¿Están sesgadas las evaluaciones de la docencia universitaria realizadas por los estudiantes?

Are student evaluations of university teaching biased?

https://doi.org/10.4438/1988-592X-RE-2024-404-621

María Castro Morera
https://orcid.org/0000-0002-2597-3621
Universidad Complutense de Madrid

Enrique Navarro-Asencio
https://orcid.org/0000-0002-3052-146X
Universidad Complutense de Madrid

Coral González Barbera
https://orcid.org/0000-0002-0016-4828
Universidad Complutense de Madrid

Resumen

Los cuestionarios que utilizan a los estudiantes como fuente de información para valorar la docencia universitaria son una herramienta habitual en los sistemas de evaluación de las universidades. Los docentes universitarios suelen cuestionarlas aludiendo a la posibilidad de que los estudiantes emitan valoraciones sesgadas, vinculadas a rasgos o acontecimientos docentes que no están relacionados con la valoración, ecuánime, de la actividad docente. El objetivo principal de este trabajo es examinar las relaciones entre las características de los estudiantes y de los profesores y las puntuaciones en el cuestionario de evaluación de la docencia aplicado a los estudiantes de la Universidad Complutense de Madrid, para detectar posibles patrones sesgados con relación a la valoración que éstos ofrecen de sus profesores. Se ha realizado un modelo jerárquico lineal de clasificación cruzada, con dos niveles, siendo el primer nivel los estudiantes y el segundo los profesores.

La muestra de este trabajo está compuesta 143.377 encuestas, respondidas por 33.071 estudiantes que supuso la evaluación de 7.885 actividades docentes y 3922 profesores en el curso 2016-17. Los resultados indican que las valoraciones que los estudiantes emiten sobre los profesores están influidas sobre todo por el interés que manifiestan por la asignatura, la edad de estudiantes y docentes y, en menor medida, la asistencia, horas de estudio y calidad investigadora. Hay que destacar que no tiene relación alguna con las valoraciones sobre la docencia el tipo de estudios de grado o máster que cursan, el rendimiento académico del estudiante, ni la categoría laboral del profesor.

Tras este análisis de los resultados, no se puede afirmar la existencia de sesgos invalidantes derivados del uso de los cuestionarios para la evaluación de la docencia universitaria respondidos por los estudiantes.

Palabras clave: evaluación del profesorado, educación superior, evaluación de la docencia por los estudiantes, calidad de la docencia, cuestionarios, modelos jerárquicos lineales, sesgos.

Abstract

Questionnaires that use students as a source of information to evaluate university teaching are a common tool in university evaluation systems. The lecturers often question their value by alluding to the possibility that students may make biased judgments, linked to teaching traits or events not related to a fair assessment of the teaching activity. The main objective of this work is to examine the relationships between the characteristics of students and lecturers and the scores on the teaching evaluation questionnaire applied to students at the Complutense University of Madrid, in order to detect possible biased patterns in the evaluation they offer of their teachers. A hierarchical linear cross-classification model was used, with two levels, taking students as the first level and the lecturers as the second. The sample of this work is composed of 143,377 surveys, completed by 33,071 students, which involved the evaluation of 7,885 teaching activities and 3,922 university teachers in the academic year of 2016-17.The results show that the students' evaluations of their lecturers are mainly influenced by their interest in the subject, the age of the students and their lecturers and, to a lesser extent, attendance, hours of study and research quality. It should be noted that the type of undergraduate or master's degree studies, student's academic performance, and the lecturer's job category are not related to the teaching evaluations.After this analysis of the results, we cannot deduce the existence of invalidating biases derived from the evaluation of university teaching by questionnaires answered by the students

Keywords: teacher evaluation, higher education, student evaluation of teaching, quality of teaching, questionnaires, hierarchical linear modeling, bias.

Recibido: 10-12-2022 Aceptado: 30-06-2023

Introducción

La adecuación y pertinencia de las valoraciones de los estudiantes sobre la actividad docente del profesorado para evaluar una parte la labor docente en la universidad es una discusión antigua y cotidiana. En cada universidad, tanto nacional como internacional, los docentes universitarios expresan dudas razonables sobre el uso de la percepción de los estudiantes para la evaluación de la docencia universitaria (Cox et al, 2021).

Los cuestionarios que utilizan a los estudiantes como fuente de información para valorar la docencia universitaria, denominados *Student Evaluation of Teaching* (SET a partir de ahora), son una herramienta habitual y generalizada en los sistemas de evaluación y de rendición de cuentas de las universidades. Estas valoraciones de la docencia comienzan formalmente en los años 20 del siglo pasado, en la Universidad de Washington (Guthrie, 1954; Kulik, 2001). El primer informe sobre SET se hizo público en 1927 por Remmers y Brandenburg. El uso de estos cuestionarios ha ido evolucionando con el paso del tiempo y con el cambio de necesidades de las universidades. En la actualidad, los SET participan en procesos de evaluación tanto formativa como sumativa (Johnson, 2000; Spooren et al., 2013).

En España, la inclusión de la evaluación de la actividad docente en los procesos de selección, promoción y estabilización del profesorado universitario (ANECA, 2017) ha aumentado enormemente el uso y demanda de los cuestionarios de evaluación de los estudiantes, siendo en algunos momentos la parte central de la evaluación de los docentes.

Los trabajos empíricos desarrollados en nuestro país se han centrado mayoritariamente en el diseño y el análisis psicométrico de instrumentos de evaluación más o menos estándar (véase Castro et al. 2020; Casero, 2008; Mayorga et al. 2016, López-Cámara et al. 2016; Molero y Ruíz, 2005, Muñoz et al. 2002) o en el análisis descriptivo de los resultados de los cuestionarios del alumnado en contextos específicos (p.e. De Juanas y Beltrán, 2014; Ordoñez y Rodríguez, 2015).

Un conjunto importante de las objeciones frecuentemente planteadas se refiere a la emisión de valoraciones sesgadas por parte de los estudiantes, vinculadas a la atribución de cierta maleabilidad en el juicio vinculada a rasgos o acontecimientos docentes que no están relacionados con la valoración, ecuánime, de la actividad docente. El

'sesgo' puede definirse como la situación en la que "una característica de un estudiante, profesor o curso afecta a las evaluaciones realizadas, ya sea positiva o negativamente, pero no está relacionada con ningún criterio de buena enseñanza, como la mejora de los aprendizajes de los estudiantes" (Centra y Gaubatz, 2000, p. 17). En el ámbito anglosajón, se encuentran hallazgos consolidados sobre los posibles sesgos en las valoraciones de los estudiantes a la docencia universitaria (Esarey y Valdés, 2020; Marsh, 1987; Spencer y Schmelkin, 2002; Spooren, 2010; Sulis et al., 2019; Wachtel, 1998). En el ámbito español, destacan los trabajos de García et al. (2011) y Gómez et al. (2013); o el trabajo de revisión de Casero (2010).

La literatura consultada muestra consenso en la idea de que las valoraciones de los estudiantes se asocian positivamente (correlaciones superiores a 0,40) con las de otras fuentes como supervisores, colegas y observadores externos (Beran y Violato, 2005; Marsh, 1987). Se puede afirmar que los estudiantes ofrecen puntuaciones similares a las que ofrecen otros evaluadores.

Además, se ha demostrado que los SET son instrumentos sólidos técnicamente hablando. Son instrumentos fiables, estables y consistentes (Marsh, 1984; Clayson, 2018). Se reconoce una estructura multidimensional del constructo (Spooren et al 2013; Spooren et al. 2014 y Lizasoain-Hernández, et al., 2017). No obstante, algunos autores apuntan la presencia de un único factor general de calidad docente percibida en los cuestionarios analizados (Castro et al. 2020). Asimismo, Spooren et al. (2017) destacan que los factores que influyen en las valoraciones de los estudiantes son cinco: calidad de la enseñanza, rigor del curso, nivel de interés, curso y capacidad de ayuda del profesor. Por tanto, la literatura apunta a que son instrumentos relativamente válidos con relación a indicadores de enseñanza efectiva y poco sensibles a sesgos.

Otro grupo de resultados de investigación sobre SET muestran que algunas características de docentes, estudiantes y asignaturas suelen estar asociadas a posibles valoraciones sesgadas de los estudiantes. Los factores estudiados son variados, por ejemplo, de los estudiantes se considera tanto el rendimiento, su interés por las materias, el tipo o rama de estudios, así como características como su edad o sexo. De los docentes, su experiencia docente o investigadora y también su edad o sexo. Y de las materias, aunque también pueden asociarse a los docentes, el curso en el que se imparte o la rama de conocimiento.

Recibido: 10-12-2022 Aceptado: 30-06-2023

El rendimiento de los estudiantes es una de las características más estudiadas en la literatura sobre SET, en tanto que indicador de las consecuencias de una enseñanza eficaz (Penny, 2003) y como procedimiento para el estudio de la validez convergente (Spooren et al., 2013). Los resultados de las investigaciones consultadas no son coincidentes. Los meta-análisis de Cohen (1980, 1981) muestran una relación positiva entre moderada y grande entre el rendimiento de los estudiantes y las valoraciones que otorgan a sus docentes en estos instrumentos, también el trabajo de Clayson (2009) encuentra hallazgos a favor de esa relación. Sin embargo, el meta-análisis de Uttl et al. (2017), muestra una clara ausencia de relación. Y también hay evidencias en contra, como los de Mohanty et al. (2005), el Stark-Woblewski et al. (2007) o el de Braga et al. (2014) y más recientemente el de Berezvai et al. (2021). Es necesario señalar también que no está clara su relación con la eficacia docente en sentido estricto, es decir, cuando ésta se mide en términos de rendimiento del alumnado. En consecuencia, autores como Hornstein (2017) y Carpenter et al. (2020) no recomiendan su uso para evaluar la capacidad docente, sobre todo para tomar decisiones sobre contratación o promoción.

El uso de las calificaciones académicas como prueba de la validez de las percepciones de los estudiantes sobre la enseñanza es objeto de debate desde los años 70 (Marsh, 1987; Griffin, 2004; Gump, 2007; Marsh y Roche, 2000). Como sintetiza Spooren (2010), la primera interpretación es que las calificaciones pueden reflejar una buena enseñanza y los SET reconocen esa calidad y, en consecuencia, los estudiantes con mejores notas valoran mejor a sus profesores. Una segunda interpretación es que los docentes pueden poner mejores notas para recibir mejores valoraciones en los SET, esto sería un caso claro de sesgo. En los datos de este trabajo, la evaluación de los docentes se lleva a cabo antes de que los estudiantes conozcan sus calificaciones, así se evita este posible sesgo. Una tercera tendencia apunta a la asociación de actitudes o percepción que un estudiante tiene sobre su aprendizaje (tales como interés en la materia o aspectos motivacionales) con la valoración que otorgan al profesor. En este sentido, Greimel-Fuhrmann y Geyer (2003) muestran que el comportamiento del docente determina en gran medida el interés del estudiante y Paswan y Young (2002) también demuestra que la interacción entre docente y estudiantes afecta a su nivel de interés. Más recientemente, Carpenter et al. (2020) argumentan que la impresión que tiene un estudiante sobre su capacidad para aprender y sobre cómo

debe ser el proceso de enseñanza determinan su valoración del docente. Los mismos autores apuntan que esa visión puede ser errónea y, por tanto, su opinión sobre la eficacia del docente también lo será.

Fjortoft (2005) relaciona la asistencia regular a las clases con un mayor interés y motivación por el aprendizaje. Los resultados de investigación que incluyen la asistencia como factor relacionado con las puntuaciones en los SET no son homogéneos, puesto que mientras unos muestran la importancia de esta relación entre asistencia a clase y valoración del profesor (Beran y Violato, 2005; Davidovitch y Soen, 2006) otros destacan su irrelevancia (Guinn y Vincent, 2006).

Se justifica, por tanto, que el rendimiento y variables que reflejen la motivación hacia el aprendizaje tengan efecto en los SET porque, en parte, están determinadas por la calidad de la docencia. Spooren et al. (2013) señalan que el estudio y esfuerzo del estudiante son indicadores de su interés y motivación, siendo estos dependientes parcialmente de la organización didáctica de la asignatura.

Si ponemos la atención en características de los estudiantes que no están vinculadas a esa calidad del proceso de aprendizaje, como el sexo o la edad, los resultados de investigación tampoco son concluyentes. Por ejemplo, el trabajo de Centra y Gaubatz (2000) o el de Spooren (2010), concluyen que la relación entre el sexo del estudiante y el SET no es determinante. Sin embargo, otras investigaciones muestran que podría haber un efecto de interacción entre el sexo de los estudiantes y el de los docentes con respecto al SET. La tendencia general en los resultados es que las profesoras reciben menores puntuaciones (Basow et al., 2006; Boring, 2017; Boring et al., 2016; Mitchell Martin, 2018 y Rivera y Tilcsik, 2019).

El trabajo de Sprinkle (2008) estudió esta interacción y también otras características del docente (edad, el sexo, el estilo de enseñanza) y concluye que la edad, el sexo y la interacción entre sexo de estudiantes y de docentes son factores relacionados con las valoraciones de los estudiantes. Sus resultados mostraron que las alumnas tienden a valorar mejor a profesoras y los alumnos a los profesores. Respecto al efecto de la edad, Spooren (2010) también halla un efecto significativo, donde estudiantes de mayor edad tienden a otorgar mejores puntuaciones a sus docentes. Aunque el tamaño del efecto, tanto del sexo como la edad, es pequeño. Wachtel (1998) apuntó que las valoraciones más elevadas que los estudiantes de más edad proporcionan puede deberse a una mayor madurez o por una mayor especialización de las materias evaluadas y que, por tanto, pueden suscitar un mayor interés.

Recibido: 10-12-2022 Aceptado: 30-06-2023

En una revisión de la literatura sobre este tema, el trabajo de Griffin (2004) muestra correlación, aunque muy débil, entre el sexo del docente y las valoraciones de sus estudiantes, siendo las profesoras las que tienen valoraciones más altas que sus colegas varones.

Otros estudios no encuentran correlaciones entre la edad y el género de los docentes y los resultados en los SET (Ting, 2000). Tampoco Spooren (2010) halló un efecto significativo de estas variables, aunque en el estudio de McPherson et al. (2009) los resultados mostraron mejores valoraciones de los profesores más jóvenes. Y en la revisión de Wachtel (1998) se observa una relación inversa entre la edad del docente y las valoraciones de los estudiantes. Los profesores de mayor edad reciben calificaciones menos favorables. Y, tal y como señalan Spooren et al (2013), la antigüedad, la productividad científica y la categoría del profesor son indicadores indirectos de las habilidades didácticas del profesor y del dominio de la materia. Por ejemplo, la experiencia docente es un factor relacionado con mejores puntuaciones en los SET (McPherson y Jewell, 2007 y McPherson et al., 2009), en cambio el número de publicaciones no tiene efecto significativo (Ting, 2000).

Finalmente, con respecto al análisis del ámbito de estudio en el que imparte docencia el profesor, Theall y Franklin (2001) indican que en el ámbito de las ciencias se reciben puntuaciones de SET más bajas que en el ámbito de humanidades. Resultados similares a los de Basow y Montgomery (2005). Asimismo, Kember y Leung (2011), mediante un modelo de ecuaciones estructurales multigrupo, concluyen que, aunque la estructura explicativa de las puntuaciones SET es equivalente entre áreas (invarianza de la configuración), los docentes de humanidades reciben puntuaciones más altas que los que imparten en ciencias exactas o en las carreras de negocio y empresa (invarianza métrica).

En síntesis, la investigación sobre los posibles sesgos de los estudiantes cuando valoran la calidad de sus docentes no proporciona evidencias concluyentes. Los resultados muestran que las características de los estudiantes, asociadas al aprendizaje, pueden tener impacto en esas valoraciones, como el rendimiento, el interés por la materia, el tempo de estudio o la asistencia. Incluso el impacto de la edad en las valoraciones puede estar asociada a un mayor interés por la materia o madurez, sobre todo si son cursos especializados, como las titulaciones de Máster. Un efecto significativo de estos factores no es, por tanto, un indicador de valoraciones sesgadas por parte de los estudiantes. Solo si el estudiante conoce su nota previamente a la valoración de la calidad

del docente se podría considerar un factor de sesgo. En cambio, los resultados de estos estudios han encontrado un efecto cruzado entre el sexo del estudiante que emite la valoración y el de los docentes, aunque el tamaño del efecto es bajo puede indicar cierto sesgo. En el mismo sentido, las características del docente que determinan su dominio de la materia, como la experiencia docente o la producción científica, son factores que pueden tener impacto en las puntuaciones SET, sin que suponga un factor de sesgo.

El objetivo principal de este trabajo es examinar las relaciones entre las características de los estudiantes y de los profesores y las puntuaciones en el cuestionario de evaluación de la docencia aplicado a los estudiantes de la Universidad Complutense de Madrid (UCM) para detectar posibles patrones sesgados con relación a la valoración que éstos ofrecen de sus profesores. Para ello, se plantean los siguientes objetivos específicos:

- Analizar el impacto de características de los estudiantes vinculadas al proceso de enseñanza y aprendizaje (notas, interés, dificultad, asistencia y horas de estudio).
- Analizar el impacto de características demográficas del estudiante (sexo y edad).
- Analizar el impacto de características de los docentes vinculadas al proceso de enseñanza y aprendizaje (categoría laboral, producción científica, experiencia docente).
- Analizar el impacto de características demográficas del docente (sexo y edad).
- Analizar el efecto cruzado del sexo de estudiantes y de los docentes.

Método

Esta investigación es un análisis secundario de la encuesta aplicada a los estudiantes de la UCM en el marco del programa Docentia implantado en esta universidad. El diseño utilizado es de carácter no experimental con propósito correlacional y exploratorio. Se ha probado de forma empírica el efecto de características de estudiantes y docentes en los resultados de los SET, pero considerando la falta de consenso en la teoría consultada es complejo probar modelos confirmatorios.

Recibido: 10-12-2022 Aceptado: 30-06-2023

Muestra

La muestra de este trabajo está compuesta por estudiantes que valoraron la actuación docente del profesor en las distintas actividades docentes en las que están matriculados. Así, 33.071 estudiantes (65,1 % mujeres y una edad media de 22 años (D.T=5,281)) respondieron un total de 143.377 encuestas que evaluaron 7.885 actividades docentes en las que están implicados un total de 3922 profesores (48 % mujeres y una edad promedio de 49 años (D.T.=7,739)) y un total de 7885 asignaturas que impartieron en grado y máster en el curso 2016-17. Es importante recordar que el marco del programa Docentia de ANECA exige evaluar al profesor en el conjunto de su actividad docente. En términos promedio, cada profesor fue evaluado por 31 estudiantes.

Los profesores que formaron parte de la muestra tenían un historial comprometido con la evaluación de su actividad docente (al menos dos evaluaciones docentes en dos cursos consecutivos) y de su excelencia en el mismo (con evaluaciones positivas). Con relación a la distribución del profesorado, el 25% de los docentes eran del área de salud, el 21,9% de ciencias experimentales, el 35,1% de ciencias sociales, el 18% de arte y humanidades.

Instrumentos y variables

El cuestionario del estudiante consta de 17 preguntas, con una escala de valoración de 0 a 10 puntos, a la que se añade la posibilidad de contestar No Sabe. Los formularios de evaluación de los estudiantes se administraron *on line* durante los dos periodos en los que se aplicaron las encuestas (diciembre y mayo) del curso académico 2016-2017.

La variable de respuesta es el promedio de las valoraciones de los estudiantes a las 17 cuestiones, expresada en una escala global de 0 a 10 (media= 7,95; D.T = 2,188) y la fiabilidad estimada a través del coeficiente α de Cronbach es de 0,98 (Castro et al. 2020). La unidimensionalidad se probó de nuevo en esta investigación mediante un análisis factorial confirmatorio, logrando valores aceptables (CFI=0,93; TLI:0,915; RMSEA=0,066 y SRMR=0,038).

En el estudio se utilizan las siguientes variables vinculadas a la emisión de valoraciones sesgadas expresadas en la literatura (ver Tabla I).

Recibido: 10-12-2022 Aceptado: 30-06-2023

TABLA I. Relación de variables de estudiantado y profesorado

Variables del estudiante	**Escala de medida**
Sexo del estudiante	0 = Varón 1= Mujer
Edad del estudiante	Variable continua. Centrada por la media del grupo
Asistencia declarada a clase	4 =Menos 20% 3 =20%-39% 2 =40%-59% 1 =60%-79% 0=80% o más
Horas de estudio semanal	4=Menos de 1h 3 =De 1 a 4 2=De 5 a 7 1=De 8 a 10 0=Más de 10
Interés declarado por la asignatura	Escala de 0 a 10.
Dificultad percibida de la asignatura	Escala de 0 a 10.
Rendimiento promedio a lo largo de toda la trayectoria universitaria del estudiante	Escala de 0 a 10.
Tipo de estudios (Grado o Máster)	0= Grados 1= Máster Oficial
Variables del profesorado	
Sexo del profesor	0 = Varón 1= Mujer
Edad del profesor	Variable continua. Centrada por la media del grupo
N° de sexenios	0 a 6
Años de experiencia docente (n° de quinquenios)	0 a 8
Categoría laboral con la universidad	PDI Funcionario PDI Laboral
Rama de estudios en las que enseña	0 = Salud 1 = CC. Experimentales 2 = CC. Sociales 3 = Artes y Humanidades
Variables entre niveles	
Sexo	0=Estudiante (Mujer) - Docente (Mujer) 1=Estudiante (Varón) - Docente(Varón) 2=Estudiante (Mujer) - Docente(Varón) 3=Estudiante (Varón) - Docente (Mujer)

Fuente: Elaboración propia.

Recibido: 10-12-2022 Aceptado: 30-06-2023

Análisis de datos

Los SET son el resultado de las percepciones de los estudiantes sobre la actividad docente, pero estas percepciones pueden estar influenciadas tanto por las características de los estudiantes como por las de los docentes, diferenciándose claramente dos niveles de variabilidad, que además comparten contexto.

En esta evaluación los estudiantes completan varias encuestas correspondientes a distintos profesores y esto conlleva que los datos no tengan una estructura completamente anidada. Si ocurriera así, cada docente sería evaluado por un alumnado diferente. En los datos utilizados, un mismo estudiante puede valorar a varios docentes y, por tanto, no son totalmente independientes. Fue necesario, por tanto, estimar los resultados empleando un modelo de regresión multinivel de clasificación cruzada (Rasbasch y Goldstein, 1994). Esto implica que la identificación del estudiante está asociada al docente evaluado. Otra consideración es que los estudiantes pueden evaluar al mismo docente, pero en asignaturas distintas. Además, un docente puede tener la valoración en una o más asignaturas. Por tanto, el primer nivel, incluye la variabilidad entre estudiantes (cruzado con el docente y la asignatura). El segundo nivel, representa la variabilidad entre la combinación del docente y la asignatura. Los modelos estiman el impacto de las características de estudiantes y docentes sobre puntuaciones totales del cuestionario Docentia (efectos fijos) y las varianzas residuales asociadas a los dos niveles de agregación de los datos (efectos aleatorios).

Para dar respuesta a los diferentes objetivos planteados se estimaron un total de 9 modelos (ver Tabla II), el primero no incluye predictores y está destinado a probar que existe suficiente varianza residual entre profesores para determinar si se debe continuar con el plan de análisis (Modelo 0). Además, sirve como referente para la comparación del resto de modelos que incluyen predictores. El resto de los modelos incorporan diferentes grupos de predictores con el propósito de recoger evidencias empíricas de su impacto sobre las evaluaciones de la calidad del docente. En la siguiente tabla se describe que características de los estudiantes y los docentes incluye cada uno:

El Modelo 1 incorporó el efecto de los predictores relacionados con el aprendizaje de los estudiantes (notas, interés, dificultad y asistencia

Recibido: 10-12-2022 Aceptado: 30-06-2023

TABLA II. Modelos estimados y predictores incluidos en cada uno

Modelo	**Predictores**
0	Nulo. Sin predictores.
1	Estudiantes: Aprendizaje (notas, interés, dificultad, asistencia, horas de estudio y tipo de titulación)
1b	Estudiantes: Rendimiento (notas)
2	Estudiantes: Aprendizaje (interés, dificultad, asistencia y horas de estudio) + factores demográficos (sexo y edad)
2b	Estudiantes: factores demográficos (sexo y edad)
3	Estudiantes: Aprendizaje (interés, dificultad, asistencia y horas de estudio) + factores demográficos (sexo y edad); + Varianza aleatoria de los predictores en el nivel 2
4	Estudiantes: Aprendizaje (interés, dificultad, asistencia y horas de estudio) + factores demográficos (sexo y edad); + Varianza aleatoria de los predictores en el nivel 2; + Docentes: factores demográficos (sexo y edad)
4b	Estudiantes: Aprendizaje (interés, dificultad, asistencia y horas de estudio) + factores demográficos (edad); + Varianza aleatoria de los predictores en el nivel 2; + Docentes: factores demográficos (edad); + sexo cruzado entre niveles
5	Estudiantes: Aprendizaje (interés, dificultad, asistencia y horas de estudio) + factores demográficos (sexo y edad); + Varianza aleatoria de los predictores en el nivel 2; + Docentes: factores demográficos (sexo y edad) + factores académicos (categoría laboral, nº de sexenios y nº de quinquenios)
5b	Docentes: factores demográficos (sexo y edad) + factores académicos (categoría laboral, nº de sexenios y nº de quinquenios)

y tipo de titulación) en la parte en la parte fija del modelo. De forma complementaria, se estimó otro modelo para comprobar el efecto del rendimiento de forma separada (Modelo 1b). El Modelo 2 añadió las variables de sexo y edad. También se probó el efecto por separado de estos predictores en el Modelo 2b. El modelo 3 incluye los efectos de esos predictores en el segundo nivel, incluyendo parámetros de varianza aleatoria. El modelo 4, comienza con la introducción de características de los docentes, primero se incluyó el sexo de los docentes y la edad. En un modelo complementario al anterior, el 4b, se sustituyó la variable sexo de estudiantes y docentes por el efecto cruzado. Finalmente, en el modelo 5 (Final) se incorporó categoría laboral, producción científica y experiencia docente. También, a modo de complemento, se estimó

Recibido: 10-12-2022 Aceptado: 30-06-2023

un modelo con las características solo de los docentes (Modelo 5b) y coeficientes aleatorios de estos predictores en el nivel 2.

Para comparar los modelos se utilizaron los estadísticos de ajuste global de verosimilitud restringida (-2 log-likelihood), y los criterios de información AIC y BIC. Un menor valor de estos índices señala mejor ajuste del modelo. Además, para comprobar si la varianza explicada por los modelos con predictores era significativa se calculó la diferencia entre los índices de verosimilitud (con distribución Chi-^{2} y con grados de libertad igual a la diferencia entre el número de parámetros de los modelos comparados). Valores significativos indican que la inclusión de los predictores explica de forma significativa una parte de la variabilidad de las puntuaciones del docente.

Para estimar la importancia de los predictores se siguieron las recomendaciones de Lorah (2018). Se estimó R^2 (Snijders y Bosker, 2012) para comprobar la cantidad de varianza de error que se ha reducido en el primer nivel.

$$R^2 = 1 - \frac{\sigma^2_{nivel\,1\,Final} + \sigma^2_{nivel\,2\,Final}}{\sigma^2_{nivel\,1\,Inicial} + \sigma^2_{nivel\,2\,Inicial}} \quad (1)$$

σ^2 es la varianza entre los niveles de anidamiento de los datos. Se compararon los resultados de dos modelos. El numerador incluye los resultados del modelo completo o final y en el denominador el modelo sin predictores. También se calculó f^2 (Cohen, 1992) para estimar el tamaño del efecto completo, considerando nivel 1 y nivel 2.

$$f^2 = \frac{R^2}{1 - R^2} \quad (2)$$

Los valores de 0,02 en adelante se consideran efectos con importancia baja, a partir de 0,15 medios y de 0,35 o superior efectos grandes. Se añadió, además, el Coeficiente de Correlación Intraclase (CCI) como una estimación de la proporción de variabilidad de los resultados en el segundo nivel del modelo, es decir, el efecto de los docentes.

Todos los análisis se realizaron con el programa estadístico IBM-SPSS 27, utilizando el módulo MIXED (Modelos lineales Mixtos).

Resultados

La Tabla III presenta los resultados de los diferentes modelos estimados, mostrando los coeficientes de la parte fija, las varianzas aleatorias de los dos niveles y, entre paréntesis, los errores típicos asociados. Con el propósito de facilitar su lectura se incluyen los modelos principales, los complementarios se destacan en la descripción de estos resultados. En la Tabla IV se incluyen los distintos índices de ajuste global, la estimación de la proporción de varianza explicada por los modelos que incluyen predictores y la correlación intraclase. La valoración del ajuste se lleva a cabo sobre los 9 modelos presentados en la sección de metodología.

Como se puede observar, el estudio de los efectos aleatorios de este modelo nulo indica que existe variación residual en el nivel de estudiantes, considerando que es un efecto que incluye una asociación entre estudiantes, docentes y cursos valorados (3,21, E.T.=0,013, p<0,005) y aleatoria en el nivel 2, que incluye la variación entre los docentes (1,58, E.T.= 0,040, p<0,005). Queda comprobado, por tanto, que existe varianza en los dos niveles. El punto de corte en este modelo (la puntuación media esperada en el SET profesorado para todos los alumnos en todos los cursos) es de 7,916 (E.T. = 0,028, p<0,005), sobre 10 puntos.

Como se muestra en la Tabla III, en el modelo final (Modelo 5), el punto de corte es 4,501 (E.T = 0,041, p<0,005). Este valor promedio representa la puntuación de calidad docente cuando los predictores que incluye el modelo son iguales a cero. También conviene recordar, con fines interpretativos, que las variables edad están centradas respecto a la media y, por tanto, el valor 0 es de 22 años para los estudiantes y 49 para los docentes.

El modelo explica el 36% (R^2=0,363) de la variabilidad de las respuestas de los estudiantes. Y, el efecto global del conjunto de predictores tiene un tamaño del efecto grande (f^2=0,571). Este modelo ajusta notablemente mejor a los datos que el modelo nulo inicial, con una diferencia muy importante de parámetros en la estimación (3 vs 23 parámetros) y una considerable reducción de los índices de ajuste Global (-2 log-likelihood, AIC y BIC). Las variables del docente (experiencia investigadora, experiencia docente, edad y sexo), por su parte, aportan el 3,5% a la explicación de la variabilidad (diferencia entre los tamaños del efecto del Modelo 5 y el Modelo 3), considerándose un tamaño del efecto pequeño. Además, como se observa en los valores de la CCI, el

Recibido: 10-12-2022 Aceptado: 30-06-2023

TABLA III. Modelos de efectos cruzados estimados

Efectos	Modelo 0	Modelo 1	Modelo2	Modelo 3	Modelo 4	Modelo 5
Intercepto	7,916 (0,022)***	4,635 (0,034)***	4,574 (0,038)***	4,57 (0,035)***	6,061 (0,081)***	4,501 (0,041)***
Sexo estudiante (Mujer)			0,052 (0,009)***	0,049 (0,010)***	0,051 (0,010)***	0,053 (0,010)***
Edad estudiante			0,14 (0,01)***	0,014 (0,001)***	0,14 (0,001)***	0,14 (0,001)***
Nota						
[Asistencia=Menos 20%]		-0,54 (0,028)***	-0,558 (0,028)***	-0,57 (0,031)***	-0,57 (0,031)***	-0,566 (0,031)***
[Asistencia=20%-39%]		-0,452 (0,026)***	-0,465 (0,026)***	-0,459 (0,028)***	-0,457 (0,028)***	-0,456 (0,028)***
[Asistencia=40%-59%]		-0,301 (0,019)***	-0,303 (0,018)***	-0,312 (0,021)***	-0,312 (0,021)***	-0,313 (0,021)***
[Asistencia=60%-79%]		-0,194 (0,013)***	-0,195 (0,013)***	-0,198 (0,016)***	-0,1987(0,015)***	-0,197 (0,015)***
[Asistencia=80% o más]		-	-	-	-	-
[Horas de estudio=Menos de 1h]		0,122 (0,026)***	0,144 (0,026)***	0,138 (0,027)***	0,138 (0,027)***	0,139 (0,027)***
[Horas de estudio=De 1 a 4]		0,217 (0,024)***	0,227 (0,026)***	0,226 (0,024)***	0,225 (0,024)***	0,225 (0,024)***
[Horas de estudio=De 5 a 7]		0,159 (0,024)***	0,170 (0,024)***	0,169 (0,024)***	0,169 (0,024)***	0,17 (0,024)***
[Horas de estudio=De 8 a 10]		0,1 (0,026)***	0,106 (0,026)***	0,106 (0,027)***	0,106 (0,027)***	0,105 (0,027)***
[Horas de estudio=Más de 10]		-	-	-	-	-
Interés		0,44 (0,002)***	0,434 (0,002)***	0,437 (0,002)***	0,437 (0,002)***	0,437 (0,002)***
Dificultad		-0,015 (0,002)***	-0,015 (0,002)***	-0,016 (0,002)***	-0,016 (0,002)***	-0,017 (0,002)***
Titulación (Máster)						
Sexo docente (Mujer)					-0,07 (0,025)**	-0,066 (0,024)**
Edad profesor					-0,028 (0,001)***	-0,04 (0,002)***
Nº sexenios						0,025 (0,1)*
Nº quinquenios						0,027 (0,1)*

(continúa)

TABLA III. Modelos de efectos cruzados estimados (continuación)

Efectos	Modelo 0	Modelo 1	Modelo2	Modelo 3	Modelo 4	Modelo 5
Varianza aleatoria						
σ2e (nivel1)	3,21 (0,013)***	2,248 (0,009)***	2,243 (0,008)***	2,097 (0,010)***	2,01 (0,010)***	2,01 (0,010)***
σ2 (nivel2)	1,58 (0,040)***	0,932 (0,025)***	0,931 (0,024)***	0,721 (0,046)***	0,663 (0,021)***	0,663 (0,021)***
σ2 (sexo estudiante)				0,094 (0,01)***	0,093 (0,01)***	0,091 (0,01)***
σ2 (asistencia)				0,168(0,1)***	0,166 (0,1)***	0,166 (0,1)***
σ2 (horas estudio)				0,034 (0,006)***	0,034 (0,006)***	0,034 (0,006)***
σ2 (interés)				0,002 (0,000)***	0,002 (0,000)***	0,002 (0,000)***
σ2 (dificultad)				0,003 (0,000)***	0,004 (0,000)***	0,0043(0,000)***

*p<0,05; **p<0,01;***p<0,005

TABLA IV. Modelos de efectos cruzados estimados

	M0	M1	M1b	M2	M2b	M3	M4	M4b	M5	M5b
-2 log.*likelihood*	553034,556	505163,436	552533,12	504916,456	552231,034	504019,01	503635,197	503634,437	503615,197	552630,194
AIC	553038,556	505167,436	552537,12	504920,456	552235,034	504033,01	503649,197	503648,437	503629,197	552634,194
BIC	553058,171	505187,051	552556,736	504940,071	552254,65	504101,662	503717,849	503717,09	503697,849	552653,809
Nº de parámetros	3	13	4	15	5	20	22	23	24	6
R^2 (nivel1)		0,337	0,013	0,338	0,007	0,349	0,362	0,362	0,363	0,028
f^2 (Total)		0,507	0,013	0,510	0,007	0,536	0,567	0,567	0,571	0,029
CCI	0,330	0,293	0,329	0,293	0,330	0,231	0,217	0,217	0,215	0,316

21,5% aproximadamente de la varianza de los resultados se mantiene en el 2º nivel de anidamiento.

También conviene resaltar que, del conjunto de factores, el interés del estudiante por la materia es la característica con mayor poder explicativo en los resultados. En el lado opuesto, el sexo de docentes y profesores o las variables de experiencia docente e investigadora, aunque significativas, son las que menos aportan. Las puntuaciones que emite el estudiante mejoran en 0,437 (0,002; p<0,005) por cada punto de aumento en el nivel de interés.

También las valoraciones de los SET mejoran si los estudiantes se sitúan por encima de la edad media (0,14 puntos por cada año). Así un estudiante que finalice sus estudios de grado o comience máster valorará mejor a sus docentes. De forma opuesta, las valoraciones de los estudiantes disminuyen a medida que el docente supera la edad media, unos 49 años, (-0,04 puntos por cada año).

Y con relación al *sexo*, las mujeres tienden a valorar 0,053 puntos más altos a (todos/as) sus docentes. Y, en términos promedios, se valora 0,04 puntos peor a las profesoras. El modelo 4b probó el efecto cruzado entre docentes y estudiantes de la variable sexo y se observó un efecto diferencial cuando los estudiantes varones evalúan a sus profesoras, con una puntuación de 0,079 puntos por debajo que las estudiantes valorando al mismo colectivo.

En cualquier caso, el impacto de estas variables, aunque significativo, como indican los tamaños del efecto de los modelos que incluyen únicamente características demográficas del estudiante (Modelo 2b) o solo características de los docentes (Modelo 5b), tienen una importancia prácticamente nula.

La asistencia a clase declarada por el estudiante es una variable ordinal con 5 categorías que expresada en porcentaje. Se ha realizado una codificación de contraste, situando el nivel máximo de asistencia (más del 80%) en el punto de corte. La relación entre la puntuación otorgada al profesor y la asistencia es lineal y positiva. Los estudiantes que dicen asistir a clase "casi siempre", proporcionan calificaciones más altas de sus profesores en comparación con aquellos que dicen asistir "casi nunca" (menos del 20%), -0,56 de diferencia entre los dos grupos.

Las horas declaradas de estudio y trabajo semanal también es una variable ordinal con 5 categorías. Se ha realizado una codificación de contraste, situando el máximo en más de 10 horas de trabajo a la semana

Recibido: 10-12-2022 Aceptado: 30-06-2023

por asignatura como punto de corte. La relación entre la puntuación otorgada al profesor y la asistencia es no lineal, alcanza un máximo en la valoración del profesor cuando el estudiante dedica entre 1 y 4 horas semanales de estudio. Estos estudiantes valoran 0,225 puntos más alto a sus docentes que el grupo que estudia más de 10 horas.

La dificultad percibida por el estudiante es una variable valorada en una escala de 0 a 10. También ha resultado ser una característica significativa con impacto negativo en la valoración de la actividad docente del profesor (-0,017).

La experiencia investigadora certificada a través de sexenios tiene un impacto significativo y positivo en la valoración que recibe el profesor (0,031, E.T.= 0,012, $p<0,05$), también la experiencia docente medida a través del número de quinquenios afecta positivamente (0,027, E.T.= 0,01, $p<0,05$).

El modelo que incluye únicamente los factores del docente (Modelo 5b) tiene una mayor capacidad explicativa que el incorpora solo las características demográficas del estudiante (Modelo 2b), un tamaño del efecto de 0,029 (bajo) frente a 0,007 (nulo).

Si comparamos la capacidad explicativa de variables del estudiante vinculadas al aprendizaje (Modelo1) con el efecto de sus factores demográficos (Modelo 2b), observamos una gran diferencia en el tamaño de sus efectos. Mientras que los primeros tienen un efecto grande (0,507), los segundos, como ya hemos indicado arriba no consiguen un tamaño suficiente para valorarlos como un efecto bajo (0,007).

Finalmente, resulta interesante observar cuáles son las características de los estudiantes y profesores que no han tenido impacto estadístico en la valoración de sus profesores. No se observan diferencias debidas a la especialización de los estudios (grado o máster). Aunque es importante señalar que esta variable está relacionada con la edad del estudiante. Aunque con precaución, se puede afirmar que el comportamiento valorativo de un estudiante es independiente del tipo de estudios que cursa.

La nota promedio obtenida a lo largo de sus estudios universitarios (entendida como una variable de calificación general de la trayectoria universitaria) no es significativa estadísticamente, lo que sugiere que los estudiantes con calificaciones más altas no proporcionan sistemáticamente calificaciones mayores a sus profesores. Tampoco resultan significativas si se incluyen con el conjunto de características del estudiante (Modelo

2). En cambio, si se analizan por separado (Modelo 1b), su tamaño del efecto es de 0,013. Una importancia mayor que las características demográficas (Modelo 2b). Asimismo, la rama de estudios y la categoría laboral del profesor tampoco muestran tener efectos.

Conclusiones

Los resultados de este trabajo de investigación, con una muestra ciertamente amplia de estudiantes, profesores y actividades docentes, han permitido probar empíricamente el efecto de diferentes factores que la literatura señala como indicadores de un posible sesgo en las valoraciones que los estudiantes emiten sobre la calidad de sus docentes. Además de establecer la relación con otras características que sí pueden estar determinadas por los procesos de enseñanza.

Considerando los resultados del modelo final, las valoraciones que los estudiantes emiten sobre los profesores están influenciadas, sobre todo, y por este orden, por el interés que manifiestan los estudiantes por la asignatura, las edades de los estudiantes y los docentes, la asistencia declarada a clase, la dificultad percibida, las horas de estudio. y por la capacidad investigadora del docente (medida como sexenios).

Si consideramos las características relacionadas con el proceso de aprendizaje del estudiante, que son un reflejo del compromiso con el estudio (interés, asistencia y horas de estudio), valoran mejor a los docentes aquellos estudiantes con más interés por las asignaturas y mayor tasa de asistencia. De hecho, precisamente el impacto del interés fue uno de los cinco factores que Spooren et al (2017) identificaron como factor de calidad docente. También Greimel-Fuhrmann y Geyer (2003) y Paswan y Young (2002) apuntaron en sus trabajos que el comportamiento del docente y sus interacciones con el estudiante determinan su nivel de interés. Y, por tanto, no puede clasificarse como un factor de sesgo.

La asistencia es otro factor que determina los resultados. Es una relación directa entre asistencia y puntuación del docente. Los estudiantes que asisten a menos del 40% y también del 20% valoran con aproximadamente medio punto menos a sus docentes que los que asisten el 80% o más. Estos resultados están en la línea de Beran y Violato (2005) y Davidovitch y Soen (2006), que destacaron su importancia y contradicen, por tanto, los de Guinn y Vincent (2006). Fjortoft (2005) relaciona la asistencia

Recibido: 10-12-2022 Aceptado: 30-06-2023

regular a las clases con un mayor interés y motivación por el aprendizaje. La pregunta que se platea aquí es si esos estudiantes que acuden a menos del 50% de las clases emiten valoraciones sin sesgo. Es una ausencia provocada por el tipo de enseñanza del docente o por falta de interés.

También influye en las evaluaciones, pero con menos fuerza, las horas de estudio declaradas, siendo una relación no lineal, en la que hay un máximo, digamos que es un tiempo de estudio razonable, pues la dedicación más allá de un determinado número de horas puede expresar otro tipo de dificultades (del estudiante, del curso, del profesor…) que se alejan del marco de la normalidad del estudio y del trabajo autónomo del estudiante. En este sentido, Spooren et al. (2013) señala que estudio y esfuerzo también son indicadores de su interés y motivación y también dependen, en parte, de la calidad de la enseñanza. Para finalizar con los factores de aprendizaje del estudiante, la dificultad percibida es un cierto inhibidor de sus valoraciones. Aunque el efecto es muy bajo (-0,016 por cada nivel de aumento de esa percepción). Esta característica, aunque significativa, tiene poca aportación a la variabilidad explicada por el modelo.

Si continuamos con aspectos del docente que pueden determinar la calidad de su enseñanza, se ha encontrado un efecto positivo de la experiencia investigadora (nº de sexenios) y la experiencia (nº de quinquenios). Con un mayor impacto del primero. Este resultado que muestra el reconocimiento específico del perfil del profesor universitario que conjuga experiencia docente y experiencia investigadora. Resultados similares se han encontrado en los trabajos de Spooren et al. (2013), que asoció estas variables a las habilidades didácticas del profesor y el dominio de la materia. Los resultados coinciden con McPherson y Jewell (2007) y McPherson et al. (2009), que demostraron que la experiencia docente es un factor relacionado con mejores puntuaciones en los SET (McPherson y Jewell, 2007 y McPherson et al., 2009), y coinciden con los de Ting (2000), que halló un efecto de la calidad de la producción científica, aunque medida con el número de citas de las publicaciones. Estos factores, aunque significativos, tienen un tamaño del efecto prácticamente nulo.

Si nos centramos en las características demográficas de estudiantes y docentes (sexo y edad), confirmamos un efecto significativo. La edad es la de mayor poder explicativo, algo mayor que factores como las horas de estudio, la dificultad y las experiencia investigadora y docente de

los profesores. De los factores docentes, su edad es la que aporta más a la explicación de la variabilidad de resultados. Los estudiantes que tienen un año más que la media, 23 años, otorgan 0,14 puntos más en promedio a sus docentes. Por tanto, emiten una mejor valoración de la calidad de la docencia al final de su formación. Este resultado coincide con Sprinkle (2008) y Spooren (2010), donde estudiantes de mayor edad tienen a otorgar mejores puntuaciones a sus docentes. Esto tampoco puede considerarse un factor de sesgo porque, como apuntó Wachtel (1998), las valoraciones más elevadas pueden estar provocadas por un mayor nivel de madurez o una mayor especialización de las materias. Aspectos que indican un mayor interés del estudiante.

También la edad del docente ha resultado significativa. Resultados similares a los encontrados por McPherson et al. (2009), donde las mejores valoraciones las obtuvieron los profesores más jóvenes. Por tanto, evidencia que apoya los resultados de la revisión de Wachtel (1998), que observó una relación inversa entre la edad del docente y las valoraciones de los estudiantes. De sus características es la que tiene mayor poder explicativo, aunque el tamaño del efecto es bajo.

El sexo también resultó significativo, las estudiantes valoran más generosamente a sus docentes y a las docentes se las juzga más severamente. Esto último coincide con gran parte de los resultados de investigación (Basow et al., 2006; Boring, 2017; Boring et al., 2016; Mitchell Martin, 2018 y Rivera Tilcsik, 2019). También se probó el efecto cruzado de esta variable entre estudiantes y docentes. Se encontró un efecto significativo y se observó que los estudiantes varones valoran peor a sus profesoras que a sus profesores, como Sprinkle (2008). Sin embargo, en este estudio las estudiantes valoran igual a sus profesoras que a sus profesores, algo que no ocurría en el trabajo de ese mismo autor. Aun así, los tamaños de los efectos de estas variables no pueden ni siquiera considerarse bajos. Además, el modelo que incluye el sexo cruzado (Modelo 4b) no aportó varianza significativa respecto al modelo que los incluyó por separado (Modelo 4). Por tanto, no puede considerarse un factor que sesgue las valoraciones de los estudiantes. No es determinante, como argumentaron Centra y Gaubatz (2000) y Spooren (2010). La relación con las puntaciones del SET tendría que considerarse extremadamente débil, como la encontrada en la revisión de Griffin (2004).

De las comparaciones entre modelos principales y complementarios se concluye que las características del estudiante son las que más importan

Recibido: 10-12-2022 Aceptado: 30-06-2023

en la explicación de la puntuación de los SET, con un tamaño del efecto grande. Las variables del docente, en conjunto, tienen un tamaño del efecto bajo, siendo la edad la que tiene mayor contribución.

El sexo, tanto de docentes como estudiantes, no tienen a penas contribución a la explicación de los resultados. Tampoco el sexo cruzado. Aunque se han hallado valores significativos, los tamaños del efecto no permiten concluir que importen.

Hay que destacar que no tiene relación alguna con las valoraciones sobre la docencia el tipo de estudios de grado o máster que cursan. No obstante, la edad puede incorporar ya este efecto porque los estudiantes de máster están por encima de la media. Tampoco el tipo de plaza que ocupa el docente (funcionario o laboral). No se ha confirmado que el profesorado con vinculación permanente obtenga valoraciones más altas que el que no lo es. Tampoco el efecto significativo del área de conocimiento, resultados que difieren de los trabajos de Theall y Franklin (2001), Basow y Montgomery (2005) y Kember y Leung (2011).

Es especialmente importante destacar la ausencia de relación con el rendimiento durante la trayectoria universitaria, pues incluso los estudiantes con peores expedientes emiten similares valoraciones sobre sus profesores que los estudiantes con buenos historiales académicos. A la luz de estos resultados, opuestos a los meta-analisis de Cohen (1980, 1981), que muestran una relación positiva entre moderada y grande, y también con el trabajo de Clayson (2009). No se puede interpretar como un factor de sesgo porque los estudiantes no conocen su calificación de forma previa. Ni tampoco se puede asociar a un premio por la buena enseñanza, como indicó Spooren (2010).

De acuerdo con nuestros datos (coincidentes con los de autores como, Mohanty, et al., 2005; Stark-Wroblewski et al., 2007; Braga et al., 2014; Uttl et al., 2017; Berezvai et al., 2021) el impacto de variables como las calificaciones son insignificantes, no contribuyendo en absoluto a la explicación de la variación existente entre las valoraciones de los profesores. Y, como apuntan Hornstein (2017) y Carpenter et al. (2020), no es recomendable el uso de esas notas para valorar la capacidad docente.

Tras este análisis de los resultados, no se observa la presencia de sesgos invalidantes derivados del uso de los SET para la valoración de la docencia. Los estudiantes universitarios juzgan en función de la experiencia universitaria que tienen, de sus intereses y de sus

necesidades para estudiar y aprender en el entorno universitario. A la vista de los resultados descriptivos, las valoraciones promedio de los docentes alcanzan el notable. Es cierto que en las aulas universitarias hay estudiantes diversos, con mucho y poco interés, con asistencia frecuente y nula asistencia, que son los perfiles de estudiantes que más diferencias producen en las evaluaciones de los docentes. Quizá estos resultados podrían orientar la inclusión en los sistemas de evaluación de la docencia universitaria de diversas ponderaciones en función de alguna de las variables que han resultado tener un peso específico alto. Aunque en términos generales, los estudiantes universitarios no parecen evaluar a los docentes desde posibles prejuicios o falta de conocimiento sobre lo que es una enseñanza universitaria de calidad.

Obviamente, no se puede concluir que las percepciones de los estudiantes sobre la enseñanza son imparciales, pero en el marco de este estudio, con esta muestra importante de estudiantes y profesores y en términos promedios, hay una tendencia empírica más próxima a la ausencia de sesgos en estas valoraciones que a lo contrario. Hay razones, apoyadas en los resultados, para creer que las percepciones de los estudiantes sobre la calidad de la enseñanza son esencialmente no sesgadas.

Como prospectiva, considerando el gran efecto que el interés del estudiante tiene en las puntuaciones de los SET, es necesario profundizar en qué factores lo producen y/o lo median. La edad del estudiante, la asistencia, las horas de estudio y la dificultad son factores que pueden estar relacionados con el interés. Es necesario definir un modelo explicativo de estas características y probarlo empíricamente mediante un análisis causal.

Una limitación importante de este estudio es el efecto de selección entre los encuestados. Los cuestionarios de evaluación se distribuyeron durante el curso (previo a la calificación del estudiante) a todo el estudiantado y no existe mayor control sobre las características del sujeto que responde voluntariamente. Si bien la muestra es tan amplia que resulta razonable pensar que los resultados de este trabajo están suficientemente avalados, o al menos, no cuestionados por sesgos en la muestra. También están seleccionados los profesores de la muestra, que tienen un historial previo de evaluación constante en el tiempo y con resultados positivos.

Los resultados de este trabajo avalan la calidad técnica de los cuestionarios destinados a evaluar la calidad de la docencia a través de

Recibido: 10-12-2022 Aceptado: 30-06-2023

las valoraciones de los estudiantes. Consideramos que las valoraciones emitidas por los estudiantes, si bien no deben constituir el único elemento para la valoración de la docencia universitaria, tienen la virtud de facilitar la cobertura de la evaluación de la actividad docente del profesorado, incorporando diversidad de valoraciones que no sería posible utilizando otros instrumentos o procedimientos, siempre que se produzcan en un marco de comportamiento esencialmente no sesgado.

Desde el punto de vista de los autores, resulta de suma importancia que los responsables de los programas de evaluación de la docencia universitaria comuniquen al profesorado los indicadores de buen funcionamiento de las herramientas que se utilizan, facilitando la confianza en los sistemas a partir de la cualificación técnica de los mismos y tratando de evitar las "sospechas" de valoraciones sistemáticamente sesgadas de los estudiantes.

Referencias bibliográficas

ANECA (2017). *Orientaciones generales para la aplicación de los criterios acreditación nacional para el acceso a los cuerpos docentes universitarios*. Recuperado el 22/10/2022 de: https://acortar.link/SMjquS

Basow, S. A. & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*, 91-106.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly, 30(1)*, 25-35. https://doi.org/10.1111/j.1471-6402.2006.00259.x

Beran, T., & Violato, C. (2005). Ratings of university teacher instruction: How much do student and course characteristics really matter? *Assessment and Evaluation in Higher Education, 30(6)*, 593–601. https://doi.org/10.1080/02602930500260688

Berezvai, Z., Lukáts, G. D. & Molontay, R. (2021) Can professors buy better evaluation with lenient grading? The effect of grade inflation on student evaluation of teaching. *Assessment Evaluation in Higher Education, 46(5)*, 793-808. https://doi.org/10.1080/02602938.2020.1821866

Boring, A. (2017). Gender biases in student evaluations of teaching. *Journal of public economics*, *145*, 27-41. https://doi.org/10.1016/j.jpubeco.2016.11.006

Boring, A, Ottoboni., K & Stark, P.B. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research*. 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Braga, M., Paccagnella, M., & Pellizzari, M. (2014). Evaluating students' evaluations of professors. *Economics of Education Review, 41*, 71-88. https://doi.org/10.1016/J.ECONEDUREV.2014.04.002

Carpenter, S. K., Witherby, A. E., & Tauber, S. K. (2020). On Students' (Mis)judgments of Learning and Teaching Effectiveness. *Journal of Applied Research in Memory and Cognition, 9(2)*, 137-151. https://doi.org/10.1016/J.JARMAC.2019.12.009

Casero, A. (2008). Propuesta de un cuestionario de evaluación de la calidad docente universitaria consensuada entre alumnos y profesores. *Revista de Investigación Educativa, 26(1)*, 25-44.

Casero, A. (2010). ¿Cómo es el buen profesor universitario según el alumnado? *Revista Española de Pedagogía, 246*, 223-242.

Castro, M., Navarro, E. & Blanco, A. (2020). La calidad de la docencia percibida por el alumnado y el profesorado universitarios: análisis de la dimensionalidad de un cuestionario de evaluación docente. *Educación XX1, 23(2)*, 41-65. https://doi.org/10.5944/educXX1.25711

Centra, J. A., & Gaubatz, N. B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*, 17–33. https://doi.org/10.1080/00221546.2000.11780814

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education, 31(1)*, 16-30. https://doi.org/10.1177/0273475308324086

Clayson, D. E. (2018). Student evaluation of teaching and matters of reliability. *Assessment Evaluation in Higher Education, 43(4)*, 666-681. https://doi.org/10.1080/02602938.2017.1393495

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112(1)*, 155-159.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: a meta-analysis of findings. *Research in Higher Education, 13(4),* 321-341.

Recibido: 10-12-2022 Aceptado: 30-06-2023

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Review of Educational Research, 51(3)*, 281-309. https://doi.org/10.3102/00346543051003281

Cox, S. R., Rickard, M. K., & Lowery, C. M. (2021). The student evaluation of teaching: let's be honest – who is telling the truth? Marketing Education Review*, 32(1)*, 82-93. https://doi.org/10.1080/10528008.2021.1922924

Davidovitch, N., & Soen, D. (2006). Class attendance and students' evaluation of their college instructors. *College Student Journal, 40(3)*, 691–703.

De Juanas, A. & Beltrán, J.A. (2014). Valoraciones de los estudiantes de ciencias de la educación sobre la calidad de la docencia universitaria. *Educación XXI, 17(1),* 59-82. https://doi.org/10.5944/educxx1.17.1.10705

Esarey, J. & Valdes, N. (2020). Unbiased, reliable, and valid student evaluations can still be unfair. *Assessment Evaluation in Higher Education*, *45(8),* 1106-1120. https://doi.org/10.1080/02602938.2020.1724875

Fjortoft, N. (2005). Students' motivation for class attendance. *American Journal of Pharmaceutical Education, 69(1),* 107–112.

García, E., Colom, X., Martínez, E., Sallarés, J. & Roca, S. (2011). La encuesta al alumnado en la evaluación de la actividad docente del profesorado. *Aula abierta, 39(3)*, 3-14.

Gómez, J. C., Gómez, M., Pérez, M. C., Palazón, A. & Gómez, J. (2013). Interacción entre las expectativas académicas del alumno y la evaluación del profesor. *Aula abierta, 41(2)*, 35-44.

Greimel-Fuhrmann, B., & Geyer, A. (2003). Students' evaluation of teachers and instructional quality-Analysis of relevant factors based on empirical evaluation research. *Assessment Evaluation in Higher Education, 28(3)*, 229-238. https://doi.org/10.1080/0260293032000059595

Griffin, B. W. (2004). Grading leniency, grade discrepancy, and student ratings of instruction. *Contemporary Educational Psychology, 29(4)*, 410–425. https://doi.org/10.1016/j.cedpsych.2003.11.001

Guinn, B., & Vincent, V. (2006). The influence of grades on teaching effectiveness ratings at a Hispanic-serving institution. *Journal of Hispanic Higher Education, 5(4)*, 313–321. https://doi.org/10.1177/1538192706291138

Gump, S. E. (2007). Student evaluation of teaching effectiveness and the leniency hypothesis: A literature review. *Educational Research Quarterly, 30(3)*, 55–68.

Guthrie, E. R. (1954). *The evaluation of teaching: a progress report*. University of Washington.

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4(1)*, https://doi.org/10.1080/2331186X.2017.1304016

Johnson, R. (2000). The authority of the student evaluation questionnaire. *Teaching in Higher Education, 5(4)*, 419–434. https://doi.org/10.1080/713699176

Kember, D. & Leung, D. Y. P. (2011). Disciplinary Differences in Student Ratings of Teaching Quality. *Research in Higher Education, 52*, 278–299. https://doi.org/10.1007/s11162-010-9194-z

Kulik, J. A. (2001). Student ratings: validity, utility and controversy. *New Directions for Institutional Research, 109*, 9-25. https://doi.org/10.1002/ir.1

Lizasoain-Hernández, L., Etxeberria-Murgiondo, J., & Lukas-Mujika, J. F. (2017). A proposal for a new questionnaire for the evaluation of teachers at the University of the Basque Country. Dimensional, differential and psychometric study. *RELIEVE, 23(2)*. https://doi.org/10.7203/relieve.23.2.10436

López-Cámara, A. B., González-López, I. & de León-Huertas, C. (2016). Un análisis factorial exploratorio para la construcción de un modelo de indicadores de evaluación docente universitaria. *Cultura y Educación, 27(2),* 337-371.

Lorah, J. (2018). Effect size measures for multilevel models: Definition, interpretation, and TIMSS example. *Large-Scale Assessments in Education, 6(1)*, 1-11.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and utility. *Journal of Educational Psychology, 76(5)*, 707-754. https://doi.org/10.1037/0022-0663.76.5.707

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues and directions for future research. *International Journal of Educational Research, 11(3)*, 253-388. https://doi.org/10.1016/0883-0355(87)90001-2

Recibido: 10-12-2022 Aceptado: 30-06-2023

Marsh, H. W., & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: Popular myth, bias, validity or innocent bystanders? *Journal of Educational Psychology, 92(1)*, 202–228. https://doi.org/10.1037/0022-0663.92.1.202

Mayorga, M. J., Gallardo, M. & Madrid, D. (2016). Cómo construir un cuestionario para evaluar la docencia universitaria. *Revista de Ciències de l'educació, 2*, 6-22. https://doi.org/10.17345/ute.2016.2.974

McPherson, M. A. & Jewell, R. T. (2007). Leveling the playing field: Should student evaluation scores be adjusted?. *Social Science Quarterly, 88(3)*, 868–881. https://doi.org/10.1111/j.1540-6237.2007.00487.x

McPherson, M. A., Jewell, R. T., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern economic journal, 35(1)*, 37-51. https://www.jstor.org/stable/20642462

Mitchell, K., & Martin, J. (2018). Gender Bias in Student Evaluations. *PS: Political Science & Politics, 51(3)*, 648-652. https://doi.org/10.1017/S104909651800001X

Mohanty, G., Gretes, J., Flowers, C., Algozzine, B., & Spooner, F. (2005). Multi- method evaluation of instruction in engineering classes. *Journal of Personnel Evaluation in Higher Education, 18*, 139-151. http://doi.org/10.1007/s11092-006-9006-3

Molero, D. & Ruíz, J. (2005). La evaluación de la docencia universitaria. Dimensiones y variables más relevantes. *Revista de Investigación Educativa, 23(1)*, 57-84.

Muñoz, J. M., Ríos de Deus, M. P. & Abalde, E. (2002). Evaluación docente vs. Evaluación de la calidad. *RELIEVE, 8(2)*.

Ordoñez, R. & Rodríguez, M. R. (2015). Docencia en la universidad: valoraciones de los estudiantes de la universidad de Sevilla. *Bordón. Revista de Pedagogía, 67(3)*, 85-101. http://doi.org/10.13042/Bordon.2015.67305

Paswan, A. K., & Young, J. A. (2002). Student evaluation of instructor: A nomological investigation using structural equation modeling. *Journal of Marketing Education, 24(3)*, 193-202. https://doi.org/10.1177/0273475302238042

Penny, A. R. (2003). Changing the agenda for research into students views about university teaching: four shortcomings of SRT research. *Teaching in Higher Education, 8(3)*, 399-411. https://doi.org/10.1080/13562510309396

Rasbasch, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and cross- classified random structures using a multilevel model. *Journal of Educational and Behavioral Statistics, 19(4),* 337–350. https://doi.org/10.2307/1165397

Rivera, L. A., & Tilcsik, A. (2019). Scaling down inequality: Rating scales, gender bias, and the architecture of evaluation. *American Sociological Review, 84(2)*, 248-274. https://doi.org/10.1177/0003122419833601

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Sage.

Spencer, K. J., & Schmelkin, L. P. (2002). Student perspectives on teaching and its evaluation. *Assessment and Evaluation in Higher Education, 27(5)*, 397-409. https://doi.org/10.1080/0260293022000009285

Spooren, P. (2010). On the credibility of the judge. A cross-classified multilevel analysis on student evaluations of teaching. *Studies in Educational Evaluation, 36(4)*, 121-131. https://doi.org/10.1016/j.stueduc.2011.02.001

Spooren, P.; Brockx, B. & Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Review of Educational Research, 83(4)*, 598-642. https://doi.org/10.3102/0034654313496870

Spooren, P.; Mortelmans, D. & Christiaens, W. (2014). Assessing the validity and reliability of a quick scan for student's evaluation of teaching. Results from confirmatory factor analysis and G Theory. *Studies in Educational Evaluation, 43*, 88-94. https://doi.org/10.1016/j.stueduc.2014.03.001

Spooren, P.; Vandermoere, F.; Vanderstraeten & Pepersmans, K. (2017). Exploring high impact scholarship in research on students evaluation of teaching (SET). *Educational Research Review, 22,* 129-141. https://doi.org/10.1016/j.edurev.2017.09.001

Sprinkle, J. E. (2008). Student Perceptions of Effectiveness: An Examination of the Influence of Student Biases. *College Student Journal, 42(2),* 276–293.

Stark-Wroblewski, K., Ahlering, R. F., & Brill, F. M. (2007). Toward a more comprehensive approach to evaluating teaching effectiveness: Supplementing student evaluations of teaching with pre-post learning measures. *Assessment & Evaluation in Higher Education, 32(4)*, 403–415. https://doi.org/10.1080/02602930600898536

Recibido: 10-12-2022 Aceptado: 30-06-2023

Sulis, I., Porcu, M. & Capursi, V. (2019). On the use of the Student Evaluation of Teaching: A longitudinal analysis combining measurement issues and implications of the exercise. *Social Indicators Research, 142*, 1305-1331. https://doi.org/10.1007/s11205-018-1946-8

Ting, K. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637–661. https://doi.org/10.1023/A:1007075516271

Theall, M., & Franklin, J. (2001). Looking for bias in all the wrong places: A search for truth or a witch hunt in student ratings of instruction? *New Directions for Institutional Research, 109*, 45–56. https://doi.org/10.1002/ir.3

Uttl, B., White, C. A., & Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: Student evaluation of teaching ratings and student learning are not related. *Studies in Educational Evaluation, 54*, 22–42. https://doi.org/10.1016/j.stueduc.2016.08.007

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23(2)*, 191–210. https://doi.org/10.1080/0260293980230207

Información de contacto: Enrique Navarro Asencio. Universidad Complutense de Madrid. Facultad de Educación, departamento de Investigación y Psicología en Educación. C/ Rector Royo Villanova 1, 28040, Madrid, España. E-mail: enriquen@ucm.es