Educ Med. 2016;xxx(xx):xxx–xxx

Educación Médica

www.elsevier.es/edumed

ORIGINAL

Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple

Juan F. del Cañizo[a,b,*], Mercedes Sanz Sánchez[b] y Pedro García-Barreno[a,b]

[a] *Departamento de Cirugía, Universidad Complutense de Madrid, Madrid, España*
[b] *Sección de Cirugía Experimental, Hospital General Universitario Gregorio Marañón, Madrid, España*

Recibido el 16 de marzo de 2016; aceptado el 17 de marzo de 2016

PALABRAS CLAVE
Nivel aceptable de resultados;
Pruebas de elección múltiple

Resumen Se presenta la experiencia de once años de evaluación de los conocimientos teóricos de la asignatura Fisiopatología Quirúrgica en el grupo docente del Hospital General Universitario Gregorio Marañón dependiente del Departamento de Cirugía de la Universidad Complutense. Las pruebas durante este tiempo han sido homogéneas y han consistido en un test de 100 preguntas de elección múltiple (PEM) con 5 respuestas posibles y solo una correcta extraídas de nuestra base de datos de preguntas.

En el curso 2015-16 se ha aplicado el mismo proceso de evaluación a los alumnos de Patología Quirúrgica de Digestivo de 4.° curso. Este mismo grupo de alumnos se evaluó el año anterior en la asignatura de Fisiopatología Quirúrgica de 3.er curso y se observa una alta concordancia de los resultados demostrando la robustez del sistema de evaluación.

En este trabajo se atiende únicamente a la evaluación de los conocimientos teóricos de los alumnos. La evaluación de las aptitudes y de las actitudes se realiza de forma diferente y no son motivo de este trabajo.

KEYWORDS
Acceptable levels of results;
Multiple choice test

Reliability of the calculation technique of acceptable level of results in multiple choice tests

Abstract The experience of eleven years of evaluation of theoretical knowledge of the subject Surgical Pathophysiology in the teaching Group of the Hospital General Universitario Gregorio Marañón belonging to the Department of Surgery of the Complutense University of Madrid is presented. Tests during this time have been homogeneous and consisted of 100 multiple choice questions with 5 possible answers and only one correct taken from our database of questions.

* Autor para correspondencia.
Correos electrónicos: canizo@hggm.es, jfcanizo@ucm.es (J.F. del Cañizo).

http://dx.doi.org/10.1016/j.edumed.2016.03.006

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006

During academic year 2015-16 the same evaluation process was applied to the students of Digestive Surgical Pathology (4th grade). This same group of students was assessed last year in the course of Surgical Pathophysiology of 3rd grade and high concordance of the results is observed demonstrating the robustness of the evaluation system.

This paper take care only to the evaluation of theoretical knowledge of students. Assessing skills and attitudes is performed in a different way and are not the aim for this work.

Introducción

La evaluación de los conocimientos teóricos de los estudiantes de medicina supone siempre un reto para los profesores, sobre todo a la hora de seleccionar el procedimiento de evaluación. El objetivo de la evaluación es certificar que los alumnos que superan las pruebas tienen los conocimientos suficientes sobre la materia evaluada, se trata por tanto de establecer un nivel de conocimientos y en ningún caso seleccionar a los alumnos por sus resultados[1].

Como el objetivo de la prueba es certificar que el alumno sabe lo necesario de la asignatura se deben utilizar pruebas de criterios absolutos, es decir, los resultados deben depender solo del nivel de conocimientos del alumno y no de la posición que ocupe en los resultados globales de la prueba. De esta forma si todos los alumnos superan el nivel establecido aprueban todos, si ninguno lo superase ninguno aprobaría.

Los métodos basados en la distribución normal («campana de Gauss») se podrían utilizar si lo que pretendiésemos es hacer una selección de alumnos, pero en el caso de la evaluación de una asignatura está claro que debemos utilizar pruebas de criterios absolutos.

La siguiente decisión que el equipo docente tiene que tomar es la del tipo de prueba que pretende utilizar para la evaluación. En nuestro caso se optó por las denominadas pruebas objetivas por preguntas de elección múltiple (los llamados «test»).

Se entiende por prueba objetiva aquel instrumento de evaluación que busca valorar el nivel instructivo del alumnado utilizando una serie de preguntas presentadas con el máximo de claridad y brevedad y cuya respuesta exige del alumno únicamente la elección entre varias respuestas posibles. Para su calificación existen normas concretas que hacen que la calificación no esté condicionada por la persona que califica, de ahí la denominación de «objetivas»[2].

Este tipo de pruebas tienen bastantes ventajas:

- Pueden explorar un amplio campo de conocimientos evaluando todos los aspectos de la asignatura.
- Se corrigen con facilidad y como la corrección es automática la objetividad está garantizada.
- El trabajo en equipo tanto en la preparación como en la corrección permite una crítica constructiva de las preguntas y su posterior análisis.
- Por tanto, en general, aumentan la fiabilidad y la validez de la prueba.

Sin embargo, su preparación exige mucho tiempo y una gran coordinación del grupo docente. Por otro lado debe eliminarse la posible influencia del azar en las respuestas y no aportan datos sobre la creatividad del estudiante. Su dificultad a la hora de prepararlas las hace poco apropiadas para grupos pequeños de estudiantes.

El último reto con que se enfrenta el grupo docente es el establecimiento del nivel de superación de la prueba con criterios absolutos. En este sentido se utilizó la técnica del cálculo del nivel aceptable de resultados (NAR) tal y como se describe en la guía pedagógica para el personal de la salud de la OMS[3].

Objetivos

Describir la experiencia de nuestro grupo docente con la evaluación de los estudiantes de tercero de Medicina en la asignatura de Fisiopatología y Propedéutica Quirúrgica durante los últimos 11 años.

Explorar si la técnica del cálculo del NAR descrita en la guía pedagógica es adecuada y fiable para establecer el nivel del aprobado en pruebas de preguntas de elección múltiple (PEM).

Métodos

Confección de los test

En todos los casos han consistido en una prueba de 100 PEM con 5 respuestas posibles y solo una correcta extraídas de nuestra base de datos de preguntas.

Nivel de suficiencia de las pruebas

Se ha calculado el nivel de suficiencia por la técnica del NAR para cada examen y cada grupo de alumnos según se describe en la Guía pedagógica para el personal de salud de la OMS[3]. Con esta técnica los fallos no descuentan puntuación.

El NAR corresponde al umbral que permite decidir (según criterios absolutos) el paso o no de un estudiante «que sabe justo lo necesario».

El NAR para una prueba es igual a la suma de los índices de aceptabilidad de cada PEM.

Para calcular el índice de aceptabilidad de una pregunta hay que determinar cuántas respuestas debe eliminar «de

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006

entrada» un estudiante que sabe justo lo necesario. El índice se calcula dividiendo uno por el número de respuestas no eliminadas.

- Si no se elimina ninguna respuesta el índice es 1/5 = 0,20.
- Si se elimina una el índice es 1/4 = 0,25.
- Si se eliminan dos el índice es 1/3 = 0,33.
- Si se eliminan tres el índice es 1/2 = 0,50.
- Si se eliminan cuatro el índice es 1/1 = 1.

Cada uno de los profesores del grupo calcula el NAR de la prueba y el nivel definitivo se calcula como la media de los NAR obtenidos.

Para calcular la nota del examen en una escala de 10 se calcula una recta de ajuste en la que el NAR corresponde al 5 y el número de respuestas máximo obtenido en el examen una vez corregido corresponde al 10.

Análisis de las preguntas

Para conseguir una base de datos de preguntas eficaz es muy importante analizar cada una las preguntas después de cada prueba. Para esto una vez corregido el examen se calculan los índices de dificultad y de discriminación de cada pregunta.

Para esto lo primero es ordenar los resultados del examen por orden de notas. Se selecciona entonces el 33% de los alumnos con notas más altas y se le denomina «grupo fuerte», después se selecciona el 33% de alumnos con notas más bajas y se le denomina «grupo débil».

Se puede entonces calcular los siguientes índices:

Índice de dificultad de una pregunta

Índice que permite determinar en qué medida una pregunta de examen es fácil o difícil. Es el porcentaje (%) de estudiantes que han respondido correctamente a una pregunta de examen; así pues, varía entre 0 y 100%.

Cálculo: se utiliza la fórmula siguiente:

$$IndDif = \frac{F + f}{N} \times 100$$

Donde: F = número de respuestas exactas en el grupo fuerte; f = número de respuestas exactas en el grupo débil; N = número total de estudiantes en los dos grupos.

Cuanto más elevado es este índice más fácil es la pregunta. Índices de dificultad entre 30 y 70% son aceptables.

Índice de discriminación de una pregunta

Cifra que permite determinar en qué medida una pregunta es lo bastante selectiva como para distinguir un grupo fuerte de un grupo débil de estudiantes. Varía entre -1 y +1. Cálculo: se utiliza la fórmula siguiente:

$$IndDisc = 2 \times \frac{F - f}{N}$$

Donde: F = número de respuestas exactas en el grupo fuerte; f = número de respuestas exactas en el grupo débil; N = número total de estudiantes en los dos grupos.

Cuanto más elevado es este índice, más posible es diferenciar entre «fuertes» y «débiles». En otras palabras, ayuda a reconocer a los mejores estudiantes.

>0,35 pregunta excelente.
>0,25 y <0,34 pregunta buena.
>0,15 y <0,24 pregunta marginal - a revisar.
<0,15 pregunta mala - probablemente a eliminar.

La información obtenida de los índices de cada pregunta se pasa a la base de datos para la evaluación de la calidad de las preguntas almacenadas.

Selección de las preguntas

El campo de los procesos intelectuales (sector cognitivo o cognoscitivo) incluye aquellos objetivos que tienen que ver con la memoria, el reconocimiento de datos y con el desarrollo de capacidades y habilidades intelectuales. En él, pueden diferenciarse tres niveles[4]:

Nivel I.- Recuerdo: la información se recupera básicamente en la misma forma en la cual se almacenó. En este nivel no se espera que los estudiantes transformen o manipulen los datos adquiridos, sino simplemente que los recuerden en la misma forma en que se les presentaron.

Nivel II.- Interpretación: implica la capacidad de transformación de la información por parte del alumno. En este nivel se requiere que los estudiantes demuestren que son capaces de modificar la información adquirida por medio de algún tipo de procesamiento antes de responder a una pregunta.

Nivel III.- Solución de problemas: requiere capacidad de análisis por parte del sujeto, va más allá de la simple comprensión de la información e implica la capacidad de descubrir cómo interactúan las diferentes partes del todo. El sujeto debe analizar la información que se le proporciona antes de abordar la resolución del problema.

A la hora de confeccionar el examen es muy importante tener en cuenta los niveles de las preguntas que se formulan y sobre todo tratar de evitar el abuso de preguntas de primer nivel y garantizar un número suficiente de preguntas de tercer nivel.

Se debe procurar también que el examen esté equilibrado desde el punto de vista del número de preguntas de cada tema a evaluar. Los índices de discriminación y de dificultad de las preguntas pueden ayudar mucho a la hora de su selección.

Las preguntas de cada examen se evalúan calculando sus índices y se incluyen en una base de datos de preguntas.

Cálculo de las calificaciones finales

Las pruebas se realizan con 100 preguntas, pero las calificaciones finales del examen se deben expresar con un número entre 0 y 10 con una cifra decimal. Para convertir el número de respuestas acertadas en calificaciones se utiliza una recta de ajuste en la que el nivel aceptable de resultados corresponde al 5 y la nota máxima obtenida en el examen corresponde al 10.

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006

Tabla 1 Resultados de los exámenes

Curso	Presentados	NAR	Media	DS	Máximo	Mínimo	% > NAR
2004_05	84	60	71,7	11,3	96	40	88,1
2005_06	93	63	69,5	7,3	83	47	79,6
2006_07	94	65	75,8	7,8	89	49	91,5
2007_08	103	65	73,6	8,7	91	47	83,5
2008_09	124	71	72,5	6,3	84	51	68,5
2009_10	140	72	80,8	5,6	91	61	93,6
2010_11	123	70	72,0	6,2	89	60	62,6
2011_12	143	70	79,4	6,8	92	50	93,0
2012_13	128	75	80,5	7,3	94	46	81,3
2013_14	126	63	76,8	7,0	90	52	93,7
2014_15	139	76	79,5	7,0	92	52	75,5
Medias	117,9	68,2	75,6	7,4	90,1	50,5	82,8
DS	20,9	5,3	4,0	1,5	3,9	6,0	10,6
Coef var	17,7	7,7	5,3	20,7	4,3	11,9	12,8
Máx.	143,0	76,0	80,8	11,3	96,0	61,0	93,7
Mín.	84,0	60,0	69,5	5,6	83,0	40,0	62,6

Resultados

De cada examen se recogieron los siguientes datos: número de alumnos presentados y NAR de la prueba. La media y la desviación estándar de las calificaciones, la nota máxima y la mínima y el tanto por ciento de alumnos con calificación por encima del NAR (aprobados). Se confeccionó también un histograma con los resultados de cada prueba (tabla 1).

Alumnos presentados

En la figura 1 se muestra la evolución del número de alumnos desde el curso 2004-05 hasta la actualidad (fig. 1 superior izda).

Se observa que el número de alumnos aumentó de forma significativa de 2004 a 2009 pasando de alrededor de 80 alumnos en 2004 a 140 en 2009, desde entonces el número de alumnos parece haberse estabilizado alrededor de los 130.

Hay que tener en cuenta que el profesorado en ese periodo no ha aumentado y en los últimos años está disminuyendo por los problemas presupuestarios.

Nivel aceptable de resultados

La media del NAR en las 11 pruebas es de 68,2 preguntas con un máximo de 76 y un mínimo de 60, la desviación típica es de 5,3.

En la figura se representa la evolución del NAR calculado en los 11 exámenes; se observa una cierta tendencia al aumento con el paso de los años a excepción dell año 13-14 que fue bastante bajo (fig. 1 superior derecha).

Calificaciones

La media de respuestas acertadas fue de 75,6 con una media máxima de 81 y una mínima de 69.

Las desviaciones estándar de las medias variaron entre 11,3 y 5,6 con una media de 7,4 respuestas lo que refleja bastante homogeneidad en los resultados. (fig. 1 inferior izda).

Tanto por ciento de alumnos con calificación superior al NAR (aprobados)

El nivel medio de aprobados en este periodo fue de 82,8% con un máximo de 93,7 y un mínimo de 62,6 (fig. 1 inferior derecha).

Estas cifras son explicables ya que nuestros alumnos tienen que obtener muy buenos resultados en las pruebas de acceso para poder ingresar en nuestra facultad y es, por tanto, esperable que el nivel de aprobados sea alto.

Comparación de los resultados de la prueba de 2015 con los de la prueba de 2016

En el año 2016 se ha comenzado a utilizar este método de evaluación con los alumnos de Patología Quirúrgica de digestivo que se imparte en el cuarto curso de la carrera lo que ha permitido, por primera vez, comparar los resultados de dos pruebas distintas en un mismo grupo de alumnos.

En el año 2015 se evaluó la asignatura de Fisiopatología Quirúrgica estableciéndose un nivel aceptable de resultados de 76 preguntas acertadas. En el año 2016 se evaluó la asignatura de Patología Quirúrgica I parte de digestivo y se estableció un nivel aceptable de resultados de 55 respuestas acertadas.

Si el método de evaluación es robusto deberemos esperar que las calificaciones obtenidas por los alumnos en 2015 no deben diferir significativamente de las obtenidas en 2016 a pesar de las diferencias entre las asignaturas, los grupos de profesores y los diferentes niveles aceptables de resultados de las dos pruebas.

Para confirmar esta hipótesis se realizó un test de Student para medidas repetidas (datos apareados) entre las

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006

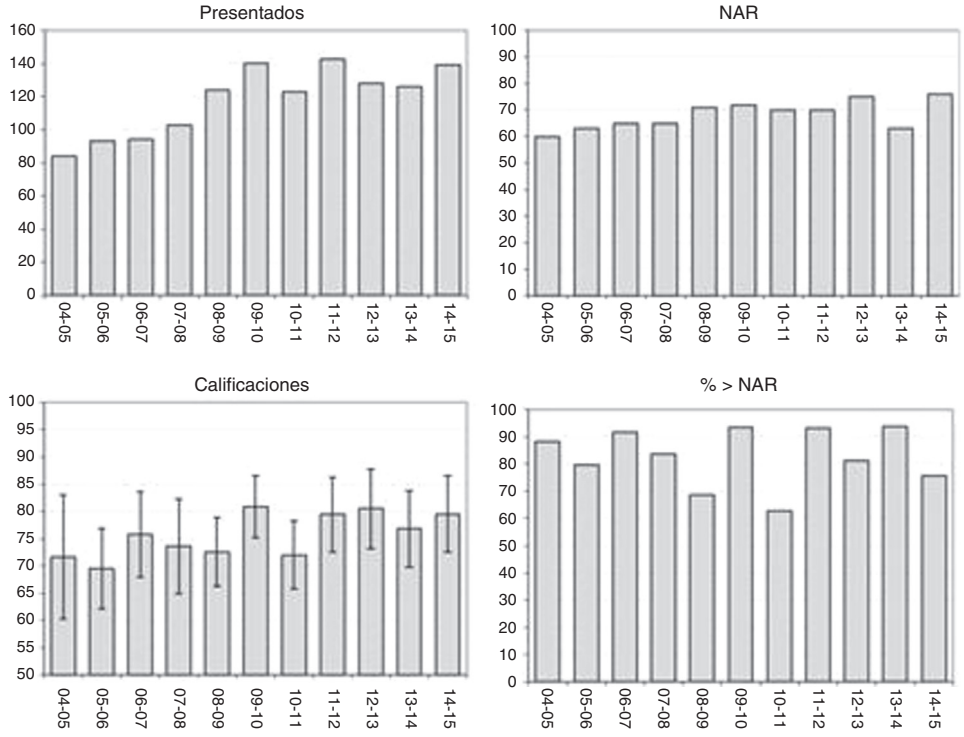

Figura 1 Gráficas del número de alumnos presentados desde el curso 04-05, del nivel aceptable de resultados de cada una de las pruebas, del número de respuestas acertadas y del tanto por ciento de alumnos que superaron el NAR en cada prueba.

calificaciones y el número de respuestas acertadas en las dos pruebas.

Se observaron diferencias estadísticamente significativas entre el número de respuestas acertadas en 2015 frente a las acertadas en 2016. Sin embargo, no se observaron diferencias significativas entre las calificaciones de 0 a 10 en los dos años.

Se calculó también la correlación entre las notas de las dos pruebas obteniéndose un coeficiente de correlación de 0,568 con una significación de $p < 0,01$.

Discusión

Nivel aceptable de resultados

Hay una correlación significativa ($p = 0,029$) entre el NAR y la media de respuestas acertadas de los exámenes, cuanto más alto el NAR mayor es la media de las respuestas correctas del examen (fig. 2). Esto es lógico ya que el NAR es una estimación del nivel para aprobar. Un NAR alto denota una mayor facilidad de las preguntas del examen.

Diferencia entre la media de las respuestas acertadas y el NAR

La diferencia entre la media de respuestas correctas en un examen y el NAR refleja de alguna forma el acierto del grupo de profesores a la hora de estimar el nivel de la prueba. Esta diferencia varió entre 13,8 y 1,5 puntos con una media de 7,5. Es evidente que esta diferencia no debe ser muy grande

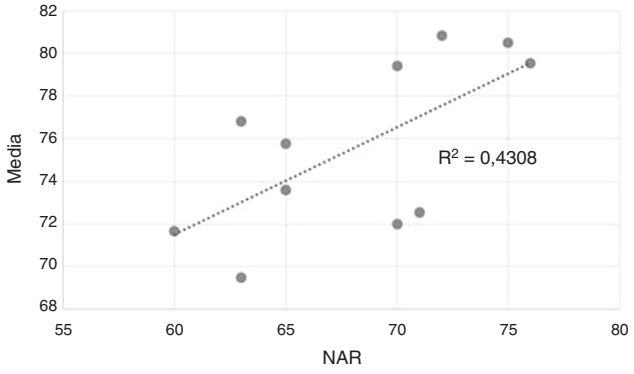

Figura 2 Correlación entre el nivel aceptable de resultados y la media de respuestas acertadas en cada prueba.

ni muy pequeña, probablemente debería estar alrededor de 7,5 que es el valor medio.

Hay una estrecha correlación muy significativa ($p < 0,001$) entre esta diferencia y el tanto por ciento de aprobados en la prueba. Cuanto mayor es la diferencia, mayor es el número de aprobados, cuando es muy alta quiere decir que el NAR de la prueba ha sido inferior al deseable (fig. 3), o bien que el grupo de alumnos es excepcionalmente bueno.

Comparación de los resultados de 2015 con los de 2016

El grupo de profesores que confeccionó el examen de 2016 no tenía experiencia previa con pruebas de elección múltiple

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006

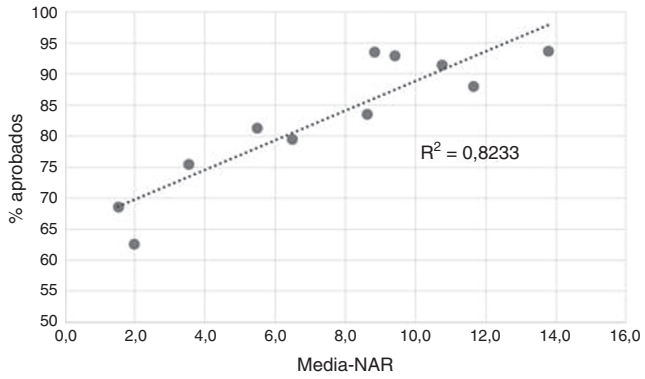

Figura 3 Correlación entre la diferencia de la media de respuestas acertadas y el nivel aceptable de resultados respecto al tanto por ciento de alumnos que superan la prueba.

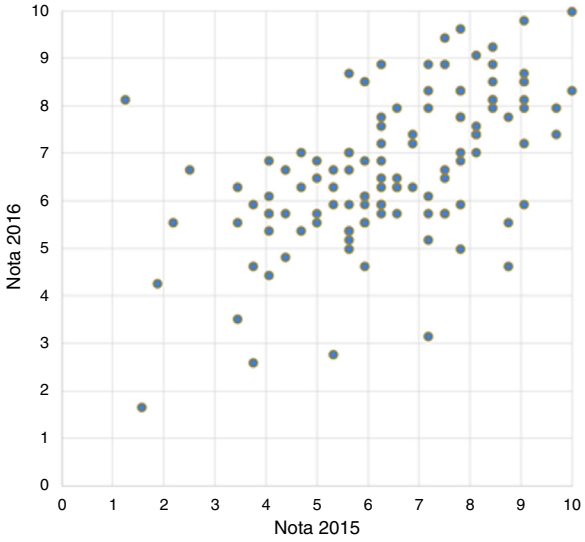

Figura 4 Correlación entre las notas obtenidas en la prueba de 2015 y las notas obtenidas en la prueba de 2016.

salvo el coordinador de la asignatura que era el mismo que en la prueba de 2015.

Como hemos comentado en el apartado de resultados, se observaron diferencias estadísticamente significativas entre el número de respuestas acertadas en 2015 frente a las acertadas en 2016. La diferencia media entre las respuestas acertadas fue de 16,5 respuestas. Este resultado era esperable ya que el NAR de 2105 era de 76 respuestas acertadas y de 55 en 2016.

Sin embargo, no se observaron diferencias significativas entre las calificaciones de 0 a 10 en los dos años, la diferencia media entre las dos pruebas fue inferior a 0,3 puntos, lo que apunta a que este método de evaluación es fiable y robusto, ya que no parece depender de la materia que se evalúa, del grupo de profesores que confecciona la prueba ni del nivel aceptable de resultados de la misma.

Además demuestra también que el ajuste de las notas por una recta normaliza los resultados, lo que permite comparar distintas pruebas aunque hayan partido de niveles aceptables diferentes.

Se comprobó además que existe una correlación estadísticamente significativa entre las calificaciones obtenidas por cada estudiante en las dos pruebas (coeficiente de correlación = 0,568 con $p < 0,01$. fig. 4).

Hemos de recordar una frase de la guía pedagógica de la OMS que dice: «Cuando varios instrumentos diferentes de medida dan un resultado concordante, a pesar de sus imperfecciones, la fiabilidad de la evaluación aumenta»[4]. Estos resultados demuestran un alto nivel de fiabilidad de nuestro sistema de evaluación.

Conclusiones

- La técnica del cálculo del nivel aceptable de resultados para una prueba PEM ha demostrado ser un método objetivo para establecer el nivel de aprobado utilizando criterios absolutos.
- Hay que realizar un esfuerzo importante a la hora del cálculo del NAR para que la diferencia entre este y la media de notas de la prueba no sea muy grande. Para una prueba de 100 preguntas esta diferencia debería estar alrededor de 7.
- El ajuste de las calificaciones utilizando una recta ha demostrado que normaliza los resultados obtenidos y permite la comparación entre pruebas.
- Este método de evaluación ha demostrado ser efectivo y sus resultados congruentes con distintas asignaturas, diferentes grupos de profesores y con diferentes niveles aceptables de resultados de las pruebas.

Conflicto de intereses

Los autores declaran no tener ningún conflicto de intereses.

Bibliografía

1. Carreño F. Enfoques y principios teóricos de la evaluación. México: Editorial Trillas; 1981.
2. Del Cañizo JF. Proyecto docente para la enseñanza de la Fisiopatología y Propedéutica Quirúrgica. Madrid: Universidad Complutense de Madrid; 2011.
3. Guilbert JJ. Guía pedagógica para el personal de salud. 5.ª edición Valladolid: Ed. O.M.S. Instituto de Ciencias de la Educación; 1989.
4. Bloom BS. Taxonomía de los objetivos de la Educación. I Ámbito del conocimiento. Alcoy: Editorial Marfil; 1975.

Cómo citar este artículo: Del Cañizo JF, et al. Fiabilidad de la técnica del cálculo del nivel aceptable de resultados en pruebas de preguntas de elección múltiple. Educ Med. 2016. http://dx.doi.org/10.1016/j.edumed.2016.03.006