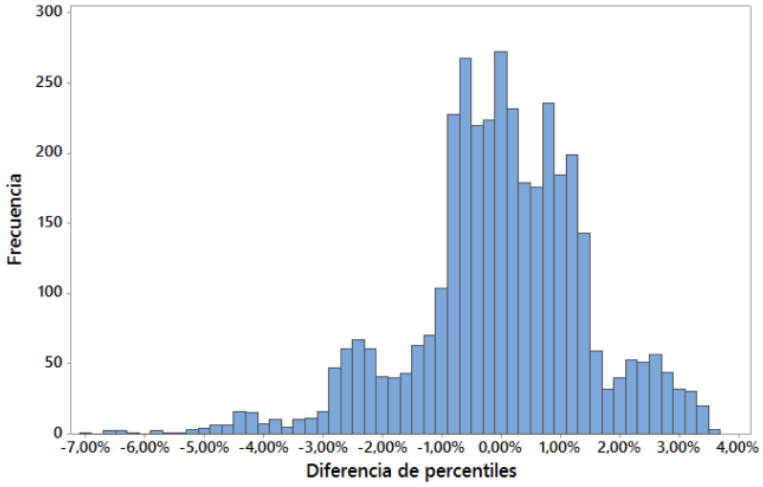


Figura 3. Diferencias de percentil en el examen MIR de 2015 de la muestra de opositores. Percentil alcanzado con las preguntas consideradas como válidas por la comisión calificadora menos el percentil alcanzado sustituyendo las preguntas que se consideran anulables según el algoritmo desarrollado en el presente artículo.



mático, se produce un incremento del área bajo la curva de un 1,24%. La figura 2b muestra la curva de información correspondiente a la prueba MIR de 2015. Se representa en azul la curva de información obtenida con las preguntas anuladas por la comisión calificadora y en rojo la correspondiente a la prueba después de la anulación de las preguntas propuestas por la comisión calificadora junto con las propuestas por la metodología que se desarrolla en el presente artículo. Si bien el incremento obtenido en el área bajo la curva es de un 1,24% y puede parecer pequeño, resulta de importancia a la hora de realizar la clasificación de los examinados, pues el cambio de un reducido número de preguntas puede llegar a ser determinante a la hora de poder elegir o no la especialidad deseada.

La figura 3 representa las diferencias de percentil existentes en el examen MIR de 2015 en la muestra de opositores de las que se dispone de sus contestaciones a las preguntas del examen. El cálculo se ha realizado restando al percentil de preguntas netas alcanzado con las preguntas consideradas como válidas por la comisión calificadora, el que obtendría cada médico si se sustituyen las preguntas que se consideran anulables, según el algoritmo desarrollado en el presente artículo, por preguntas de reserva. Así, tal y como se puede observar en la figura 3, habría examinados que llegarían a disminuir su percentil de número de orden en un 7%, mientras que otros lo aumentarían hasta un 3,6%. En valores absolutos, resulta que el cambio de unas preguntas por otras supondría en un 22,54% de los examinados una modificación del percentil de entre una y dos unidades, mientras que en el 18,51% de los opositores el cambio de percentil debido a esto superaría las dos unidades.

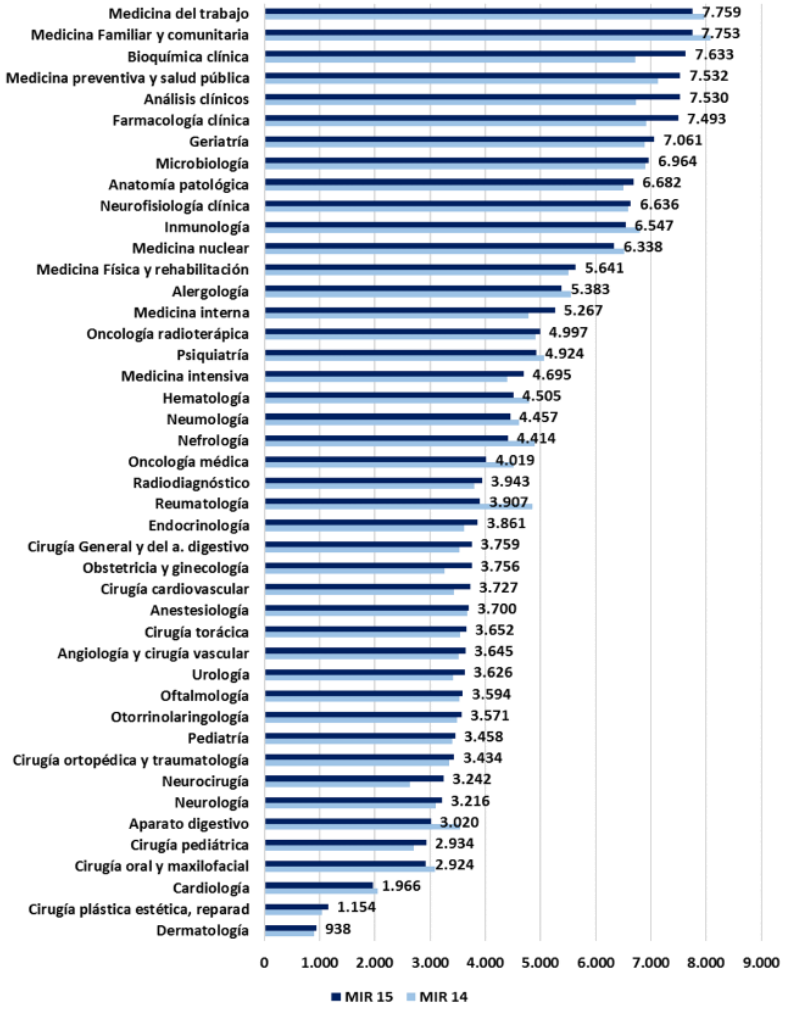
Los resultados de las pruebas MIR de los últimos años han evidenciado que en los números de orden comprendidos entre el 3.000 y el 4.000, una diferencia de una pregunta acertada representa entre 100 y 125 números de orden, y esto, tal y como ya se ha indicado con anterioridad, puede suponer la diferencia entre poder escoger o no cierta especialidad.

La figura 4 muestra los últimos números con los que se agotaron las plazas de las diferentes especialidades en la prueba MIR de 2015. Del total de 44 especialidades, el agotamiento de 17 de ellas se produjo con números de orden comprendidos entre el 3.000 y el 4.000. Aunque el último número que escoge cada especialidad varía de una convocatoria a otra, hay una tendencia a que se mantenga en un entorno de valores similares, si el número de plazas ofertadas de esa especialidad no se modifica notablemente. En el caso concreto de esta prueba MIR de 2015, a igualdad de baremo académico, la diferencia de preguntas netas entre el médico en el puesto 3.000 y el que obtuvo el puesto 4.000 fue de menos de nueve preguntas netas; téngase en cuenta que una neta equivale a una pregunta acertada, mientras que cada pregunta fallada resta un tercio de neta. Con un baremo de 1,75 (baremo mediano de la población de presentados al MIR) se necesita-

ron 138,67 respuestas netas para alcanzar el número de orden 3.000, 137,67 respuestas netas para el 3.100, 136,67 respuestas netas para el 3.200, 136 respuestas netas para el 3.300, 135 respuestas netas para el 3.400, 134 respuestas netas para el 3.500, 133,33 respuestas netas para el 3.600, 132,33 respuestas netas para el 3.700, 131,67 respuestas netas para el 3.800, 130,67 respuestas netas para el 3.900 y 129,67 respuestas netas para el 4.000. Si se tiene en cuenta que anular una pregunta que un examinado ha fallado por otra de reserva que ha acertado supone modificar su puntuación en 1,33 preguntas netas (suma de 0,33 por la que ha dejado de fallar y de una pregunta neta adicional por la pregunta válida de reserva que la sustituye), se comprende la importancia de cada una de las preguntas que se anulen dado que pueden suponer un cambio en el número de orden del opositor que afecte a sus probabilidades de poder escoger o no su especialidad objetivo. Sirva como ejemplo el agotamiento en la convocatoria 2015 de ocho especialidades en un intervalo de puntuaciones con tan sólo 1,33 respuestas netas de diferencia (los cálculos están realizados para un baremo de 1,75): Obstetricia y ginecología (número 3.756, 132 respuestas netas), Cirugía cardiovascular (número 3.727, 132,33 respuestas netas), Anestesiología (número 3.700, 132,33 respuestas netas), Cirugía torácica (número 3.652, 132,67 respuestas netas), Angiología y cirugía vascular (número 3.645, 133 respuestas netas), Urología (número 3.626, 133 respuestas netas), Oftalmología (número 3.594, 133,33 respuestas netas) y Otorrinolaringología (número 3.571, 133,33 respuestas netas). Con estas referencias, queda constatado que cada pregunta neta puede ser clave a la hora de alcanzar o no la especialidad objetivo.

El razonamiento efectuado, así como los números de orden de la figura 4, son una referencia para los examinados no afectados por el cupo de extranjeros porque, en caso de estar afectados por dicho cupo, el agotamiento de la mayoría de las especialidades coincidirá con el número de orden de agotamiento del cupo de extranjeros. En el caso del MIR de 2015 se agotó en el número de orden 4.547 para el 4% de las plazas ofertadas, es decir, 244 plazas.

Figura 4. Últimos números con los que se escogieron plazas de las diferentes especialidades en las pruebas MIR de 2015. Nota: en los cálculos se excluyen los centros privados que requieren conformidad previa a la elección y las últimas plazas asignadas por el turno de discapacidad igual o superior al 33%.



Discusión

Causas de anulación de las preguntas

Los motivos por los que una pregunta del examen MIR puede ser anulada se pueden agrupar, de manera simplificada, en tres apartados:

- *Defectos en 'qué' es preguntado:* no se cumple que una y sólo una de las opciones de respuesta sea correcta.
- *Defectos en 'cómo' está redactada la pregunta:* por ejemplo, defectos tipográficos en el cuadernillo de preguntas que puedan llevar a que una opción correcta deje de serlo, o defectos de impresión en el cuadernillo de imágenes que impidan su interpretación correcta.
- *Defectos en 'cuánto' cumple la pregunta su objetivo dentro del examen:* no es otro que conseguir discriminar entre los distintos niveles de conocimiento de los médicos examinados.

Anulaciones por defectos en el contenido médico de la pregunta

La comisión calificadora no comunica públicamente los criterios seguidos para anular las preguntas, por lo que sólo se puede realizar un análisis de los probables motivos de anulación por criterio experto *a posteriori* de las anulaciones. Después de analizar las 30 preguntas de las convocatorias MIR 2010 a 2015 anuladas por la comisión calificadora entre la publicación de las plantillas de respuesta provisionales y definitivas, los autores estiman que la comisión calificadora de la prueba utiliza casi mayoritariamente un criterio de anulación médico (analizando 'qué' se pregunta). Trece de las preguntas anuladas (el 43% del total) probablemente se anularon porque más de una de las opciones de respuesta podían ser válidas, 13 preguntas (43%) probablemente porque ninguna de las opciones de respuesta podía considerarse completamente válida, una pregunta (3%) por error tipográfico en su redacción en el cuadernillo de preguntas y cuatro preguntas (13%) por otros criterios.

Anulaciones por defectos en la redacción de la pregunta

Un trabajo publicado recientemente [16] se ha centrado en el estudio de la calidad técnica de la redacción de las preguntas de los exámenes MIR de las convocatorias entre 2009 y 2013 (el cómo se pregunta). Los defectos técnicos en la redacción deben ser conocidos por las personas responsables de la elaboración y revisión de las preguntas del examen para que la dificultad asociada sea únicamente derivada de 'qué' se pregunta (el concepto médico) y no de 'cómo' se pregunta (dificultades irrelevantes o defectos de redacción que premien más el conocimiento de las técnicas de test que el conocimiento médico que se pretende evaluar). En nuestra opinión, no sólo las preguntas con errores tipográficos deberían de ser consideradas por la comisión calificadora para su posible anulación, sino también las que tienen dificultades irrelevantes, asociadas, por ejemplo, a una redacción ambigua que dificulte la interpretación de la pregunta o de las opciones de respuesta, o que posibiliten contestar con el conocimiento de 'técnicas de test', que no es el que se pretende medir con el examen.

Anulaciones por deficiente calidad psicométrica de la pregunta

En el presente artículo se propone una metodología psicométrica de ayuda a la detección de preguntas candidatas a su anulación en la prueba MIR (tercer criterio de anulación de preguntas o evaluación de la capacidad de la pregunta de conseguir su objetivo de discriminar dentro del examen al que sabe del que no sabe). Las preguntas anulables según este criterio se caracterizan por no ser discriminativas en absoluto, o por presentar curvas de probabilidad en las que la posibilidad de acertar la respuesta correcta disminuye a medida que aumenta el nivel de conocimiento del individuo examinado. Si bien estos criterios no pueden ni deben ser los únicos que se tengan en cuenta a la hora de anular preguntas, desde el punto de vista de los autores se considera que pueden ser de gran utilidad como criterios de apoyo al trabajo que realiza la comisión calificadora de la prueba.

Para hacer 'el examen del examen MIR', en su evaluación de calidad métrica, hay que estudiar el comportamiento de los examinados en cada pregunta. No sólo hay que estudiar globalmente el examen, sino también cada pregunta en particular, y la calidad métrica del examen MIR es el resultado de la calidad métrica del conjunto de sus preguntas no anuladas. Eso justificaría que las preguntas se anularan por su falta de capacidad para medir lo que han de medir (los distintos niveles de conocimiento médico), y para eso la comisión calificadora de la prueba debería disponer de un estudio psicométrico de todas y cada una de las preguntas, una vez realizado el examen, para poder tomar decisiones al respecto de la anulabilidad o no de cada una de ellas en la sesión en la que se aprueba la plantilla de respuestas definitivas del examen.

En el supuesto de tener que anular un número de preguntas superior al disponible de reserva, entendemos que se podría acomodar ligeramente la escala final del examen. No hay nada escrito sobre ello en las normas que rigen las pruebas selectivas y, por lo tanto, entendemos que no existe un factor administrativo limitante que obligue a que el número final de preguntas válidas sea 225.

La lógica muestra que la tendencia de los examinados es a reclamar sólo preguntas que han fallado, de las que han encontrado bibliografía con argumentos que apoyen que la respuesta considerada como válida no lo es, o bien que existe más de una opción de respuesta válida. Por ello, probablemente, las preguntas que concentran un mayor número de reclamaciones son las preguntas con una dificultad alta, donde la mayoría de los examinados han errado en la elección de la respuesta considerada como válida en la plantilla de corrección. El volumen de reclamaciones recibidas sobre una pregunta puede no estar en consonancia con la calidad de

los argumentos esgrimidos para la anulación de dicha pregunta.

Las respuestas netas de examen cambian con la entrada en acción de las anulaciones y las preguntas de reserva por taxativo orden de éstas. La supuesta mejora global de las puntuaciones netas del conjunto de todos los examinados con la entrada en acción de las preguntas de reserva es sólo global y no necesariamente individual, ya que, si bien casi todas las anuladas son preguntas difíciles que han errado la mayor parte de los examinados, no siempre es así, y se da el caso de preguntas anuladas que no deberían haberlo sido por su calidad métrica a pesar de haberlo sido por otros criterios. Dependiendo de las distintas combinaciones individuales de preguntas válidas, erróneas y no contestadas entre los pares de pregunta anulada y pregunta de reserva activada en la plantilla, cada anulación puede suponer, a nivel individual, un cambio en la puntuación neta de examen que oscile entre −1,33 respuestas netas (si tiene acertada la pregunta anulada y errónea la de reserva que la sustituye) y +1,33 respuestas netas (si tiene fallada la pregunta anulada y acertada la pregunta de reserva que la sustituye).

Tratar con rigor métrico las posibles preguntas que se van a anular es un asunto obligado para atender a los principios de igualdad, mérito y capacidad que rigen en las pruebas de selección a cargo de la Administración pública, y ésta debe utilizar los medios técnicos a su alcance para conseguir la aplicación de esos principios.

Dado que las sucesivas curvas de la función información de las diferentes pruebas MIR muestran claramente cómo esta prueba resulta menos discriminativa para los examinados de conocimientos más altos, cobra gran relevancia poder anular todas las preguntas que no sean adecuadas desde el punto de vista psicométrico.

Limitaciones del estudio

En relación con las limitaciones del trabajo, debe tenerse en cuenta que no se han podido analizar los resultados obtenidos por todos los examinandos de cada una de las convocatorias MIR consideradas, sino que se ha tenido acceso para cada año estudiado a una muestra de la población formada por los médicos examinados que decidieron introducir sus respuestas en la página web habilitada a tal efecto. Esta información presenta un 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. Debido a la existencia de ese sesgo, no podemos excluir que la discriminación del examen y de sus preguntas mejorase si, en lugar de una muestra de 13.984 examinados, hubiéramos dispuesto del análisis de las respuestas de todos los presentados al MIR en las convocatorias estudiadas.

Otra de las debilidades que se le pueden achacar a la metodología propuesta es que, tal y como se ha podido comprobar en los exámenes MIR analizados, en algunas ocasiones no bastaría con disponer de 10 preguntas adicionales como reserva, pues para cumplir los criterios de anulabilidad que se proponen podría ser necesario anular un número ligeramente mayor de preguntas. Así, en el examen MIR de 2013, la comisión calificadora propuso la anulación de ocho preguntas y, según la metodología expuesta, sería necesaria también la anulación de otras siete preguntas adicionales. Sin embargo, desde el punto de vista de los autores, esto simplemente supondría que el examen MIR pasase a tener un número menor de preguntas válidas y, dado que la misión de la prueba es la de clasificar a los médicos examinados de cierta convocatoria, sin tener relación alguna con los de convocatorias previas o posteriores, no es tan importante que el número de preguntas empleadas para el cálculo de las puntuaciones de los examinandos sea siempre 225, como que entre ellas no se incluyan preguntas que no discriminen y cuya probabilidad de acertar o fallar sea casi aleatoria, y no relacionada con el nivel de conocimiento médico del aspirante.

Así, si se aplicase esta metodología en el caso del MIR de 2013, resultaría que finalmente habría 220 preguntas en la plantilla de respuestas definitiva, pero se conseguiría incrementar el área bajo la curva de información en un 3,42% con respecto a la curva que resulta teniendo en cuenta sólo las preguntas anuladas por la comisión calificadora. Por tanto, desde el punto de vista de los autores, sería posible mejorar la calidad psicométrica de las preguntas constitutivas de la plantilla definitiva de cada prueba MIR, así como la eficiencia relativa si se aplicara la metodología expuesta en el presente artículo.

Utilidad como apoyo a la comisión calificadora de la prueba MIR

Dado que es necesario disponer de las contestaciones a todas las preguntas tanto para el cálculo del coeficiente biserial puntual como para el cálculo del coeficiente de discriminación del modelo 2PL, la detección de las preguntas que se deben anular aplicando esta metodología no se podría realizar el mismo día de la prueba (en la primera reunión de la comisión calificadora el día del examen), sino que

se tendría que hacer con posterioridad, cuando se dispusiera de las contestaciones de los candidatos a las preguntas.

También podría considerarse la posibilidad, en el caso de que no fuera necesario el uso de todas las preguntas de reserva, de que todas ellas fueran empleadas eliminando las preguntas de peores valores de discriminación y correlación biserial puntual, consiguiendo así incrementar todavía más la eficiencia relativa. Es decir, subir el punto de corte de la discriminación, en cada examen, hasta poder utilizar la totalidad de preguntas de reserva que tuvieran una calidad psicométrica suficiente.

Comparación con otros países

Aunque la propuesta de una metodología de anulación automatizada de preguntas pueda ser considerada por algunos como un cambio de paradigma, nos gustaría señalar que en la prueba Examen Único Nacional de Conocimientos de Medicina (EUNACOM) de Chile [17] se realiza un proceso automático de anulación de preguntas. Dicho proceso consiste en el cálculo del comportamiento de cada pregunta para el total de examinados, y se determina el porcentaje que respondió correctamente cada pregunta, así como la capacidad de discriminación de cada pregunta a través del coeficiente de correlación biserial puntual. Con anterioridad a la realización del examen, el consejo estadístico del EUNACOM fija los valores de corte de ambos parámetros (porcentaje de examinados que respondieron correctamente a la pregunta y capacidad de discriminación). Así, para la versión de 2012, el consejo estadístico determinó considerar preguntas válidas las que tenían una proporción de aciertos por parte de los examinados superior al 20% y con un coeficiente de correlación biserial puntual superior a 0,15 (considerablemente más exigente que el punto de corte propuesto en ese artículo). En el caso de las preguntas con un coeficiente de correlación biserial puntual superior a 0,15, pero con una proporción de aciertos inferior al 20% de los alumnos, se realiza una revisión por parte del comité de contenidos, quien será el que decida si finalmente la pregunta se anula o no. Por tanto, el número final de preguntas válidas del examen varía de un año a otro. En el examen chileno, a cada alumno, además de la puntuación obtenida, se le proporciona el percentil alcanzado, lo que hace posible la comparación de resultados entre distintas convocatorias. Esto es importante, ya que EUNACOM es un examen con un objetivo acreditador de los conocimientos médicos de los aspirantes, que no busca, como el MIR, generar una lista de candidatos para realizar una elección ordenada y secuencial de las plazas ofertadas de formación sanitaria especializada.

Nótese que, si en la prueba MIR se optase por eliminar las preguntas con una correlación biserial puntual por debajo de 0,15, se anularía un número de preguntas muy superior al de reserva disponible. En los años estudiados, 182 de las 1.410 preguntas (12,9%) tienen un valor de correlación biserial puntual inferior a 0,15. Así, en el caso de las convocatorias consideradas, el número de preguntas MIR que se deberían anular según el criterio seguido en EUNACOM sería, en 2010, de 33; en 2011, de 29; en 2012, de 27; en 2013, de 27; en 2014, de 30; y en 2015, de 36, y además hay que tener en cuenta que podrían producirse anulaciones de más preguntas debido a otros criterios.

En el caso de sumar el criterio sobre la dificultad de las preguntas, 99 de las 1.410 preguntas (7%) de las convocatorias analizadas no alcanzarían el punto de corte de dificultad utilizado por el EUNACOM al haber sido acertadas por menos del 20% de los examinados incluidos en la muestraa. El MIR contiene, por lo tanto, algunas preguntas de dificultad muy elevada, que no se permiten en el EUNACOM.

Comparación con otros estudios del Ministerio de Sanidad

En este trabajo se ha empleado como uno de los dos criterios básicos para conocer la capacidad discriminativa el índice de correlación biserial puntual. En un trabajo previo publicado por el Ministerio de Sanidad y Consumo [8] se utilizaba para tal propósito el índice de discriminación. Dicho índice, en su formulación, no hace uso de toda la información disponible, sino que utiliza únicamente los datos relativos al número de respuestas correctas de los grupos fuerte y débil de examinados. Se define como grupo fuerte el formado por el 27% de los examinados que obtuvieron las mejores puntuaciones, y el grupo débil es el formado por el mismo porcentaje de examinados que obtuvieron las peores puntuaciones. Desde el punto de vista de los autores, y aunque en un trabajo previo [4] usamos ambos a efectos comparativos con los estudios publicados por el Ministerio de Sanidad, en este trabajo se ha preferido usar el coeficiente de correlación biserial puntual por tres motivos: por una parte, porque los resultados obtenidos recogen más fielmente los del conjunto de la población (y no únicamente del 54% con resultados extremos); por el elevado grado de correlación entre la correlación biserial puntual y el coeficiente de discriminación

del modelo de dos parámetros (R^2 de 0,82 en el MIR de 2015, 0,76 en el MIR de 2014, 0,78 en el MIR de 2013, 0,83 en el MIR de 2012, 0,77 en el MIR de 2011 y 0,79 en el MIR de 2010); y, finalmente, porque este índice se utiliza en el análisis de una prueba similar al MIR, como es el EUNACOM.

En definitiva, como ya concluimos en los trabajos anteriores [4,5], el examen MIR es un examen objetivo, estructuralmente válido, y con dificultad y discriminación adecuadas. En el presente análisis se propone la utilización de una metodología de apoyo a las comisiones calificadoras de las pruebas de acceso a la formación sanitaria especializada que permitiese, con la anulación adicional de un pequeño número de preguntas con baja calidad psicométrica, mejorar la discriminación en ellas. Si, además de las preguntas anuladas por la comisión calificadora, se anulasen las preguntas con discriminación negativa (medida por los criterios propuestos), los valores promedios de correlación biserial puntual hubieran pasado en el caso del estudio realizado a 0,3151, con una mejora adicional de la discriminación del 2,08%, y los valores promedios de coeficiente de discriminación en el modelo 2PL hubieran pasado a 0,802, lo que supone una mejora adicional de la discriminación del 2,13%. Pequeñas mejoras, pero importantes en una prueba con la trascendencia del examen MIR en el futuro de los jóvenes graduados en Medicina, porque se traducen en cambios en el número de orden con el que se eligen las plazas y la posibilidad de acceder o no a la especialidad deseada, con una mayor correlación con la capacidad de la prueba de discriminar entre los distintos niveles de conocimientos médicos de los examinados.

Bibliografía

1. Real Decreto 639/2014, de 25 de julio, por el que se regula la troncalidad, la reespecialización troncal y las áreas de capacitación específica, se establecen las normas aplicables a las pruebas anuales de acceso a plazas de formación y otros aspectos del sistema de formación sanitaria especializada en Ciencias de la Salud, y se crean y modifican determinados títulos de especialista. Boletín Oficial del Estado n.º 190, de 6 de agosto de 2014. p. 63130-67.
2. Orden SSI/1892/2015, de 10 de septiembre, por la que se aprueba la oferta de plazas y la convocatoria de pruebas selectivas 2015 para el acceso en el año 2016 a plazas de formación sanitaria especializada para médicos, farmacéuticos, enfermeros y otros graduados/licenciados universitarios del ámbito de la Psicología, la Química, la Biología y la Física. Boletín Oficial del Estado n.º 224, de 18 de septiembre de 2015, p. 82031-318.
3. Ministerio de Sanidad, Servicios Sociales e Igualdad. Formación sanitaria especializada. URL: http://sis.msssi.es/fse/Default.aspx?MenuId=QE-00. [10.12.2016[.
4. 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.
5. Baladrón J, Curbelo J, Sánchez-Lasheras F, Romeo-Ladrero JM, Villacampa T, Fernández-Somoano A. El examen MIR 2015 desde el punto de vista de la teoría de respuesta al ítem. FEM 2017; 20: 29-38.
6. Guilbert JJ. Educational handbook for health personnel (offset publication n.º 35). 1 ed. Geneva: WHO; 1977.
7. Crocker L, Algina, J. Introduction to classical and modern test theory. 1 ed. Orlando, FL: Holt, Rinehart & Winston; 1986.
8. Pruebas selectivas para el acceso a plazas de formación de médicos especialistas (1982-1992). Madrid: Ministerio de Sanidad y Consumo; 1993.
9. 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.
10. Ebel RL, Frisbie DA. Essentials of education measurement. 5 ed. Englewood Cliffs, NJ: Prentice Hall; 1990.
11. Álvarez-Antón JC, García-Nieto PJ, De Cos-Juez FJ, Sánchez-Lasheras F, Blanco-Viejo C, Roqueñí-Gutiérrez N. Battery state-of-charge estimator using the MARS technique. IEEE Transactions on Power Electronics 2013; 28: 3798-805.
12. Guzmán D, De Cos-Juez FJ, Myers R, Guesalaga A, Sánchez-Lasheras F. Modeling a MEMS deformable mirror using non-parametric estimation techniques. Optics Express 2010; 18: 21356-69.
13. 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.
14. 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.
15. Muñiz J. Teoría clásica de los tests. Madrid: Pirámide; 2002.
16. Rodríguez-Díez MC, Alegre M, Díez N, Arbea L, Ferrer M. Technical flaws in multiple-choice questions in the access exam to medical specialties ('examen MIR') in Spain (2009-2013). BMC Med Educ 2016; 16: 47.
17. EUNACOM. Examen Único Nacional de Conocimientos de Medicina. URL: http://www.eunacom.cl/resultados/resultados-actuales.html. [09.12.2016.