Metodologías y prácticas docentes en la enseñanza superior

Metodologias e práticas de ensino no ensino superior

Editores
Pedro Membiela
María Isabel Cebreiros

Educación Editora

Metodologías y prácticas docentes en la enseñanza superior

Metodologias e práticas de ensino no ensino superior

Pedro Membiela y María Isabel Cebreiros
(editores)

Educación Editora

Edita Educación Editora
Roma 55, Barbadás 32930 Ourense
email: educacion.editora@gmail.com
ISBN: 978-84-15524-52-6
Año de publicación: 2024

34. Estudio del sesgo de severidad o benevolencia en una evaluación entre iguales con herramientas TIC del alumnado de Biología y Geología del Máster de Secundaria

Gregorio Jiménez Valverde[*], Genina Calafell i Subirà, Mireia Esparza Pagès y Hortensia Duran Gilabert

Grupo de Innovación Docente Consolidado EduCiTS, Facultad de Educación, Universitat de Barcelona

[*]gregojimenez@ub.edu

Resumen

Se realizó un estudio estadístico del sesgo de severidad o de benevolencia de los estudiantes de la especialidad de Biología y Geología del Máster de Formación del Profesorado de Secundaria cuando evaluaron las presentaciones orales de sus compañeros utilizando herramientas TIC (tecnologías de la información y de la comunicación).

Palabras clave

Evaluación entre iguales, máster de secundaria, MOARS, MFRM, formación del profesorado.

Introducción

Uno de los objetivos fundamentales de la educación superior es el desarrollo del pensamiento crítico en los estudiantes, tanto en sus prácticas profesionales como en su proceso de aprendizaje. La evaluación entre iguales, en la que cada estudiante evalúa la calidad de las producciones de sus compañeros (Topping, 2018), es una estrategia educativa que puede dotar de mayor protagonismo al estudiante en el proceso educativo y fomentar su pensamiento crítico, al brindarle la oportunidad de reflexionar sobre su propio aprendizaje y recibir comentarios específicos sobre sus fortalezas y debilidades. Así, el alumnado puede mejorar su capacidad para identificar y analizar información, resolver problemas y tomar decisiones fundamentadas, lo que conlleva un pensamiento crítico más desarrollado y una mejor calidad de aprendizaje en general.

Sin embargo, la evaluación entre iguales puede estar sujeta a sesgos y subjetividades (Topping, 2018), como la influencia de la relación personal con el estudiante evaluado o la percepción que se tenga de su habilidad en la materia. Esta circunstancia es especialmente importante en el caso del profesorado, ya que si se es demasiado generoso en la evaluación, los estudiantes pueden recibir calificaciones más altas de lo que merecen, lo que puede generar una percepción exagerada de su nivel real de logro y, por tanto, dificultar su progreso académico. Por otro lado, si se es demasiado severo, los estudiantes pueden sentirse desmotivados al recibir constantemente calificaciones bajas, lo que puede afectar negativamente su autoestima y confianza en su capacidad para aprender.

En este sentido, el análisis de las medidas de múltiples facetas de Rasch (*Many-facet Rasch Measurement*, MFRM) propuesto por Linacre (1989) puede ayudar a identificar patrones de sesgo en una evaluación entre iguales, ya que este modelo descompone la variabilidad en las calificaciones de una prueba en diferentes fuentes o facetas, como la habilidad de los evaluados, los diferentes criterios de evaluación utilizados por los evaluadores y los sesgos de estos últimos. El MFRM resulta especialmente útil para la evaluación de respuestas subjetivas, como las que se encuentran en campos como la medicina, la psicología o la educación. Al proporcionar una medida precisa de la habilidad del evaluador, el modelo puede incluso eliminar los sesgos de severidad o benevolencia encontrados en los evaluadores y determinar cuál hubiera sido la calificación "justa" en ausencia de estos sesgos. No obstante, es importante destacar que en el modelo MFRM la severidad o benevolencia del evaluador se define de manera relativa, lo que implica que el cálculo se realiza en comparación con otros evaluadores que también participan en la misma actividad evaluativa (Anthony et al., 2021).

A pesar de la existencia de sólidos fundamentos teóricos que respaldan la evaluación entre iguales, es esencial comprender cómo los futuros docentes perciben esta práctica, ya que estas percepciones pueden influir significativamente en su práctica docente futura (Darling-Hammond, 2017). Se ha hallado que los docentes en formación inicial suelen tener una comprensión parcial y limitada de las estrategias de evaluación, como señala Maclellan (2004). Con frecuencia, los aspirantes a docentes perciben la evaluación meramente como una forma de determinar si los estudiantes han captado un concepto o no, como indica Otero (2006).

En un estudio realizado por Jiménez (en prensa) con docentes en formación inicial del área de ciencias, se encontró que un 20,5 % de los participantes no sabían qué es la evaluación entre iguales y entre los que afirmaron saber qué es, muchos tenían una concepción exclusivamente calificadora, sin aludir al valor formativo de esta actividad. Solo un 31,8 % había tenido experiencias previas con la evaluación entre iguales antes de cursar el Máster de Secundaria, y principalmente las describieron como simples actividades de calificación. Sin embargo, el 95 % consideró que se puede aprender a través de la evaluación de sus

compañeros y, de hecho, el 87 % afirmó que se deberían realizar más evaluaciones entre pares. No obstante, la evaluación entre iguales no está exenta de reticencias: casi el 30 % del alumnado participante creía que señalar los errores cometidos por sus compañeros podía molestarles y un 20,5 % afirmó sentirse incómodo si tuviera que evaluar a sus compañeros. En cualquier caso, reconocieron la importancia de ser objetivos y justos en sus evaluaciones en su futura labor docente y, de hecho, valoraron de forma unánime la conveniencia de saber si eran demasiado estrictos o benevolentes evaluando el trabajo de los demás, en comparación con otros docentes.

Descripción de la experiencia

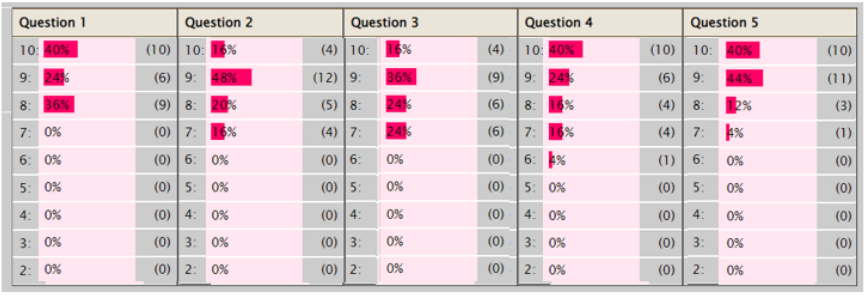
La experiencia se ha desarrollado durante el curso 2022-2023 con los estudiantes de la asignatura Didáctica de la Biología y la Geología de la misma especialidad del Máster de Formación del Profesorado de Secundaria de la Universitat de Barcelona. En total, participaron 30 estudiantes (18 mujeres y 12 hombres), correspondientes al único grupo de la asignatura.

La actividad de evaluación entre iguales consistió en la realización de una exposición oral apoyada por una presentación con diapositivas, en grupos de tres o cuatro estudiantes sobre un tema relacionado con la parte de Didáctica de la Geología de la asignatura. Los temas abarcaban el sistema solar, la tectónica de placas, el vulcanismo y magmatismo, los combustibles fósiles, las energías renovables, la dinámica marina, el tiempo geológico y la dinámica atmosférica. Los criterios evaluables de cada presentación fueron el nivel científico, aspectos didácticos, actividades y recursos educativos, presentación gráfica y presentación oral. La docente consensuó con los estudiantes una rúbrica de evaluación para la actividad, que fue entregada con antelación al inicio de la misma.

Al finalizar cada presentación oral, los estudiantes del resto de grupos y la docente realizaron la evaluación de la presentación utilizando la aplicación MOARS (Jiménez, 2021). La calificación se otorgó a los cinco criterios mencionados anteriormente en una escala de puntuación de 2 a 10, de acuerdo con la rúbrica mencionada (MOARS permite un máximo de nueve valores para calificar un ítem, lo que explica la elección de la escala de evaluación).

Una vez concluidas todas las evaluaciones, MOARS ofreció el resultado de las evaluaciones entre iguales de las presentaciones orales de cada grupo en forma de gráficos de barras (figura 1).

Para realizar el análisis estadístico según el modelo MFRM, se exportaron los datos recogidos en MOARS y se importaron en el programa FACETS (v. 3.85.0, www.winsteps.com), que devuelve un archivo con el resultado del análisis MFRM y que incluye diferentes estadísticos, gráficos y tablas.



Question 1		Question 2		Question 3		Question 4		Question 5	
10: 40%	(10)	10: 16%	(4)	10: 16%	(4)	10: 40%	(10)	10: 40%	(10)
9: 24%	(6)	9: 48%	(12)	9: 36%	(9)	9: 24%	(6)	9: 44%	(11)
8: 36%	(9)	8: 20%	(5)	8: 24%	(6)	8: 16%	(4)	8: 12%	(3)
7: 0%	(0)	7: 16%	(4)	7: 24%	(6)	7: 16%	(4)	7: 4%	(1)
6: 0%	(0)	6: 0%	(0)	6: 0%	(0)	6: 4%	(1)	6: 0%	(0)
5: 0%	(0)	5: 0%	(0)	5: 0%	(0)	5: 0%	(0)	5: 0%	(0)
4: 0%	(0)	4: 0%	(0)	4: 0%	(0)	4: 0%	(0)	4: 0%	(0)
3: 0%	(0)	3: 0%	(0)	3: 0%	(0)	3: 0%	(0)	3: 0%	(0)
2: 0%	(0)	2: 0%	(0)	2: 0%	(0)	2: 0%	(0)	2: 0%	(0)

Figura 1. Detalle de las puntuaciones recibidas por uno de los grupos. Cada barra indica el número (y porcentaje) de estudiantes que otorgaron esa calificación en un criterio de evaluación (*question*) determinado

Resultados y discusión

Para el análisis del sesgo de severidad o benevolencia conviene fijarse en uno de los gráficos que genera FACETS: el mapa de Wright, que ofrece un panorama general del análisis llevado a cabo. La figura 2 muestra la sección de dicho mapa referida a la variable relacionada con el grado de severidad/benevolencia de los evaluadores. En dicho gráfico, los estudiantes han sido representados por un código de tres cifras y la docente por la palabra PROFE, y todos ellos han sido ordenados en orden decreciente según su severidad como evaluadores. Los valores más positivos indican mayor severidad y valores más negativos indican mayor benevolencia, siendo estadísticamente significativo el sesgo cuando su valor absoluto es superior a 1. Tres estudiantes (095, 140 y 184) mostraron un sesgo significativo de severidad mientras que dos de ellos (220 y 842) lo mostraron de benevolencia. De hecho, el modelo estadístico indica que ha encontrado cuatro niveles diferenciados de severidad entre los estudiantes, siendo el nivel con mayor severidad el que agrupa a los estudiantes 095, 140 y 184.

Además, el modelo permite calcular el *outfit* (*outlier-fit mean-square*) e *infit* (*inlier-fit mean-square*) de cada evaluador. El *infit* se refiere a la medida de ajuste de las respuestas observadas a las respuestas esperadas, considerando el grado de dificultad del ítem y el nivel de habilidad del evaluador. Por otro lado, el *outfit* se enfoca en evaluar el ajuste de las respuestas observadas a las respuestas esperadas, sin tener en cuenta el grado de dificultad del ítem y el nivel de habilidad del evaluador. Estas medidas proporcionan una evaluación del grado de ajuste del modelo MFRM a los datos observados, lo que se utiliza para evaluar la calidad de las respuestas de los evaluadores y la adecuación del modelo a los datos. Según Linacre (2002), se produce un "ajuste útil" de los datos al modelo estadístico cuando ambos parámetros se encuentren entre el rango 0,50-1,50. Cuando estos valores superan el 2, se considera una degradación del modelo estadístico. En nuestro caso solo dos estudiantes obtuvieron valores de *infit* y *outfit* superiores a 2, lo que indica que sus respuestas tienen un deficiente grado

de ajuste al modelo MFRM y ello puede ser debido a que evaluaron de manera inconsistente y errática a sus compañeros o a que tuvieron dificultades para entender la rúbrica de evaluación. De hecho, el valor de confiabilidad obtenido (0,89, siendo el máximo de 1.00) y la similitud entre los acuerdos esperados (27,6 %) y los observados (28,0 %) confirman la consistencia, confiabilidad y alineación de las evaluaciones realizadas.

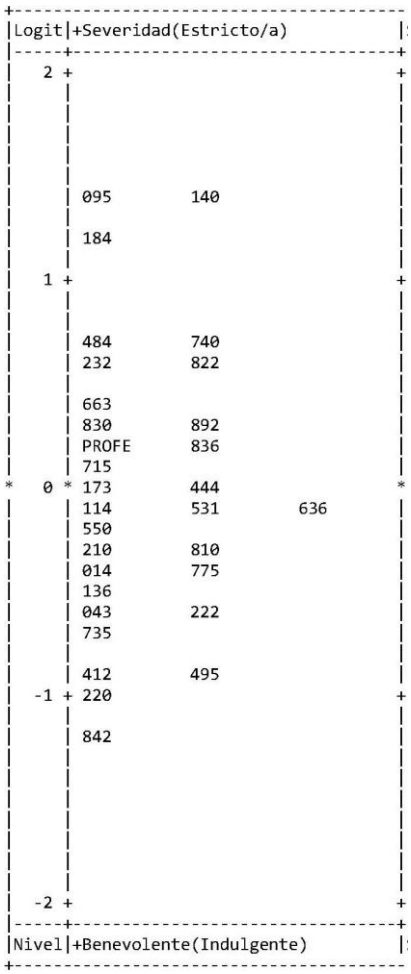


Figura 2. Extracto del mapa de Wright, con los estudiantes (evaluadores) ordenados según su severidad

En una sesión posterior de la asignatura, la docente compartió con su alumnado el resultado de la evaluación entre iguales, incluyendo los detalles presentados en la figura 1 para cada grupo. Además, presentó públicamente la figura 2

y explicó su significado, destacando las posiciones de los estudiantes que habían mostrado sesgos significativos de severidad o benevolencia. También proporcionó un *feedback* privado a aquellos estudiantes que habían mostrado un comportamiento errático o inconsistente en sus evaluaciones. Todo esto generó un enriquecedor debate sobre la evaluación entre iguales y su valor formativo, impulsando la reflexión del alumnado sobre la subjetividad presente en todos los actos de evaluación, incluso cuando se cree que se está siendo completamente objetivo.

Conclusión

La evaluación entre iguales no solo es una herramienta eficaz para fomentar la reflexión y el aprendizaje de los estudiantes, sino que también puede ayudar a desarrollar habilidades clave como el pensamiento crítico y el *feedback* formativo. Además, esta modalidad de evaluación formativa puede ser especialmente útil en la formación inicial del profesorado ya que brinda la oportunidad de practicar la evaluación en un entorno seguro y de recibir información sobre su propia práctica como docentes, de tal manera que puedan mejorar su capacidad de evaluar de manera justa y rigurosa.

Agradecimientos

El presente texto nace en el marco del proyecto "Evaluación entre iguales con herramientas web 2.0 y TIC con profesorado en formación inicial" y se hace constar la colaboración del Vicerrectorado de Docencia y del programa RIMDA de la Universitat de Barcelona en la difusión de este trabajo.

Referencias

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*, 51 (1), 25-39.

Darling-Hammond, L. (2017). Teacher education around the world: What can we learn from international practice? *European Journal of Teacher Education*, 40 (3), 291-309.

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.

Jiménez, G. (en prensa). Estudio de las concepciones iniciales del profesorado en formación inicial sobre la evaluación entre iguales. En *Educación Siglo XXI: nuevos retos, nuevas soluciones*, volumen 3. Dykinson.

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

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

Maclellan, E. (2004). Initial knowledge states about assessment: Novice teachers' conceptualizations. *Teaching and Teacher Education*, 20 (5), 523-535.

Otero, V. K. (2006). Moving beyond the 'get it or don't' conceptions of formative assessment. *Journal of Teacher Education*, 57 (3), 247-255.

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