Andalucía se mueve con Europa

Innovación e investigación educativa para la formación docente

Francisco Javier Hinojo Lucena
Salvador Mateo Arias Romero
María Natalia Campos Soto
Santiago Pozo Sánchez

Dykinson, S.L.

Junta de Andalucía

UNIÓN EUROPEA
Fondo Europeo de Desarrollo Regional

Editorial DYKINSON, S.L. Meléndez Valdés, 61 - 28015 Madrid
Teléfono (+34) 91 544 28 46 - (+34) 91 544 28 69
e-mail: info@dykinson.com
http://www.dykinson.es
http://www.dykinson.com
Consejo Editorial véase www.dykinson.com/quienessomos

El contenido de este libro ha sido sometido a un proceso de revisión y evaluación por pares ciegos.

ISBN: 978-84-1122-023-1

CAPÍTULO 29.
OBTENCIÓN DE CALIFICACIONES JUSTAS EN UNA EVALUACIÓN ENTRE IGUALES CON PROFESORADO DE SECUNDARIA EN FORMACIÓN INICIAL

Gregorio Jiménez Valverde, Genina Calafell i Subirà y Mireia Esparza Pagès.

1. INTRODUCCIÓN

Uno de los principales objetivos de la educación superior es ayudar a los estudiantes a desarrollar el pensamiento crítico en sus prácticas profesionales y ser más reflexivos en su proceso de aprendizaje (Falchikov y Goldfinch, 2000). A pesar de que el estudiante cada vez se involucra más en el proceso de aprendizaje, con metodologías activas que lo alejan de un mero receptor de información en las clases (Mintzes y Walter, 2020), lo cierto es que su papel en la evaluación continúa siendo discreto. Varios autores abogan por dar mayor protagonismo al estudiante en el proceso de evaluación (Rodríguez, Ibarra y García, 2013). Una de las maneras de involucrar más al estudiante en el proceso de evaluación es a través de la evaluación entre iguales, en la que cada uno evalúa la calidad de las producciones de sus compañeros (Topping, 2018). Es justamente el análisis de las producciones de sus iguales lo que dota a esta evaluación del carácter formador, pues permite al estudiante realizar mejores juicios de su propio trabajo, e identificar avances y dificultades o errores en su aprendizaje. De hecho, Wilian y Leahy (2015) consideran que la autoevaluación es más significativa cuando previamente el estudiante ha evaluado el trabajo de los compañeros y ha recibido un feedback de su propia producción.

Si bien Rasch (1960) postulaba que el acierto en un ítem de evaluación, en clave dicotómica, dependía exclusivamente de la habilidad del examinando y de la dificultad del ítem, lo cierto es que, a la hora de evaluar, el examinador no es neutro, sino que está sujeto a la percepción y subjetividad humana, a su propia experiencia y a sesgos y errores. Estas circunstancias personales pueden afectar a la calidad de su evaluación, es decir, evaluar o calificar una actividad es un proceso subjetivo (Wu, 2017) y, por tanto, estas otras variables debidas al evaluador también pueden tener incidencia en la evaluación de la producción del estudiante.

ISBN: 978-84-1122-023-1

Uno de los sesgos personales que puede causar mayor distorsión en una evaluación entre iguales de tipo cuantitativo es el sesgo de severidad o benevolencia que puedan tener algunos estudiantes (Myford y Wolfe, 2003), especialmente en las evaluaciones en las que los estudiantes son evaluados únicamente por algunos de sus compañeros. Cuando todos los estudiantes evalúan a todos, el posible sesgo de severidad o benevolencia de alguno de ellos queda difuminado o diluido, ya que todas las evaluaciones que realice se verán impregnadas de este sesgo y, por tanto, afectará por igual a todos los compañeros, con lo cual todas las calificaciones se verán incrementadas o disminuidas en la misma proporción. Sin embargo, cuando se organiza una evaluación entre iguales en la que las producciones de los estudiantes son evaluadas por un número reducido de compañeros (por ejemplo, cada producción es evaluada por uno o dos), la calificación final de un estudiante será diferente según si ha sido evaluado por un compañero con un sesgo significativo de severidad o benevolencia en comparación con otro evaluador que no lo tenga. En estas situaciones se hace necesario disponer de algún mecanismo que identifique y cuantifique el sesgo de severidad o benevolencia de los evaluadores, de tal manera que pueda ser eliminado y haga posible obtener unas calificaciones más justas u objetivas.

Linacre (1989) ha propuesto una extensión del modelo de Rasch, en el que, además de la habilidad del examinando y de la dificultad del ítem, se tienen en cuenta otras variables, llamadas "facetas", que también pueden influir en la calificación final que obtiene el estudiante, como el sesgo de severidad o benevolencia. Además, al incorporar las consideraciones realizadas por Andrid (1978) y Masters (1982), la propuesta de Linacre es aplicable no solo a evaluaciones dicotómicas (como el modelo original de Rasch), sino también a aquellas realizadas de acuerdo a una escala ordinal, lo cual permite aplicar el modelo cuando la escala de calificación se ha basado en los diferentes niveles de logro o calidad de una rúbrica de evaluación o se han usado escalas ordinales, como la escala de Likert, para la evaluación. El modelo de Linacre, llamado *Many-facet Rasch Measurement* (MFRM), permite, por tanto, estudiar el sesgo de severidad o benevolencia en una evaluación cuantitativa entre iguales en la que se haya utilizado una rúbrica con escala ordinal (Engelhard y Wind, 2017). Es importante recalcar que el modelo MFRM define la severidad (o benevolencia) del evaluador en términos relativos, es decir, el cálculo que hace de la severidad o benevolencia es en comparación con los otros evaluadores que participan en la misma actividad (Anthony et al., 2021).

ISBN: 978-84-1122-023-1

El objetivo de este estudio es detectar y eliminar los sesgos de benevolencia y severidad de un grupo estudiantes del Máster de Formación del Profesorado de Secundaria cuando estos actúan como evaluadores y poder obtener, por tanto, la calificación justa que correspondería en ausencia de estos sesgos.

Los estudiantes de este máster probablemente han tenido una limitada y escasa experiencia en actividades de evaluación entre iguales en los grados universitarios que cursaron con anterioridad a dicho máster. Siendo esto ya motivo suficiente para que participen en experiencias de evaluación de este tipo durante su formación inicial como docentes, lo cierto es que es probable que algunos de ellos tengan sesgos de severidad o benevolencia sin ser conscientes de ello, lo cual es un aspecto importante que hay que tener en cuenta en su propia formación inicial, puesto que cuando ejerzan como profesores de Secundaria realizarán continuas evaluaciones a su alumnado.

2. MÉTODO

La experiencia que se presenta en este trabajo se ha llevado a cabo durante el curso 2019-2020 con estudiantes de la asignatura "Didáctica de la Química" del Máster de Formación del Profesorado de Educación Secundaria Obligatoria, Bachillerato y Formación Profesional de la Universitat de Barcelona. En total, participaron 27 estudiantes (11 mujeres, 16 hombres), correspondientes al único grupo de la asignatura.

La actividad objeto de la evaluación entre iguales consistía en la confección de la programación de una unidad didáctica de Química, de un tema de su elección, para cualquiera de los cursos de la Educación Secundaria Obligatoria (ESO), siguiendo el modelo de programación competencial del Departamento de Educación de la Generalitat de Catalunya (2020). Para ello, los estudiantes se distribuyeron en grupos de dos, salvo un grupo, que se constituyó con tres estudiantes, en una actividad de evaluación entre iguales. La rúbrica de evaluación de las programaciones (tabla 1) se consensuó con el alumnado antes de empezar la actividad y, finalmente, estuvo compuesta por doce ítems, cada uno de ellos con una escala de valoración con cuatro niveles de logro (1=no alcanzado, 2=satisfactorio, 3=notable y 4=sobresaliente).

ISBN: 978-84-1122-023-1

Tabla 1

Ítems de la rúbrica de evaluación.

Número de ítem (I)	Descripción del ítem
1	Plantea un título a la unidad didáctica en relación con los contenidos a trabajar
2	Concreta materia, curso y número de horas
3	La cabecera de la unidad didáctica contiene todos los elementos curriculares que conforman una unidad didáctica: dimensión, competencias del ámbito científico-tecnológico y de los ámbitos transversales, criterios de evaluación curriculares, contenidos clave y curriculares
4	Las situaciones de aprendizaje se plantean con preguntas o como problemas contextualizados que hay que resolver
5	Responde a una secuencia didáctica lógica (ciclo de aprendizaje), es decir, organiza de forma secuencial la exploración de ideas previas, la introducción de nuevos conocimientos, la estructuración de conocimientos y la aplicación y transferencia de conocimientos
6	Diversifica recursos, espacios y/o materiales aplicables
7	Formula los objetivos, los criterios de evaluación y los indicadores de evaluación graduados en 3 niveles
8	Contiene competencias de los ámbitos transversales (digital y personal y social)
9	Se fomenta tanto el trabajo individual como el colectivo del alumnado, incentivando la autonomía

ISBN: 978-84-1122-023-1

10	Incluye evaluación formativa y formadora. Además, se incluyen procedimientos de feedback, autoevaluación y evaluación entre iguales (coevaluación)
11	Incluye medidas de atención a la diversidad. Tiene en cuenta las pautas del diseño universal de aprendizaje
12	Valoración global

Una vez los grupos de estudiantes habían realizado sus programaciones, estas fueron enviadas al docente quien, después de anonimizarlas, las repartió al alumnado. Puesto que se decidió previamente que cada programación sería evaluada por un grupo reducido de estudiantes, se tuvo que prestar especial atención a la asignación de qué estudiantes tendrían que evaluar qué programaciones. Ello es debido a que el análisis estadístico posterior debe poder comparar dos medidas y debe permitir discernir si la mayor calificación de una programación respecto de otra se debe a que, efectivamente, es una programación de mayor calidad o si esta diferencia es debida a la mayor benevolencia del evaluador la calificó. En otras palabras, es necesario que, como mínimo, 2 estudiantes evalúen la misma programación y que cada programación comparta, como mínimo, un evaluador con otra programación, de tal manera que ninguna programación quede "desconectada" y, por tanto, se pueda comparar el sesgo de severidad de todos los evaluadores. La tabla 2 muestra el esquema de evaluación de las programaciones (P) entre los estudiantes/evaluadores (Ev), identificados con un número. Cada estudiante recibió dos programaciones para evaluar (identificadas en la tabla 2 con OB). Opcionalmente, y tras la evaluación de estas dos primeras programaciones, los estudiantes que quisieron pudieron evaluar una tercera programación (identificadas en la tabla 1 con vo).

Tabla 2

Diseño del reparto de las programaciones (P) a los evaluadores (Ev): cada programación fue evaluada por 5 evaluadores y cada estudiante evaluó 2-3 programaciones.

	P1	**P2**	**P3**	**P4**	**P5**	**P6**	**P7**	**P8**	**P9**	**P10**	**P11**	**P12**	**P13**
Ev01	OB												OB

ISBN: 978-84-1122-023-1

Ev05	OB	OB											
Ev13	OB	OB											vo
Ev15		OB	OB										
Ev03		OB	OB										
Ev09			OB	OB			vo						
Ev25			OB	OB									
Ev07				OB	OB							vo	
Ev23				OB	OB								
Ev04					OB	OB							
Ev19					OB	OB							
Ev18						OB	OB			vo			
Ev26						OB	OB						
Ev12							OB	OB					
Ev17					vo		OB	OB					
Ev06								OB	OB				
Ev20								OB	OB		vo		
Ev22		vo							OB	OB			
Ev02									OB	OB			
Ev11			vo							OB	OB		
Ev16						vo				OB	OB		
Ev21											OB	OB	
Ev27											OB	OB	
Ev14												OB	OB
Ev10									vo			OB	OB
Ev08	OB												OB
Ev24	vo			OB				OB					
TOTAL	5	5	5	5	5	5	5	5	5	5	5	5	5

Nota. “OB” significa evaluación obligatoria, “vo” significa evaluación voluntaria (adicional).

Como puede verse en dicha tabla, las evaluaciones de los estudiantes Ev13 y Ev15 estarán conectadas, porque ambos evalúan la programación P2. Sin embargo, el estudiante Ev13 no está conectado directamente con el estudiante Ev09 (no evalúan

ISBN: 978-84-1122-023-1

ninguna programación en común), sin embargo, están indirectamente conectados a través de las evaluaciones que realice el estudiante Ev03, ya que este evaluará la programación P2 (compartida con Ev13) y la P3 (compartida con Ev09). De esta manera es posible comparar las severidades de todos los estudiantes entre ellas, bien sea por conexiones directas o indirectas, de otra forma el modelo estadístico no funcionaría. Las evaluaciones voluntarias se fueron distribuyendo entre todas las programaciones para reforzar las conexiones directas e indirectas entre todos los evaluadores. Finalmente, todas las programaciones fueron evaluadas por 5 estudiantes.

2.1. Evaluación con MOARS

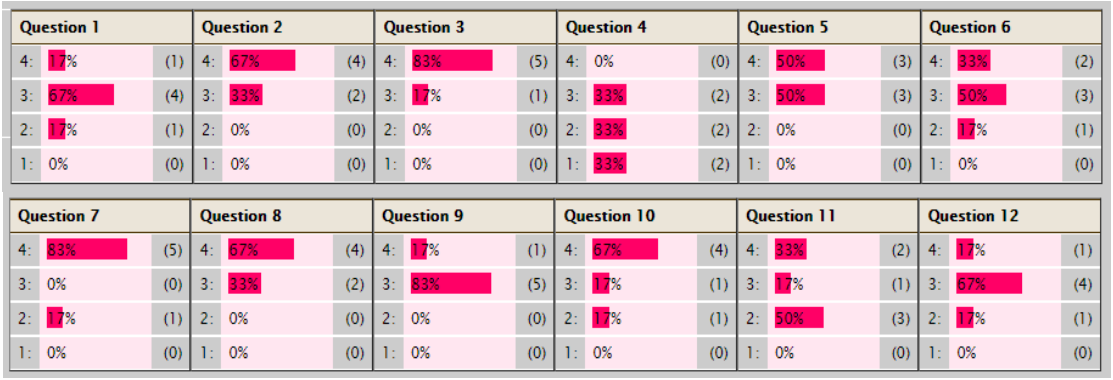
MOARS (*MObile Audience Response System*) es un software gratuito, que puede usarse en cualquier dispositivo que tenga un navegador de Internet, por ejemplo, los propios dispositivos móviles del alumnado. Para poder usarlo en esta experiencia, previamente tuvimos que descargarlo (junto con su módulo *Peer Assessment*) de la web www.moars.com y, a continuación, instalarlo en un servidor web con PHP5 y MySQL. Una vez instalado, se creó un curso dentro de MOARS y se añadieron los estudiantes, generándose para cada uno de ellos un nombre de usuario y contraseña, que luego el docente tuvo que compartir con cada estudiante. A continuación, fue necesario crear la actividad de evaluación entre iguales en MOARS e introducir la rúbrica de evaluación que posteriormente utilizarían nuestros estudiantes. Puede encontrarse más información sobre el programa MOARS y su funcionamiento en Jiménez (2021).

Después de que los estudiantes hubieran realizado las evaluaciones de las programaciones asignadas, el docente activó la opción en MOARS para que cada estudiante pudiera consultar el resultado de la evaluación de la programación en la que había participado: lo que veían era una serie de gráficos de barras (uno para cada ítem de la rúbrica) en el que se indicaba el número de compañeros (y el porcentaje que ello representaba) que eligió cada uno de los cuatro niveles de logro de la rúbrica (figura 1). Estos gráficos, junto con los resultados individualizados y tablas clasificatorias de los mismos, están disponibles también para el docente, a través de la opción “classroom results” de MOARS.

Figura 1

Visualización de las valoraciones otorgadas a una programación en MOARS

ISBN: 978-84-1122-023-1



Question 1			Question 2			Question 3			Question 4			Question 5			Question 6		
4:	17%	(1)	4:	67%	(4)	4:	83%	(5)	4:	0%	(0)	4:	50%	(3)	4:	33%	(2)
3:	67%	(4)	3:	33%	(2)	3:	17%	(1)	3:	33%	(2)	3:	50%	(3)	3:	50%	(3)
2:	17%	(1)	2:	0%	(0)	2:	0%	(0)	2:	33%	(2)	2:	0%	(0)	2:	17%	(1)
1:	0%	(0)	1:	0%	(0)	1:	0%	(0)	1:	33%	(2)	1:	0%	(0)	1:	0%	(0)

Question 7			Question 8			Question 9			Question 10			Question 11			Question 12		
4:	83%	(5)	4:	67%	(4)	4:	17%	(1)	4:	67%	(4)	4:	33%	(2)	4:	17%	(1)
3:	0%	(0)	3:	33%	(2)	3:	83%	(5)	3:	17%	(1)	3:	17%	(1)	3:	67%	(4)
2:	17%	(1)	2:	0%	(0)	2:	0%	(0)	2:	17%	(1)	2:	50%	(3)	2:	17%	(1)
1:	0%	(0)	1:	0%	(0)	1:	0%	(0)	1:	0%	(0)	1:	0%	(0)	1:	0%	(0)

Para realizar el análisis estadístico según el modelo MFRM, se utilizó el software comercial Facets (v. 3.83.3), del cual existe también una versión gratuita, totalmente funcional, aunque limitada a 2000 datos, llamada Minifac. Ambos programas, que funcionan bajo Windows, pueden conseguirse en la página www.winsteps.com. La introducción de los datos de las evaluaciones en Facets resulta muy fácil a partir de MOARS ya que este programa tiene una opción, "Research data", que exporta los resultados de las evaluaciones en un formato directamente compatible con Facets, lo cual simplifica el proceso. Una vez importados los datos a analizar, se ejecutó Facets.

3. RESULTADOS Y DISCUSION

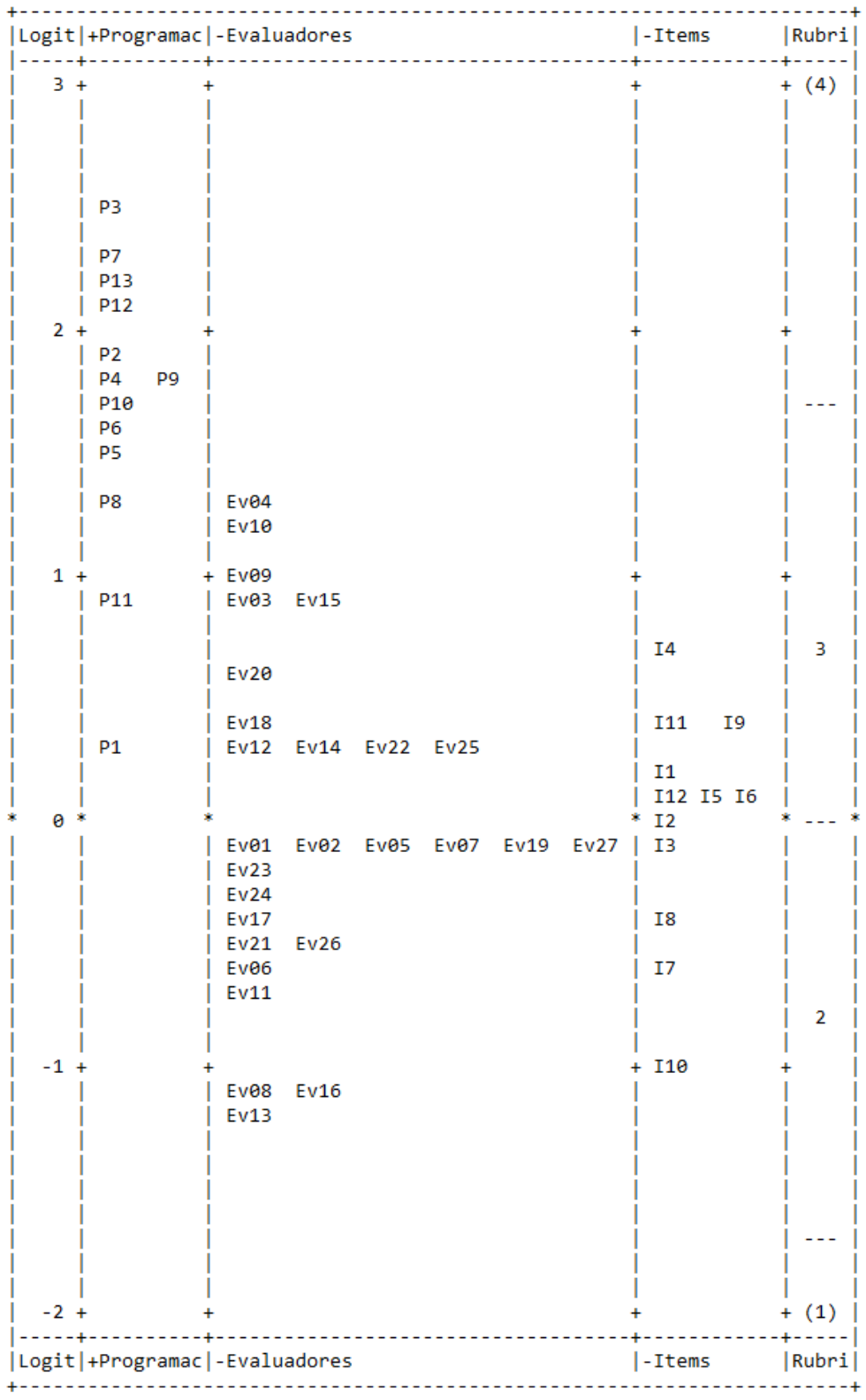
Al ejecutar Facets, el sistema devuelve un fichero con el resultado del análisis MFRM, que incluye diferentes estadísticos, tablas y gráficos. La figura 2 presenta uno de estos gráficos, el mapa de Wright o mapa de la medida de las facetas analizadas, que da una visión general del análisis estadístico realizado, así como de la dispersión entre los diferentes elementos de cada faceta.

La primera columna de esta tabla ("Logit") muestra la escala común en la que se han medido todas las facetas: el lógito, o logaritmo del cociente entre la probabilidad de que un estudiante reciba, en un ítem de la rúbrica, un nivel de logro determinado (por ejemplo, 2) y la probabilidad de que reciba el nivel de logro inmediatamente inferior (1). Esta escala puede oscilar entre 0 (fijado en el nivel medio de las facetas) y $\pm\infty$.

ISBN: 978-84-1122-023-1

Figura 2

Mapa de Wright o de las medidas de las facetas analizadas



La segunda columna (“Programac”) ordena las programaciones de los estudiantes, según la primera faceta analizada: el rendimiento de los estudiantes, expresado en forma

ISBN: 978-84-1122-023-1

de las calificaciones justas que recibieron sus programaciones. Cuanto más alta es la posición de la programación en esta segunda columna, mayor es su valor en lógitos y, por tanto, mayor es la calificación de la programación. Pero además de en lógitos, esta faceta (y solo ésta) también está expresada en la escala original de la rúbrica, es decir, la escala de 1 a 4 que los estudiantes utilizaron para evaluar las programaciones: es la información que se muestra en la quinta columna ("Rubri"). Los valores numéricos de las calificaciones recibidas y de los lógitos de las programaciones se muestran en la tabla 3.

Tabla 3

Calificaciones recibidas, calificaciones justas y lógitos de las programaciones, ordenadas por calificación justa/lógitos)

Programación	Calificación recibida	Calificación justa	Lógitos
3	3,53	3,73	2,47
7	3,60	3,69	2,32
13	3,62	3,66	2,20
12	3,53	3,63	2,13
2	3,42	3,56	1,90
4	3,42	3,51	1,78
9	3,35	3,51	1,78
10	3,52	3,47	1,67
6	3,38	3,46	1,67
5	3,30	3,40	1,49
8	3,32	3,31	1,32
11	3,27	3,12	0,95
1	3,00	2,70	0,27

A la vista de los datos de la tabla 3, P3 es la programación que ha recibido una mejor evaluación de sus compañeros, con una calificación justa de 3,73, según la escala original de la rúbrica, lo que corresponde a 2,47 lógitos. Sin las correcciones del sesgo de severidad/benevolencia, esta programación habría obtenido una calificación de 3,53 y habría sido superada por P13 (ahora en tercera posición) con una calificación de 3,62. Estas dos programaciones, junto con P7 y P12, son las mejor valoradas por los

ISBN: 978-84-1122-023-1

estudiantes, con valores superiores a los 2 lógitos. Las dos programaciones que peores calificaciones han recibido de sus compañeros son P11 y, especialmente, P1, ambas por debajo de 1 lógito. En los dos casos, el recálculo para obtener calificación justa les ha perjudicado: P11 había recibido una calificación inicial de 3,27, en contraposición con una calificación justa de 3,12; mientras que P1 había recibido una calificación inicial de 3,00, siendo 2,70 la calificación final, una vez eliminado el sesgo de benevolencia del que ambas programaciones se habían beneficiado. Además, si la calificación final se hubiese dado en números enteros, en coherencia con la escala ordinal de valoración de los ítems de la rúbrica, las programaciones P2, P4 y P9 se habrían visto perjudicadas de no haberse corregido los sesgos de severidad/benevolencia: en estos tres casos, la calificación justa es de 4, pero sin eliminar los sesgos de severidad y benevolencia sería de 3. Contrariamente, la programación P10 se habría visto beneficiada, ya que su calificación justa redondeada es de 3, mientras que este valor sin eliminar sesgos sería de 4.

La tercera columna ("Evaluadores") ordena a los estudiantes, según la segunda faceta analizada: su severidad como evaluadores. Valores positivos indican mayor severidad, valores negativos indican mayor generosidad o benevolencia y este sesgo es significativo, en uno u otro sentido, cuando su valor absoluto es igual o superior a 1. La severidad media se sitúa en 0 lógitos. Se observa una cierta dispersión entre los valores extremos: los más severos fueron los estudiantes Ev04 y Ev10, ambos con lógitos superiores a 1: 1,26 y 1,17, respectivamente, lo que indica que el sesgo de severidad fue significativo en estos dos casos. En el polo opuesto, los estudiantes Ev08, Ev16 y Ev13 mostraron un sesgo significativo de benevolencia, con lógitos por debajo de -1.

Encontramos que los estudiantes que muestran un sesgo significativo de severidad (Ev04 y Ev10) son estudiantes con peores rendimientos académicos: Ev10 fue coautor de la P1, la segunda peor valorada por sus compañeros. El estudiante Ev04, si bien la programación de la que es coautor (P4) se sitúa en una posición intermedia, suspendió la asignatura. Este hallazgo concuerda con estudios anteriores (Jiménez y Llitjós, 2006), en los que comprobamos que estudiantes que recibían calificaciones bajas de sus compañeros solían ser, a su vez, muy severos calificando el trabajo de los demás. Y, curiosamente, uno de los coautores (Ev13) de la programación mejor valorada (P3) ha resultado ser uno de los estudiantes con sesgo de benevolencia significativo.

Además del grado de severidad o benevolencia de los evaluadores, también podemos estudiar el nivel de consistencia de sus evaluaciones. El outfit (*outlier fit mean-square*) es un parámetro sensible a las calificaciones puntual e inesperadamente altas, mientras

ISBN: 978-84-1122-023-1

que el infit (*inlier fit mean-square*) es un parámetro que proporciona una estimación de la consistencia con la que un evaluador determinado utiliza la escala de valoración a lo largo de los ítems evaluables y de las programaciones evaluadas. Ambos parámetros, que pueden variar de 0 a infinito, tienen un valor esperado de 1. Cuando sus valores son superiores a 1 indican una mayor variación de la esperada en sus evaluaciones (el desajuste sería muy severo a partir de 2, lo que implicaría una distorsión o degradación del modelo estadístico), mientras que valores inferiores a 1 indican menor variación de la esperada. Linacre (2002) ha sugerido que se produce un "ajuste útil" de los datos al modelo estadístico cuando ambos parámetros se encuentren entre el rango 0,50-1,50. En nuestro caso, los valores medios de outfit y de infit de todos los evaluadores son 1,03 y 1,05, respectivamente, valores muy próximos a 1 y que están dentro del rango óptimo propuesto por Linacre para dar por bueno el ajuste. En cuanto a los valores individuales de estos dos parámetros, solo un estudiante, Ev11, supera estos límites, con un infit de 1,82 y un outfit de 1,88 (pero aún inferiores a 2). Superar el límite de 1,50 propuesto por Linacre puede implicar que el estudiante ha mostrado un comportamiento errático o aleatorio a la hora de realizar sus evaluaciones, es decir, muestra ciertas dificultades para utilizar de forma consistente la rúbrica de evaluación. Esto queda confirmado en la tabla que genera Facets con las "respuestas inesperadas": de un total de 780 valoraciones (13 programaciones, cada una evaluada 5 veces y cada evaluación implica 12 ítems de valoración), el análisis ha identificado 9 valoraciones "inesperadas", dos de las cuales tienen al estudiante Ev11 como evaluador.

La cuarta columna ("Items") ordena la tercera faceta analizada, esto es, los ítems que conforman la rúbrica, representados por I+número de ítem en la rúbrica, en función de su dificultad: cuanto más alta es su posición en la columna y, por tanto, mayor es su valor en lógitos, más difícil resultó ser ese ítem, es decir, más difícil fue que las programaciones tuvieran una valoración alta en dicho ítem. La dificultad media se sitúa en 0 lógitos. El ítem que ha resultado ser el más difícil fue el número 4, referido a la contextualización de las actividades de la programación ("las situaciones de aprendizaje se plantean con preguntas o como problemas a resolver contextualizados), con 0,71 lógitos (y una calificación media justa de 3,14). En cambio, el ítem en el que globalmente han obtenido mejores valoraciones es el I10, que hace referencia a la evaluación ("Incluye evaluación formativa y formadora. Además, se incluyen procedimientos de feedback, autoevaluación y coevaluación), con -1,03 lógitos (y una calificación media justa de 3,78).

ISBN: 978-84-1122-023-1

Una vez la actividad de evaluación entre iguales hubo concluido, el profesor mostró al alumnado el mapa de Wright (figura 2) y explicó el significado de cada columna y de la posición de los diferentes elementos en ellas, enfatizando las posiciones de los evaluadores que indicaban sesgos significativos de severidad o de benevolencia. Además, informó privadamente al estudiante Ev11 del carácter errático e inconsistente de su evaluación. Con todo ello, se abrió un debate en clase alrededor de la evaluación y del componente subjetivo que siempre lleva asociada y se animó a los estudiantes a que incluyeran sus reflexiones en el portafolios de la asignatura. En la revisión de los portafolios del alumnado la mayoría de las reflexiones hacían referencia al carácter innovador de la experiencia y al carácter formativo y formador de la misma ("recibir feedback es la mejor forma de mejorar", "esta actividad nos ha permitido analizarnos a nosotros mismos como evaluadores"), aunque algunos comentarios todavía reflejaban ciertas reticencias hacia las calificaciones otorgadas por sus compañeros ("puede que algunos compañeros hayan sido un poco injustos en sus calificaciones"). Finalmente, se pidió al alumnado que valorara globalmente esta experiencia de evaluación. Del total de respuestas recogidas (N=22), 11 (la mitad del alumnado) indicaron que la experiencia había sido "positiva", mientras que la otra mitad (11) la valoraron como "muy positiva".

4. CONCLUSIONES

El análisis estadístico, según el modelo MFRM ha permitido identificar y cuantificar los sesgos de severidad y de benevolencia que ha mostrado nuestro alumnado en una evaluación entre iguales, en la que "todos evalúan a algunos" y, puesto que los datos de las evaluaciones entre iguales se ajustan bien a dicho modelo, se han podido obtener unas calificaciones más justas de la actividad realizada, es decir, unas calificaciones sin sesgos de benevolencia o severidad.

Adicionalmente, los valores individuales de ajuste al modelo nos han permitido identificar qué evaluadores han mostrado un comportamiento errático o aleatorio a la hora de evaluar las programaciones de sus compañeros.

El proceso seguido, además de corregir los sesgos en las calificaciones de los estudiantes, también facilita la reflexión en torno a la subjetividad presente en todos sus actos de evaluación, incluso cuando creen que están siendo totalmente objetivos.

ISBN: 978-84-1122-023-1

REFERENCIAS

Andridch, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Anthony, C. J., Styck, K. M., Cooke, E., Martel, J. R. y Frye, K. E. (2021). Evaluating the impact of rater effects on behavior rating scale score validity and utility. *School Psychology Review*. DOI: 10.1080/2372966X.2020.1827681.

Departamento de Educación de la Generalitat de Catalunya (2020). *Programar per competències a l'educació secundària obligatòria*, 2ª ed. Gabinet Tècnic del Departament d'Educació.

Engelhard, G. y Wind, S. A. (2017). *Invariant measurement with raters and rating scales: Rasch models for rater-mediated assessments*. Routledge.

Falchikov, N. y Goldfinch, J. (2000). Student peer assessment in higher education: A meta-analysis comparing peer and teacher marks. *Review of educational research, 70*(3), 287-322.

Jiménez, G. y Llitjós, A (2006). Deducción de calificaciones individuales en actividades cooperativas: una oportunidad para la coevaluación y la autoevaluación en enseñanza de las ciencias. *Revista Eureka sobre enseñanza y divulgación de las Ciencias, 3*(2), 172-187.

Jiménez, G. (2021). Evaluación entre iguales representativa e inmediata con dispositivos móviles en el aula de ciencias: MOARS. En *29 Encuentros de Didáctica de las Ciencias Experimentales* (pp. 28-35). Universidad de Córdoba y APICE.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. MESA Press.

Linacre, J. M. (2002). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16*, 878.

Masters, G. N. (1982). A Rasch model por partial credit scoring. *Psychometrika, 47*, 149-174.

Mintzes, J.J. y Walter, E. M. (Eds) (2020). Active Learning in College Science. Springer.

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

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. The University of Chicago Press.

ISBN: 978-84-1122-023-1

Rodríguez, G., Ibarra, M.S. y García, E. (2013). Autoevaluación, evaluación entre iguales y coevaluación: conceptualización y práctica en las universidades españolas. *Revista de Investigación en Educación, 11*(2), 198-210.

Topping, K. J. (2018). *Using peer assessment to inspire reflection and learning*. Routledge.

Wilian, D. y Leahy, S. (2015). *Embedding formative assessment*. Learning Sciences International.

Wu, M. (2017). Some IRT-based analyses for interpreting rater effects. *Psychological Test and Assessment Modeling, 79*(4), 453–470.

ISBN: 978-84-1122-023-1