EDSUP
NÚCLEO MILENIO EXPERIENCIA ESTUDIANTIL EN EDUCACIÓN SUPERIOR

¿Cuándo es sesgo y cuándo diferencias de género en la evaluación de la docencia universitaria?*

Edgar Valencia, María Conzuelo Aguilar y Tamara Zúñiga.

Síntesis

- El sesgo de género en evaluaciones de docencia universitaria puede influir dramáticamente en decisiones de selección, desarrollo y promoción académica incrementando las brechas que afectan a las mujeres en la academia.
- La mera diferencia en puntajes entre hombres y mujeres no es indicativa de sesgo en una evaluación de la docencia. Primero hay que demostrar que el funcionamiento de la evaluación no introduce sesgo para dimensionar la brecha de genero de forma defendible.
- El estudio ilustra el uso del análisis de funcionamiento diferencial del ítem (DIF) para examinar si el significado que las y los estudiantes atribuyen a un ítem cambia en función del género del docente, sugiriendo sesgo estadístico.
- Para ilustrar el uso de DIF, el estudio examina las respuestas de un cuestionario de evaluación docente contestado por estudiantes de postgrado de dos departamentos en una Escuela de Educación en Norteamérica. Los resultados refuerzan la idea de que el sesgo de género no es una propiedad de la evaluación de la docencia en si misma sino dependiente de participantes y contexto.
- El estudio anima a los profesionales involucrados en el análisis y uso de las evaluaciones de la docencia universitaria a utilizar las mejores herramientas disponibles para diseñar y minimizar sesgo de género, reforzando la credibilidad y utilidad de estas evaluaciones, y contribuyendo a visibilizar y reducir las brechas que afectan a las mujeres en la academia.

Introducción

La literatura sobre evaluación de la docencia universitaria señala correlaciones irrelevantes con variables no relacionadas calidad de la docencia incluyendo características personales como el género (Andersen & Miller, 1997; Basow & Martin,2012; Feldman,1993; Stark & Freishtat, 2014; Boring et al., 2016). La diferencia entre hombres y mujeres en su evaluación docente puede obedecer a diversas razones.

El concepto ***techo de cristal*** refiere a la falta de acceso de las mujeres a mejores salarios, poder y oportunidades en comparación con los hombres (Bertrand, 2017). Una explicación de esta diferencia está relacionada con la inversión en capital humano. Las mujeres en comparación a los hombres invierten menos en educación, formación y experiencias laborales, acumulando menor capital ya que suelen buscar un equilibrio entre el trabajo y obligaciones familiares, estando en desventaja a la hora de solicitar puestos académicos, de titularidad, liderazgo y alta dirección (Gómez Cama et al., 2016; Weisshaar,2017). Por lo tanto, diferencias en la evaluación de la docencia podrían reflejar brechas reales en oportunidad de desarrollar habilidades para la docencia.

Dos mecanismos que juegan un rol decisivo a la hora de perpetuar la subrepresentación de mujeres en la academia son los estereotipos de género y expectativas. Los estereotipos refieren a una creencia que da forma a juicios que hace un individuo sobre los miembros de un grupo sobre la base de la pertenencia al grupo. En la evaluación docente, los estudiantes juzgan a sus profesores basados en estereotipos (Arbuckle & Williams, 2003; Cundiff et al.,2018, Mitchell & Martin, 2018; Bavishi et al., 2010; Rivera & Tilcsik, 2019), dando puntuaciones más altas a los grupos que se

*Este policy brief está basado en el Artículo: Valencia, E. Gender-biased evaluation or actual differences? Fairness in the evaluation of faculty teaching. High Educ 83, 1315–1333 (2022). https://doi.org/10.1007/s10734-021-00744-1.

asemejan a los criterios utilizados en la evaluación. Este criterio a menudo se relaciona con la "competencia", favoreciendo a los hombres sobre las mujeres (Rivera & Tilcsik, 2019). Estos estudios son coherentes con el hecho que las características del método de evaluación, incluyendo instrumento (con sus dimensiones de la calidad docente, contenido de los ítems, formato de respuesta) participantes, y condiciones de administración pueden desencadenar cambios en las respuestas y sesgo (Rivera & Tilcsik, 2019; Zipser & Mincieli, 2018), afectando la validez (Messick,1995).

En segundo lugar, están las expectativas y la violación de estas expectativas. Esto ocurre cuando los estudiantes tienen creencias que no son coherentes con el comportamiento de su profesor (MacNell et al., 2015). Por ejemplo, los estudiantes pueden creer que la disponibilidad es prototípica de un género específico (mujeres), en alguna profesión como, por ejemplo, enfermería, trabajo social, educación. Si una docente vulnera la expectativa (mostrarse accesible), los estudiantes podrían anclar su evaluación asignando un peor puntaje.

¿Cómo tradicionalmente se examina sesgo de género en evaluación docente?

Sesgo en evaluación docente universitaria se produce cuando "una característica del estudiante, del profesor o del curso afecta a la evaluación realizada, ya sea positiva o negativamente, pero no está relacionada con ningún criterio de buena enseñanza" (Centra, 2003, p. 498). Métodos típicos para determinar este impacto positivo o negativo son el análisis de correlación, el análisis de regresión y el ANOVA. Todas estas técnicas estadísticas indican una diferencia entre hombres y mujeres. Cualquier resultado estadísticamente significativo es tradicionalmente interpretado como evaluación sesgada. Sin embargo, la mera diferencia a partir de datos de evaluación docente ya recolectados (observacionales) no es sinónimo de evaluación sesgada pues puede reflejar diferencias reales en el desempeño docente entre hombres o mujeres. Incluso cuando no se observa una diferencia, todavía podría existir sesgo explicando la atenuación de esas diferencias a partir del análisis de diferencias con datos observacionales.

¿A qué nos referimos en medición por ecuanimidad?

La **ecuanimidad** comprende la teoría y métodos que ayudan a examinar si una determinada **medición** como la evaluación docente, produce diferencias injustas entre un grupo subrepresentado (o focal) y un grupo de referencia. Desde la perspectiva de ecuanimidad, el sesgo refiere a "subrepresentación del constructo o varianza irrelevante en una prueba que afecta de forma diferencial el rendimiento de distintos grupos de examinados" (AERA, APA, y NCME, 2014, p. 40). Por ejemplo, el contenido de cierto ítem con la palabra "competente" podría desencadenar un proceso cognitivo irrelevante en la respuesta de los estudiantes al evaluar la docencia activando el "estereotipo de competencia", impactando la interpretación de la pregunta entre estudiantes que depende del género del profesor. Cómo la "competencia" está asociada a hombres, es más fácil para estos docentes recibir puntajes más altos debido al estereotipo.

Ecuanimidad distingue el impacto diferencia como la diferencia en desempeño entre grupos de examinados y no implica sesgo, y sesgo estadístico, que varianza irrelevante que cambia el funcionamiento de un ítem en función del grupo (como en el ejemplo). DIF (diferential item functioning) es una de muchas técnicas de análisis estadísticos disponibles que permite responder a la pregunta: ¿está este ítem midiendo el mismo atributo en dos grupos relativo a los otros ítems? Es decir, el ítem es invariante (significa lo mismo) entre dos grupos. DIF junto con otras técnicas para prevenir e identificar sesgo estadístico son un gran paso conceptual y metodológico para examinar sesgo de género en evaluación docente comparado con el análisis de diferencias con datos observacionales.

Metodología

El estudio examina las respuestas de estudiantes de postgrado a un cuestionario de evaluación docente online en una universidad norteamericana para dos departamentos académicos (A y B). El cuestionario contiene ocho ítems con afirmaciones que apuntan a la experiencia de aprendizaje de los estudiantes.

Análisis

El estudio examina las respuestas de los alumnos estimando las dificultades de los ítems del cuestionario, que refiere a cuán difícil es atribuir una respuesta favorable a un docente. Luego, el estudio compara si la dificultad de cada ítem cambia dependiendo el género, es decir, si esa dificultad es distinta para docentes hombres y mujeres **con el mismo nivel de calidad docente**.

Si las dificultades de los ítems cambian según género, habría evidencia de DIF, es decir, evidencia de que el cuestionario podría estar favoreciendo a un grupo. De no haber DIF, uno podría comparar los resultados de la evaluación docente por género del docente e interpretarlos como diferencias en su calidad de enseñanza (impacto diferencial). El impacto diferencial refleja diferencias reales que podrían estar explicadas por la oportunidad de desarrollar habilidad docente (techo de cristal). Este análisis fue realizado para cada departamento por separado.

Resultados, Conclusiones y Recomendaciones

En el caso del departamento “B” no hay evidencia de DIF por género del docente, lo que significa que el cuestionario no fue interpretado de forma distinta para docentes hombres y mujeres por los estudiantes. En el departamento “A” hay evidencia de DIF por género del docente: las mujeres tienden a recibir una actitud significativamente más favorable de los estudiantes en algunos ítems en comparación a sus pares hombres **con el mismo nivel de calidad docente**. El análisis en el departamento “A”, revela la existencia de DIF en cinco ítems de ocho ítems, en los que cuatro ítems eran atributos más difíciles de atribuir a docentes hombres que mujeres. Es decir, el estudio sugiere DIF, pero no problemas de ecuanimidad.

El estudio también reporta diferencia de género en calidad docente. En el departamento “B” no hay una diferencia (impacto diferencial) que afecte a las docentes mujeres. El análisis de dos departamentos dio lugar a conclusiones diferentes, lo que refuerza la noción de que validez de una medición y la presencia de sesgo de género son más bien atributos locales, es decir, hay que examinarlos en su contexto (para cada instrumento, participantes y condiciones de administración).

Los profesionales en Educación Superior deben conocer, comprender e implementar formas de elaborar, testear y demostrar la neutralidad del contenido y del procedimiento de evaluación docente, utilizando las mejores herramientas tanto cualitativas como cuantitativas para examinar sesgo de género, así como para reforzar la credibilidad de las evaluaciones docentes. De este modo, también se podrá fortalecer la toma de decisiones sobre el desarrollo y la promoción académica de los docentes para que no afecten a mujeres u otros grupos que se encuentren subrepresentados.

Referencias:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Andersen, K., & Miller, E. D. (1997). Gender and Student Evaluations of Teaching. *PS: Political Science and Politics*, 30(2), 216. https://doi.org/10.2307/420499.

Arbuckle, J., & Williams, B. D. (2003). *Students' perceptions of expressiveness: Age and gender effects on teacher evaluations. Sex Roles*, 49(9–10),

507–516.
https://doi.org/10.1023/A:1025832707002.

Basow, S. A., & Martin, J. L. (2012). Bias in student evaluations. In *Effective evaluation of teaching: A guide for faculty and administrators* (pp. 40–49). Society for the Teaching of Psychology

Bavishi, A., Madera, J. M., & Hebl, M. R. (2010). The effect of professor ethnicity and gender on student evaluations: Judged before met. *Journal of Diversity in Higher Education*, 3(4), 245–256. https://doi.org/10. 1037/a0020763.

Bertrand, M. (2017). The glass ceiling. Becker Friedman Institute for Research in Economics Working Paper No. 2018-38, https://doi.org/10.2139/ssrn.3191467

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

Centra, J. A. (2003). Will teachers receive higher student evaluations by giving higher grades and less course work? *Research in Higher Education*, 44(5), 495–518.
https://doi.org/10.1023/A:1025492407752.

Cundiff, J. L., Danube, C. L., Zawadzki, M. J., & Shields, S. A. (2018). Testing an intervention for recognizing and reporting subtle gender bias in promotion and tenure decisions. *The Journal of Higher Education*, 89(5), 611–636. https://doi.org/10.1080/00221546.2018.1437665.

Feldman, K. A. (1993). College students' views of male and female college teachers: Part II—Evidence from students' evaluations of their classroom teachers. *Research in Higher Education*, 34(2), 151–211. https://doi. org/10.1007/BF00992161

Gómez Cama, M., Larrán, M. J., & Andrades Peña, F. J. (2016). Gender differences between faculty members in higher education: A literature review of selected higher education journals. *Educational Research Review*, 18, 58–69. https://doi.org/10.1016/j.edurev.2016.03.001.

MacNell, L., Driscoll, A., & Hunt, A. N. (2015). What's in a name: Exposing gender bias in student ratings of teaching. *Innovative Higher Education*, 40(4), 291–303. https://doi.org/10.1007/s10755-014-9313-4.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741–749. https://doi. org/10.1037/0003-066X.50.9.741.

Mitchell, K. M. W., & Martin, J. (2018). Gender Bias in Student Evaluations. *PS: Political Science & Politics*, 51 (03), 648–652. https://doi.org/10.1017/S104909651800001X.

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.

Stark, P., & Freishtat, R. (2014). An evaluation of course evaluations. *ScienceOpen Research* https://www.
scienceopen.com/document/id/ad8a9ac9-8c60-432a-ba20-4402a2a38df4.

Weisshaar, K. (2017). Publish and perish? An assessment of gender gaps in promotion to tenure in academia. *Social Forces*, 96(2), 529–560. https://doi.org/10.1093/sf/sox052.

Zipser, N., & Mincieli, L. (2018). Administrative and structural changes in student evaluations of teaching and their effects on overall instructor scores. *Assessment & Evaluation in Higher Education*, 43(6), 995–1008.