anales de psicología, 2016, vol. 32, nº 2 (mayo), 484-491
http://dx.doi.org/10.6018/analesps.32.2.212261

ISSN edición impresa: 0212-9728. ISSN edición web (http://revistas.um.es/analesps): 1695-2294

Aprendizaje de las competencias de investigación en el Grado de Medicina: análisis y evaluación de las calificaciones de los estudiantes en el Trabajo de Final de Grado

Luis González-de Paz[1,2], Marta Elorduy-Vaquero[1], Montserrat Virumbrales[1], Jordi Real[1,3], Xisca Sureda[1,4,5], Alicia Borrás-Santos[1,6,7], José Mª Martínez-Sánchez[1,4,5]

[1] *Facultad de Medicina y Ciencias de la Salud. Universitat Internacional de Catalunya, Sant Cugat del Vallés, España.*
[2] *CAP Les Corts. Grupo Transversal en Investigación en Atención Primaria, Institut d'Investigació Biomèdica August Pi i Sunyer (IDIBAPS), Barcelona (España).*
[3] *Institut d'Investigació en Atenció Primària Jordi Gol (IDIAP Jordi Gol), Barcelona (España).*
[4] *Unitat de Control del Tabaquisme, Programa de Prevenció i Control del Càncer, Institut Català d'Oncologia - ICO, Hospitalet del Llobregat (España).*
[5] *Grup de Prevenció i Control del Càncer, Institut d'Investigació Biomèdica de Bellvitge – IDIBELL, Hospitalet del Llobregat (España).*
[6] *Centre de Recerca en Epidemiologia Ambiental (CREAL), Barcelona (España).*
[7] *Centro de Investigación Biomédica en Red Enfermedades Respiratorias (CIBERES). Madrid (España).*

Resumen: La adaptación al Plan Bolonia del Grado de Medicina en España ha introducido en los planes de estudio la realización del Trabajo de Final de Grado (TFG). A continuación se presenta un modelo de TFG y se estudia la validez de un sistema de evaluación de las calificaciones para medir los resultados del aprendizaje de las competencias de investigación y pensamiento crítico.
Se evaluaron 62 trabajos. Para su realización los estudiantes contaron con un equipo de tutores expertos en investigación clínica y en metodología. El trabajo fue evaluado por tres profesores, cada uno completó una rúbrica de 10 ítems. La validez del sistema de evaluación se analizó mediante el modelo de múltiples facetas de Rasch, que permite estudiar las diferencias en la severidad de los evaluadores, los resultados de los estudiantes y la dificultad de los ítems. Se calificaron los trabajos de 62 estudiantes, totalizando 186 rúbricas. El 58% de los estudiantes mostraron una habilidad superior a la severidad de los evaluadores y a la dificultad de los ítems de la rúbrica. No se observaron diferencias significativas en la severidad de los evaluadores, ni en las calificaciones finales según el tipo de evaluador. El ítem más difícil para los estudiantes fue la selección del diseño epidemiológico del protocolo. El método de análisis propuesto para el TFG es eficaz para evaluar la adquisición de las competencias relacionadas con la investigación y el pensamiento crítico en los estudiantes del último curso del Grado de Medicina.
Palabras clave: Grado de Medicina, Currículo, Educación Médica, Competencia Profesional, Estándares Médicos, Europa, España.

Title: Learning to perform research in the degree of medicine: analysis and evaluation of the student ratings of the final degree dissertation.
Abstract: The Bologna Plan applied to the Medicine Degree in Spain has included a final degree dissertation into the coursework. In this manuscript, we analyzed the validity of a rubric and the grading criteria used to assess the students skills in research and their critical thinking.
A total of 62 final degree dissertations were evaluated. Each student was supervised by a clinical or epidemiologist researcher. Each dissertation was rated by three academic assistants using a 10-item rubric. The validity of the scores and the differences in the severity of the examiners were analyzed with a multi-faceted Rasch model, which allowed examining severity within evaluators, ability of the students and difficulty of items. 62 final dissertation were evaluated with a total of 186 rubrics. 58% of the students had higher ability than the severity of the evaluators and the difficulty of the items of the evaluation rubric. No significant differences were observed within the severity of the professors, neither within the final scores. The most difficult item for students was selection of epidemiological design. This study showed that the methodology proposed to evaluate the final degree dissertation is effective to assess the skills of conducting research and critical thinking in medicine students.
Keywords: Degree of Medicine, Curriculum, Medical Education, Education of Medical standards, Professional Competence, Europe, Spain.

Introducción

En el año 1999, con la declaración de Bolonia, comenzó el proceso de construcción del Espacio Europeo de Educación Superior (EEES) (Declaración Conjunta de los Ministros Europeos de Enseñanza, 1999; Nogales Espert, 2010) mediante 12 objetivos que buscan dinamizar la función de las universidades europeas en el contexto global (Lobato, Lagares, Alén, y Alday, 2010). En España, el denominado Plan Bolonia ha transformado las antiguas diplomaturas y licenciaturas en los nuevos grados universitarios. En la carrera de Medicina, las reformas han modificado profundamente la estructura de los planes de estudio; destacan los cambios en los criterios de evaluación de los procesos de aprendizaje, la adaptación de los contenidos al sistema de medida de los nuevos créditos (el "European Credit Transfer System", ECTS), la estimulación de la movilidad de los estudiantes entre los diferentes países del EEES, y la mejora de la educación en investigación médica (Arnalich Fernández, 2010).

Una de las diferencias del proceso de adaptación al EEES en la carrera de medicina es que los estudiantes pueden acceder al programa de doctorado al finalizar el grado, sin necesidad de realizar un máster universitario oficial (Real decreto 99/2011). Esta característica requiere que los planes de estudio incorporen procesos de aprendizaje eficaces relacionados con las competencias de pensamiento crítico e investigación (Prat et al., 2004). La evaluación de estas competencias se puede realizar mediante actividades como el Trabajo de Final de Grado (TFG), que es otra de las novedades normativas introducidas a partir del Plan Bolonia. Sin embargo, el marco normativo en España permite a cada Facultad de Medicina establecer el contenido docente del TFG y el método de evaluación, y en la actualidad no hay un modelo estándar (Real decreto 1393/2007; ANECA, 2005).

En la Facultad de Medicina y Ciencias de la Salud de la Universidad Internacional de Cataluña, el TFG es una asignatura con una carga lectiva de 8 créditos ECTS. La finalidad es que los estudiantes consoliden las competencias relacionadas con la investigación, aprendidas a lo largo de todo el

*** Dirección para correspondencia [Correspondence address]:**
Dr. José M. Martínez Sánchez. Facultad de Medicina y Ciencias de la Salud. Universitat Internacional de Catalunya, Sant Cugat del Vallés, Barcelona (España). E-mail: jmmartinez@uic.es

grado. En la asignatura el alumno tiene que desarrollar un protocolo o proyecto original de investigación clínica o de salud pública, a partir de una pregunta de investigación generada por él mismo. El formato de la memoria escrita del TFG que se ha de presentar es similar al requerido por los proyectos candidatos al Fondo de Investigación Sanitaria. Cuatro profesores de la Facultad son los responsables de guiar a cada alumno mediante una serie de tutorías. La evaluación de la memoria escrita del TFG se realiza mediante una rúbrica calificada por tres evaluadores: el tutor del alumno, un profesor del Departamento -también tutor del TFG-, y un experto clínico en la temática del TFG de los Hospitales Universitarios con los que colabora la UIC.

La evaluación mediante rúbricas compuestas de ítems está ampliamente establecida en la evaluación de resultados educativos. El análisis se realiza habitualmente con el promedio ponderado de las puntuaciones de cada evaluador. Sin embargo, este método no tiene en cuenta elementos de los que puede depender la calificación final del alumno, como la severidad o benevolencia de los evaluadores, la dificultad propia de cada uno de los ítems que componen la rúbrica, y/o la habilidad de los estudiantes (Eckes, 2011; Iramaneerat, Yudkowsky, Myford, y Downing, 2008). En algunos casos, la variabilidad interna de los elementos, especialmente cuando se refieren a las calificaciones procedentes de los evaluadores más severos o más benevolentes, puede llegar a comprometer la totalidad del sistema de evaluación (Lunz, Wright, y Linacre, 1990). Un sistema de análisis eficaz para detectar y cuantificar la variabilidad en la severidad de los evaluadores y la competencia de los estudiantes es el Modelo de Múltiples Facetas Rasch (MFR) (Eckes, 2011; Linacre y Wright, 2002).

El MFR es una extensión del Modelo de Rasch para escalas de calificación. Se aplica cuando existen diversos elementos (facetas) que pueden contribuir al error de la medida y facilita analizar de manera independiente la contribución de cada elemento. Los resultados se expresan mediante *lógits*, o el logaritmo del cociente entre la probabilidad de que un alumno reciba una calificación en un ítem y la probabilidad de que reciba la calificación inmediatamente inferior. Esta transformación de la medida ordinal original permite ubicar a cada elemento en un continuo real. En comparación con otros análisis de concordancia, el MFR es un método robusto y estable cuando existen valores ausentes, puesto que no todos los evaluadores evalúan a todos los estudiantes, tal y como sucede en este caso. Sin embargo, la utilización de este modelo no suele trascender más allá de los departamentos de psicometría y de metodología de evaluación médica.

La evaluación de las competencias de investigación y pensamiento científico mediante el TFG es una novedad en la educación médica elemental en España (Real decreto 1393/2007). Sin embargo, el sistema de evaluación y análisis no es homogéneo y depende del criterio de cada Facultad de Medicina (Lobato et al., 2010; Saz Pérez, 2013). El objetivo de este trabajo es estudiar la validez de la rúbrica de calificaciones del TFG mediante el MFR para evaluar las competencias de investigación y pensamiento crítico en los estudiantes/as del sexto curso del Grado de Medicina.

Método

Se trata de un estudio de evaluación de la rúbrica de calificación de la memoria escrita del TFG presentada por los estudiantes de 6º curso del Grado de Medicina de la Universidad Internacional de Cataluña.

Participantes y desarrollo del TFG: el protocolo de investigación (memoria escrita)

Se evaluaron las calificaciones de todos los estudiantes/as de sexto curso de la primera promoción de médicos de la Universidad Internacional de Cataluña (n=62). Todos cursaron la asignatura de TFG durante el curso académico 2013/2014.

Al inicio de la asignatura, todos los estudiantes completaron un curso de formación de 10 horas de duración (6 horas presenciales y 4 virtuales). El objetivo del mismo era recordar los conocimientos teóricos necesarios para la redacción de la memoria escrita del TFG. Entre los contenidos impartidos en este curso destacaban: cómo generar una idea de investigación, formulación de una hipótesis operativa, los diseños epidemiológicos en investigación, aspectos de metodología del protocolo de investigación, normas de redacción científica y cuestiones relativas a requerimientos éticos. Durante los 8 meses siguientes, cada alumno asistió a 5 tutorías (3 individuales y 2 grupales) con un profesor-tutor asignado al azar de los cuatro posibles. Cada tutoría tuvo una duración de 1 a 2 horas, cada mes o mes y medio. Previamente a la cita con el tutor, los estudiantes tenían que enviar la parte del trabajo realizada. En las tutorías, el alumno exponía los avances, preguntaba cuestiones específicas y el tutor daba pautas de guía para continuar trabajando individualmente hasta la siguiente tutoría. El contenido de las 5 tutorías se describe a continuación.

1ª tutoría (grupal): El objetivo era que el alumno estableciera la idea, pregunta de investigación y el diseño adecuado a dicha pregunta. En grupos de 4 a 6 estudiantes, cada uno de ellos expuso la idea de su proyecto; a continuación los otros compañeros preguntaron cuestiones relativas a la idea y a las variables a estudiar, debatiendo posteriormente acerca del mejor diseño epidemiológico para llevar a cabo la idea de la investigación. Siempre orientados y dirigidos por el tutor asignado.

2ª, 3ª y 4ª tutoría (individual): El objetivo fue realizar un seguimiento pormenorizado de la evolución de los distintos apartados de la memoria escrita del TFG: introducción, metodología, organización, aspectos éticos, bibliografía, etc. En la 2ª tutoría la tarea específica fue entregar el borrador de la introducción. En la 3ª, parte de los métodos, y en la 4ª, parte de las cuestiones relativas a la organización de la información y análisis, incluyendo también un borrador avanzado de la memoria escrita final del TFG.

5ª tutoría (grupal): Esta tutoría se realizó en grupos de 4 a 6 estudiantes y se programó dos semanas antes de la defensa del TFG ante un tribunal. El objetivo era trabajar la comuni-

cación científica del proyecto de manera oral, para ello los estudiantes presentaron el TFG a los compañeros y al tutor y realizaron preguntas, simulando el tribunal de lectura y defensa del TFG.

Evaluadores de la memoria escrita

La memoria escrita del TFG - el protocolo/proyecto de la investigación- se envió a tres evaluadores: un profesor del área correspondiente a la temática del trabajo (especialista clínico o de salud pública), un profesor de la asignatura de TFG, y el profesor-tutor del alumno que presentaba el trabajo. Cada uno evaluó el TFG de manera independiente y mediante la misma rúbrica de evaluación de competencias y resultados de aprendizaje.

Instrumento de evaluación

El diseño de la rúbrica para evaluar las competencias y resultados de aprendizaje lo realizó un equipo multidisciplinar de profesores de la Facultad de Medicina y Ciencias de la Salud de la UIC con experiencia en investigación clínica, salud pública y educación médica, participaron también los profesores-tutores de la asignatura del TFG. El diseño se realizó en dos fases: una primera consistente en la revisión de las competencias y resultados de aprendizaje propuestos para el TFG, la forma de desarrollarlos y la selección del método de evaluación que permitiera demostrar su adquisición. Para ello se siguieron los consejos de los organismos de evaluación de la calidad docente para el TFG en el Grado de Medicina (ANECA, 2005; Prat et al., 2004). A continuación, se elaboró un primer borrador de los ítems de la rúbrica. En la segunda fase se debatió la primera propuesta de rúbrica de evaluación, se introdujeron modificaciones y se consensuó el contenido final. Se determinó que la rúbrica final estaría compuesta por 10 ítems que agrupaban los resultados de aprendizaje de las 5 competencias relacionadas con la investigación y pensamiento crítico. Los ítems de la rúbrica final utilizaban una escala de calificación para la evaluación de 5 categorías (0 - .25 - .50 - .75 - 1). En la Tabla 1 se detallan los ítems de la rúbrica final para evaluar la memoria escrita del TFG.

Análisis estadístico de la rúbrica

Se utilizó el modelo de Múltiples Facetas de Rasch para escalas de calificación para estudiar la contribución de los elementos que componen la calificación del TFG y para detectar diferencias importantes en el criterio utilizado por los evaluadores (Myford y Wolfe, 2003). La estimación se realizó mediante el programa FACETS (Linacre, 2014). El modelo de MFR de escalas de calificación se formula:

$$\log (P_{nijk} / P_{nij(k-1)}) = B_n - D_i - C_j - F_k$$

Donde, P_{nijk} es la probabilidad de que el estudiante *n* sea calificado en el ítem *i* por el evaluador *j*, en la categoría *k*. B_n es la habilidad del estudiante *n*; D_i es la dificultad del ítem *i*; C_j es la severidad del evaluador *j*; F_k es el umbral de Rasch-Andrich de la categoría *k* en la escala de calificación, es la localización de la variable latente (con relación a la dificultad de los ítems) donde las categorías k y k-1 son igualmente probables.

Tabla 1. Items de la rúbrica final de evaluación de las competencias y resultados de aprendizaje del TFG del Grado de Medicina de la Universidad Internacional de Cataluña.

1.	El marco conceptual del problema de salud está bien definido y justificados en la introducción.
2.	La bibliografía utilizada es adecuada, actualizada y está bien referenciada siguiendo las normas de Vancouver.
3.	Realiza una apropiada valoración crítica de la literatura científica utilizada justificando la idoneidad del trabajo.
4.	El objetivo e hipótesis están expresados con claridad de manera que orienten a desarrollar el trabajo.
5.	El diseño metodológico es apropiado para alcanzar los objetivos.
6.	La población diana, estudio y muestra (si procede) están definidos correctamente en el trabajo.
7.	El/la alumno/a define correctamente los aspectos éticos del estudio.
8.	El estudio es viable y factible dentro del marco propuesto por el/la alumno/a.
9.	El análisis estadístico propuesto es el adecuado para alcanzar los objetivos del TFG.
10.	El TFG contemplado y justifica correctamente las potenciales limitaciones y amenazas a la validez interna y externa

El ajuste del modelo de MFR se estudió mediante diferentes indicadores para verificar la funcionalidad de la escala de calificación y la bondad de ajuste; este análisis contribuyó a aportar evidencia de la validez interna de la rúbrica (Smith y Smith, 2004). La funcionalidad de la escala de los ítems de la rúbrica se estudió examinando tres indicadores: 1) la frecuencia de las puntuaciones en todas las categorías: en cada una de las categorías debe haber al menos un 10% de observaciones; 2) los promedios de dificultad de cada una de las categorías de la escala: se esperaba que las estimaciones de dificultad fueran incrementándose a medida que la categoría fuera superior y 3) el estadístico de media cuadrática *Outfit*, que indica desajustes extremos en las calificaciones, por ejemplo en las que se han utilizado mayoritariamente categorías muy altas o muy bajas, este indicador debía estar en un rango .7-1.4 (Linacre, 2002b). En caso de que alguno de los tres criterios no se cumpliera, se valoró la posibilidad de colapsar las categorías con la adyacente superior.

El estudio de los indicadores de bondad de ajuste de cada uno de los elementos (severidad de los evaluadores, la habilidad de los estudiantes y la dificultad de los ítems) incluyó el estudio de los índices *Infit* y *Outfit*. Valores entre .7 y 1.4 indican que no hubo desajustes extremos ni internos en la idiosincrasia de los evaluadores cuando calificaron a los estudiantes (Linacre, 2002a; Linacre, 2002b). Se analizó la fiabilidad de cada uno de los elementos mediante los índices de separación, comparables al coeficiente alfa de Cronbach (va-

lores superiores a .7 indicaron una correcta discriminación de cada una de las escalas).

Adicionalmente, se realizaron algunas comprobaciones y análisis específicos para cada uno de los elementos. En la estimación de severidad de los evaluadores, se agregaron las puntuaciones de los evaluadores especialistas. El motivo es que en este grupo la frecuencia de evaluaciones por alumno se consideró insuficiente, ya que algunos especialistas sólo participaron en la evaluación de una memoria de TFG. Se examinó la condición de independencia mediante el índice de *Rasch-Kappa*: valores cercanos a 0 probaron la independencia estadística (Eckes, 2011).

En la escala de habilidad de los estudiantes se comprobó si había diferencias estadísticamente significativas para el total de las calificaciones relativas a los tres tipos de evaluadores. Para ello se realizó una prueba ANOVA, o la prueba de Kruskal-Wallis en caso de no cumplirse las condiciones de aplicación. Adicionalmente, se examinó si había diferencias en las estimaciones de habilidad relativas a cada uno de los estudiantes por grupo de evaluadores. Para ello se realizaron comparaciones por parejas para los tres tipos de evaluadores (tutor-especialista-profesor) detectando así la diferencia de la habilidad. Se consideró la significación estadística mediante una serie de pruebas de *t* de *Student* para muestras apareadas con un nivel de confianza del 95%. En la escala de dificultad de los ítems se examinaron los presentaban mayores discordancias.

La unidad continua de los resultados mediante el modelo de MFR - el lógit- oscila entre $0 \pm \infty$; en este caso, para facilitar la interpretación se trasformaron linealmente a una escala 0-10. Esta transformación se realizó una vez finalizado el análisis mediante el programa informático FACETS (Linacre, 2014). Para esta trasformación se tiene en cuenta el rango de lógits obtenidos entre los estudiantes con la calificación más baja y más alta; a continuación se establece el factor de conversión del lógit a la escala 0-10; esta constante la utiliza el programa para establecer el nuevo origen de la escala 0-10 con el que se pueden re-escalar todos los valores de habilidad de los estudiantes, severidad de los evaluadores y dificultad de los ítems son trasformados a la nueva escala de 0-10.

Aspectos éticos y conflictos de intereses

El presente trabajo cumple con las normas éticas de publicación de trabajos científicos. En todo momento se respetaron la confidencialidad de los estudiantes y de los profesores. La finalidad del estudio era la investigación para la mejora de la evaluación de los procedimientos docentes y la divulgación para la comunidad científica y académica. Los autores no reportan ningún conflicto de interés.

Resultados

Todos los estudiantes matriculados en la asignatura del TFG presentaron la memoria escrita del trabajo (*n*=62). El trabajo de cada alumno fue evaluado por el tutor (*n*=4), un profesor del departamento (*n*=4) y un especialista con experiencia en investigación clínica y salud pública, (*n*=25). En total se analizaron 3 rúbricas por alumno, resultando un total de 186 rúbricas, cada una con 10 ítems, resultando un total de 1860 ítems, que evaluaban las competencias y resultados de aprendizaje.

El análisis de la funcionalidad de la escala reveló que la frecuencia de las puntuaciones en las categorías 0 y .25 fue inferior al 10%, además las estimaciones de dificultad de las categorías inferiores eran superiores a las de las categorías superiores, violando el criterio de elevación progresiva de la dificultad en cada categoría superior. Siguiendo el criterio establecido se decidió colapsar las categorías 0 y .25 y explorar a continuación los indicadores de calidad, bondad de ajuste al modelo y resultados de las mediciones. Los 3 criterios de calidad y funcionalidad de la escala de calificación se mostraron adecuados al colapsar las 3 categorías inferiores: 0, .25 y .5. Esta acción aumentó la efectividad de las mediciones (separación), la estabilidad de las estimaciones y contribuyó a la mejora de los indicadores de bondad de ajuste.

La Figura 1 muestra el mapa de elementos que permite visualizar y comparar las escalas de severidad de los evaluadores, la habilidad de los estudiantes y la dificultad de los ítems de la rúbrica de manera conjunta. Los resultados se expresan con la escala de lógits original trasformada a 0-10. Se observa que la rúbrica fue utilizada de manera similar por los evaluadores (segunda columna). La distribución de la habilidad de los estudiantes (tercera columna) muestra un patrón casi normal. Finalmente, en la cuarta columna se presenta la distribución de la dificultad de los ítems de la rúbrica. El mapa de elementos permite realizar comparaciones: El 58% de los estudiantes mostró una habilidad que superaba la dificultad de los 10 ítems.

Evaluación de la severidad, fiabilidad e independencia de los evaluadores

La tabla 2 muestra los indicadores de severidad de los evaluadores. El promedio de severidad fue de 0± .44 lógits con un rango de 1.19 lógits, (5.02± .15, en la escala 0-10, rango de .96). Los indicadores de bondad de ajuste *Infit* y *Outfit* no evidenciaron que los evaluadores calificaron de manera coherente unos respecto a los otros. El índice *Rasch-Kappa* de cada evaluador estuvo próximo a 0, sugiriendo la ausencia de influencias entre los mismos evaluadores o por elementos externos. El índice de fiabilidad de .90 indicó que la variabilidad en la severidad de las calificaciones era aceptable.

Evaluación de los resultados de habilidad de los/as estudiantes/as del TFG

El promedio de los resultados de los estudiantes (habilidad) fue .97± .63 lógits (6.17± .36 en la escala de 0-10). El alumno con mayor habilidad tuvo una puntuación de 2.29 lógits (7.74 en la escala 0-10) y el de menor habilidad de 4.63 (-.32 lógits); rango =2.61 lógits, (3.11 en la escala 0-10). En la

escala de 0-10 cuatro estudiantes (6.45%) obtuvieron puntuaciones inferiores al 5 sobre 10, nota que marca tradicionalmente el aprobado. La fiabilidad de la escala de habilidad fue moderada (0.75). La figura 2 muestra la distribución de las calificaciones de cada alumno según el tipo de evaluador. No se encontraron diferencias estadísticamente significativas para el total de calificaciones de los estudiantes entre los tres tipos de evaluadores (*Kruskal Wallis*=2.228, p= .328). Las habilidades individuales por grupo de evaluadores revelaron diferencias estadísticamente significativas (p<0.05), en 10 casos (13%) entre los profesores y los especialistas; en 6 casos (10%) entre los tutores y los especialistas y en un solo caso (2%) entre los profesores y los tutores.

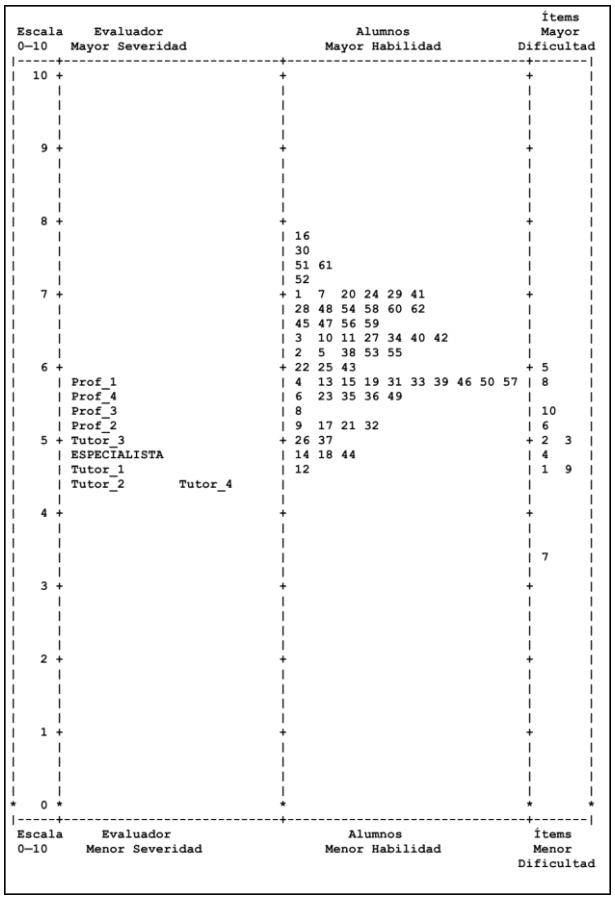


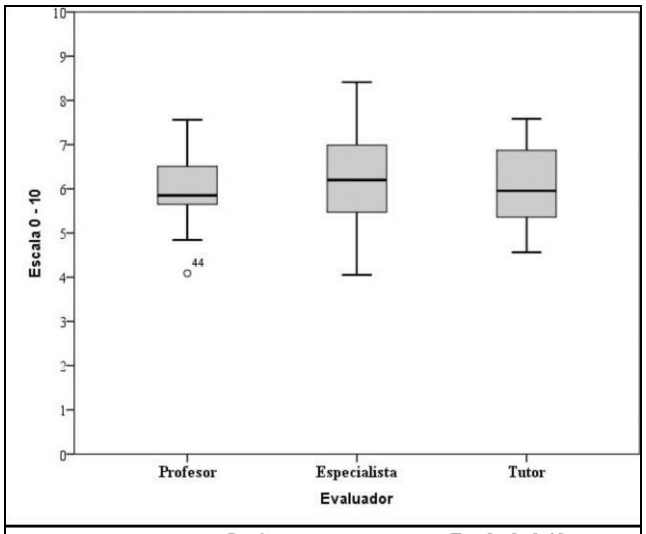
Figura 1. Mapa ítems. Es el resultado de la estimación mediante el modelo de MFR. En parte izquierda está la escala transformada de lógits, a la escala 0-10. La segunda columna es la distribución de severidad de los evaluadores, de más, en la parte superior, a menos severidad, en la parte inferior, la categoría especialista agrupa los profesores externos al departamento (n=25). En la tercera columna se haya la distribución de la habilidad de los estudiantes ordenada de mayor competencia, en la parte superior, a menor, en la parte inferior, cada número corresponde a un estudiante. Finalmente, la cuarta columna, presenta la distribución de la dificultad de los ítems de la rúbrica, ordenados de manera similar. Puesto que todos los elementos utilizan la misma escala se pueden comparar entre ellos.

Tabla 2. Indicadores de ajuste y estimación de la severidad de los evaluadores ordenados de mayor a menor severidad.

Evaluador	Estimación de la severidad				
	Lógits (*D.E.*)	**Escala 0 - 10**	***Infit***	***Outfit***	***Rasch-Kappa***
Prof_1	.59 (.12)	5.72	.98	1.07	-.02
Prof_2	.52 (.14)	5.30	.89	.85	-.03
Prof_3	.27 (.12)	5.34	1	.96	-.02
Prof_4	.23 (.12)	5.64	1.17	1.17	.04
Especialista	.03 (.13)	4.76	.98	.93	.10
Tutor_1	-.21 (.07)	4.65	1.05	1.02	-.05
Tutor_2	-.31 (.14)	4.39	.86	1.05	-.03
Tutor_3	-.53 (.14)	5.06	1.09	1.01	-.01
Tutor_4	-.6 (.16)	4.31	.78	.67	.08

Fiabilidad inter-evaluadores: .90, *Rasch-Kappa* Total= -.01

Prof. denomina a un profesor del departamento de la asignatura de TFG, *Tutor* denomina a un mentor del alumno que evalúa el trabajo con la misma rúbrica. La numeración de prof y tutor corresponde a la mima persona con distinto rol de evaluación. *Especialista* es el conjunto de profesores externos (n=25) que evaluó la memoria escrita del TFG.



Evaluador	Logits		Escala de 0-10	
	Promedio (DE)	**Mediana**	**Promedio (DE)**	**Mediana**
Especialista	1.14 (1.04)	1.24	6.38 (1.23)	6.49
Profesor	1.03 (.72)	.86	6.24 (.85)	6.04
Tutor	1.09 (.83)	.98	6.31 (.98)	6.17

Figura 2. Distribución de las calificaciones por alumno en la escala 0 -10 por tipo de evaluadores. En la parte inferior las cifras exactas de los estadísticos de centralidad y la escala no transformada de lógits.

Evaluación de la dificultad de los ítems de la rúbrica

En la Tabla 3 se muestran los indicadores de la evaluación de la dificultad de los ítems. El promedio de dificultad en la escala fue 2.13± .6 lógits, (5.02± .09 en la escala de 0-10). El ítem con mayor dificultad fue la selección del diseño del protocolo con una medida de dificultad de .86 lógits (6.04 en la escala 0-10), seguido de la viabilidad para realizar la propuesta del estudio. Los indicadores de bondad de ajuste *Infit* y *Outfit* y la fiabilidad de la escala (.95) indicaron que las estimaciones de la escala eran óptimas.

Tabla 3. Indicadores de ajuste y estimación de la dificultad de los ítems ordenados de mayor a menor dificultad

Items	Estimación		Estadísticos de ajuste al modelo	
	Lógits (*ES*)	Escala 0-10	*Infit*	*Outfit*
5. Diseño metodológico apropiado para los objetivos.	.86 (0.11)	6.04	.93	.93
8. El estudio es viable y factible.	.70 (0.11)	5.85	1.18	1.22
10. Justificación de limitaciones y validez.	.33 (0.11)	5.4	.86	.82
6. Población diana definida.	.2 (0.11)	5.25	1	.96
2. Bibliografía adecuada.	.05 (0.12)	5.08	1.19	1.22
3. Valoración crítica de la literatura científica.	.05 (0.12)	5.08	.93	.92
4. Objetivo e hipótesis expresados con claridad.	-.24 (0.13)	4.73	.96	.91
1. Marco conceptual definido y justificado.	-.31 (0.13)	4.65	.88	.81
9. Análisis estadístico adecuado.	-.36 (0.13)	4.59	1	.99
7. Define los aspectos éticos del estudio.	-1.27 (0.17)	3.51	1.15	1.1

Fiabilidad de la escala, índice de separación: 0.95

Discusión

La rúbrica de calificación utilizada para evaluar la memoria escrita del TFG demostró una alta efectividad para valorar las competencias y resultados de aprendizaje relacionados con la investigación y pensamiento crítico en los estudiantes del Grado de Medicina. El análisis realizado permitió comprobar la variabilidad de los elementos que contribuyen a la calificación final: la severidad de los evaluadores, la habilidad de los estudiantes y la dificultad de los ítems de la rúbrica, y comparándolos entre ellos utilizando una métrica común.

Los resultados de este estudio demuestran la efectividad del modelo adoptado en la Universitat Internacional de Catalunya, consistente en designar a un grupo de profesores multidisciplinar y con carga docente exclusiva para conducir la asignatura de TFG. Este sistema tiene la ventaja de que los estudiantes cuentan con un referente, experto en metodología e investigación clínica, durante toda la asignatura. Asimismo, los estudiantes se encuentran con una situación cercana a la realidad en investigación clínica, donde los facultativos tienen que trabajar con equipos multidisciplinares. Sin embargo, uno de los inconvenientes hallados durante las tutorías es la imposibilidad de que los profesores designados tengan experiencia en todos los temas que los estudiantes han pretendido desarrollar. Por ello, una de las mejoras del TFG, que se tendrá en cuenta en futuros cursos, es facilitar un listado de temas propuesto por los mismos tutores de TFG para que los estudiantes los desarrollen. Creemos que esta es una limitación menor porque los resultados de severidad de los evaluadores especialistas, externos a la asignatura de TFG, fue inferior a la de los propios tutores, por lo que se demuestra que los estudiantes trasladaron los conocimientos de metodología científica a la redacción de un proyecto de investigación clínica o salud pública, correctamente planteado.

Para poder llevar a cabo la estimación de las calificaciones mediante el modelo de MFR se tuvieron que colapsar tres categorías de las 5 iniciales: las que corresponden a 0, .25 y .5, indicando que los evaluadores sólo discriminaron entre tres posibles categorías de calificación (por ej.: d*eficiente, adecuado, excelente)*. La escasa utilización de estas dos categorías inferiores podría indicar una comprensión inexacta del significado de cada una (Wolfe y Smith, 2007). Por esta razón, en la próxima convocatoria de evaluación del TFG se enviará a los evaluadores información que oriente acerca de cómo utilizar la escala de calificación, con ejemplos prácticos de qué calificación corresponde a cada categoría. Este tipo de orientaciones ("*scoring rubric notes*") permite reducir la variabilidad y los sesgos en las calificaciones de los evaluadores y son habituales en situaciones similares en el campo de la medicina (Boulet, Rebbecchi, Denton, McKinley, y Whelan, 2004; Peeters, Sahloff, y Stone, 2010; Till, Myford y Dowell, 2013).

La escasa variabilidad en la severidad de los evaluadores, evidenciada en que la severidad de los evaluadores no difirió en más de un punto de la escala de 0 a 10 entre el evaluador más severo y el más benevolente, demostró que los evaluadores utilizaron una idiosincrasia similar cuando calificaron a los estudiantes. Sin embargo, el análisis reveló que los evaluadores más estrictos fueron los profesores del departamento, seguido de los especialistas y, finalmente, los más benevolentes, los tutores. El hecho de que los tutores realizaron el seguimiento individual de los estudiantes pudo influir en la benevolencia de éstos a la hora de evaluar. Este sesgo al alza en las puntuaciones podría solventarse en próximas convocatorias mediante la adicción de ítems específicos en la rúbrica del tutor y así poder evaluar exclusivamente el proceso de aprendizaje, excluyendo al tutor de la evaluación de la memoria escrita del TFG de los estudiantes a los que ha realizado seguimiento.

Los ítems más difíciles se referían al diseño del estudio y a la factibilidad del mismo. En lo que se refiere al primero, la selección del diseño del estudio es probablemente el punto más controvertido ante un problema de investigación, puesto que siempre puede haber más de una opción metodológica; aun cuando el alumno lo hubiera justificado de manera adecuada, un experto siempre puede tener otra opinión apropiada. En cuanto al segundo, los estudiantes presentaron propuestas de investigación que cubrían desde diseños observacionales hasta ensayos clínicos con fármacos; en todos ellos la posibilidad de llevar a cabo el estudio, siendo los investigadores los estudiantes del último curso de medicina era controvertida. Como aspecto a mejorar en las próximas con-

vocatorias del TFG se insistirá a los estudiantes en que propongan diseños de investigación acordes con sus posibilidades de realización durante la residencia o al menos como investigadores noveles. En este sentido, el TFG podría ser un punto de partida para el futuro desarrollo de una posible tesis doctoral durante la realización de la especialidad.

El impulso del aprendizaje de las competencias de investigación en la carrera de medicina ha sido una novedad más del Plan Bolonia. Sin embargo, no está claro en qué momento del proceso los estudiantes deben demostrar su aprendizaje. En algunos países, como Alemania, el último año del grado está centrado exclusivamente en las habilidades clínicas (Nikendei, Krautter, Celebi, Obertacke y Junger, 2012), en otros países se incluye la posibilidad de que el alumno escoja un itinerario opcional en el que la investigación puede tener cabida (Van den Akker, Dornan, Scherpbier, Oude Egbrink y Snoeckx, 2012); en Suiza, por ejemplo, la formación en investigación de los futuros médicos ha de estar al mismo nivel que en otras carreras del ámbito biomédico (Michaud, 2012). En España, la regulación ha llevado a que todas las facultades de medicina incorporen el TFG en los planes de estudio, con lo que se posibilita desarrollar una evaluación de las competencias en investigación de manera clara (Real decreto 1393/2007). La controversia reside en que la legislación no ha venido acompañada de directrices ni reglamentos prácticos. Los motivos de esta debilidad apuntan a la falta de conexión entre el regulador y los órganos rectores de las Facultades de Medicina (Arnalich Fernández, 2010; Baños y Bosch, 2008; Nogales Espert, 2010). En el caso de Cataluña, la agencia de evaluación de la calidad ha publicado diferentes cuadernos metodológicos dedicados a los TFG en el área de las ingenierías (Valderrama Vallés, Prades Nebot, y Rodríguez Espinar, 2009) y las ciencias jurídicas y sociales (Mateo Andrés, Rodríguez Espinar y Prades Nebot, 2009). Aunque estas recomendaciones podrían ser válidas para el TFG de medicina, las características propias de esta titulación deberían ser consideradas para elaborar un cuaderno metodológico del TFG específico en medicina, o al menos para los Grados de Ciencias de la Salud. Nuestro estudio, ha sido realizado a partir de la legislación vigente, siguiendo las indicaciones que propone el proyecto DISSENY (Prat et al., 2004), a partir del proyecto TUNING para medicina (Ross et al., 2014) y clarifica un modelo de TFG así como su evaluación.

Una de las limitaciones que presenta este estudio es el número de estudiantes con los que se ha realizado la evaluación. Un tamaño de muestra mayor podría haber generado estimaciones más precisas. Sin embargo, los indicadores de fiabilidad y de calidad del modelo MFR fueron óptimos y sugieren que la interpretación de los resultados es adecuada. Las calificaciones de la competencia y resultados de aprendizaje que se presentan en este estudio corresponden exclusivamente a la evaluación del TFG calificada por un tribunal compuesto de tres miembros mediante una rúbrica, esta calificación supone el 45% de la nota final de la asignatura, el 55% de la nota restante se evaluaba midiendo otros aspectos como: la presentación oral del TFG, asistencia a las tutorías, presentación de las tareas o evaluación del trabajo de un compañero. Sin embargo, el efecto de esta potencial limitación es moderado porque la correlación entre la nota estimada mediante el modelo MFR y la nota final tuvo una correlación positiva (r=0,884, p<0,001). Finalmente, una limitación a tener en cuenta en la evaluación de la variabilidad de los profesores es la agregación de los 25 especialistas en una sola categoría; sin embargo, la evidencia que la distribución de las calificaciones por alumno respecto a las calificaciones de los otros dos evaluadores fue similar, con lo que el efecto de agregar a los 25 especialistas tiene poca relevancia.

En conclusión, el análisis de la rúbrica del TFG y de las calificaciones de los estudiantes evidencia que un TFG en la carrera de medicina consistente en la realización de un protocolo de investigación clínica, o salud pública, en el que el alumno es guiado por un profesor experto en metodología de investigación, permite evaluar la adquisición de las competencias de pensamiento crítico e investigación. El análisis de la rúbrica verifica que el modelo de TFG puede ser útil para otras facultades de medicina a la hora de implementar esta novedad del Plan Bolonia.

Agradecimientos.- Los autores queremos agradecer a los estudiantes que finalizaron la primera promoción de medicina de la Universitat Internacional de Catalunya la posibilidad de realizar este estudio. Especial consideración merecen el Dr. Albert Balaguer, decano de la Facultad de Medicina y Ciencias de la Salud de la Universitat Internacional de Catalunya, y la Dra. María Dolores Navarro-Rubio del Institut Albert J. Jovell de Salut Pública i Pacients de la Universitat Internacional de Catalunya, por los comentarios y sugerencias realizados durante la redacción del manuscrito.

Referencias

ANECA (Agencia Nacional de Evaluación de la Calidad y Acreditación). (2005). *Libro blanco. Título de grado en medicina*. Disponible en: http://www.aneca.es/var/media/150312/libroblanco_medicina_def.pdf

Arnalich Fernández, F. (2010). Adaptación del nuevo grado en medicina al espacio europeo de educación superior. ¿Cuál ha sido la aportación de bolonia? *Revista Clínica Española, 210*(9), 462-7.

Baños, J. E.y Bosch, F. (2008). ¿Ha llegado, por fin, el lobo? la adaptación de los planes de estudio de medicina al espacio europeo de educación superior: Consideraciones en torno a una orden ministerial. *Medicina Clínica (Bar), 131*(12), 457-459.

Boulet, J. R., Rebbecchi, T. A., Denton, E. C., McKinley, D. W. y Whelan, G. P. (2004). Assessing the written communication skills of medical school graduates. *Advances in Health Sciences Education: Theory and Practice, 9*(1), 47-60.

Declaración conjunta de los ministros europeos de enseñanza. *Espacio europeo de enseñanza superior.* Bolonia, Italia: 19 de Junio de 1999

Eckes, T. (2011). *Introduction to many-facet rasch measurement*. Frankfurt am Main: Peter Lang.

Iramaneerat, C., Yudkowsky, R., Myford, C. M. y Downing, S. M. (2008). Quality control of an OSCE using generalizability theory and many-faceted rasch measurement. *Advances in Health Sciences Education: Theory and Practice, 13*(4), 479-493.

Linacre, J. (2014). *Facets computer program for many-facet Rasch measurement.* v. 3.71.4. Beaverton, Oregon: Winsteps.com.

Linacre, J. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*(2), 878. http://www.rasch.org/rmt/rmt162f.htm

Linacre, J. (2002b). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. y Wright, B. (2002). Construction of measures from many-facet data. *Journal of Applied Measurement, 3*(4), 486-512.

Lobato, R., Lagares, A., Alén, J. y Alday, R. (2010). El desarrollo del proceso de "Bolonia" y el Grado de Medicina. situación actual y expectativas para su implantación definitiva. *Neurocirugia (Astur), 21*(2), 146-56.

Lunz, M., Wright, B. y Linacre, J. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-345.

Mateo Andrés, J., Rodríguez Espinar, S. y Prades Nebot, A. (2009). *Guia per a l'avaluació de competències en el treball de final de grau en l'àmbit de les ciències socials i jurídiques*. Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya. Disponible en: http://www.aqu.cat/doc/doc_95455311_1.pdf

Michaud, P. (2012). Reforms of the pre-graduate curriculum for medical students: The bologna process and beyond. *Swiss Medical Weekly, 17*, 142:w13738.

Myford, C. y Wolfe, E. (2003). Detecting and measuring rater effects using many-facet rasch measurement: Part I. *Journal of Applied Measurement, 4*(4), 386-422.

Nikendei, C., Krautter, M., Celebi, N., Obertacke, U. y Junger, J. (2012). Final year medical education in germany. *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen, 106*(2), 75-84.

Nogales Espert, A. (2010). Reflections about the implementation of the european higher education area in a school of medicine. [Reflexiones sobre la implantacion del "espacio europeo" en una facultad de medicina] *Anales De La Real Academia Nacional De Medicina, 127*(1), 89-98.

Peeters, M. J., Sahloff, E. G. y Stone, G. E. (2010). A standardized rubric to evaluate student presentations. *American Journal of Pharmaceutical Education, 74*(9), 171.

Prat, J., Carreras, J., Branda, L., Miralles, R., Fenoll, M., Rodríguez, S. y Grifoll, J. (2004). *Competències professionals bàsiques comunes dels llicenciats en medicina formats a les universitats de catalunya* Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya. http://www.aqu.cat/doc/doc_73693838_1.pdf .

Real decreto 1393/2007, de 29 de octubre, por el que se establece la ordenación de las enseñanzas universitarias oficiales. BOE, Martes 30 de Octubre de 2007. Disponible en http://www.boe.es/boe/dias/2007/10/30/pdfs/A44037-44048.pdf.

Real decreto 99/2011, de 28 de enero, por el que se regulan las enseñanzas oficiales de doctorado. BOE, Jueves 10 de Febrero de 2011. Disponible en http://www.boe.es/boe/dias/2011/02/10/pdfs/BOE-A-2011-2541.pdf.

Ross, M., Nikolić, N., Peeraer, G., Murt, A., Kroiča, J., Elcin, M., Cumming, A. (2014). Report of the MEDINE2 bachelor of medicine (bologna first cycle) tuning project. *Medical Teacher, 36*(4), 314-21.

Saz Pérez, J. (2013). Bolonia: Una oportunidad perdida. la heterogeneidad de los estudios de medicina en españa. *Revista Clínica Española, 213*(9), 440-1.

Smith, E. V. y Smith, R. M. (2004). *Introduction to rasch measurement: Theory, models and applications*. Maple Grove, Minn.: JAM Press.

Till, H., Myford, C. y Dowell, J. (2013). Improving student selection using multiple mini-interviews with multifaceted rasch modeling. *Academic Medicine, 88*(2), 216-223.

Valderrama Vallés, E., Prades Nebot, A. y Rodríguez Espinar, S. (2009). *Guia per a l'avaluació de competències als treballs de final de grau i de màster a les enginyeries*. Barcelona: Agència per a la Qualitat del Sistema Universitari de Catalunya. Disponible en: http://www.aqu.cat/doc/doc_21214293_1.pdf

Van Den Akker, M., Dornan, T., Scherpbier, A., Oude Egbrink, M. G. y Snoeckx, L. H. (2012). Easing the transition: The final year of medical education at Maastricht university. *Zeitschrift Fur Evidenz, Fortbildung Und Qualitat Im Gesundheitswesen, 106*(2), 92-97. doi:10.1016/j.zefq.2012.02.013 [doi]

Wolfe, E. W.y Smith, E. V.,Jr. (2007). Instrument development tools and activities for measure validation using rasch models: Part I - instrument development tools. *Journal of Applied Measurement, 8*(1), 97-123.

(Artículo recibido: 11-11-2014; revisado: 08-02-2015; aceptado: 03-03-2015)