COLECCIÓN CONOCIMIENTO CONTEMPORÁNEO

Innovación e investigación docente en educación: experiencias prácticas

Coordinadoras
Carmen Romero García
Olga Buzón García

Dykinson, S.L.

INNOVACIÓN E INVESTIGACIÓN DOCENTE EN EDUCACIÓN:
EXPERIENCIAS PRÁCTICAS

Diseño de cubierta y maquetación: Francisco Anaya Benítez

Madrid - 2021
N.º 31 de la colección Conocimiento Contemporáneo
1ª edición, 2021

ISBN 978-84-1377-593-7

CÁLCULO DEL SESGO DE SEVERIDAD O BENEVOLENCIA EN UNA EVALUACIÓN ENTRE IGUALES CON HERRAMIENTAS TIC

GREGORIO JIMÉNEZ VALVERDE
Grupo de Innovación Docente EduCits. Universitat de Barcelona

1. INTRODUCCIÓN

Diversos autores han señalado la importancia de incluir diferentes modalidades de evaluación en la enseñanza, que vayan más allá de la realizada exclusivamente por el docente y en la que la evaluación sea entendida como una parte más del proceso de aprendizaje del alumnado, con un enfoque que permita identificar los avances y las dificultades o errores del proceso educativo y que involucre al estudiante activamente (Anker-Hansen y Andrée, 2019; Lui and Andrade, 2014). Una de estas modalidades es la evaluación entre iguales (*peer assessment*), según la cual "los estudiantes realizan un análisis y valoración sobre las actuaciones y/o producciones desarrolladas por algún estudiante o grupo de estudiantes de su mismo estatus o nivel" (Rodríguez, Ibarra y García, 2013) y permite que, al igual que en la autoevaluación, el alumnado adopte un papel más activo en su aprendizaje (Orsmond, Merry y Reiling, 1996; Reese-Durham, 2005), promoviendo además la autorregulación de su aprendizaje y la capacidad metacognitiva de pensar más críticamente sobre su propio trabajo (Nicol et al., 2014).

En España, las disposiciones legales vigentes en materia educativa incorporan cada vez más, de manera más o menos explícita, la evaluación entre iguales (y/o la autoevaluación) en los currículos oficiales (por ejemplo, el artículo 4.3 de la orden ENS/108/2018, que regula la evaluación de la ESO en Cataluña), y en trece países europeos las directrices oficiales para la evaluación recomiendan la autoevaluación o la evaluación entre iguales durante el periodo de escolarización obligatoria

(Ministerio de Educación, Cultura y Deporte, 2012). Autores como Ibarra, Rodríguez y Gómez (2012) o Mogessie (2015) reclaman también mayor presencia de la evaluación entre iguales en la enseñanza universitaria.

Sin embargo, la implementación de la evaluación entre iguales no está exenta de dificultades. Por ejemplo, el enorme número de datos numéricos que se genera cuando esta evaluación es cuantitativa: si un docente desea llevar a cabo una evaluación entre iguales en una clase de 25 estudiantes utilizando una rúbrica de 5 ítems, en la que "todos evalúan a todos", se generan 3000 datos, lo cual puede desalentar a que dicho docente acabe llevando a cabo la evaluación, debido al esfuerzo y tiempo que puede representar analizar toda la información generada. Para salvar esta dificultad, el docente puede organizar la evaluación simplificando el número de trabajos que tienen que evaluar sus estudiantes, es decir, que cada estudiante evalúe un número determinado (normalmente bajo) de trabajos de sus compañeros, como la experiencia que se describe en Custodio, Máquez y Sanmartí (2015). En este tipo de casos, donde "todos evalúan a algunos", se pierde representatividad, ya que el trabajo de un estudiante no es evaluado por todos sus compañeros y, además, si alguno de los estudiantes tiene un sesgo de severidad a la hora de actuar como evaluador, los estudiantes evaluados por él recibirán calificaciones significativamente más bajas que las que hubieran recibido si hubiesen sido evaluados por otros compañeros sin este sesgo. El sesgo de severidad o benevolencia ha sido señalado como uno de los sesgos que pueden mostrar los estudiantes cuando evalúan a sus compañeros (Myford y Wolfe, 2003).

Las tecnologías de la información y la comunicación (TIC) y las herramientas web 2.0 podrían, no obstante, ayudar a minimizar o superar estas dos dificultades de las evaluaciones entre iguales. Además, el uso de estas herramientas en las evaluaciones entre iguales podría estar asociada a mejoras en el aprendizaje respecto de las evaluaciones que se realizan de manera tradicional, en papel, de acuerdo con el estudio de Li et al. (2020).

2. OBJETIVOS

- Desarrollar una metodología que permita realizar una evaluación entre iguales, en la que "todos evalúan a todos" y en la que sea ágil y sencilla la gestión de todos los datos de evaluación generados.
- Detectar y cuantificar el sesgo de benevolencia o de severidad de los estudiantes, cuando estos actúan como evaluadores.

3. METODOLOGÍA

La experiencia descrita en este trabajo se ha llevado a cabo durante el curso 2019-2020 con estudiantes de la asignatura "Técnicas fisicoquímicas y químicas de análisis de aguas" del Ciclo Formativo de Química Ambiental, en el Institut Mercè Rodoreda, de L'Hospitalet de Llobregat (Barcelona). En total, participaron 34 estudiantes (21 mujeres y 13 hombres), correspondientes a dos grupos-clase y de edades comprendidas entre los 17 y los 43 años.

La actividad tuvo lugar al inicio del curso y, puesto que los estudiantes no se conocían entre ellos, se minimizó el problema de la posible falta de objetividad en sus evaluaciones (sesgo de amistad), tal y como menciona Chen (2010) en una experiencia similar de evaluación entre iguales. La actividad, de carácter voluntario, consistió en elegir una noticia relacionada con el agua y realizar una presentación oral, de entre 4 y 6 minutos de duración, utilizando un soporte digital (tipo PowerPoint o Prezi), en la que tenían que destacar los aspectos más importantes de la noticia y hacer un comentario crítico sobre la misma.

Después de cada presentación oral y utilizando sus propios dispositivos móviles, los estudiantes tuvieron que realizar dos evaluaciones: una cualitativa, usando la aplicación PollEveywhere y otra cuantitativa, con el programa MOARS. Estas dos evaluaciones fueron anónimas, siguiendo las conclusiones del estudio de Panadero y Alqassab (2019), quienes señalaron que una evaluación entre iguales anónima brinda unos comentarios más críticos por parte de los evaluadores y apunta a una ligera tendencia a un mejor desempeño.

Los resultados recogidos en la evaluación cuantitativa fueron analizados estadísticamente por el docente con el software Minifac, que permite detectar y cuantificar el sesgo de severidad o benevolencia de los evaluadores. A continuación, se explica con más detalle cada una de las tres herramientas digitales mencionadas.

3.1. Evaluación entre iguales cualitativa: PollEverywhere

PollEverywhere (www.polleverywhere.com) es una tecnología web 2.0 que permite obtener un feedback inmediato del alumnado, tanto en preguntas de respuesta múltiple como en preguntas abiertas y, por defecto, lo hace manteniendo el anonimato de los evaluadores. Además de los planes de pago, PollEverywhere dispone de un plan gratuito –que es el que se ha utilizado en este trabajo- que permite recoger hasta 40 respuestas para una pregunta determinada, sin límite de preguntas.

Para que los estudiantes pudieran realizar las evaluaciones cualitativas, también llamadas *peer feedback* (Gielen et al., 2020), el docente creó previamente una pregunta abierta para cada estudiante (34 en total) y les pidió que, usando sus dispositivos móviles y PollEverywhere, indicaran un aspecto positivo y una propuesta constructiva de mejora de cada presentación, adaptación del método de "dos estrellas y un deseo" propuesto por Wiliam y Leahy (2015) para dar un feedback formativo. No fue necesario que el docente facilitara los 34 enlaces individuales para evaluar a cada estudiante, sino que PollEverywhere permite simplificar este proceso: el alumnado se dirige a la URL de la cuenta del docente en PollEverywhere (www.pollev.com/*nombre-de-usuario*) y, sin que los evaluadores tengan que realizar ninguna acción adicional (ni siquiera actualizar la página web cada vez que presenta un nuevo compañero), van apareciendo en los dispositivos móviles de los estudiantes cada una de las evaluaciones cualitativas que tienen que realizar, a medida que el docente va activando, desde su panel de control, cada una de ellas. También se pedía a quien acababa de realizar la presentación oral, el evaluado, que realizara una autoevaluación o un comentario crítico sobre su propia actuación, y para que el docente pudiese identificar este comentario del resto, este estudiante tenía que escribir tres asteriscos al final de su autoevaluación.

3.2. Evaluación entre iguales cuantitativa: 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*, PA) 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 de ese curso, 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 una "encuesta" en MOARS, que es la rúbrica de evaluación de la actividad, indicando las preguntas (PR) y las opciones posibles de respuesta. En nuestro caso, las preguntas consistían en 8 afirmaciones sobre diferentes aspectos de la presentación oral de las que los estudiantes-evaluadores tenían que indicar su grado de acuerdo, según una escala de Likert del 1 al 7, siendo 1=" totalmente en desacuerdo" y 7="totalmente de acuerdo". Las afirmaciones eran:

- La primera diapositiva contenía título y nombre del estudiante, la presentación ha durado entre 4 y 6 min, la última diapositiva contenía un resumen de la presentación.
- Las diapositivas contienen frases cortas y fáciles de leer (tamaño y color de la letra adecuados, con un fondo que facilita la lectura), hay un buen equilibrio entre imágenes y texto.
- Se ha utilizado un vocabulario simple y preciso, usando sus propias palabras. El texto no presentas faltas de ortografía
- La presentación es coherente y está bien organizada. Sigue un orden lógico.
- La exposición oral ha sido clara y concisa, con un tono de voz alto y relajado, postura corporal correcta y sin detenerse a leer las diapositivas.

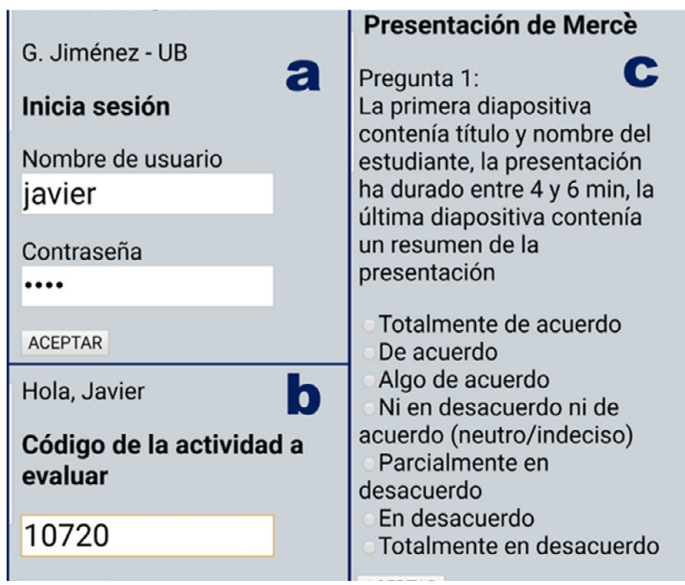
– En la primera parte de la presentación, la noticia elegida ha sido explicada y descrita con rigor y claridad.

– En la segunda parte de la presentación, el análisis crítico de la noticia es adecuado, riguroso y con conexiones a otros temas

– La presentación ha sido interesante, tanto por el tema elegido como por la manera de exponerla, y el orador ha conseguido atraer la atención del público.

Para acabar de configurar la rúbrica en MOARS, el docente tuvo que abrir el módulo PA y seleccionar la "encuesta" correspondiente a esta actividad. El sistema entonces asignó automáticamente un código numérico (llamado "actividad a evaluar") a cada una de las presentaciones, que es el código que el docente tenía que facilitar a los evaluadores justo antes de cada presentación oral, para que pudieran realizar la evaluación (figura 1). Además, el profesor también activaba la pregunta abierta de PollEverywhere correspondiente a esa presentación.

***FIGURA 1**. Ejemplo de uso de MOARS por parte de un estudiante-evaluador. El usuario "javier" inicia sesión con su nombre de usuario y contraseña (a), luego introduce el código a evaluar de la presentación de Mercè (b) y ya puede iniciar la evaluación (c).*

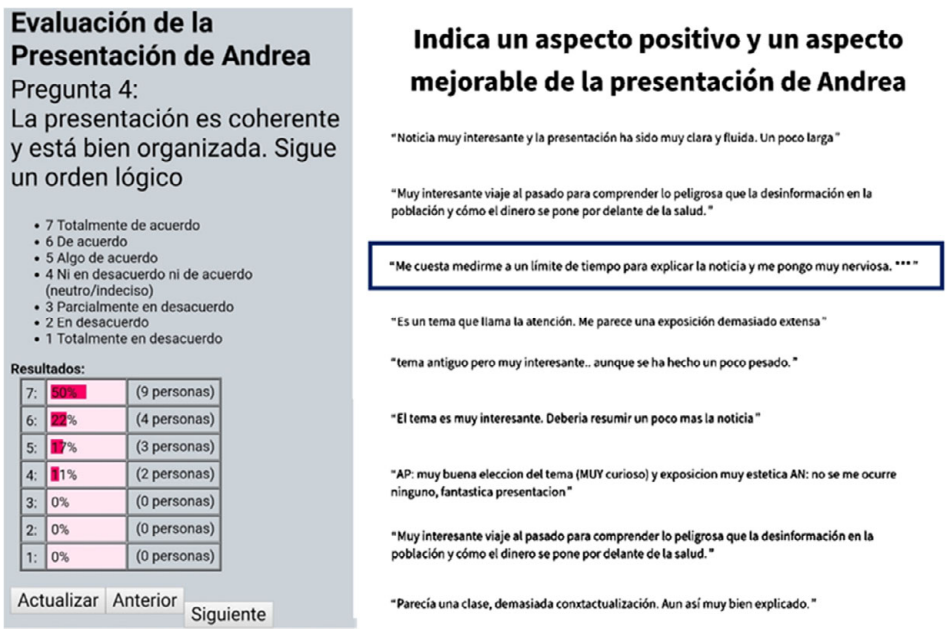


Fuente: autoría propia (a partir de www.moars.es)

Al finalizar cada presentación, los evaluadores realizaban la evaluación cualitativa y cuantitativa y el evaluado únicamente realizaba su autoevaluación en PollEverywhere (MOARS no permite la autoevaluación), disponiendo todos de unos minutos para hacerlo. El docente también fue evaluando las presentaciones de sus estudiantes, tanto cualitativa como cuantitativamente, y para el evaluado, la valoración del profesor era indistinguible de las que recibía de sus compañeros.

Una vez concluidas todas las presentaciones orales, el docente activó la posibilidad de que cada estudiante pudiera consultar, en MOARS, el feedback cuantitativo recibido: lo que veían era una serie de gráficos de barras (uno para cada una de las preguntas de la rúbrica, PR), en el que se indicaba el número de compañeros (y porcentaje) que eligió cada una de las siete opciones posibles de la escala Likert. Además, facilitó a cada estudiante el enlace de su pregunta abierta de PollEverywhere en el que podían consultar el feedback cualitativo recibido (figura 2).

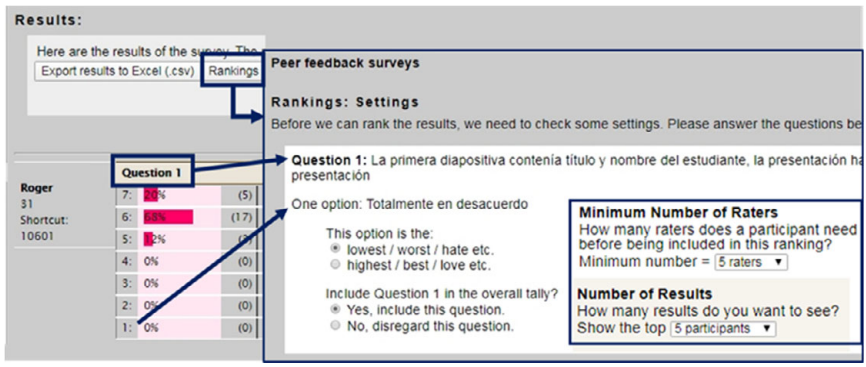
FIGURA 2. *A la izquierda, resultado de la evaluación cuantitativa recibida por una estudiante en una de las preguntas (PR4). A la derecha, extracto de la evaluación cualitativa recibida por esta misma estudiante (en el recuadro, su comentario autoevaluativo).*



Fuente: autoría propia a partir de www.moars.es y www.polleverywhere.com

Desde la perspectiva del docente, los resultados de la evaluación recogidos por MOARS están disponibles en dos formatos. En el primero de ellos, "classroom results", el docente ve los resultados de cada estudiante en forma de gráficos de barras (un histograma para cada pregunta de la rúbrica, PR), con la posibilidad de exportarlos a una hoja de cálculo, o puede ver los resultados en formato de rankings, con diferentes opciones de configuración (figura 3). El segundo formato, "research data", exporta los datos de las evaluaciones para ser analizados estadísticamente por Minifac o Facets.

FIGURA 3. *"Classroom results" de la evaluación entre iguales. Puede verse el resultado obtenido en la primera pregunta de la rúbrica, PR1, por el alumno Roger ("actividad a evaluar" 10601), con el número de compañeros (y porcentaje) que eligieron cada una de las posibles respuestas. A la derecha, menú que aparece si se selecciona la opción "ranking" para mostrar los resultados y en la que es necesario indicar qué preguntas se desea que se tengan en cuenta para confeccionar dicho ranking y si la respuesta 1 de cada pregunta corresponde a la peor o a la mejor valoración.*



Fuente: autoría propia (a partir de www.moars.es)

3.3. Cálculo del sesgo de severidad o benevolencia: MINIFAC

Como ya se ha indicado, es adecuado realizar un análisis estadístico si se desea identificar y cuantificar la severidad/benevolencia de cada estudiante cuando actúa como evaluador. Nosotros aplicamos el modelo de Rasch de múltiples facetas (*Many-Facet Rasch Measurement*, MFRM) a los datos recogidos por MOARS. Este modelo analiza simultáneamente diferentes variables (llamadas "facetas") que pueden afectar a la calificación final de un estudiante, entre ellas el grado de

severidad o benevolencia de los compañeros que le evaluaron (Eckes, 2015). Para llevar a cabo este análisis estadístico utilizamos el programa gratuito Minifac (www.winsteps.com/minifac.htm), que es una versión limitada de un software comercial llamado Facets. Ambos funcionan únicamente bajo Windows y permiten realizar el análisis MFRM, pero la versión gratuita solo permite procesar 2000 datos, aunque mantiene todas las funcionalidades de la versión comercial. Precisamente, cuando MOARS exporta los datos con la opción "Research data" lo hace generando un archivo directamente compatible con cualquiera de estos dos programas, lo cual simplifica dicho análisis.

Al ejecutar Minifac el sistema genera un archivo con los resultados del análisis estadístico. Entre ellos, el mapa de las medidas de las facetas analizadas o mapa de Wright, que es una tabla generada por Minifac en la que se ofrece un resumen visual de los resultados del análisis MFRM. Debido a la limitación anteriormente comentada de Minifac solo se pudieron analizar simultáneamente 5 de las 8 preguntas de la rúbrica de evaluación, que es a lo que corresponde el mapa de Wright de uno de los dos grupos-clase que se muestra en la figura 4.

La primera columna de esta tabla ("Measr") 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 una puntuación en una pregunta (por ejemplo, 4) y la probabilidad de que reciba la puntuación inmediatamente inferior (3). Esta escala puede oscilar entre 0 (fijado en el nivel medio de las facetas) y $\pm\infty$.

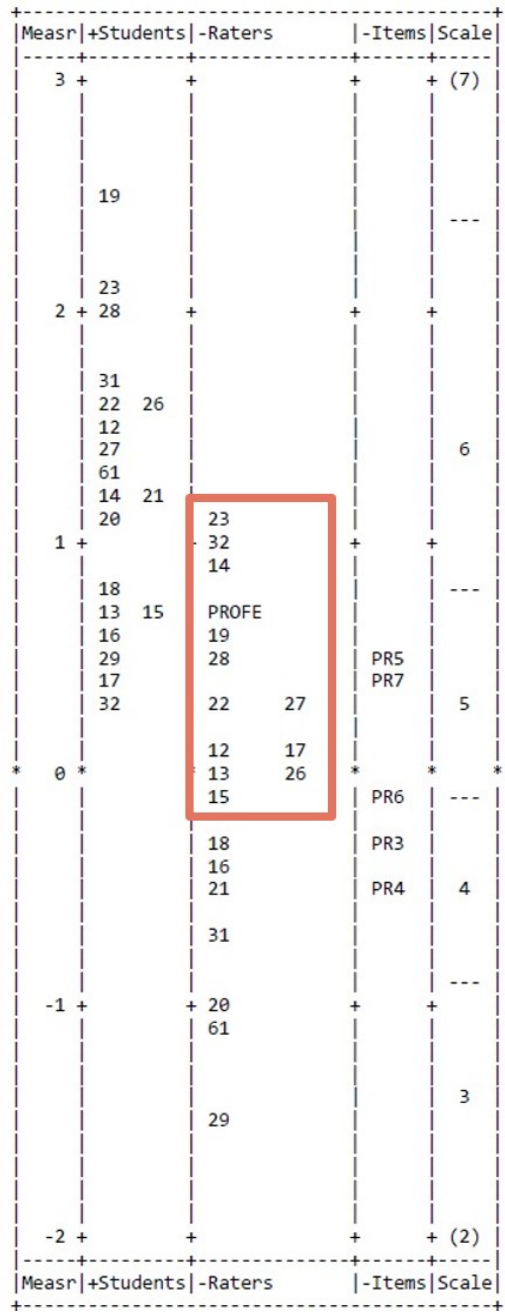
La segunda columna ("Students") distribuye a los estudiantes según la primera faceta analizada: su rendimiento global en la actividad, en este caso, es la calificación global de la presentación oral. Valores superiores o inferiores a 0 lógitos indican mayor o menor rendimiento de los estudiantes. Pero además de en lógitos, esta faceta (y solo esta) también está expresada en la escala original de la rúbrica, es decir, la escala Likert del 1 al 7: es la información que se muestra en la quinta columna ("Scale"). A la vista de los datos, el estudiante que ha mostrado mayor rendimiento en la actividad ha sido el número 19, con 2,52 lógitos (y 6,48 en la escala original), seguido por un segundo grupo formado por los estudiantes 23 y 28; un tercer grupo, formado por 9 estudiantes con

lógitos entre 1,67 y 1,10 y, por último, un cuarto grupo con los estudiantes con menor rendimiento (lógitos inferiores a 1), cerrado por el estudiante número 32, con solo 0,28 lógitos (4,98 en la escala de Likert original).

La tercera columna ("Raters") 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. Se observa una cierta dispersión entre los valores extremos: el más severo fue el estudiante número 23, con 1,06 lógitos y el más generoso fue, con diferencia, el número 29, con -1,50 lógitos. Vemos también que el estudiante con peor rendimiento (32) fue el segundo más severo a la hora de evaluar a sus compañeros, algo que 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 estrictos calificando el trabajo de los demás. Por último, cabe añadir que, como el profesor también participó en la evaluación de las presentaciones de los estudiantes usando MOARS, puede verse su posición en esta columna ("PROFE", con 0,92 lógitos) y, por tanto, comprobar qué estudiantes no mostraron un grado de severidad significativamente diferente a la de él (son aquellos dentro del rectángulo de esta columna).

La cuarta columna ("Items") ordena la tercera faceta analizada, esto es, las preguntas de la rúbrica (PR), en función de su dificultad: cuanto mayor es el valor en lógitos, más difícil resultó ser ese ítem, es decir, más difícil fue que los evaluados tuvieran una calificación alta en dicho ítem. La afirmación PR5 ("la exposición oral ha sido clara y concisa, con un tono de voz alto y relajado, postura corporal correcta y sin detenerse a leer las diapositivas") fue el ítem de evaluación considerado más difícil, al ser el que recibió la puntuación más alta (0,50 lógitos) y la afirmación PR4 ("la presentación es coherente y está bien organizada. Sigue un orden lógico") fue el ítem considerado más fácil (-0,51 lógitos). Es en esta faceta donde encontramos la menor dispersión entre los valores mayor y menor.

FIGURA 4. *Mapa de Wright de uno de los dos grupos-clase analizados. Los estudiantes están identificados con un código numérico.*



Fuente: autoría propia, a partir de Minifac.

Para la obtención de la calificación final de la actividad de un estudiante determinado (que le serviría como bonificación en la puntuación del primer examen de la asignatura), se tuvo en cuenta la calificación otorgada por el docente y por sus compañeros de su presentación oral. Con el objetivo de que se tomaran en serio su papel de evaluadores y para evitar que dieran calificaciones arbitrarias, sin sentido o injustificadas, se les advirtió que la calificación final de la actividad podría sufrir una penalización si mostraban sesgos significativos de benevolencia o severidad, de si su posición en la columna "Raters" en el mapa de Wright quedaba muy alejada de la del docente o de si de la inspección de sus datos como evaluadores se observaban calificaciones al azar, aleatorias o incoherentes. El alumnado conocía estos criterios de evaluación, así como la rúbrica de la evaluación entre iguales cuantitativa, con anterioridad al inicio de las presentaciones.

4. RESULTADOS

La valoración de esta experiencia por parte de los estudiantes se llevó a cabo a través de una encuesta que estos completaron al finalizar la actividad (N=34). Dicha encuesta constaba de dos partes: en la primera tenían que indicar cómo valoraban diferentes dimensiones de la actividad y en la segunda se les pidió que indicaran algún aspecto positivo y negativo de la experiencia.

Con respecto a la primera parte, los estudiantes consideraron positivo o muy positivo tanto poder recibir un feedback cualitativo (91,1%) como una evaluación cuantitativa (82,4%). Sin embargo, estos porcentajes descendieron sensiblemente, especialmente el cuantitativo, cuando se les preguntó acerca de ser ellos los evaluadores. Así, el porcentaje del alumnado que consideró positivo o muy positivo poder evaluar cuantitativamente a sus compañeros bajó hasta el 70,6%, mientras que la bajada en el caso de la evaluación cualitativa no fue tan pronunciada, quedándose en el 85,3%. Poder elegir el tema de la presentación fue valorado como positivo o muy positivo por casi la totalidad del alumnado (95,6%), mientras que hubo menor nivel de consenso cuando se les preguntó acerca de que en la calificación global de la presentación se

tuviera en cuenta, además de la del profesor, la valoración de sus compañeros: un 64,7% lo consideró como positivo o muy positivo, pero un 35,3% lo valoró como "neutro". Si bien todos los estudiantes consideraron que hubo tiempo suficiente para realizar la presentación oral, un 38,2% manifestó que el tiempo disponible para realizar las evaluaciones (entre presentación y presentación) fue insuficiente.

En cuanto a la utilidad de las evaluaciones recibidas de cara a realizar mejores presentaciones en el futuro, un 58,9% se mostró de acuerdo o muy de acuerdo en que la evaluación cuantitativa recibida les sería útil en futuras presentaciones, mientras que el porcentaje para esa misma cuestión referida a la evaluación cualitativa fue del 76,5%. De hecho, cuando se les preguntó cuál de las dos evaluaciones les resultaría más útil de cara al futuro, el 17,6% de los estudiantes consideró que el feedback cuantitativo (MOARS) les sería más útil para mejorar en futuras presentaciones, mientras que el 47,1% opinó que el feedback cualitativo (PollEverywhere) es el que les sería más útil. El resto (35,3%) consideró que los dos tipos de feedback son igualmente útiles de cara a mejorar en futuras presentaciones. Sobre la evaluación cualitativa cabe destacar que un 70,6% del alumnado consideró que los comentarios que habían recibido eran "realistas" y en cuanto a la rúbrica, el 88,2% del alumnado manifestó que había entendido qué se tenía que evaluar en todas las preguntas de esta.

Por lo que concierne a la segunda parte, algunos de los comentarios positivos recibidos con mayor frecuencia hacían referencia a la utilidad del feedback recibido ("te ayudan a mejorar algunos aspectos para siguientes presentaciones", "me pareció muy bien que los compañeros pudieran opinar y aportar críticas constructivas sobre las presentaciones de los otros"), incluso señalando que al no conocerse entre ellos, el feedback recibido era más objetivo ("saber qué opinan mis compañeros de mi manera de hacer una presentación objetivamente, ya que no me conocían de nada"). También se registraron comentarios positivos en relación con la ejecución de la actividad: que era dinámica ("es dinámica y moderna"), innovadora ("me gusta que se utilice el móvil para evaluar y no hacerlo a mano"), o que, al tener que evaluar a otros compañeros, les hacía prestar más atención ("al evaluar a los compañeros

prestamos más atención a lo que se está explicando y como se explica y se aprende más"). También hubo quien destacó como aspecto positivo la posibilidad de subir nota con la actividad ("ayuda a subir nota, lo que está muy bien para los que quieren subirla") y la posibilidad de elegir libremente el tema de la exposición ("libertad para poder elegir el tema").

En cuanto a los aspectos negativos, algunos comentarios lamentaban que la exposición fuera individual ("propondría hacerlo por grupos", "la presentación, mejor en parejas"), que los evaluadores pudieran no ser siempre objetivos o no estuvieran atentos ("algunas personas no han sido objetivas a la hora de evaluar a los compañeros", "a veces los compañeros evalúan sin haber estado atentos a la presentación"), que se tenga en cuenta la evaluación de los compañeros en la nota final ("no dar tanta relevancia a la valoración de los compañeros a la hora de calificar definitivamente la nota de la presentación"), o que la escala de valoración era demasiado amplia ("el rango de valoración era bastante amplio") o que había poco tiempo para evaluar, ("tener algo más de tiempo para responder a las encuestas entre presentación y presentación"). Este último aspecto ya había sido identificado en la primera parte de la encuesta.

Finalmente, se les preguntó "¿Cuál es tu valoración global hacia esta actividad?" y un 94,1% del alumnado contestó con "positiva" o "muy positiva".

5. DISCUSIÓN

Hemos presentado una metodología que nos ha permitido realizar una actividad de evaluación entre iguales (tanto cualitativa como cuantitativa), en la que "todos evalúan a todos" y en la que ha sido fácil y ágil la gestión del elevado número de datos numéricos generados.

A la luz de los resultados obtenidos en la encuesta final, hemos podido constatar que nuestro alumnado valora positivamente poder recibir feedback de sus compañeros, especialmente el de tipo cualitativo. Sin embargo, no se siente igual de cómodo a la hora de tener que evaluar a sus compañeros, especialmente cuando esta evaluación es de tipo

cuantitativa. Destaca también el sentimiento de algunos estudiantes de que la evaluación pueda no ser subjetiva y, por tanto, consideran positivo que la actividad se realizara al inicio de curso, cuando aún no se conocían entre ellos, ya que eso evita sesgos de amistad. Valoramos positivamente que los estudiantes hayan considerado innovadora esta actividad, así como que consideren que han aprendido más sobre el tema en cuestión y hayan estado más atentos a las presentaciones de sus compañeros. A la vista de los comentarios recogidos en la encuesta, y de cara a futuras experiencias similares, es importante que los estudiantes dispongan de suficiente tiempo para realizar la evaluación de sus compañeros y tal vez sea adecuado reducir la amplitud de la escala de valoración.

Distintos autores han señalado la necesidad de que el alumnado asuma mayor implicación en el proceso de evaluación (Sanmartí y Jorba, 1995) y, cada vez de una manera más explícita, los currículos educativos incluyen la evaluación entre iguales. En este sentido, MOARS y PollEverywhere han demostrado ser dos recursos digitales que facilitan la evaluación entre iguales, puesto que han permitido procesar estas evaluaciones de una manera rápida, lo cual ayuda a reducir la carga de trabajo del docente y la sensación de incomodidad que puede mostrar parte del alumnado a la hora de tener que participar en ellas, ya que incluso algún estudiante menciona expresamente como un aspecto positivo el poder realizar las evaluaciones con el móvil, en lugar de con papel.

Por último, el análisis estadístico realizado con Minifac, aplicando el modelo MFRM, ha permitido poner de relieve qué estudiantes mostraron un sesgo de severidad o, especialmente, de benevolencia, incluso pudiéndolo llegar a cuantificar. Además del sesgo propiamente debido al estudiante, es posible que la dispersión en esta faceta pueda deberse en parte a la inexperiencia del alumnado en este tipo de actividades o a la ambigüedad en la redacción de las preguntas de la rúbrica. En futuras experiencias de este tipo se tendrá en cuenta la elaboración conjunta de la rúbrica de evaluación con el alumnado o la realización de algún ejercicio previo de entrenamiento, como sugieren Dochy et al. (1999).

6. CONCLUSIONES

MOARS ha podido gestionar el gran volumen de datos de la evaluación cuantitativa en la modalidad "todos evalúan a todos", haciendo de dicha evaluación un proceso sencillo e innovador para el alumnado.

La evaluación cualitativa recibida con PollEverywhere ha permitido concretar y ampliar el feedback cuantitativo. El alumnado ha mostrado una ligera preferencia por este tipo de feedback sobre el cuantitativo, tanto a la hora de darlo como de recibirlo.

El análisis MFRM realizado con Minifac ha permitido detectar y cuantificar el sesgo de severidad o benevolencia de los estudiantes, además de ordenarlos según su rendimiento global.

7. REFERENCIAS

Anker-Hansen, J. y Andrée, M. (2019). Using and rejecting peer feedback in the science classroom: a study of students' negotiations on how to use peer feedback when designing experiments. Research in Science & Technological Education, 37 (3), 346-365.

Chen, C. H. (2010). The implementation and evaluation of a mobile self- and peer-assessment system. Computers & Education, 55 (1), 229–236.

Custodio, E., Márquez, C. y Sanmartí, N (2015). Aprender a justificar científicamente a partir del estudio del origen de los seres vivos. Enseñanza de las Ciencias, 33 (2), 133-155.

Dochy, F., Segers, M. y Sluijsmans, D. (1999). The use of self-, peer and co-assessment in higher education: a review. Studies in Higher Education, 24 (3), 331-350.

Eckes, T. (2015). Introduction to Many-Facet Rasch Mesurement, 2ª ed. Peter Lang.

Gielen, S., Peeters, E., Dochy, F., Onghena, P. y Struyven, K. (2010). Improving the effectiveness of peer feedback for learning. Learning and Instruction, 20 (4), 304-315.

Ibarra, M. S., Rodríguez, G. y Gómez, M. A. (2012). La evaluación entre iguales: beneficios y estrategias para su práctica en la universidad. Revista de Educación, 359, 206-231.

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.

Li, H., Xiong, Y., Hunter, C., Guo, X. y Tywoniw, R. (2020). Does peer assessment promote student learning? A meta-analysis. Assessment & Evaluation in Higher Education, 45 (1), 193-211.

Lui, A. y Andrade, H. Student Peer Assessment (2014). En R. Gunstone (Ed.), Encyclopedia of Science Education. Springer.

Myford, C.M. y Wolfe, E. W. (2003). Detecting and Measuring Rater Effects Using Many-Facet Rasch Mesurement: Part I. Journal of Applied Measurement, 4 (4), 386–422.

Ministerio de Educación, Cultura y Deporte (2012). La enseñanza de las ciencias en Europa: políticas nacionales, prácticas e investigación. Secretaría General Técnica. DOI: https://dx.doi.org/10.2797/90921

Mogessie, M. (2015). Peer-assessment in higher education – twenty-first century practices, challenges and the way forward. Assessment & Evaluation in Higher Education, 42 (2), 226-251.

Nicol, D., Thomson, A. y Breslin, C. (2014). Rethinking feedback practices in higher education; A peer review perspective. Assessment & Evaluation in Higher Education, 39 (1), 102-122.

Orsmond, P., Merry, S. y Reiling, K. (1996). The importance of marking criteria in the use of peer assessment. Assessment & Evaluation in Higher Education, 21 (3), 239-250.

Panadero, E. y Alqassab, M. (2019). An empirical review of anonymity effects in peer assessment, peer feedback, peer review, peer evaluation and peer grading. Assessment & Evaluation in Higher Education, 44 (8), 1253-1278.

Reese-Durham, N. (2005). Peer evaluation as an active learning technique. Journal of Instructional Psychology, 32 (4), 338-343.

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.

Sanmartí, N. y Jorba, J. (1995). Autorregulación de los procesos de aprendizaje y construcción de conocimientos. Alambique, 4, 59-77.

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