Fiabilidad de un examen clínico objetivo estructurado para predecir la futura competencia profesional de estudiantes de medicina

César CHUNG, Luis NAVARRETE, Jesús FLORIDO (†), Alberto SALAMANCA, Pablo TORNE, Rafael LÓPEZ, Luis ALBENDÍN, Francisco DEL MORAL, José M. PEINADO

Introducción. Cuando con los resultados de un examen clínico objetivo estructurado (ECOE) se decide sobre la futura competencia profesional de estudiantes de medicina, la fiabilidad de dicha prueba debe adecuarse a esta finalidad.

Objetivo. Calcular la fiabilidad (alfa de Cronbach) de una serie de ECOE y su relación con la duración, número de participantes, estaciones, ítems y evaluadores.

Sujetos y métodos. Se analizan 14 ECOE realizados a 2.995 estudiantes de cuarto y quinto curso de la Facultad de Medicina de Granada desde 2004 a 2013.

Resultados. La fiabilidad fue ≥ 0,70 en el 92,84% de los ECOE. También fue significativamente ≥ 0,70 cuando la duración total fue ≥ 60 minutos (p = 0,042), el número de estaciones ≥ 10 (p = 0019), el número de ítems ≥ 50 (p = 0,018) y el número de evaluadores ≥ 6 (p = 0,018). No se observaron diferencias con el número de estudiantes ni con las opciones al ítem utilizadas.

Conclusiones. Los ECOE cuyos resultados se utilicen para aprobar asignaturas de la carrera de medicina deben tener una fiabilidad ≥ 0,70. Para alcanzar dicha fiabilidad o mayor, el formato debe constar de al menos 10 estaciones, durar ≥ 60 minutos, tener ≥ 50 ítems y ≥ 6 evaluadores.

Palabras clave. Duración. ECOE. Estaciones. Evaluadores. Fiabilidad. Ítems.

Reliability of an objective structured clinical examination to assess the future professional competence of medical students

Introduction. When the future professional competence of medical students is decided based on results of an objective structured clinical examination (OSCE), the reliability of this test should be adequate to this purpose.

Aim. To calculate the reliability (Cronbach's alpha) of each one of OSCEs we performed and its relationship with the duration, number of participants, stations, items and evaluators.

Subjects and methods. Fourteen OSCE tests performed to 2995 medical students of 4th and 5th year of the Faculty of Medicine of Granada between 2004 to 2013 were analyzed.

Results. The reliability was ≥ 0.70 in 92.84% of the OSCEs. It was also significant ≥ 0.70 with a total duration ≥ 60 minutes (p = 0.042), and a number of stations ≥ 10 (p = 0.019), a number of items ≥ 50 (p = 0.018) and a number of evaluators ≥ 6 (p = 0.018). No differences with the number of students, neither with the options to the item were observed.

Conclusions. The OSCEs carried out in centers which results are used to approve subjects of the medical career, must have a reliability ≥ 0.70. To achieve this reliability or greater, the format should consist of at least: 10 stations, a duration ≥ 60 minutes, and having ≥ 50 items and ≥ 6 evaluators.

Key words. Duration. Evaluators. Items. OSCE. Reliability. Stations.

Introducción

Desde los primeros exámenes clínicos objetivos estructurados (ECOE) realizados en España en 1994 [1], esta prueba se ha implementado en un número cada vez mayor de instituciones, facultades y asociaciones científicas [2]. Desde 2016 se ha establecido como prueba obligatoria en la mayoría de las facultades de medicina españolas antes de la graduación [3].

Departamento de Obstetricia y Ginecología. Facultad de Medicina. Universidad de Granada. Granada, España.

Correspondencia:
Prof. César Chung. Departamento de Obstetricia y Ginecología. Facultad de Medicina. Universidad de Granada. Avenida de las Ciencias. Parque Tecnológico. E-18016 Granada.

E-mail:
chung@ugr.es

Agradecimientos:
A todos los profesores y residentes de los departamentos de Obstetricia y Ginecología, de Cirugía y de las asignaturas de oftalmología, medicina, psiquiatría, otorrinolaringología y dermatología, que intervinieron como evaluadores en los ECOE; al personal de secretaría y técnico de los departamentos de Obstetricia y Ginecología y de Cirugía, por su colaboración en el montaje de los escenarios.

Recibido:
27.02.19.

Aceptado:
14.03.19.

Conflicto de intereses:
No declarado.

Competing interests:
None declared.

ISSN: 2014-9832
ISSN (ed. digital): 2014-9840

Tabla I. Distribución de ECOE por años, curso, asignatura, número de alumnos, nota media (± desviación estándar) y porcentaje de aprobados.

	Curso	Asignaturas	N.º de alumnos	Nota media	Aprobados
Año 2004	4.º	Ginecología	201	66,92 ± 10,39	86,10%
	5.º	Obstetricia	154	65,98 ± 9,27	85,70%
Año 2005	4.º	Ginecología	214	67,51 ± 9,44	84,57%
	5.º	Obstetricia	183	70,84 ± 13,34	84,15%
Año 2006	4.º	Obstetricia y ginecología	208	65,84 ± 8,26	83,17%
	5.º	Obstetricia	209	67,69 ± 11,13	84,02%
Año 2007	4.º	Conjunta: medicina, psiquiatría, cirugía, oftalmología, y obstetricia y ginecología	196	66,69 ± 9,36	85,71%
	5.º	Conjunta: medicina, dermatología, oftalmología, cirugía y otorrinolaringología	195	59,72 ± 7,94	85,12%
Año 2008	4.º	Conjunta: medicina, psiquiatría, cirugía, oftalmología, y obstetricia y ginecología	223	59,87 ± 12,92	84,75%
Año 2009	4.º	Obstetricia y ginecología	211	59,77 ± 8,77	85,30%
Año 2010	4.º	Obstetricia, ginecología y cirugía	229	71,37 ± 7,48	85,58%
Año 2011	4.º	Obstetricia, ginecología y cirugía	217	73,71 ± 6,92	84,03%
Año 2012	4.º	Obstetricia, ginecología y cirugía	248	73,05 ± 6,65	88,30%
Año 2013	4.º	Obstetricia, ginecología y cirugía	267	72,44 ± 6,64	87,26%
Total			2.995	67,24 ± 9,18	

El ECOE puede utilizarse con fines diferentes y las decisiones a tomar con las calificaciones obtenidas deben adecuarse a ellas [4]. Cuanto más importante sean sus consecuencias, mayor debe ser su fiabilidad. Si un ECOE no resulta fiable, no lo será tampoco la calificación que se otorgue a los participantes [5].

El formato original descrito por Harden y Gleeson [6] ha experimentado numerosas modificaciones. La mayoría de estos formatos carece de estudios sobre su fiabilidad que avalen la conveniencia de utilizar uno u otro con el fin de tomar decisiones responsables.

El objetivo de este estudio es analizar la fiabilidad de diversos formatos de ECOE y su relación con el número de estudiantes, estaciones y evaluadores, su duración, número de ítems y sus opciones de respuesta.

Sujetos y métodos

Se analizan 14 ECOE realizados desde 2004 a 2013 a un total de 2.955 alumnos matriculados en cuarto o quinto curso en la Facultad de Medicina de Granada. De ellos, siete ECOE fueron solo sobre obstetricia y ginecología; cuatro, conjuntos entre ginecología, obstetricia y cirugía; dos, conjuntos con todas las asignaturas de cuarto curso (medicina interna, psiquiatría, oftalmología, ginecología y obstetricia y cirugía), y uno, conjunto con todas las asignaturas de quinto curso (medicina interna, cirugía, dermatología y otorrinolaringología) (Tabla I). La nota obtenida constituyó el 10-20% de la nota final para aprobar las asignaturas. Para los cálculos estadísticos se utilizó el programa SPSS v. 20 [7,8] y se empleó, tanto para los ítems politómicos como dicotómicos, el α de Cronbach y la fórmula KR20 de Kuder-Robinson [9], considerando que ambos coeficientes tienen equivalencia matemática [10] y que es válido utilizar el comando del SPSS para computar el α de Cronbach en variables dicotómicas.

El número de alumnos por ECOE osciló entre 154 y 229. La nota media total fue de 67,24, con un porcentaje de aprobados del 84,02% o mayor (Tabla I).

Las pruebas se realizaron en grandes aulas donde las estaciones con simuladores y evaluadores, así como las pictoriales o 'de silla', se separaban con biombos. Los pacientes simulados estaban en consultas contiguas.

El número de estaciones osciló entre 8 y 11, con una duración individual de 4-8 minutos y una duración total de 33-64 minutos. Se realizaron dos rondas simultáneas en horario de mañana y tarde durante 4-5 días. El número de evaluadores, incluyendo los pacientes estandarizados, fue de 2 a 10, y el número total de ítems, entre 26 y 142. Dichos ítems se calificaron asignando a cada uno un valor numérico. Cuando las estaciones eran pictoriales o 'de silla', los alumnos contestaban a preguntas tipo test, que luego en la hoja de cálculo se consignaban como ítems dicotómicos (explícitos) de dos opciones: 0, cuando no era correcta, o 1, cuando lo era. En el resto de las estaciones, en las que intervenían evaluadores, las opciones fueron politómicas y oscilaban entre 3 y 10.

Durante los primeros años se utilizaron simultáneamente en una misma prueba varias opciones politómicas que variaban entre 3 y 10, pero en los últimos seis años, en dos ECOE se utilizaron cinco opciones, y en los últimos cuatro años, tres opciones. La opción 0 reflejaba la calificación más baja (p. ej., 'no realiza') y el número mayor, la más alta ('realiza adecuadamente') (Tabla II).

Tabla II. Número de alumnos, estaciones, duración individual y total, ítems, opciones, evaluadores y α de Cronbach.

	Curso	Alumnos	Estaciones	Duración de la estación (h)	Duración total (h)	Items totales	Opciones del ítem	Evaluadores	α de Cronbach
Año 2004	4.º	201	10	4	40	36	3 a 10	3	0,742
	5.º	154	11	4	33	40	3 a 10	3	0,680
Año 2005	4.º	214	8	8	64	26	3 a 10	2	0,726
	5.º	183	8	7	56	28	3 a 10	4	0,716
Año 2006	4.º	208	9	6	54	29	3 a 10	4	0,705
	5.º	209	9	7	63	33	3 a 10	4	0,713
Año 2007	4.º	196	10	6	60	65	3 a 10	7	0,834
	5.º	195	9	6	54	47	3 a 10	5	0,788
Año 2008	4.º	223	10	6	60	56	5	8	0,890
Año 2009	4.º	211	9	6	54	51	5	6	0,768
Año 2010	4.º	229	10	6	60	103	3	9	0,855
Año 2011	4.º	217	10	6	60	97	3	10	0,824
Año 2012	4.º	248	10	6	60	133	3	10	0,782
Año 2013	4.º	267	11	6	60	142	3	10	0,834

Resultados

Fiabilidad (consistencia interna) de todos los ítems de un ECOE

En la tabla II se observa que el α de Cronbach de los ECOE realizados osciló entre 0,680 y 0,890, y se aprecia una correlación significativa con el año de su realización (coeficiente de Pearson: $r = 0,624$; $p = 0,0017$).

Considerando esta fiabilidad en cuatro grupos, ningún ECOE obtuvo una fiabilidad ≥ 0,90; el 35,7% ($n = 5$) obtuvo fiabilidades entre 0,80 y 0,89; el 57,14% ($n = 8$), de 0,70 a 0,79, y solo el 7,14% ($n = 1$) obtuvo una fiabilidad entre 0,60 y 0,69 (Tabla II).

Duración total de la prueba

El número de estaciones y su duración individual condiciona la duración total de la prueba. De esta manera, ocho ECOE (57,1%) duraron más de 60 minutos, y seis (42,9%), menos. Cuando la duración total del ECOE fue ≥ 60 minutos, la proporción de ECOE con una fiabilidad ≥ 0,70 fue del 57,1%, significativamente mayor que cuando fue menor ($\chi^2 = 6,344$; $p = 0,042$) (Tabla III).

Número de estaciones

En nuestro estudio, el número de estaciones fue de ocho en dos ocasiones; nueve, en cuatro; diez, en seis, y once, en dos ocasiones. Al analizar la fiabilidad según el número total de estaciones se observa que, cuando el número de estaciones fue ≥ 10, la proporción de ECOE con una fiabilidad > 0,80 fue significativamente mayor cuando el número de estaciones era menor ($\chi^2 = 0,785$; $p = 0,019$) (Tabla III).

Número de ítems

Al analizar la fiabilidad del ECOE con el número de ítems, se observa que existe una correlación significativa con un mayor número de ítems (coeficiente de Pearson: $r = 0,577$; $p = 0,031$). Igualmente, cuando el número de ítems fue ≥ 50, la proporción de ECOE con una fiabilidad ≥ 0,70 fue significativamente mayor ($\chi^2 = 8,000$; $p = 0,018$) (Tabla III).

Figura 1. Correlación entre fiabilidad (α de Chronbach) y número de evaluadores (correlación de Pearson: 0,777; p = 0,001).

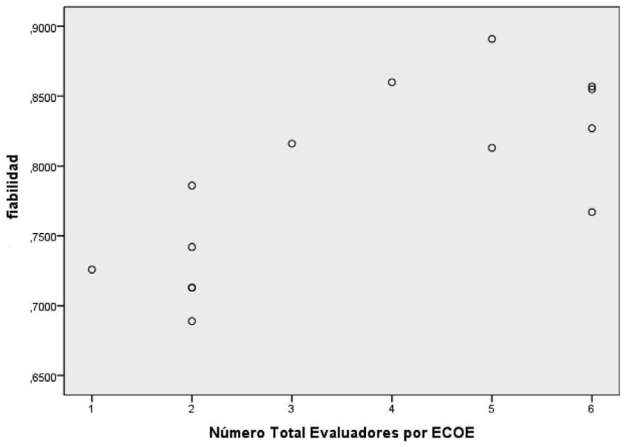

Tabla III. Fiabilidad y duración total, número de estaciones, ítems y evaluadores.

		α de Chronbach				
		0,60-0,69	0,70-0,79	0,80-0,89	Total	χ^2
Duración total	≥ 60 min	0	3	5	8 (57,1%)	6,344 (p = 0,042)
	< 60 min	1	5	0	6 (42,9%)	
	Total	1 (7,1%)	8 (57,1%)	5 (35,7%)	14	
Nº de estaciones	≥ 10	1	2	5	8 (57,1%)	0,785 (p = 0,019)
	< 10	0	6	0	6 (42,9%)	
	Total	1 (7,1%)	8 (57,1%)	5 (35,7%)	14	
N.º de ítems	≥ 50	0	2	5	7 (50%)	8,000 (p = 0,018)
	< 50	1	6	0	7 (50%)	
	Total	1 (7,1%)	8 (57,1%)	5 (35,7%)	14	
N.º de evaluadores	≥ 6	0	2	5	7 (50%)	8,000 (p = 0,018)
	< 6	1	6	0	7 (50%)	
	Total	1 (7,1%)	8 (57,1%)	5 (35,7%)	14	

Número de evaluadores

El número de evaluadores por ECOE, incluyendo a los pacientes simulados, osciló entre 2 y 10. La fiabilidad se correlacionó de forma significativa con un mayor número de evaluadores (coeficiente de Pearson: r = 0,777; p = 0,001) (Fig. 1). Cuando el número de evaluadores fue ≥ 6, la fiabilidad fue significativamente > 0,70 que cuando el número fue menor (χ^2 = 8,000; p = 0,018) (Tabla III).

Número de participantes

El número total de participantes por ECOE osciló entre 154 y 267. Se observa una correlación no significativa entre el número de participantes y la fiabilidad alcanzada (r = 0,507; p = 0,064). Interpretamos que se debe a un número parecido de participantes por ECOE.

Opciones del ítem

En nuestro estudio, en ocho ECOE utilizamos simultáneamente ítems dicotómicos y politómicos, al que denominamos grupo variable; en cuatro ECOE, todos los ítems fueron de tres opciones, y en dos, de cinco opciones. Al analizar la fiabilidad media de acuerdo con el número de opciones al ítem utilizado se observa que ésta es mayor, aunque no de forma estadísticamente significativa (p = 0,289) cuando se utilizaron cinco opciones que cuando fueron tres opciones o variable (3, 5, 7, 9 o 10 opciones) (Tabla IV; Fig. 2).

Discusión

Con la introducción del ECOE, se ha cumplido el principio que dice que, 'la forma de evaluación determina la forma en la que el estudiante aprende' [11], con ello se ha conseguido que los estudiantes se preocupen en adquirir las competencias de las cuales van a ser evaluados. Igualmente, también ha supuesto un cambio en la metodología docente en las facultades.

Fiabilidad

Cuanto mayor sea el número de participantes, de estaciones, así como de ítems y evaluadores, los resultados del ECOE son más fiables. Esto conlleva más tiempo y dinero, que no toda facultad o departamento puede soportar. El ECOE se ha puesto de moda y el formato utilizado depende del presupuesto, propiciando pruebas con menos recursos y menor fiabilidad.

Si los resultados se utilizaran para aprobar, se debe averiguar si el formato utilizado es fiable para evitar perjudicar a los estudiantes. Los niveles de fiabilidad (α de Cronbach) de un ECOE deben estar

de acuerdo con el nivel de la finalidad para la cual se ha realizado. En nuestra opinión, se pueden considerar cuatro grupos de ECOE: aquellos con una fiabilidad ≥ 0,90**,** cuando vamos a otorgar el título de médico o de una especialidad. En segundo lugar, pruebas con una fiabilidad entre 0,80 a 0,89**,** como superar un año de formación universitaria. En tercer lugar, pruebas con un rango entre 0,70 y 0,79**,** como puede ser la superación de una asignatura. Y, por último, pruebas con una fiabilidad entre 0,60 y 0,69, en el caso de pruebas cortas sobre un desempeño concreto en talleres de simulación. Los exámenes con una fiabilidad inferior a 0,60 no son apropiados para tomar decisiones con sus puntuaciones. Los valores por encima de 0,95 se consideran excesivos y no siempre se alcanzan porque seguramente se trate de ítems demasiado fáciles [12]. Los diferentes formatos utilizados en este estudio, y cuya calificación representaba el 10-20% de la nota final de las diferentes asignaturas participantes, fueron todos fiables para el fin propuesto, excepto uno, que obtuvo una fiabilidad inferior a 0,69.

Duración total y número de estaciones del ECOE

La duración total de un ECOE depende del número de estaciones y el tiempo asignado a cada una. Cuando se utilizan pocas estaciones, podemos comprobar que 'una conducta concreta es médicamente competente'; sin embargo, no podemos predecir una actuación competente de ese candidato en situaciones diferentes, ya que las pruebas con pocas estaciones son poco fiables. Para obtener una fiabilidad aceptable se necesita realizar muchas tareas, casos o situaciones diferentes, lo que condiciona pruebas de larga duración. Según Arnau-Figueras y Martínez-Carretero [12], el ECOE para llegar a una fiabilidad aceptable debe durar como mínimo 3-5 horas y tener al menos 15 estaciones. Sin embargo, no es lo mismo un ECOE de 20 estaciones con una duración de 6 minutos por estación que otro con menos estaciones de 10 minutos. Hawkins [13] opina que las estaciones más cortas son tan fiables como las más largas, y se deben preferir dependiendo de los medios y el tiempo total asignado para la prueba. Igualmente, el tiempo asignado a cada estación debe ser adecuado a la tarea profesional que se solicita, obteniéndose resultados más fiables cuando las tareas evaluadas son básicas y no complejas [14]. Para ECOE con una duración total fija previamente establecida, debe preferirse la inclusión de más estaciones que permitan la evaluación de competencias básicas en lugar de alargar el tiempo de cada estación añadiendo ítems no pertinentes [15].

Figura 2. Fiabilidad media y opciones al ítem (estadístico de Levene: 1,395; p = 0,289).

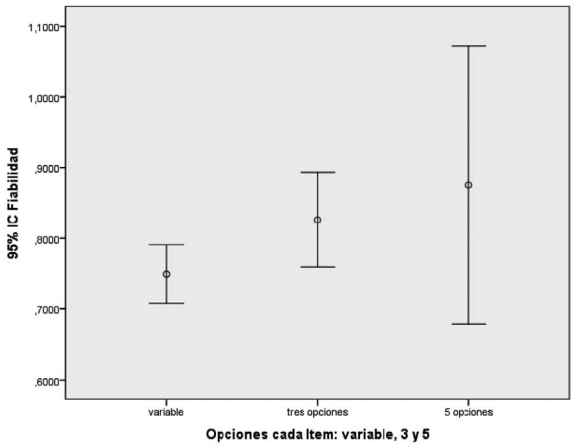

Tabla IV. Opciones al ítem y fiabilidad media (estadístico de Levene: 1,393; p = 0,289).

Opciones	N.º de ECOE	Fiabilidad media	Desviación típica	Error típico
Variable	8	0,749750	0,0488723	0,0172790
Tres	4	0,826500	0,0419643	0,0209821
Cinco	2	0,875500	0,0219203	0,0155000
Total	14	0,789643	0,0652753	0,0174456

En general podemos establecer que, para mejorar la fiabilidad de las puntuaciones, lo mejor es aumentar el número de estaciones.

Si un ECOE pretende crear un escenario parecido a la realidad –al representar, por ejemplo, la consulta de un centro de salud–, la duración de una estación no debería superar el tiempo que realmente disponen los médicos (6-8 minutos). Se necesitan entre 2 y 12 horas de duración para obtener un coeficiente generalizado de 0,80 [16]. Consideramos que, para ECOE conjuntos de todas las asignaturas de un curso de medicina o de una asignatura, con una duración total de al menos 60 minutos, se puede obtener una fiabilidad superior a 0,70.

Pero, ¿qué número mínimo de estaciones es necesario para alcanzar una fiabilidad aceptable? En el caso de tareas profesionales básicas, Stillman et al [17] consideran que son necesarias entre 6 y 10 estaciones. Para evaluaciones más complejas, con mayor responsabilidad [11], se precisan al menos 25 estaciones. Siguiendo la fórmula de Spearman-Brown [18], calculan una fiabilidad total de 0,70 para un

ECOE de seis estaciones y de 0,80 si se añaden cinco estaciones más, y sugieren que en un ECOE debe haber al menos 10 estaciones. Consideramos que un ECOE para un curso o asignatura de medicina con una fiabilidad superior a 0,70 precisa al menos 10 estaciones. Desde 2016 [3] se ha establecido el ECOE como una prueba obligatoria en la mayoría de las facultades de medicina españolas y el formato debe constar de al menos 20 estaciones y durar más de cuatro horas.

Número de ítems

La fiabilidad de un ECOE también se debe calcular sobre el número de ítems. La fiabilidad de una prueba se incrementa a medida que aumentan los ítems. Ello no significa que cualquier prueba pueda convertirse en un instrumento fiable simplemente aumentando el número de ítems. Es relativamente fácil colocar muchos ítems dicotómicos en una estación sin que se modifique en gran medida el tiempo de examen, pero se deben utilizar ítems pertinentes sobre el desempeño de la actividad a desarrollar en cada estación. Consideramos, según nuestro estudio, que para obtener una fiabilidad > 0,70, un ECOE debe tener al menos 50 ítems.

Número de participantes

La fiabilidad de un test también depende del tamaño y tipo de la muestra de sujetos a los que se aplique, por lo que un mismo test no tiene un coeficiente de fiabilidad fijo y puede variar si se aplica a otra muestra. Un gran número de participantes minimiza los posibles errores debidos al azar. Consideramos que si el número de participantes se encuentra entre 150 a 260, unido a un número de estaciones mayor de 10 y más de 50 ítems, la fiabilidad es mayor de 0,70.

Opciones de cada ítem

El desempeño de los participantes durante el ECOE se califica asignando a los ítems un valor numérico. Esta valoración se suele realizar de forma explícita, cumplimentando una lista de verificación dicotómica que informa si se realiza o no la tarea según un determinado protocolo. Con la calificación dicotómica, nos limitamos a comprobar si el candidato sigue los pasos para realizar la actividad requerida, no se juzga la calidad del proceso. La otra forma de calificar es la implícita, menos utilizada; en lla se emite un juicio sobre la calidad del proceso utilizando ítems politómicos, tipo Likert, en los cuales la respuesta presenta un escalonamiento con números impares con valores entre 3, 7, 9 y 11 opciones. En esta escala, el número 1 suele reflejar la categoría más baja, correspondiente a la etiqueta de 'muy en desacuerdo' o 'no realiza', y la categoría más alta como 'realiza adecuadamente'. Según Bond y Fox [19], en lugar del número 1, como es costumbre en el formato tipo Likert, no existe ningún problema en utilizar el número 0, ya que los números ejercen su función solo nominalmente y además el 0 indica, mejor que cualquier otra etiqueta, la ausencia/disconformidad total del atributo. Cuando en un ítem aumenta el número de alternativas, se recoge una mayor variabilidad en las puntuaciones, y cuanto más objetiva sea la puntuación, más alta será la fiabilidad [20].

Determinar cuál es el número de opciones a un ítem suficientes para dar un informe fiable que discrimine a los candidatos es un punto sobre el que se ha investigado pero no hay conclusiones definitivas [21]. La utilización de cinco opciones disminuye los errores debidos a la clemencia de los evaluadores porque solo hay una apreciación desfavorable, tales como: malo-mediano-bueno-muy bueno-excelente. Cuando son tres las alternativas, la correcta y dos alternativas incorrectas, se discrimina e informa mejor en el centro de la distribución: los mejores y los peores alumnos quedan mejor diferenciados entre sí en la parte alta y baja, respectivamente [21]; tres es el número óptimo de respuestas. Cuando se usan dos alternativas, se discrimina mejor en la parte alta a los candidatos que lo hacen adecuadamente, mientras que en la parte inferior se puede englobar tanto a los que no lo hacen adecuadamente como a los que simplemente no hacen nada.

La fiabilidad esperada para 100 ítems dicotómicos, de tres alternativas y de cinco, tiende a subir al aumentar el número de alternativas, pero a partir de ahí el incremento de la fiabilidad es mínimo e insignificante [20,22,23]. En un estudio sobre tres y cinco alternativas, Owen y Froman [24] llegan a la conclusión de que la única diferencia importante en utilizar tres opciones reside en el ahorro de tiempo. En nuestro estudio, la fiabilidad media entre ECOE con opciones al ítem variables (3, 5, 7, 9 y 11), tres alternativas o cinco, no fue significativamente mayor. Al no encontrar diferencias, aconsejamos las tres opciones porque, además de facilitar el aumento de fiabilidad si incrementamos el número de ítems, supone un menor tiempo para elegir la opción adecuada por parte del evaluador, lo cual puede ser oportuno en un ECOE de larga duración.

Evaluadores

El juicio para evaluar la competencia de una persona es complejo y siempre puede haber riesgo de error. Por diversos motivos, ningún evaluador de un ECOE puede asegurar que todas sus calificaciones están libres de error humano. Un examen puede ser poco fiable a pesar de ser perfectamente objetivo si sus resultados se basan en la objetividad de pocos evaluadores. Por ello, la evaluación conjunta realizada por un grupo de evaluadores es menos susceptible de ser errónea que la hecha por uno solo. La única manera de confiar en sus estimaciones es aumentar su número, así como el de estaciones, hasta poder confiar en que nos aproximamos a la puntuación auténtica del candidato. Si el número de evaluadores y de estaciones es suficientemente grande, la cantidad de errores debidos a la subjetividad de los evaluadores puede compensarse y, por tanto, el examen puede resultar fiable. En nuestro estudio, el número de evaluadores ha ido aumentando con los años. Contando como evaluadores a los pacientes simulados, establecemos que para una fiabilidad > 0,70 el número de evaluadores debe ser mayor de seis.

En conclusión, para realizar ECOE a muestras de estudiantes de medicina con una fiabilidad ≥ 0,70, que permita tomar decisiones importantes con las calificaciones obtenidas, el formato utilizado debe tener un mínimo de 10 estaciones, una duración total de al menos 60 minutos y contar con al menos 50 ítems y seis evaluadores. Si el número de participantes es menor de 150, se debe aumentar el número de estaciones, ítems y observadores. Para evitar errores inevitables por cansancio o rutina, se aconsejan utilizar ítems de tres opciones.

Bibliografía

1. Martínez-Carretero JM. Los métodos de evaluación de la competencia profesional: la evaluación clínica objetiva estructurada (ECOE). Educ Med 2005; 8: 18-22.
2. Ruiz E, Florensa E, Cots JM, Sellarés J, Iruela A, Blay C, et al, Primeras experiencias en evaluación de la competencia clínica de los médicos de familia de Catalunya. Aten Primaria 2001; 28: 105-9.
3. Conferencia Nacional de Decanos de Facultades de Medicina. Prueba nacional de habilidades. ECOE-CNDFME (7 de abril de 2016). URL: http://www.cndmedicina.com/wp-content/uploads/2017/01/07-04-2016_Documento-ECOE-nacional.pdf.
4. Downing S. Reliability on the reproducibility of assessment data. Med Educ 2004; 38: 1006-12.
5. Subkoviak MJ. A practitioner's guide to computation and interpretation of reliability indices for mastery test. Journal of Educational Measurement 1988; 25: 47-55.
6. Harden R, Glesson F. Assessment of clinical competence using an objective structured clinical examination (OSCE). Med Educ 1979; 13: 41-54.
7. Chung C ¿Debemos aceptar los resultados de un ECOE? Fiabilidad, validez y viabilidad en la simulación clínica. Granada: Apoyo Docente CHG; 2015.
8. Camacho J. Fiabilidad (*reliability*). In Camacho J. Estadística con SPSS versión 9 para Windows. Madrid: RA-MA; 2000.
9. Kuder GF, Richardson MW. The theory of the estimation of test reliability. Psychometrika 1937; 2: 151-60.
10. Hebson RK. Understanding internal consistency reliability estimates: a conceptual primer on coefficient alpha. Meas Eval Couns Dev 2001; 34: 177-89.
11. Newble DI, Jaeger K. The effects of assessments and examinations on the learning of medical students. Med Educ 1983; 17: 165-71.
12. Arnau-Figueras J, Martínez-Carretero JM. Comparativa de instrumentos de evaluación de la competencia. Evaluación de la competencia clínica: análisis comparativo de dos instrumentos (ECOE versus portafolio). Barcelona: AATRM; 2007.
13. Hawkins R, Boulet J. Direct observation: standardized patient. In Holmboe E, Hawkins R, eds. Practical guide to the evaluation of clinical competence. Philadelphia: Mosby Elsevier; 2008.
14. Chambers KA, Boulet JR, Gary N. The management of patient encounter time in a high-stakes assessment using standardized patients. Med Educ 2000; 34: 813-7.
15. Shatzer JH, Darosa D, Colliver JA, Barkmeier L. Station-length requirements for reliable performance-based examination scores. Acad Med 1993; 68: 224-9.
16. Van der Vleuten C, Swanson DB. Assessment of clinical skills with standardized patients: state of the art. Teach Learn Med 1990; 2: 58-76.
17. Stillman PL, Swanson DB, Smee S, Stillman AE, Ebert TH, Emmel VS, et al. Assessing clinical skills of residents with standardized patients. Ann Intern Med 1986; 105: 762-71.
18. Hodges B, Mcllroy JH. Analytic global OSCE ratings are sensitive to level of training. Med Educ 2003; 37: 1012-6.
19. Bond TG, Fox CH. Applying the Rasch model: fundamental measurement in the human sciences. 3 ed. New York: Routledge; 2015.
20. Bruno JE, Dirkwager A. Determining the optimal number of alternatives to a multiple-choice test item: an information theoretic perspective. Educ Psychol Meas 1995; 55: 959-66.
21. Morales P. Las pruebas objetivas: normas, modalidades y cuestiones discutidas. Madrid: Universidad Pontificia Comillas; 2006. URL: http://www.salgadoanoni.cl/wordpressjs/wp-content/uploads/2015/09/Univ.-Comillas_Pruebas-objetivas.pdf.
22. Haladyna TM, Downing SM, Rodriguez MC. A review of multiple-choice item-writing guidelines for classroom assessment. Applied Measurement in Education 2002; 15: 309-34.
23. Rodríguez MC. Three options are optimal for multiple-choice items: a meta-analysis of 80 years of research. Educational Measurement: Issues and Practice 2005; 24: 3-13.
24. Owen S, Froman R. What's wrong with three-option multiple choice items? Educ Psychol Meas 1987; 47: 513-22.