

ORIGINAL

Análisis psicométrico de las preguntas asociadas a imágenes radiológicas en el examen para médico interno residente en España



E. Murias Quintana^{a,*}, F. Sánchez Lasheras^b, S.M. Costilla García^c,
M. Cadenas Rodríguez^d, J. Calvo Blanco^e y J. Baladrón Romero^f

^a Doctor en Medicina y Cirugía, Profesor Asociado de Radiología, Universidad de Oviedo. Especialista en Radiología, Hospital Universitario Central de Asturias. Profesor de Radiología, Curso Intensivo MIR Asturias, Oviedo, Asturias, España

^b Doctor en Técnicas de Análisis de Datos e Inteligencia Artificial, Profesor en el Departamento de Construcción e Ingeniería de Fabricación, Universidad de Oviedo, Oviedo, Asturias, España

^c Doctor en Medicina y Cirugía, Profesor Titular de Radiología, Universidad de Oviedo. Jefe de Sección de Radiología, Hospital Universitario Central de Asturias, Oviedo, Asturias, España

^d Graduada en Enfermería, Hospital Universitario Central de Asturias, Oviedo, Asturias, España

^e Doctor en Medicina y Cirugía, Profesor de Radiología, Universidad de Oviedo. Especialista en Radiología, Hospital Universitario Central de Asturias, Oviedo, Asturias, España

^f Doctor en Medicina y Cirugía. Director de la Academia de alto rendimiento Curso MIR Asturias, Oviedo, Asturias, España

Recibido el 13 de diciembre de 2018; aceptado el 23 de abril de 2019

Disponible en Internet el 23 de junio de 2019

PALABRAS CLAVE

MIR;
Psicometría;
Radiología;
Educación;
TEST

Resumen

Antecedentes y objetivo: La psicometría es una técnica sencilla e intuitiva que se utiliza en el campo de la docencia y en el de los cuestionarios de respuesta múltiple. El examen MIR incluye desde el año 2009 preguntas asociadas a imagen radiológica. El objetivo de este trabajo es mostrar los resultados del análisis psicométrico de estas preguntas con objeto de comparar el grado de dificultad, la capacidad de discriminación y la estructura interna respecto al resto de preguntas.

Material y métodos: Se seleccionaron todas las preguntas del examen desde el año 2009 y se clasificaron en casos clínicos sin y con imagen radiológica, casos clínicos con imagen no radiológica, preguntas test y preguntas negativas. Se utilizó la teoría clásica de test y la teoría de respuesta al ítem para valorar la dificultad y el grado de discriminación de las preguntas.

Resultados: Sobre 225 preguntas, los exámenes incluyen entre un 11% y un 15% de preguntas asociadas a imágenes. Las preguntas asociadas a imagen radiológica son más difíciles (grado de dificultad corregida [IDc] 0,51), con una menor capacidad de discriminación. El aumento de la

* Autor para correspondencia.

Correo electrónico: emuriass@hotmail.com (E. Murias Quintana).

La dificultad de la pregunta radiológica se relaciona con una menor discriminación, sobre todo si la información clínica no es adecuada, o es contraria al concepto radiológico, o el concepto nunca ha sido preguntado a lo largo de la historia del MIR.

Conclusiones: Para poder igualar los estándares del examen MIR, es necesario mantener una adecuada estructura en la confección de las preguntas de radiología, en términos de entorno clínico, un adecuado uso de distractores y un menor nivel de dificultad, que se puede lograr mediante el uso de imágenes con hallazgos radiológicos típicos.

© 2019 SERAM. Publicado por Elsevier España, S.L.U. Todos los derechos reservados.

KEYWORDS

MIR;
Psychometrics;
Radiology;
Education;
Test

Psychometric analysis of questions associated with radiological images in the competitive examination for access to residency programs in Spain

Abstract

Background and objective: Psychometrics is a simple, intuitive approach used in educational research and in multiple-choice questionnaires. Since 2009, the competitive examination through which access to residency programs in Spain is determined (MIR) has included questions related to radiological images. The objective of this paper is to show the results of the psychometric analysis of these questions with the aim of comparing their degree of difficulty, discriminative capacity, and internal structure with respect to those of the other questions on the examination.

Material and methods: We analyzed all questions on the examination since 2009, classifying them as clinical cases with and without radiological images, clinical cases with and without non-radiological images, multiple choice questions, and negative questions. We used classical test theory and item response theory to assess the difficulty and degree of discrimination of the questions.

Results: Of 225 questions, between 11% and 15% of the questions included in the examinations were associated with images. The questions associated with radiological images were more difficult (corrected difficulty index, 0.51) and had worse discriminative capacity. The increased difficulty of radiological questions was associated with worse discriminative capacity, especially if the clinical information provided was inadequate or if the clinical information was contrary to the radiological concept or if there had never been any questions about the concept in previous MIR examinations.

Conclusions: To equalize the standards of the MIR examination, it is necessary to maintain an appropriate structure in devising radiology questions, with terms from the clinical context, appropriate use of distracters, and a lower level of difficulty, which could be achieved by using radiological images with typical radiological findings.

© 2019 SERAM. Published by Elsevier España, S.L.U. All rights reserved.

Introducción

La psicometría es una herramienta sencilla e intuitiva que puede ser utilizada en el campo de la docencia. Existen desde aplicaciones de las técnicas psicométricas utilizadas en investigación educativa¹, que se asocian a competencias en los procesos de enseñanza y aprendizaje mediante medios informáticos² y que se incluyen en las nuevas tecnologías que se integran en la formación del profesorado y son citadas desde los años noventa³, hasta trabajos actuales sobre las nuevas tecnologías informáticas como recursos didácticos en el entorno universitario^{4,5}.

En el entorno de los exámenes multirrespuesta, las técnicas psicométricas se han utilizado para la valoración de la calidad de las preguntas, analizar su coherencia interna, identificar su capacidad de discriminación de conocimientos y su dificultad. En los últimos años hemos utilizado y

ajustado estas técnicas psicométricas, basadas en el análisis informático de las respuestas, para implementar la formación de nuestros alumnos durante la preparación del examen para médico interno residente. Nuestro objetivo es conseguir la mayor calidad de los test de preparación para el examen, detectar lagunas formativas o buscar conceptos que debamos fomentar en esta formación postuniversitaria.

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^{6,7}.

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. Dicho examen se compone de 225 preguntas tipo test, más 10 preguntas de reserva, de respuesta múltiple, que versan sobre cualquier

campo de la medicina, y deben contestarse en un máximo de 5 horas. Cada pregunta acertada suma tres puntos y cada pregunta fallada resta uno. La nota obtenida en el examen (el 90% de la nota final), junto con la valoración del baremo o expediente académico (el 10% de la misma en la última convocatoria), permite clasificar en orden decreciente de puntuación total a todos los presentados. 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 MIR⁷⁻¹².

Desde el examen MIR de 2009 se han incluido en el mismo preguntas asociadas a una o varias imágenes (a partir de este momento nombraremos el examen como hace el Ministerio, es decir, por el año en el que se convoca, y no por el año en el que se realiza). Esta imagen puede ser radiológica, considerando radiológica todas las pruebas diagnósticas basadas en la imagen, o no radiológica (resto de imágenes que incluyen las clínicas, histologías, diagramas, espirometrías, electrocardiogramas, etc.)¹³.

En este trabajo mostramos un ejemplo de nuestro trabajo sobre el análisis psicométrico de las preguntas asociadas a imágenes radiológicas, confrontándolas con el resto de las preguntas del examen a médico interno residente, con objeto de comparar su grado de dificultad y capacidad de discriminación y para comprobar su utilidad como instrumento de medida del grado de conocimiento médico de los opositores.

Material y métodos

Se han seleccionado las preguntas correspondientes a las convocatorias del examen MIR desde la del año 2009 hasta la de 2017. Es decir, un total de 2.025 preguntas. Se han escogido estas convocatorias ya que incluyen todos los exámenes con preguntas asociadas a imagen (a principios de los ochenta se utilizaron imágenes de electrocardiogramas en convocatorias aisladas, que dejaron de incorporarse al examen, probablemente por la deficiente calidad de impresión de la época). No hemos incluido preguntas asociadas a conceptos radiológicos si la pregunta no se asociaba a una imagen radiológica.

Para su análisis, las preguntas de elección múltiple se han clasificado en los siguientes subtipos^{13,14}:

1. Casos clínicos *sin* imagen: preguntas tipo test con un enunciado largo, no vinculadas a una imagen, en las que se plantea el diagnóstico diferencial de una enfermedad, un tratamiento, el manejo diagnóstico o terapéutico de un paciente según datos de la anamnesis, exploración clínica, analíticos y de pruebas complementarias. Si incluyen información sobre pruebas de imagen, lo hacen como descripción en el texto de la pregunta.
2. Casos clínicos *con* imagen *no* radiológica: preguntas tipo caso clínico, cuyo enunciado está vinculado a una imagen o diagrama no radiológico.
3. Casos clínicos *con* imagen radiológica: preguntas tipo caso clínico que se asocian a una prueba de imagen.
4. Pregunta negativa: ítems en los que se solicita al opositor que identifique la opción falsa entre las diferentes alternativas de respuesta que ofrece la pregunta.

5. Pregunta de tipo test: resto de preguntas con un enunciado corto, que no se consideran casos clínicos ni preguntas negativas. Normalmente estas preguntas son directas y solicitan reconocer la opción correcta entre las alternativas de respuesta ofrecidas.

La psicometría es la disciplina que se encarga del conjunto de métodos, técnicas y teorías implicadas en medir y cuantificar las variables psicológicas del psiquismo humano. La psicometría engloba la teoría y la construcción de pruebas, test y otros procedimientos de medición válidos y confiables. Incluye la elaboración y aplicación de procedimientos estadísticos que permitan determinar si una prueba o test es válido o no para la medición de una variable psicológica previamente definida.

En su vertiente adaptada a valorar la respuesta de preguntas tipo test, existen varios modelos psicométricos que tienen en común la relación matemática de las características latentes (no observables) de los ítems de una prueba y de las personas que contestan ese ítem, con el fin de obtener modelos de acierto de cada sujeto respecto a cada ítem, en función de su nivel de conocimiento. Para la valoración del examen MIR, hemos utilizado dos modelos matemáticos: la teoría clásica de los test y la teoría de la respuesta al ítem (TRI), variables ya utilizadas y contrastadas en trabajos previos para este examen¹³⁻¹⁸.

Teoría clásica de los test: es una forma de medir la dificultad, discriminación y calidad de las preguntas en función del número de personas que las contestan y de su nivel de conocimiento, expresado por la nota final que obtienen en el conjunto de la prueba test a valorar¹⁹⁻²². Dentro de este modelo se utilizan las siguientes herramientas:

1. Cálculo del grado de *dificultad* (ID): representa el porcentaje de opositores que aciertan o fallan la pregunta. De esta manera se pueden dividir las preguntas como:
 - a) *Fáciles*, aquellas que aciertan al menos el 66% de la muestra.
 - b) *Regulares*, si las aciertan entre el 33% y el 66% de la muestra.
 - c) *Difíciles*, aquellas que aciertan menos del 33% de la muestra.
2. Cálculo del grado de *dificultad corregida* (IDc): representa la corrección del porcentaje de alumnos que aciertan una pregunta por el factor derivado del acierto por azar. De esta manera, en un test de tres respuestas falsas y una verdadera (en el que cada acierto suma un punto) existe un 25% de posibilidades de contestar correctamente por azar. Debido a que el error en la respuesta otorga un valor de $-0,33$ puntos, una pregunta puede variar desde valores de $-0,33$ hasta 1. Este valor es una forma más precisa para estudiar la dificultad de una pregunta con penalización por fallar, como las del examen MIR. En el presente trabajo clasificamos a las preguntas en la siguiente escala:
 - a) Muy difíciles, aquellas con valores desde $-0,33$ hasta 0.
 - b) Difíciles, entre 0 y 0,33.
 - c) Óptimas, entre 0,33 y 0,66.
 - d) Fáciles, entre 0,66 y 0,80.
 - e) Muy fáciles, por encima de 0,80, hasta 1.
3. Cálculo del grado de *discriminación*: es la correlación que existe entre la puntuación de los opositores en el total de

la prueba y su puntuación en una determinada pregunta de test. Como el examen MIR es una prueba única, se considera que aquellos opositores con una mayor puntuación en el examen tienen mayor nivel de conocimientos médicos generales y aquellos con una puntuación más baja tienen un menor nivel de dichos conocimientos. Para el cálculo de este valor se utiliza en el presente trabajo el *índice de correlación biserial puntual* (r_{pbis}), cuyo objetivo es medir la calidad discriminativa de las preguntas. A mayor valor en el r_{pbis} , mayor será la relación entre obtener una puntuación alta en el test y el hecho de haber contestado correctamente esa pregunta en concreto. Este índice permite clasificar la discriminación de las preguntas como:

- a) Excelentes: si el valor obtenido del coeficiente de correlación biserial puntual es mayor o igual que 0,40.
- b) Buenas: si el valor obtenido es mayor o igual que 0,30 y menor que 0,40.
- c) Regulares: si el valor obtenido es mayor o igual que 0,20 y menor que 0,30.
- d) Pobres: si el valor obtenido es mayor o igual que 0 y menor que 0,20.
- e) Pésimas: en los casos en los que el coeficiente de correlación biserial puntual es negativo.

Teoría de respuesta al ítem (TRI): esta teoría psicométrica se utiliza para predecir cómo contestarían los opositores a las preguntas según su nivel de conocimientos. Para ello, se proponen unos modelos de probabilidad que estiman con qué probabilidad un individuo es capaz de responder de manera correcta a cierta pregunta. En el presente trabajo se ha utilizado un modelo conocido como logístico de dos parámetros (2 PL), para establecer la relación entre la probabilidad de acertar la pregunta de los examinados y su nivel de conocimiento médico demostrado en la totalidad de la prueba MIR. De acuerdo con el modelo propuesto, la probabilidad de obtener una respuesta correcta depende, por una parte, de los parámetros de cada pregunta (dificultad y discriminación), y por otra, del nivel de conocimiento del sujeto. Existen dos valores a considerar en este modelo²³⁻²⁸:

1. **Dificultad TRI:** representa el índice de dificultad de una pregunta corregido por el nivel de conocimiento del opositor, cuyo valor se encuentra normalmente entre -4 a +4. Un valor más alto nos indica que la pregunta es más difícil y un valor más bajo nos indica que la pregunta es más fácil.
2. **Discriminación TRI (DC-R):** representa la capacidad de discriminación de una pregunta corregido por el nivel de conocimiento de los opositores. Siguiendo la escala propuesta por los autores en un trabajo anterior¹⁵, se pueden clasificar las preguntas de forma similar a la utilizada para los valores del índice de correlación biserial puntual:
 - a) **Excelente:** si el valor del coeficiente de discriminación es mayor que 1.
 - b) **Buena:** si el valor del coeficiente de discriminación es mayor o igual que 0,70 y menor o igual que 1.
 - c) **Regular:** si el valor del coeficiente de discriminación es mayor o igual que 0,40 y menor que 0,70.

- d) **Pobre:** si el valor del coeficiente de discriminación es mayor o igual que 0 y menor que 0,40.
- e) **Pésima:** si el valor del coeficiente de discriminación es negativo.

Mediante las variables de la teoría de respuesta al ítem, se puede generar la curva de probabilidad de contestar una determinada pregunta de test en función del conocimiento del opositor, lo que nos muestra no solamente el grado de discriminación de la pregunta, sino a qué nivel de conocimiento se produce la máxima discriminación.

Estas variables psicométricas se han utilizado para valorar todos los tipos de preguntas en las cuales hemos clasificado los exámenes MIR. Las preguntas asociadas a imagen radiológica se comparan con el resto de tipo de preguntas del examen, en relación con las variables psicométricas descritas.

Resultados

El examen MIR suele estar compuesto, desde el año 2009, aproximadamente por un 45% de preguntas clasificadas como casos clínicos, un 15% de preguntas negativas, un 25% de preguntas test y alrededor de un 15% de preguntas asociadas a imagen. La mitad de las preguntas asociadas a imagen se corresponden con preguntas radiológicas y el resto son preguntas asociadas a imagen *no* radiológica (imágenes de anatomía patológica, fotografías clínicas, espirometrías, electrocardiogramas, etc.). Las preguntas con imagen radiológica oscilan entre 11 y 22; el examen con mayor número de preguntas vinculadas a imágenes radiológicas fue el examen MIR de 2010 (9,78% del total de preguntas) y el examen con menor número de preguntas vinculadas a imágenes radiológicas fue el de la convocatoria de 2016 (4,89% del total de preguntas) (fig. 1).

Las preguntas con imagen radiológica son las que tienen en el examen MIR un valor promedio de dificultad mayor, con una media del 64% de probabilidad de acierto (fig. 2). Considerando un valor óptimo de dificultad entre el 33% y el 66%, el examen más difícil para este tipo de preguntas fue el de la convocatoria de 2014, con un promedio del 39,8% de aciertos, y los exámenes más fáciles fueron los de las convocatorias de 2012 y 2017, superando el 71% de aciertos. Respecto a las preguntas con imagen *no* radiológica, la convocatoria con mayor promedio de dificultad es la de 2015, y la de menor dificultad la de 2011, siendo más difíciles las preguntas con imagen radiológica en todas las convocatorias, salvo en 2010, 2012 y 2015.

Con respecto al promedio de los valores de índice de dificultad con corrección del azar, las preguntas con imagen radiológica son, en promedio, las más difíciles del examen MIR, con una media de 0,51 (fig. 2). Considerando un valor óptimo de dificultad entre 0,33 y 0,66, el examen más difícil para este tipo de preguntas fue el de la convocatoria de 2014 con un valor promedio de 0,23, y los exámenes más fáciles fueron los de las convocatorias de 2012 y 2017, superando el valor promedio de 0,62. Respecto a las preguntas con imagen *no* radiológica, la convocatoria con mayor dificultad es la de 2015 y la de menor dificultad la de 2011, siendo más difíciles las preguntas con imagen radiológica en todas las convocatorias, salvo en 2010, 2012 y 2015. En las

Convocatoria	Casos clínicos		Casos clínicos con imagen						Preguntas negativas		Preguntas tipo test		Total
	N.º	(%)	N.º no radiológica	(%)	N.º radiológica	(%)	Subtotal	(%)	N.º	(%)	N.º	(%)	N.º
2009	92	40,89	11	4,89	14	6,22	25	11,11	35	15,56	73	32,44	225
2010	112	49,78	8	3,56	22	9,78	30	13,33	22	9,78	61	27,11	225
2011	102	45,33	11	4,89	18	8,00	29	12,89	27	12,00	67	29,78	225
2012	109	48,44	14	6,22	20	8,89	34	15,11	32	14,22	50	22,22	225
2013	111	49,33	18	8,00	16	7,11	34	15,11	29	12,89	51	22,67	225
2014	105	46,67	20	8,89	14	6,22	34	15,11	28	12,44	58	25,78	225
2015	115	51,11	12	5,33	18	8,00	30	13,33	26	11,56	54	24,00	225
2016	104	46,22	21	9,33	11	4,89	32	14,22	37	16,44	52	23,11	225
2017	94	41,78	16	7,11	17	7,56	33	14,67	42	18,67	56	24,89	225
Total	944	46,62	131	6,47	150	7,41	281	13,88	278	13,73	522	25,78	2.025

Figura 1 Proporción de casos clínicos, casos clínicos con imagen (radiológica y no radiológica), preguntas negativas y preguntas de test de la prueba MIR en las convocatorias comprendidas entre 2009 y 2017. Los colores indican la distribución de la proporción de preguntas, en rojo los años con menor volumen, en amarillo con un volumen intermedio y en verde los años con mayor volumen de preguntas.

	Casos clínicos		Imagen no radiológica		Imagen radiológica		Preguntas negativas		Preguntas de test		Total	
	Media	Desviación estándar	Media	Desviación estándar	Media	Desviación estándar	Media	Desviación estándar	Media	Desviación estándar	Media	Desviación estándar
Dificultad	0,68	0,2	0,68	0,2	0,64	0,23	0,65	0,21	0,67	0,22	0,67	0,21
Dificultad corregida	0,58	0,27	0,55	0,27	0,51	0,3	0,53	0,27	0,55	0,29	0,55	0,28
Discriminación	0,32	0,16	0,33	0,15	0,29	0,16	0,32	0,17	0,34	0,18	0,32	0,17
Coefficiente de correlación punto biserial	0,31	0,13	0,28	0,11	0,25	0,12	0,29	0,11	0,3	0,12	0,3	0,12
Dificultad TRI	-0,81	7,55	-0,46	3,64	-0,23	4,64	-0,51	0,72	-0,63	9,01	-0,66	7,27
Discriminación TRI	0,79	0,45	0,69	0,37	0,58	0,33	0,58	0,33	0,81	0,43	0,76	0,43

Figura 2 Valor medio y desviación estándar para las variables dificultad, índice de dificultad con corrección de los efectos del azar, índice de discriminación, coeficiente de correlación punto biserial, índice de dificultad según la teoría de respuesta al ítem (TRI), e índice de discriminación según la teoría de respuesta al ítem de las preguntas de los ejercicios de examen de las pruebas MIR entre 2009 y 2017. La escala de colores muestra los valores de las variables, en rojo los valores más bajos, en amarillo los intermedios y en verde los más altos. Nótese que las preguntas asociadas a imagen son las que presentan mayor dificultad y menos capacidad de discriminación.

figuras 3 y 4 se puede observar la comparación del promedio de índice de dificultad corregido de las preguntas con imagen radiológica y las preguntas con imagen no radiológica a lo largo de la serie temporal analizada.

El r_{pbis} es una medida de la discriminación de las preguntas, que permite clasificar como preguntas buenas o excelentes aquellas que tienen un valor superior 0,30. En el examen MIR, los valores promedio de este índice para las preguntas de tipo test, las preguntas negativas y los casos clínicos se aproximan al 0,30 (buena discriminación) (fig. 2), mientras que los valores promedio para las preguntas con imagen radiológica son de 0,25 (discriminación regular). En las convocatorias de los años 2011 y 2012, el promedio de los valores del índice superó el 0,30, coincidiendo con años de preguntas más fáciles. Las preguntas asociadas a imagen no radiológica tienen valores promedio de este índice de discriminación mayores que los de las preguntas asociadas a imagen radiológica, salvo en las convocatorias de 2010, 2012 y 2015 (figs. 3 y 4).

Respecto a la capacidad de discriminación de las preguntas utilizando la métrica de la teoría de respuesta al

ítem, consideramos una pregunta como buena si supera el valor de 0,70 y excelente si supera el 1. El examen MIR presenta una buena capacidad de discriminación, superando en promedio el 0,70, siendo las preguntas test y los casos clínicos las preguntas con mayor valor promedio de discriminación TRI. Las preguntas con imagen radiológica y las preguntas negativas son las que tienen peor valor promedio de discriminación TRI, 0,58 (fig. 2). En la evolución temporal, los años con valores promedio de discriminación TRI mayores para las preguntas vinculadas a imagen radiológica fueron las convocatorias de 2011 y 2012, que superaron el 0,74. Las preguntas asociadas a imagen no radiológica tienen valores promedio de discriminación TRI mayores que los de las preguntas asociadas a imagen radiológica, salvo en las convocatorias de 2010, 2012 y 2015 (figs. 3 y 4).

En el análisis estadístico de la frecuencia de las opciones de respuesta múltiple en las preguntas asociadas a imágenes radiológicas se observó una significativa predilección por las respuestas dos y tres de las opciones mostradas, ya que se encontró que la respuesta 3 era correcta casi en el 50% de las preguntas.

Convocatoria	índice de dificultad		índice de dificultad corregida		índice de discriminación		índice de correlación punto biserial		dificultad TRI		discriminación TRI	
	Imagen no radiológica	Imagen radiológica	Imagen no radiológica	Imagen radiológica	Imagen no radiológica	Imagen radiológica	Imagen no radiológica	Imagen radiológica	Imagen no radiológica	Imagen radiológica	Imagen no radiológica	Imagen radiológica
2009	0,65	0,56	0,51	0,41	0,30	0,23	0,24	0,20	-0,85	1,31	0,47	0,39
2010	0,59	0,70	0,41	0,57	0,33	0,37	0,24	0,29	0,09	-1,18	0,51	0,64
2011	0,78	0,70	0,65	0,58	0,46	0,31	0,40	0,30	-1,09	-0,40	1,11	0,75
2012	0,71	0,74	0,59	0,63	0,28	0,36	0,27	0,33	-1,75	-1,24	0,65	0,75
2013	0,63	0,59	0,46	0,44	0,45	0,34	0,33	0,28	0,12	0,27	0,85	0,69
2014	0,70	0,40	0,58	0,23	0,34	0,16	0,29	0,15	-0,87	4,09	0,68	0,32
2015	0,57	0,62	0,37	0,47	0,26	0,26	0,18	0,20	3,18	-0,02	0,38	0,46
2016	0,69	0,64	0,57	0,50	0,28	0,29	0,27	0,25	-0,42	-0,82	0,68	0,56
2017	0,76	0,72	0,68	0,62	0,29	0,22	0,30	0,23	-1,86	-2,80	0,79	0,57

Figura 3 Evolución del índice de dificultad, índice de dificultad corregida, índice de discriminación, índice de correlación punto biserial, dificultad y discriminación según la teoría de respuesta al ítem (TRI) de las preguntas con imagen, radiológicas y no radiológicas a lo largo de las convocatorias. La escala de colores indica los años con los valores más bajos, en rojo y los años con los valores más altos.

Discusión

La psicometría es una técnica objetiva y con una amplia base científica que ha demostrado de forma sistemática su valor para el análisis de los métodos de evaluación aplicados en un grupo de sujetos sometidos a una prueba test^{17,21,23}. De una forma simplificada, los exámenes de respuesta múltiple, como el examen MIR, se pueden analizar mediante la "teoría clásica de test"²⁰ y la "teoría de respuesta al ítem (TRI)"¹⁵, que permite conocer la probabilidad de acertar o fallar una pregunta en función de los conocimientos del sujeto.

En un examen de respuesta múltiple, el conocimiento de un alumno viene representado por su puntuación total en la prueba. Por lo tanto, la principal limitación de estas técnicas, en un examen como el MIR, viene representado por ser una única medición, en un único momento, de los conocimientos del sujeto, lo que puede generar errores de validez externa^{15,20}. Idealmente, sería necesario valorar a este sujeto en múltiples pruebas a lo largo de varios días para disminuir este error. Para contrarrestar parcialmente esta limitación, el Ministerio incluye la valoración conjunta del examen junto con el expediente académico, cuyo peso en la nota final ha variado a lo largo de las convocatorias.

Por otro lado, la sistemática de trabajo de los centros de preparación del examen MIR hace que un concepto preguntado en una convocatoria sea incluido con posterioridad en el material de estudio, por lo que, si se repitiese dicho concepto en otra convocatoria, cambiarían significativamente los valores psicométricos obtenidos por la pregunta en la que fuera evaluado el concepto por segunda vez. Desde el inicio de las preguntas con imagen radiológica y no radiológica en el MIR, todos los centros de preparación incluyen bases de datos y preguntas asociadas a imagen, dando menor o mayor importancia a la asignatura según sus criterios e incluyendo o no profesores específicos de radiología.

Los exámenes MIR del período analizado (2009-2017) constituyen globalmente una prueba de respuesta múltiple bien estructurada, con una dificultad corregida por el azar (IDc) en un nivel óptimo (0,55), con una calidad discriminativa media de sus preguntas medida por el r_{pbis} buena (0,30) y con una capacidad de discriminación medida por el DC_R buena (0,76). Similares valores se han obtenido en

estudios previos al analizar una única convocatoria^{15,20} o al analizar en global un número de convocatorias¹⁴.

Si se dividen las preguntas según subgrupos basados en el enunciado; las preguntas con mayor calidad y capacidad de discriminación son las basadas en casos clínicos (aquellas que tratan acerca de una historia clínica de un enfermo simulado, y en las que se pregunta por su diagnóstico y/o tratamiento), y también las preguntas de test directas de un concepto médico (ambas superan en DC_R el 0,75 y en r_{pbis} el 0,30). Las preguntas con enunciado negativo, en las que las opciones contienen tres supuestos verdaderos y uno falso que hay que identificar, tienen un promedio de discriminación menor.

Desde la convocatoria de 2009, los exámenes incluyen entre un 11% y un 15% de preguntas vinculadas a una o varias imágenes, tanto radiológicas como no radiológicas (histopatología, electrocardiogramas, espirometrías, fotografías clínicas, etc.). Estos dos subgrupos de preguntas comparten las características de obligar al opositor a consumir más tiempo para interpretarlas y contestarlas, ya que implican asociar a la lectura del enunciado el análisis de una imagen impresa en papel. En teoría, añadir estas preguntas asociadas a imagen al examen tienen como objetivo acercar las preguntas a la práctica clínica habitual.

El estudio psicométrico de las preguntas vinculadas a imagen radiológica señala que, como promedio, son más difíciles (IDc de 0,51, manteniéndose en un nivel óptimo), pero tienen, como promedio, una menor capacidad de discriminación en el examen (r_{pbis} de 0,25 y DC_R de 0,60, ambos valores correspondientes en sus escalas a una discriminación regular). En su promedio, las preguntas vinculadas a imágenes no radiológicas son ligeramente más fáciles y más discriminativas que las vinculadas a imágenes radiológicas, pero menos discriminativas que los casos clínicos y las preguntas directas tipo test.

Si analizamos la evolución de los valores psicométricos de las preguntas asociadas a imagen radiológica a lo largo de la serie temporal estudiada, el promedio de discriminación aumenta en aquellas convocatorias en las cuales la dificultad de las preguntas se acerca a la dificultad media del examen MIR (años 2010 y 2011) y la discriminación disminuye en aquellas en las que existe un aumento significativo de la dificultad de estas, como en la convocatoria de 2014.

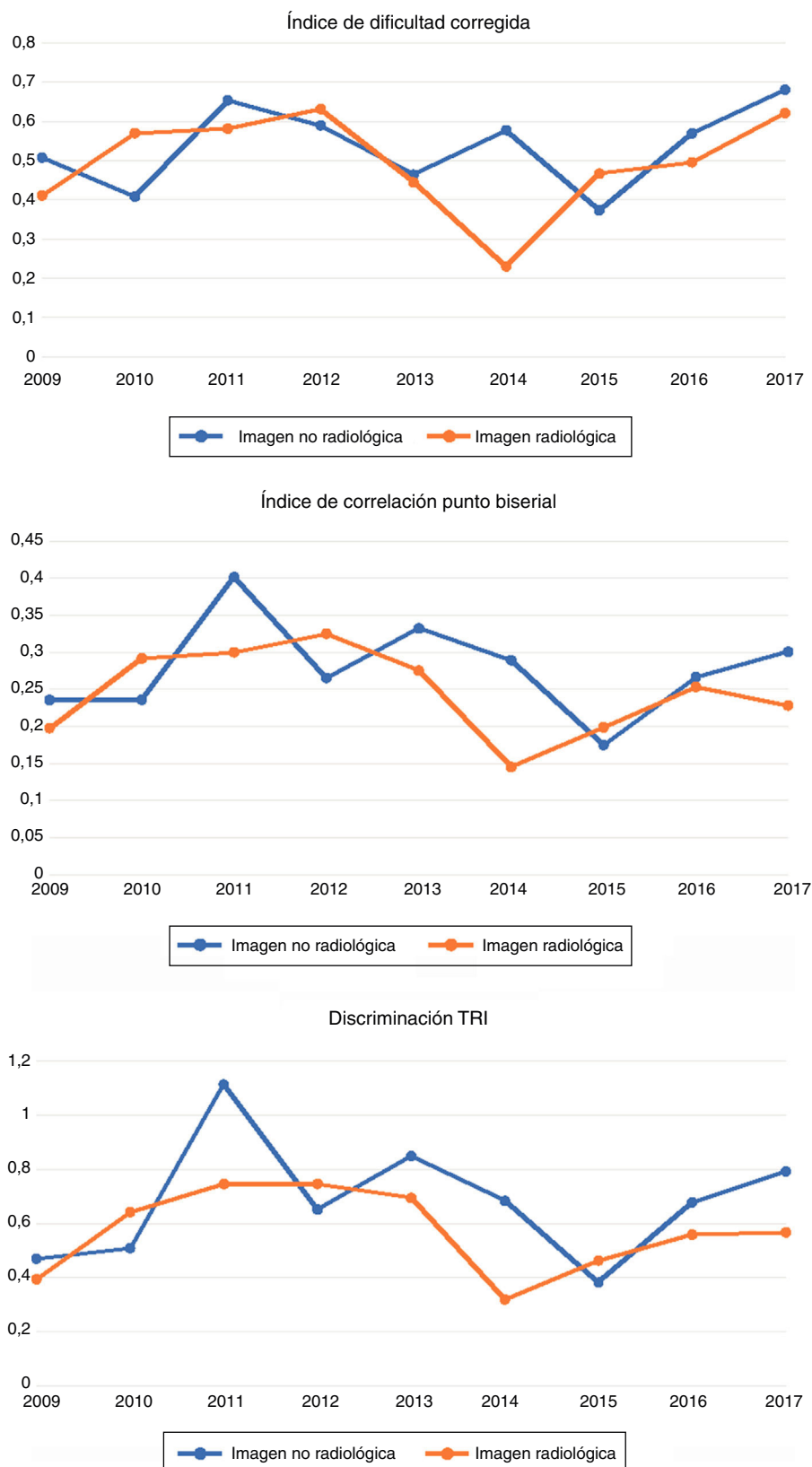


Figura 4 Comparación de la evolución de los principales parámetros psicométricos entre preguntas asociadas a imagen radiológica y preguntas asociadas a imagen no radiológica. Índice de dificultad con corrección del azar, cuando más alto más difícil de contestar correctamente es la pregunta. Índice de correlación punto biserial que representa una medida de la calidad de las preguntas, considerando preguntas buenas o excelentes aquellas que superan 0.30. Discriminación TRI representa la capacidad de discriminación de una pregunta para los distintos niveles de conocimiento de los opositores, se considera que una pregunta es buena si supera el 0.70 y excelente si supera el 1.

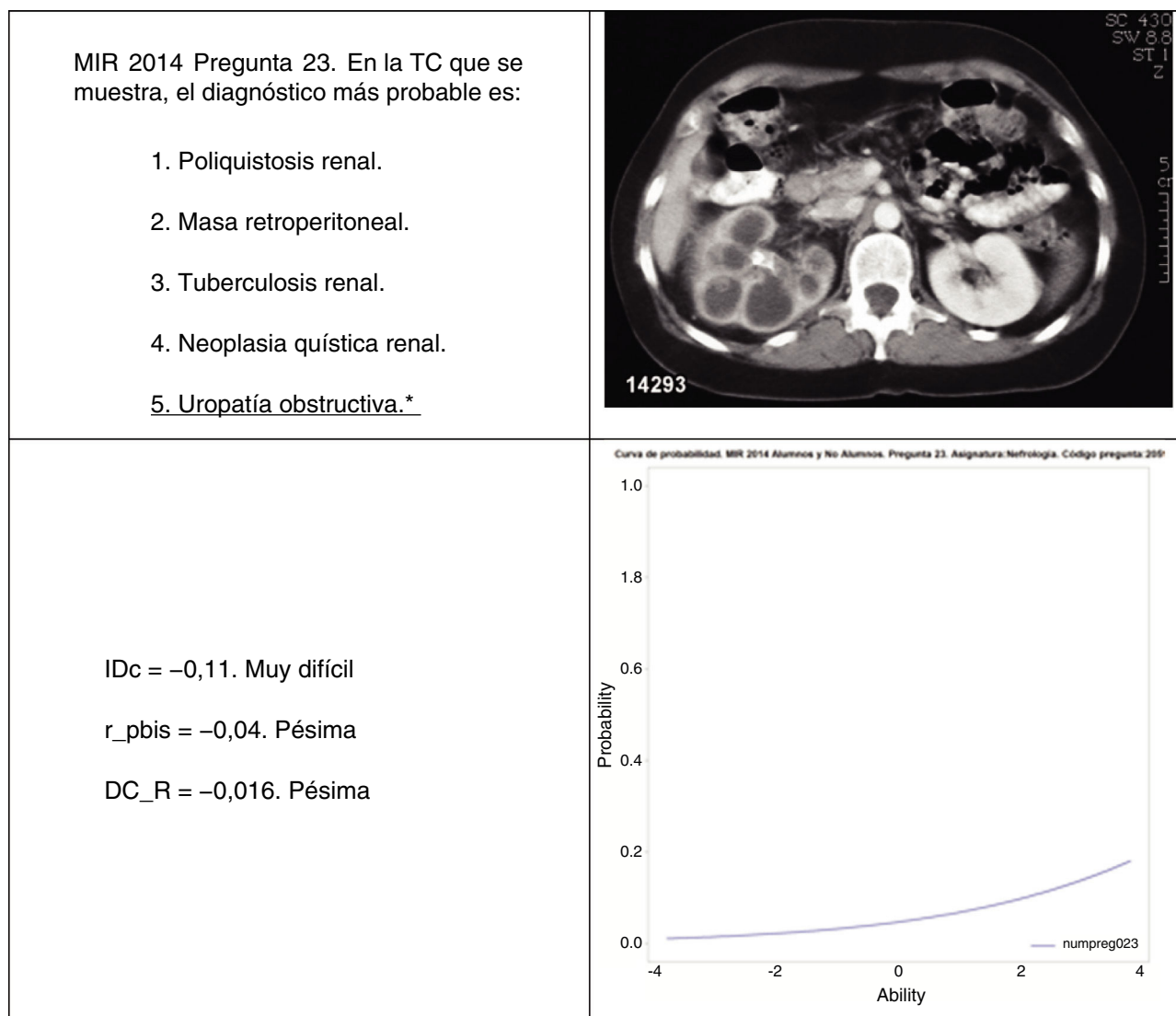


Figura 5 Ejemplo de pregunta de alta dificultad y baja discriminación. La discriminación de la pregunta es pésima. En la curva de probabilidad de acierto, los opositores con mayores puntuaciones en el examen tienen más posibilidades de acertarla que los alumnos con puntuaciones bajas en el MIR, aunque las diferencias entre ambos son pequeñas. La pielonefritis xantogranulomatosa tiene la imagen característica de cálculo más abscesos renales y probablemente es una enfermedad conocida por los opositores al caer en exámenes previos. Además, estos conocen la imagen de la litiasis renal obstructiva por tomografía computarizada y ecografía. Al hacer una pregunta sin clínica característica (cólico renal), con una imagen que no es típica de dilatación de la pelvis renal se consigue confundir tanto a los opositores del grupo fuerte como a los del grupo débil. Al intentar aumentar la dificultad de la pregunta disminuyendo la información clínica suministrada y con una imagen no característica, se consigue el efecto indeseable de disminuir la capacidad de discriminación de la pregunta.

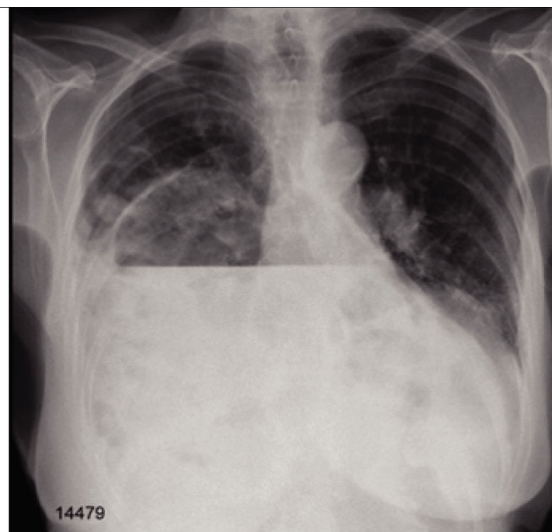
El análisis individual de las preguntas demuestra que los opositores fallan más, y por lo tanto resultan más difíciles, aquellas preguntas radiológicas en las que el concepto preguntado representa la capacidad de visualizar un hallazgo radiológico sin información clínica asociada, las preguntas en las cuales el hallazgo radiológico cambia la respuesta que se intuye de la información clínica suministrada, o las preguntas en las cuales el hallazgo radiológico no es un concepto clínico importante (por su rareza o excepcionalidad) o que no ha sido preguntado en la historia del MIR. Estos tres factores provocan un aumento significativo de la dificultad de la pregunta, y una disminución de su

discriminación, al provocar que opositores con un menor nivel de conocimientos tengan más probabilidades de acertar la pregunta que aquellos con un mayor nivel de conocimientos (figs. 5, 6, 7 y 8).

En cambio, las preguntas asociadas a imagen radiológica de mayor calidad discriminativa son aquellas que se encuentran integradas en un caso clínico de un paciente simulado, son coherentes con la información clínica suministrada y en las que el concepto radiológico preguntado es significativo por su relevancia clínica. Estas preguntas suelen tener un nivel de dificultad óptimo-fácil (figs. 9, 10, 11 y 12). En estos casos, aunque no se ha analizado en profundidad, hay

MIR 2015 Pregunta 23. Mujer de 42 años de edad que acude a urgencias por aumento de la disnea en las últimas semanas acompañando leve dolor torácico punzante que aumenta con los movimientos respiratorios. No ha presentado fiebre ni aumento de la tos. A la auscultación pulmonar se objetiva disminución del murmullo vesicular en base derecha. Se realiza la siguiente radiografía del tórax que demuestra la existencia de:

1. Tumor broncogénico
2. Hernia de Bochdalek.*
3. Absceso pulmonar.
4. Neumopericardio



IDc = 0,06. Muy difícil

r_pbis = 0,03. Pésima

DC_R = 0,04. Pésima

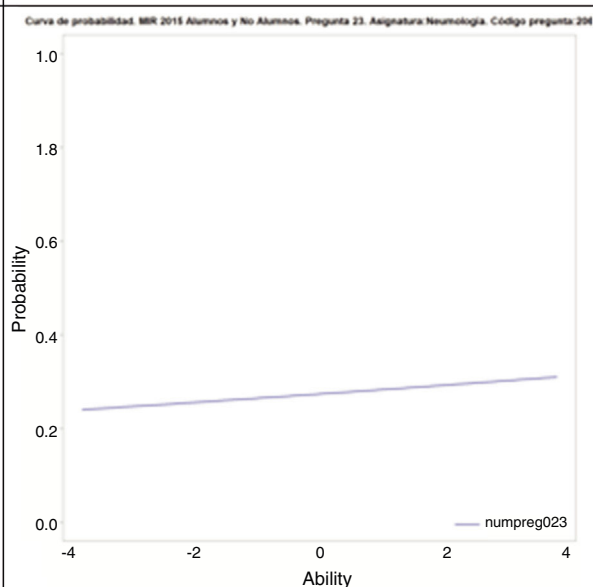
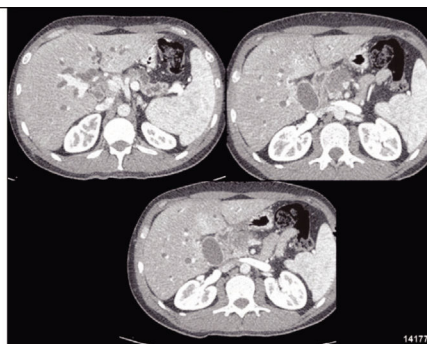


Figura 6 Ejemplo de una pregunta en la que se aborda un concepto radiológico básico (hernia diafragmática, hernia de Bochdalek), pero en la que la imagen muestra una presentación atípica (hernia en el lado derecho, de gran tamaño y con herniación gástrica). Al intentar aumentar la dificultad de la pregunta, usando una forma atípica de la enfermedad, y a pesar de un entorno clínico adecuado, la pregunta tiene una discriminación baja. La curva de probabilidad, prácticamente horizontal, indica que los opositores con más conocimientos tienen la misma probabilidad de acertarla que los opositores con menos conocimientos.

MIR 2013 Pregunta 2. Hombre de 40 años sin antecedentes de interés. Desde hace 4 semanas, refiere cuadro de deposiciones diarreicas, asociado a ictericia de piel y mucosas, coluria, hipocolia y dolor en hipocondrio derecho, acompañado de pérdida de peso (10 kg aprox.) y prurito intenso. En la analítica destaca: Bilirrubina total: 15,3 mg/dl. Alanina-aminotransferasa: 70 U/l

Aspartatoaminotransferasa: 85 U/l.

Serología para hepatitis negativa y elevación del CA 19-9. Se realiza TC abdominal. Se muestran 3 secciones axiales. En relación a los hallazgos de la prueba radiológica, ¿cuál de los siguientes tratamientos está indicado?:



1. Drenaje biliar por vía endoscópica retrógrada y quimioterapia sistémica.
2. Duodenopancreatectomía cefálica con resección vascular.
3. Embolización hepática con esferas cargadas de quimioterapia.
4. Drenaje biliar transparieto-hepático y quimioterapia sistémica.
5. Tratamiento con corticoides.

IDc = 0,36. Difícil
r_pbis = -0,01. Pésima
DC_R = -0,05. Pésima

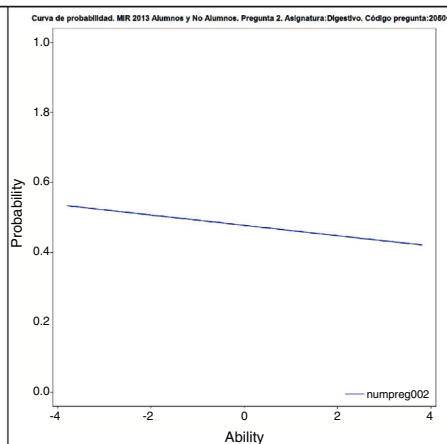


Figura 7 Ejemplo de pregunta con discriminación negativa. Existe una mayor probabilidad de acierto en los opositores que han sacado peor puntuación en el MIR que en los alumnos con mejores puntuaciones. Se ha buscado que la respuesta al concepto radiológico solicitado, identificar una neoplasia pancreática con dilatación de la vía biliar, sea de mayor dificultad al preguntar los criterios de irresecabilidad de una neoplasia pancreática. Al preguntar un concepto que se encuentra actualmente en discusión

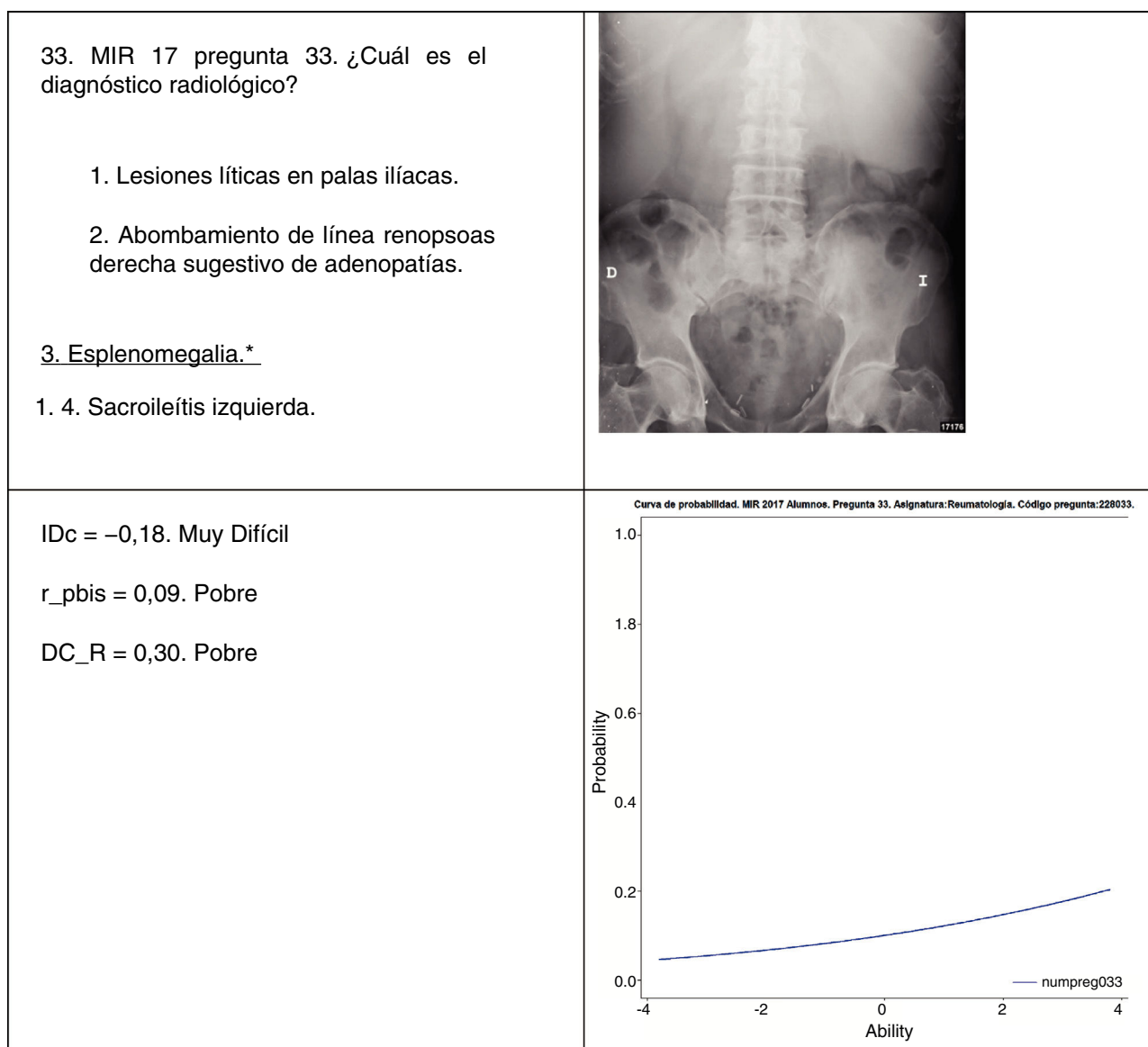


Figura 8 Ejemplo de pregunta psicométricamente no válida, en la que el nivel de dificultad es tan alto que provoca una nula capacidad de discriminación. En este caso el problema principal de la pregunta deriva de usar una imagen radiológica que induce a error, al incluir entre los distractores el diagnóstico de “sacroileítis izquierda” y aportar una imagen en la que la articulación sacroilíaca izquierda no es visible. La dificultad se incrementa al preguntar por una patología (esplenomegalia) cuyo diagnóstico se suele hacer con otra prueba de imagen (ecografía) y no incluir en el enunciado ningún dato de la exploración clínica del paciente.

indicios de que las imágenes de tomografía computarizada obtienen mejor nivel de dificultad que las de radiología simple.

Por lo tanto, y desde un punto de vista psicométrico, sería recomendable que las preguntas asociadas a imagen radiológica en el examen MIR mejorasen su calidad discriminativa para adecuarse a los estándares del examen y a los valores promedios de este, lo que implica que deben cumplir los

estándares básicos de las preguntas de respuesta múltiple ya comentados en varios trabajos asociados a la radiología²⁹.

Por otro lado, se observó un significativo sesgo de ocultación de respuesta (las respuestas más frecuentes son la tres y la dos, y las más infrecuentes la primera y la última). En las preguntas con imagen radiológica del último MIR, la respuesta 3 era correcta casi en el 50% de las preguntas. Probablemente esto se debe a no realizar una aleatorización

(la arteria no se encuentra infiltrada, pero si el eje venoso mesentérico), provoca que la respuesta correcta sea discutible, y disminuye la capacidad de discriminación.

30 ABENICAL SYSTEM
 Target Center Rep. 2013.08
 1. 11054
 2. 4
 3. 1054
 4. 140.85 Ax
 1000.40 Scan
 101 F

1402208
 2013.08.08
 12:50:50 PM
 572.1012
 1000.40
 101 F

R 120
 11054
 140.85 Ax
 1000.40 Scan
 101 F

03698
 11054
 140.85 Ax
 1000.40 Scan
 101 F

1402208
 2013.08.08
 12:50:50 PM
 572.1012
 1000.40
 101 F

Figura 9 Ejemplo de una pregunta radiológica con un buen índice de discriminación. El diagnóstico diferencial que se solicita es la patología del colon, por lo que los distractores son adecuados con una adecuada explicación radiológica que justifica la respuesta. El entorno clínico es adecuado y la imagen radiológica muestra correctamente la patología sin otras imágenes que puedan llevar a errores de interpretación.

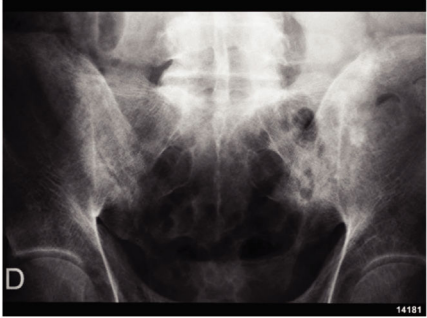
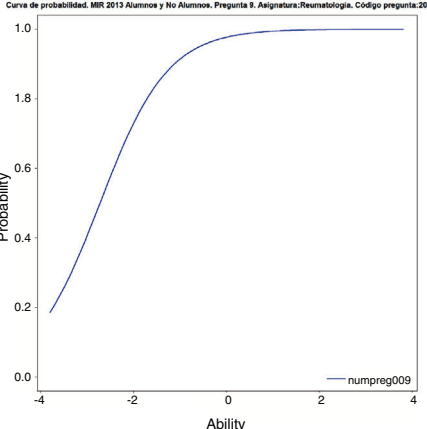
MIR 2013 Pregunta 9. Hombre de 41 años remitido por presentar desde hace más de 5 años dolor glúteo bilateral alternante de predominio nocturno, que interfiere el sueño y alivia con la movilización. Se acompaña de rigidez de más de 1 hora tras la inactividad. Niega dolor, tumefacción o deformidad a otro nivel. Aporta analítica elemental sin alteraciones excepto PCR 11,5 mg/dl ($N < 5$) y la radiografía adjunta. La alteración radiológica objetivada y el diagnóstico más probable sería: 1. Fractura de Looser-Osteomalacia. 2. Sacroileítis bilateral-Espondilitis anquilosante.* 3. Coexistencia de lesiones líticas y blásticas -Enfermedad de Paget. 4. Lesiones blásticas- Adenocarcinoma prostático con metástasis óseas. 5. Osteoesclerosis-Linfoma.	
IDc = 0,52. Óptima r_pbis = 0,42. Excelente DC_R = 1,02. Excelente	
	Curva de probabilidad. MIR 2013 Alumnos y No Alumnos. Pregunta 9. Asignatura: Reumatología. Código pregunta: 20 

Figura 10 Ejemplo de concepto radiológico correctamente preguntado en el examen MIR. El entorno clínico es adecuado, la imagen radiológica muestra el hallazgo claramente y los distractores son adecuados.

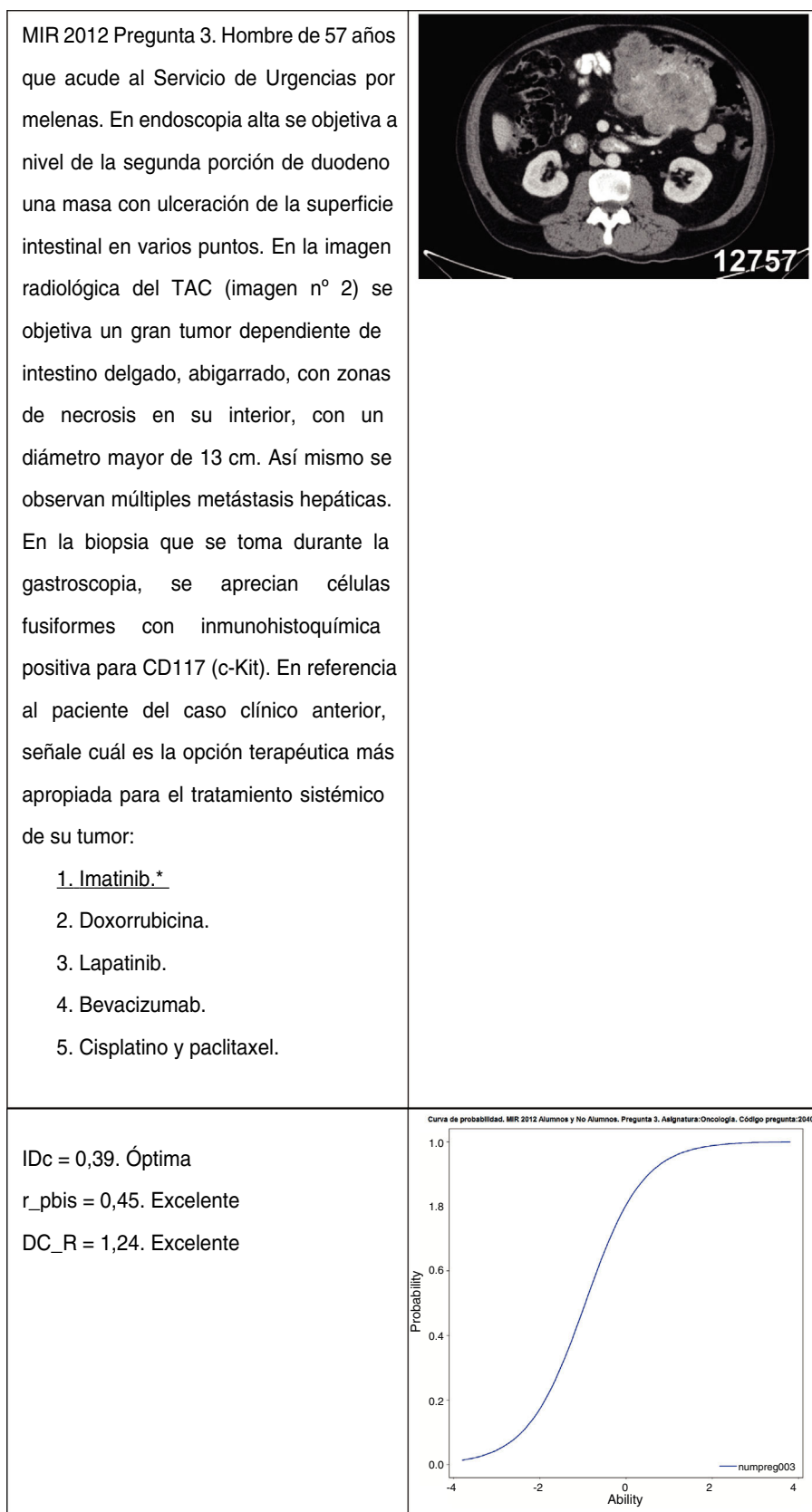


Figura 11 Ejemplo de pregunta con una capacidad de discriminación excelente, que es máxima aproximadamente en el nivel de conocimiento mediano de la muestra. Esta pregunta busca que el opositor sea capaz de identificar los distintos tipos de cáncer en el intestino delgado y su adecuado tratamiento quimioterápico. La posibilidad de discutir el diagnóstico radiológico se corrige mediante la introducción de un adecuado entorno clínico y una presentación inmunohistoquímica adecuada.

MIR 2015 Pregunta 21. Hombre de 71 años. Fumador de 2-3 cigarrillos al día hasta hace 7 años. Ha trabajado como empleado en empresa de embalaje, sin exposición a humos o polvo.

Hipertensión arterial y cardiopatía

hipertensiva. Hace 2 años fue

diagnosticado de una fibrilación auricular paroxística y recibió tratamiento con amiodarona durante 18 meses a dosis de 200 mg/día y posteriormente 400 mg/día en los últimos 6 meses. Hace 5 meses presentó un cuadro de tos, fiebre y disnea. Se realizó una radiografía de tórax (imagen A) y recibió tratamiento antibiótico y esteroideo con mejoría. El control radiológico posterior fue normal.

Acude a consulta por presentar astenia, fiebre y tos de 2 semanas de evolución.

En la analítica destaca hemograma normal, proteína C reactiva 9,5 mg/dL (normal 0-1), VSG 39 mm (1ª hora). Se realiza una radiografía de tórax (imagen B) y se decide ingreso. Se realiza una broncoscopia con lavado broncoalveolar, con el siguiente recuento celular: Linfocitos 74% (cociente CD4/CD8 = 0,9), macrófagos 18%, neutrófilos 4%. La citología se informa con presencia de macrófagos espumosos. Se indica realizar

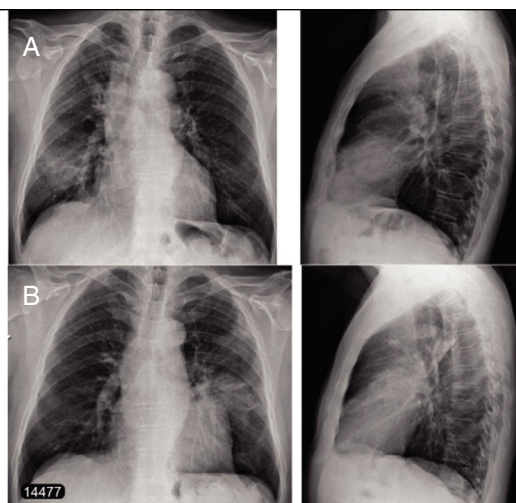


Figura 12 Ejemplo de pregunta con buen nivel de discriminación a pesar de un nivel de dificultad elevado. El nivel de dificultad es alto debido a la necesidad de valorar cuatro imágenes y al preguntar por una enfermedad relativamente infrecuente, la neumonía organizada. El adecuado entorno clínico de la pregunta y la ausencia de distractores que puedan discutir significativamente la opción correcta (solicitan diagnóstico diferencial con la sarcoidosis, la fibrosis pulmonar y el pulmón de amiodarona) hacen que la pregunta sea adecuada desde el punto de vista psicométrico.

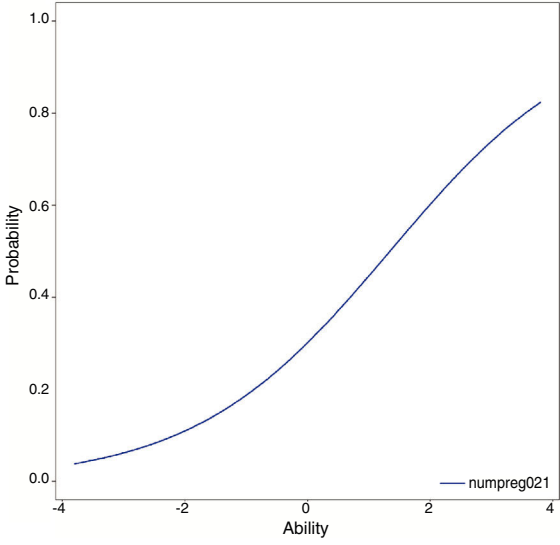
una biopsia por videotoracosopia. Ante los hallazgos clínicos, radiológicos y analíticos con los que ya contamos podemos concluir: <u>1. El cuadro anatomopatológico más probable es una neumonía organizada.*</u> 2. Una neumonitis por amiodarona parece poco probable dado que el paciente ha tomado el fármaco durante 2 años. 3. El cociente límfocitario del lavado broncoalveolar sugiere un cuadro de sarcoidosis. 4. El cuadro sugiere una neumonía de tipo fibrótico 5.	
IDc = 0,28. Difícil r_pbis = 0,32. Buena DC_R = 0,70. Buena	Curva de probabilidad. MIR 2016 Alumnos y No Alumnos. Pregunta 21. Asignatura:Neumología. Código pregunta:201 

Figura 12 (Continued)

del orden de respuestas, principalmente por las preguntas de anatomía radiológica que hacen referencia a números incluidos en la imagen.

Los conceptos preguntados deben ser básicos y representar los conocimientos de diagnóstico por imagen atribuibles a un médico no especialista, deben estar rodeados de un entorno clínico adecuado y los distractores utilizados en la pregunta deben contener la suficiente información para evitar la discusión del concepto radiológico. La imagen debe ser de buena calidad y no debe contener hallazgos que pueden interferir en los distractores utilizados. Las preguntas asociadas a imagen con mayor capacidad de discriminación en la historia del MIR son aquellas con un nivel medio-bajo de dificultad y cuya redacción es adecuada.

En los exámenes analizados, al aumentar la dificultad de las preguntas, la calidad discriminativa de estas disminuye, probablemente porque desde un punto de vista radiológico se asocia un aumento de la dificultad con hallazgos radiológicos raros, atípicos o que lleven a un diagnóstico diferencial más extenso o complejo. No es lo mismo un diagnóstico clínico difícil realizado en una paciente real, que hacer una pregunta multirrespuesta difícil en el entorno de una prueba tipo MIR. Si el objetivo que se persigue es aumentar la dificultad de la pregunta, este se puede conseguir preguntando por enfermedades menos frecuentes o con hallazgos más difíciles de identificar en la imagen, siempre y cuando se mantenga la estructura fundamental de la pregunta que incluye un entorno clínico adecuado y el uso de distractores que no puedan ser discutidos al presentar hallazgos parecidos en la imagen.

Se podría considerar que aportar más información clínica permite al opositor contestar las preguntas radiológicas infiriendo la respuesta del texto y no de la imagen; esto, por su puesto, influye en la dificultad de la pregunta. En el caso de las preguntas con imagen no radiológica debería ocurrir de la misma manera, pero estas preguntas presentan una dificultad y una discriminación más adecuada que las radiológicas, incluso incluyendo aquellas con poca información clínica, como ocurre en las de anatomía patológica o en las imágenes clínicas. Por lo tanto, en nuestra opinión, apoyados en los datos psicométricos, preguntar sobre un hallazgo radiológico excepcional o en ausencia de un contexto clínico adecuado es la causa principal de la baja capacidad de discriminación de las preguntas asociadas a imagen radiológica. Por supuesto no afirmamos en este trabajo que la imagen radiológica sea poco importante o sea prescindible y por eso no se puedan hacer preguntas directas sin información clínica. Lo que sí se puede afirmar con los resultados de este trabajo es que, si se hacen preguntas sin información clínica o con información clínica engañosa, o imágenes puramente radiológicas, la discriminación de la pregunta baja (en el entorno actual de la muestra de los opositores al examen) y no llega a los estándares del resto de las preguntas en el MIR.

No hemos referido por parte de los opositores (en nuestras encuestas de calidad), ni hemos detectado por nuestra parte, que la calidad de la impresión de la imagen radiológica o el tamaño de esta se asocien a una mayor dificultad al contestar la pregunta, por lo que no las hemos considerado variables que deban tenerse en cuenta en la valoración de la validez interna del examen.

Conclusiones

Las variables psicométricas son un conjunto de indicadores de evaluación de las pruebas de respuesta múltiple que se deberían estudiar de forma sistemática ante cualquier examen realizado en un entorno académico o de evaluación de conocimientos. Permiten distinguir preguntas inútiles o que introducen ruido en las calificaciones de la prueba, identificar preguntas mal realizadas técnicamente y ayudan a identificar lagunas en el aprendizaje. En nuestra opinión, la evaluación psicométrica de las respuestas a las preguntas del MIR se debería emplear como criterio de apoyo a la labor de anulación de preguntas que realiza la Comisión Calificadora de la prueba¹⁴.

Las preguntas con imagen radiológica se encuentran por debajo de los valores medios del examen MIR en discriminación, y resultan de una mayor dificultad para los opositores. Si se comparan con las preguntas asociadas a imagen no radiológica del examen MIR, también se observa en estas una menor capacidad de discriminación, aunque de forma menos acusada.

Para poder igualar los estándares del examen MIR, es necesario mantener una adecuada estructura en la confección de los ítems, en términos de entorno clínico, adecuado uso de distractores y menor nivel de dificultad, que se puede lograr mediante el uso de imágenes con hallazgos radiológicos típicos.

Autoría

1. Responsable de la integridad del estudio: EMQ.
2. Concepción del estudio: EMQ, FSL.
3. Diseño del estudio: EMQ, FSL, SMCG, JBR, JCB.
4. Obtención de los datos: EMQ, FSL, MCR, JBR, JCB.
5. Análisis e interpretación de los datos: EMQ, FSL, MCR, JBR, JCB.
6. Tratamiento estadístico: EMQ, EFL, MCR.
7. Búsqueda bibliográfica: EMQ.
8. Redacción del trabajo: EMQ.
9. Revisión crítica del manuscrito con aportaciones intelectualmente relevantes: EMQ, SMCG, JBR.
10. Aprobación de la versión final: EMQ, SMCG, MCR, JBR.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Rodríguez A, Martínez F. Aplicaciones informáticas de psicometría en investigación educativa. *Comunicar*. 2003;21:163-6.
2. Imbrenón F, Silva P, Guzmán C. Competencias en los procesos de enseñanza-aprendizaje virtual y semipresencial. *Comunicar*. 2003;36:107-14.
3. Guerra S, González N, García-Ruiz R. Utilización de las TIC por el profesorado universitario como recurso didáctico. *Comunicar*. 2010;35:141-8.
4. Sieiro P. Profesorado y nuevas tecnologías. *Comunicar*. 1994;3:154-5.
5. Marcelo C, Yot C, Mayor-Ruiz C. Enseñar con tecnologías digitales en la universidad. *Comunicar*. 2015;45:117-24.

6. Real Decreto 127/1984, de 11 de enero, por el que se regula la formación médica especializada y la obtención del título de médico especialista. BOE núm. 26, de 31 de enero de 1984; p. 2524-8.
7. Programas de Formación Sanitaria Especializada. Ministerio de Sanidad, Servicios Sociales e Igualdad. Disponible en: <https://fse.mscbs.gob.es/fseweb/view/index.xhtml> (acceso 28 de septiembre de 2018).
8. Curbelo J, Fernández-Somoano A, Romeo JM, Villacampa T, Sánchez-Lasheras F, Baladrón J. La elección de la especialidad medicina intensiva: análisis de los últimos 10 años. *Med Intensiva*. 2018;42:65-8.
9. Curbelo J, Galván-Román JM, Sánchez-Lasheras F, Romeo JM, Fernández-Somoano A, Villacampa T, et al. Aparato digestivo: evolución de la elección de la especialidad en los últimos años. *Rev Esp Enf Dig*. 2017;109:614-8.
10. Curbelo J, Romeo JM, Galván-Román JM, Vega-Villar J, Martínez-Lapiscina EH, Jiménez-Fonseca P, et al. Popularidad de Neurología en España: análisis de la elección de la especialidad. *Neurología*. 2017 Dec 23; pii: S0213-4853(17)30361-4.
11. Murias E, Sánchez-Lasheras F, Fernández-Somoano A, Romeo JM, Baladrón J. Análisis de la elección de la especialidad de radiodiagnóstico en el examen MIR desde el año 2006 hasta 2015. *Radiología*. 2017;59:232-46.
12. Curbelo J, Romeo JM, Fernández-Somoano A, Sánchez Lasheras F, Baladrón J. Endocrinología y nutrición: evolución de la elección de la especialidad en los últimos años. *Endocrinol Diabetes Nutr*. 2017;64:329-31.
13. Baladrón J, Sánchez-Lasheras F, Romeo JM, Curbelo J, Fonseca P. Evolución de los parámetros dificultad y discriminación en el ejercicio de examen MIR. Análisis de las convocatorias de 2009 a 2017. *FEM*. 2018;21:181-93.
14. Baladrón J, Sánchez-Lasheras F, Villacampa T, Romeo-Ladrero JM, Fernández-Somoano A. Propuesta metodológica para la detección de preguntas susceptibles de anulación en la prueba MIR. Aplicación a las convocatorias 2010 a 2015. *FEM*. 2017;20:161-75.
15. Baladrón J, Sánchez-Lasheras F, Villacampa T, Romeo-Ladrero JM, Fernández-Somoano A. El examen MIR 2015 desde el punto de vista de la teoría de respuesta al ítem. *FEM*. 2017;20:29-38.
16. 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.
17. 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.
18. 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.
19. Gemp F. Fuentealba R. El error estándar de medida y la puntuación verdadera de los tests psicológicos: Algunas recomendaciones prácticas. *Terapia Psicológica*. 2006;24:117-29.
20. Guttman L. A basis for scaling qualitative Data. *Am Sociological Rev*. 1944;9:139-50.
21. Lord F. A theory of test scores (Psychometric Monographs no. 7). Richmond, VA: Psychometric Corporation; 1952.
22. Rasch G. Probabilistic models for some intelligence and attainment tests. Chicago, IL: University of Chicago Press; 1980.
23. 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.
24. Álvarez E, Arcos A, González S, Muñoz JF, Rueda M. Estimating population proportions in the presence of missing data. *J Computational Applied Mathematics*. 2013;237:470-6.
25. Embretson SE, Reise SP. Item response theory for psychologists. Hillsdale, NJ: Erlbaum; 2000.
26. Lord FM. Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum; 1980.
27. Birnbaum A. Some latent trait models and their use in inferring an examinee's ability. En: Lord FM, Novick MR, editores. *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley; 1968. p. 397-472.
28. Burnham KP, Anderson DR. Model selection and multimodel inference: a practical information-theoretical approach. 2.ª ed. New York: Springer-Verlag; 2002.
29. Soler Fernández R, Méndez Díaz C, Rodríguez García E. Continuing medical education: how to write multiple choice questions *Radiología*. 2013;55 Suppl 1:S28-36.