https://doi.org/10.21865/RIDEP66.5.06

El Problema del Análisis de la Evaluación de la Satisfacción Estudiantil en el Ámbito Universitario: Un Estudio de Simulación

The Problem of the Analysis of the Evaluation of Student Satisfaction in the University Environment: A Simulation Study

L. Catheryne Lancheros-Florián[1], Eduar S. Ramírez[2] y Jesús M. Alvarado[3]

Resumen

La medición de la satisfacción estudiantil ha suscitado grandes retos debido a la complejidad de esta medición. Actualmente, los modelos multinivel se han acogido como una alternativa de análisis. Sin embargo, sus restricciones hacen que su uso práctico sea difícil de lograr. Una alternativa es la aplicación de correcciones de sesgo que se enfoquen en controlar variables que disminuyen la validez de los métodos usualmente utilizados. Un estudio antecedente mostró que liberar los datos del sesgo proporcionado por las diferencias entre las puntuaciones de los profesores hace que emerjan estructuras factoriales diferentes. En el presente estudio se realizó una simulación para comprobar esos primeros hallazgos y observar qué tanto influyen las diferencias de calificaciones entre los profesores, respecto a las estimaciones de los modelos factoriales. Se observó, que aumentar las diferencias de calificación entre los profesores, generó un incremento paradójico en la calidad de los índices de ajuste.

Palabras clave: satisfacción estudiantil, sesgo, validez, estructura factorial, índice de ajuste, simulación

Abstract

Student´s satisfaction measured has raised great challenges since data processing involves the use of models capable of explaining the complexity of this measurement. In recent decades, multilevel models have been accepted as an alternative assessment for this phenomenon. However, its restrictions make its practical use difficult to achieve. An alternative is the application of bias corrections on controlling variables that will reduce the validity of the normative methods used. An antecedent study reveals that freeing the data from the bias provided by the differences between the inclinations of the teachers causes factorial structures to emerge different from those estimated, without using said procedures. In the present study, a simulation was carried out to check these first conclusions and observe how much it influences the difference in grades between teachers with respect to the estimates of the factorial models. When the difference between teachers' qualification increased, it was noted a paradoxical increase in the quality of the fit indices.

Keywords: student satisfaction, teacher, bias, validity, factor structure, fit index, simulation

[1] Magíster en Psicología y Máster. Metodología de las Ciencias del Comportamiento y de la Salud. Doctoranda en Psicología. Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid, España. Tel.: +57 3015500973. Correo: ladycala@ucm.es (Autora de correspondencia)

[2] Máster en Metodología de las Ciencias del Comportamiento y de la Salud. Doctorando en Psicología. Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid, España. Tel.: +34 611 198 942. Correo: edrami01@ucm.es

[3] Doctor en Psicología. Catedrático de Universidad. Facultad de Psicología, Universidad Complutense de Madrid. Campus de Somosaguas, 28223 Pozuelo de Alarcón, Madrid. Despacho 2106-B, España. Tel.: +34 913943055. Correo: jmalvara@ucm.es

ISSN: 1135-3848 print /2183-6051online

Introducción

La evaluación del desempeño de los profesores (SET, por sus siglas en inglés) tiene como objetivo asegurar la calidad de la educación (Stake et al., 2017; Muñoz et al., 2011), aunque cada universidad establece sus propios criterios acerca de calidad de la docencia (Oermann et al., 2018). Esto ha generado una proliferación de cuestionarios (Spooren et al., 2013), así como estudios sobre sus características psicométricas (Turull & Buxarrais, 2018). Sin embargo, estos instrumentos han sido objeto de diferentes controversias por sus limitaciones de tipo teórico y práctico, el uso de los resultados, el modelo de profesor ideal y la baja tasa de respuesta de los estudiantes (Turull & Buxarrais, 2018).

Estas mediciones se realizan habitualmente mediante encuestas que recogen las opiniones de los estudiantes sobre la eficacia de los docentes (Pascual-Gómez, 2007). Diferentes investigaciones han identificado las variables que afectan estas evaluaciones, algunas centradas en las características del profesor, tales como el entusiasmo del docente (Gruber et al., 2012), el atractivo físico (Wolbring & Riordan, 2016), los efectos de la simpatía y el interés previo en el tema pueden ser posibles efectos de sesgo en la valoración por parte de los estudiantes (Feistauer & Richter, 2018). De igual forma, otras variables que se han analizado se relacionan con la tendencia del estudiante a responder favorablemente o con aquiescencia (Spooren et al., 2013), el género y la disciplina del estudiante, entre otros (Boring et al., 2016). Además, se han encontrado efectos relacionados con variables como el género del docente, el tamaño de la clase, el tipo de asignatura (Aramburo & Luna, 2013; Luna et al., 2010), lo que se convierten en factores de sesgo (DeFrain, 2016) que requieren otro abordaje (Acevedo & Olivares, 2010). Estos sesgos afectan las valoraciones, al punto que los profesores con mayor desempeño pueden obtener puntajes más bajos que los profesores menos efectivos (Boring et al., 2016).

Por su parte, el análisis de los datos de estas evaluaciones se suele realizar utilizando modelos univariados, que asumen independencia de las variables (Marsh, 2001; Garduño, 2000), es decir, que, con este tratamiento a los datos, se ignora la relación jerárquica que existe entre las variables, y, por tanto, su estructura multinivel. Esta práctica común respecto al análisis de los métodos de estimación tradicional es poco precisa, dado que uno de los errores más frecuentes, es el uso e interpretación de puntajes promedio brutos o naturales obtenidos directamente de las respuestas de los estudiantes (Franklin, 2001).

Lo descrito anteriormente también se refleja en que la mayoría los estudios revisados sobre evaluación docente (Alsarhan, 2017; Lang & Kersting, 2007), han descuidado esta estructura anidada y jerárquica de los datos, lo cual se constituye en un elemento fundamental para que las técnicas estadísticas y las interpretaciones de los datos adquieran sentido respecto al objeto evaluado (Park & Yu, 2016). Para análisis mas coherentes con este fenómeno se han desarrollado los Modelos Multinivel, también conocidos como modelos jerárquicos, en donde los parámetros estadísticos varían en más de un nivel (Hox et al., 2017). La variable dependiente se mide en el nivel más bajo y las independientes se miden en todos los niveles disponibles (De Leeuw et al., 2008) como la facultad, las clases, el profesor, la edad, entre otros (Pascual, 2007). Estos modelos permiten abordar este tipo de estructuras jerárquicas, ya que se centran en la covarianza existente entre los resultados de los examinados (Pardo et al., 2007). Además, han sido reconocidos como los más adecuados para representar estos datos porque pretenden identificar grupos o mezclas que expliquen la variabilidad de estos. Especificar adecuadamente estos modelos tiene efectos significativos en el análisis de las estructuras factoriales de los instrumentos para la medición de la satisfacción de los estudiantes.

Por esta razón, hay investigaciones centradas en analizar este fenómeno, dado que las calificaciones de los estudiantes presentan una estructura jerárquica, en la que las calificaciones están anidadas en cursos que a su vez están agrupados en profesores (Pekka, 2013; Rampichini et al., 2004). Por lo general, en los sujetos que pertenecen al mismo subgrupo se ha identificado que no existe independencia entre sí, generando un incumplimiento en el supuesto básico del modelo lineal general: la independencia entre observaciones. Por ejemplo, en el estudio de

Bacci y Caviezel (2011) se presentaron las varianzas de los efectos aleatorios de segundo y tercer nivel para cada factor de la prueba a analizar, estos efectos resultaron ser estadísticamente significativos, lo que indica que la jerarquía de la estructura de los datos tiene un efecto significativo en la medición de satisfacción de los estudiantes. Esto permite justificar el uso de un modelo multinivel, y además la especial atención que debe prestarse a las comparaciones entre la enseñanza sobre la base de residuos de tercer nivel. Este estudio resalta que cuando se ignora la estructura de datos y se realiza el análisis partiendo de la hipótesis de independencia, sin considerar esta estructura multinivel, se está perdiendo la información respecto al aporte de las variables en el estudio de la varianza de los datos.

Sin embargo, una problemática interesante que se ha venido estudiando se refiere a un problema de validez, no solamente del instrumento utilizado, sino también de cómo algunas fuentes de variabilidad no controladas pueden conducir a estimaciones incorrectas en la medición (Bacci & Caviezel, 2011). En el caso de la evaluación del profesor, la variabilidad en las puntuaciones de los profesores podría generar sesgos que conduzcan a conclusiones erróneas en los análisis de las propiedades psicométricas de estas escalas. Estas diferencias se dan, por ejemplo, cuando un profesor imparte una asignatura más interesante en sus contenidos o bien tiene un grupo de alumnos más "benévolos" en sus valoraciones, frente a otro profesor que debe impartir una asignatura de contenido complejo y/o es valorado por un grupo de alumnos más críticos. En este caso, sabemos que el estudio de la validez del instrumento de medida se debe hacer luego de controlar estas fuentes de variabilidad (March & Hattie, 2002: Toland & de Ayala, 2005).

Lancheros-Florián et al. (2022) aplicaron una corrección para mejorar el escalamiento de los profesores, los resultados mostraron que, una vez corregido el sesgo, producido por las diferencias entre los profesores, con un método de alineación de puntuaciones, las estructuras factoriales se veían modificadas. De manera que, si la corrección de los datos puede dar lugar a modificaciones en la estructura factorial de la escala, asumiríamos que estamos ante un problema de validez de la medida (Messick, 1998) y de cómo las interpretaciones que se pueden hacer del constructo (Borsboom et al., 2004) estarían siendo afectadas. Específicamente, el estudio se centró en la fuente de evidencia estructural, ya que se resaltó cómo la estructura factorial y la fiabilidad de la medida pueden estimarse de manera incorrecta, cuando no se han controlado las fuentes de varianza que afectan a las puntuaciones. Esta situación revela que una estructura factorial incorrecta puede tener consecuencias sobre la estimación real del nivel de aptitud de los evaluados. Situación que afectaría tanto a la red nomológica del constructo a evaluar (relación con otras variables), como a las consecuencias (posibles sesgos). Por estas razones es especialmente relevante establecer cuál es la estructura correcta del instrumento.

El estudio mencionado pretendió mostrar un procedimiento donde quedaran expuestas consideraciones que no se habían tenido en cuenta en estudios previos analizados. En este estudio se identificó que antes de medir la eficiencia del profesor primero se requiere centrar las puntuaciones, controlando la variabilidad motivada por características psicológicas y motivacionales tanto de los alumnos como de las asignaturas.

En consecuencia, se hizo necesario establecer de qué manera las diferencias entre las calificaciones de los profesores pueden ser una limitación en el proceso de validez de este tipo de instrumentos. Con el fin de contrastar estos resultados, el presente estudio realizó una simulación que permitiera incluir niveles progresivos de diferencias entre profesores. Estas diferencias se refieren a las distintas calificaciones que puede obtener un profesor, es decir, se presume que en la medición de las habilidades docentes los profesores obtienen diferentes puntajes, de acuerdo con la valoración de sus estudiantes. En este sentido, se consideró relevante observar el comportamiento de los índices de ajuste absolutos cuando se estimaban modelos que podrían explicar estas diferencias.

Acorde con lo anterior, el objetivo principal de este estudio consistió en mostrar que las diferencias obtenidas en las puntuaciones de un test pueden dar lugar a especificaciones incorrectas del modelo final. Esto porque

compromete la validez de medida, específicamente la referida a la fuente de validez estructural. Nuestra investigación se centró en los efectos estructurales al ampliar la ratio entre la varianza explicada por las variables de nivel inferior (ítems del instrumento que evalúan tres dimensiones de contenido), en presencia de variabilidad de un nivel superior. En dicho nivel, están anidadas las puntuaciones que podrían darse en el contexto escolar como ocurre en los cuestionarios de satisfacción con el ejercicio docente de un determinado profesor.

Para lograr este fin, el presente estudio de simulación buscó evaluar, en primer lugar, la bondad de ajuste de cuatro modelos: (a) modelo original, (b) modelo multinivel, (c) modelo bifactor y (d) modelo unifactorial. Nos centramos en observar cómo afecta la estructura multinivel a la recuperación del modelo simulado. En segundo lugar, se observó la sensibilidad de los índices de bondad de ajuste para detectar la especificación incorrecta de los modelos.

Método

El principal interés consistió en observar cómo afecta la diferencia entre profesores a los índices de ajuste absolutos, pues las observaciones empíricas muestran que, sin la debida corrección de las diferencias, las estructuras factoriales pueden ser erróneamente especificadas (Lancheros-Florián et al., 2022).

Los parámetros de los datos se muestrearon creando un modelo de 9 ítems anidados en tres factores. Se asumió que todos los ítems tenían un peso factorial estandarizado de .7 y no se generaron correlaciones entre factores. Para la creación del modelo y de los datos se usó una simulación Monte Carlo desde el paquete Lavaan (Rosseel, 2012), soportado en el software R core Team (2020), siguiendo las recomendaciones de Lee (2015) para estudios de simulación.

Este proceso de generación de datos recrea un contexto donde los alumnos de una clase califican a su profesor. Por ende, se simularon clases de 20 alumnos y 50 profesores en total. Esto asegura que se tuvieron en cuenta condiciones reales donde los profesores suelen tener diferentes tipos de calificaciones en función de la asignatura que imparten o de sus habilidades.

Respondiendo a la pregunta de interés principal, las muestras se dividieron en partes. La primera incluía datos donde no se asumían diferencias entre los profesores, en la segunda muestra se incluyó un rango de una desviación típica para diferenciar las calificaciones de los profesores, finalmente, las demás muestras incluyeron rangos de diferencias de 2 y 3 desviaciones típicas. Se usó una muestra de 10.000 sujetos y se realizaron 100 réplicas. La elección de las réplicas usadas se hizo bajo el criterio de precisión requerida propuesto por Cohen et al. (2001) y las simulaciones se hicieron usando un chip M1 con una RAM de 8 GB. Se estableció una semilla (*set.seed:* 242) para poder establecer condiciones estables, la cual puede observarse en el código libre ofrecido por el equipo.

Una vez generados los datos, se comparó la estructura original con un modelo multinivel, un modelo bifactor y un modelo unidimensional. Estas comparaciones se hicieron al considerar que de las puntuaciones diferenciales de los profesores emergen modelos multinivel. Igualmente, la comparación con modelos unidimensionales se realizó teniendo en cuenta que los estudios aplicados han mostrado que a mayor cantidad de diferencia entre profesores los datos tienden a organizarse en estructuras de un solo factor (Lancheros-Florián et al., 2022).

Análisis de datos

Para analizar los resultados y ver los cambios entre los parámetros verdaderos y estimados se compararon los índices de ajuste de los modelos. Se estimó la prueba de chi-cuadrado, la raíz cuadrada media residual estandarizada (SRMR), la raíz del error cuadrático medio de aproximación (RMSEA) y los índices de ajuste comparativo CFI y TLI.

Según Kelloway (1998) y Hu y Bentler (1999), los valores de RMSEA de .08 representan un buen ajuste, y los valores inferiores a .05 representan un muy buen ajuste a los datos. Para el SRMR, los valores por debajo de .08 representan un ajuste razonable y los valores por debajo de .05 indican un buen ajuste. Con respecto al CFI y TLI, los valores por encima de .95 representan un ajuste muy bueno a los datos.

Tabla 1. Valores de los Índices de Ajuste en los modelos estimados

Modelo recuperado					
Condición	*P* (x^2)	SRMR	RMSEA	CFI	TLI
$Z = 0$	.491	.017	.000	1.000	1.000
$Z = [-1, 1]$	.424	.015	.005	1.000	1.000
$Z = [-2, 2]$	.686	.008	.000	1.000	1.000
$Z = [-3, 3]$	.939	.004	.000	1.000	1.000
Modelo multinivel					
Condición	*P* (x^2)	SRMR	RMSEA	CFI	TLI
$Z = 0$	.997	.457	.000	1.000	1.017
$Z = [-1, 1]$	1.000	.026	.000	1.000	1.020
$Z = [-2, 2]$	.996	.020	.000	1.000	1.010
$Z = [-3, 3]$	.999	.016	.000	1.000	1.011
Modelo Bifactor					
Condición	*P* (x^2)	SRMR	RMSEA	CFI	TLI
$Z = 0$	.921	.020	.000	1.000	1.007
$Z = [-1, 1]$	.653	.009	.000	1.000	1.002
$Z = [-2, 2]$	.724	.005	.000	1.000	1.001
$Z = [-3, 3]$	.994	.003	.000	1.000	1.001
Modelo Unidimensional					
Condición	*P* (x^2)	SRMR	RMSEA	CFI	TLI
$Z = 0$	.000	.183	.225	.332	.109
$Z = [-1, 1]$	.000	.138	.234	.574	.432
$Z = [-2, 2]$	.000	.094	.245	.754	.671
$Z = [-3, 3]$	.000	.050	.170	.856	.808

Nota. Z: Diferencias en puntuaciones típicas anadidas a cada grupo; χ^2 $p \geq .05$**:** Chi cuadrado; SRMR≤.05/.08: *The Standardized Root Mean Square;* RMSEA≤.07/.03: *Root Mean Square Error of Approximation*; CFI≥.95: *Comparative Fit Index;* TLI ≥.95: *Tucker- Lewis Index.*

Resultados

Como lo muestra la Tabla 1, los resultados se dividieron en cuatro apartados, cada uno representando un modelo. Las comparaciones de los índices de ajuste se hicieron con relación a los diferentes niveles de puntuaciones típicas imprimidas al simular diferencias en las calificaciones docentes.

Los parámetros verdaderos corresponden a los datos del modelo recuperado sin diferencias entre los docentes, en el resto, se observan modelos con diferencias entre las puntuaciones de los profesores y modelos diferentes al recuperado.

Al revisar el ajuste de los datos, vemos que el modelo recuperado, el modelo multinivel y el modelo bifactor alcanzaron los criterios de calidad en todos los índices estimados. El modelo de un factor no alcanzó un adecuado nivel de ajuste, sin embargo, los datos de la condición *z=[-3,3]* si consiguieron valores aceptables en el índice SRMR.

Al centrarnos en los escenarios simulados, con respecto al sesgo introducido, y al aumentar las puntuaciones típicas, se observó que no hay diferencias sustantivas entre los índices de ajuste de las condiciones. No obstante, paradójicamente, cuando se genera el escenario de *z[-2,2]* y *z=[-3,3]* los índices de ajuste no detectaron el sesgo e incluso mostraron un mejor comportamiento. Un caso interesante es el observado en el modelo multinivel, puesto que en los datos con diferencia de una desviación típica *z=[-1,1]* mostraron una mejora sustantiva con respecto al modelo sin sesgo *z=0* en el índice SRMR. Esto puede observarse en la Tabla 1 al apreciar los valores de los índices de SRMR, TLI y CFI.

Discusión

El objetivo de la investigación fue evaluar las consecuencias de ignorar la estructura multinivel en la recuperación del modelo teórico. De igual forma, estudiar la sensibilidad de los índices de bondad de ajuste habitualmente utilizados en el contexto del análisis factorial confirmatorio, con el fin de detectar el problema de la especificación correcta. Los efectos estructurales se estudiaron ampliando la ratio de la varianza explicada por las variables de nivel inferior, en presencia de la variabilidad de un nivel superior en el que estas están anidadas.

Con base en lo anterior, es viable asumir que la variabilidad en los dos niveles jerárquicos genera sesgo. Por lo tanto, el estudio de simulación se usó para poner énfasis en la

importancia de los procedimientos de control de sesgo potencial (March & Hattie, 2002; Toland & de Ayala, 2005), ya que, aunque existen modelos más específicos para la evaluación de este tipo de datos, estos no se centran en este fenómeno y tienen la particularidad de ser inviables en muchos ambientes aplicados.

Ahora bien, el interés principal se centraba en explorar este fenómeno, teniendo en cuenta que cuando las estructuras factoriales pueden ser modificadas por las fuentes de variabilidad de la medida estamos ante problemas de validez (Messick, 1998), especialmente de la fuente estructural.

Dentro de nuestra propuesta global, un primer nivel en la solución de las limitaciones mencionadas se expuso en una serie de pasos para el control de sesgo (Lancheros-Florián et al., 2022). Allí se mostró que, al igualar las puntuaciones de los examinados por un método de alineación de puntuaciones emergía la estructura correcta tridimensional (ítems agrupados en los tres dominios de contenido muestreados), lo que condujo a su vez a mejorar la bondad de ajuste y las estimaciones del modelo factorial. Esto está explicado por la variabilidad expresada en los dos niveles y las diferencias en los tamaños de los grupos. Situación ya reportada en las investigaciones de intersección aleatoria en modelos multinivel (Brown, 2015; Eber et al., 2021).

En el estudio de Lancheros-Florián et al., (2022), se observó adicionalmente que cuanto mayor eran las diferencias entre los puntajes de los examinados el modelo tendía a ser unidimensional. En el presente estudio, se identificó que los índices de bondad de ajuste RMSEA, CFI y TLI no han sido sensibles en la detección de sesgo en los modelos mal especificados como el bifactor y el unidimensional. De igual forma, en el modelo multinivel, los índices de ajuste mostraron un criterio adecuado cuando no se generaron diferencias en los datos *z=0*, pero también cuando se adicionaban diferencias *z=[-1,1], z=[-2,2] y z=[-3,3]*, solamente el índice SRMR mostró cierta utilidad en el escenario sin diferencias. Sin embargo, en el modelo unidimensional el índice SRMR tendía a mejorar indebidamente cuando se adicionó variabilidad en los puntajes. Este comportamiento del SRMR junto con los malos resultados del resto de índices para detectar sesgo, pueden provocar confusión en los investigadores aplicados.

Aunque no se observó la misma situación formulada por Lancheros-Florián et al., (2022), donde era el modelo de un factor el que mejor presentaba ajuste cuando los datos tenían más diferencias, la simulación de este estudio evidenció que a mayores diferencias entre la calificación de los profesores, el índice SRMR alcanzó un criterio aceptable, lo que sugiere que varianza explicada por las variables de nivel inferior si podrían modificar la estructura original de los datos. De igual manera, estos resultados evidencian que los índices de bondad de ajuste tienen una pobre capacidad para detectar problemas de sesgo por lo que deberían examinarse con precaución cuando se sabe que existe variabilidad en los datos.

De igual forma, es importante señalar que una limitación en este estudio es que en la generación de los datos los pesos factoriales de todos los ítems fueron de .7, sin correlaciones entre factores, escenario no convencional en las observaciones empíricas, por lo que futuros estudios pueden fijarse en el fenómeno cuando dichas condiciones cambian.

Por lo anterior, los resultados de este estudio abren el panorama para una línea de investigación poco explorada que requiere atención de los investigadores, sobre todo, para establecer nuevos criterios que permitan asegurar la validez de las mediciones. Esta investigación se centró en mostrar que las cualidades métricas de la medición de la calidad docente pueden verse afectadas si no se controla el sesgo característico en estas evaluaciones. De igual forma, es importante señalar que no es suficiente que estos procesos se hagan de forma correcta, sino que también se complementen con evaluaciones del compromiso académico del estudiante (Martínez et al., 2021). Lo anterior, teniendo en cuenta que tanto la calidad docente como el compromiso académico son fenómenos unitarios que determinan el proceso del aprendizaje, y aunque sabemos que estas evaluaciones no están desprovistas de limitaciones, estos primeros pasos pueden ser un buen inicio.

Referencias

Acevedo R., & Olivares, M. (2010). Fiabilidad y validez en la evaluación docente universitaria. *Actualidades Investigativas en Educación, 10*(1), 1- 38. https://doi.org/10.15517/aie.v10i1.10089

Alsarhan, A. A. M. (2017). Alternative methods of estimating the degree of uncertainty in student ratings of teaching (*Doctoral dissertation, Brigham Young University*). Theses and Dissertations. 6939.

Arámburo Vizcarra, V., & Luna Serrano, E. (2013). La influencia de las características del profesor y del curso en los puntajes de evaluación docente. *Revista Mexicana de Investigación Educativa, 18*(58), 949-968. https://doi.org/10.15366/riee2018.11.2.001

Bacci, S., &a Caviezel, V. (2011). Multilevel IRT models for the university teaching evaluation. *Journal of Applied Statistics*, *38*(12), 2775-2791. https://doi.org/10.1080/02664763.2011.570316

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research, 0*(0), 1-11. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AETBZC.v1

Borsboom, D., Mellenbergh, G. J., & Van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061. https://doi.org/10.1037/0033-295X.111.4.1061

Brown, T. A. (2015). *Confirmatory factor analysis for applied research*. Guilford publications.

Cohen, A. S., Kane, M. T., & Kim, S.-H. (2001). The precision of simulation study results. *Applied Psychological Measurement, 25*(2), 136-145. https://doi.org/10.1177/01466210122031966

DeFrain, E. (2016). *An analysis of differences in non-instructional factors affecting teacher-course evaluations over time and across disciplines*. (Doctoral dissertation, University of Arizona.)

De Leeuw, J., Meijer, E., & Goldstein, H. (2008). *Handbook of multilevel analysis*. Springer.

Feistauer, D., & Richter, T. (2018). Validity of students' evaluations of teaching: Biasing effects of likability and prior subject interest. *Studies in Educational Evaluation, 59*, 168-178. https://doi.org/10.1016/j.stueduc.2018.07.009

Eber, F. J., Holtmann, J., & Eid, M. (2021). A Monte Carlo simulation study on the influence of unequal group sizes on parameter estimation in multilevel confirmatory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(6), 827-838. https://doi.org/10.1080/10705511.2021.1913594

Franklin, J. (2001). Interpreting the numbers: Using a narrative to help others read student evaluations of your teaching accurately. *New Directions for Teaching and Learning, 87*, 85-100. https://doi.org/10.1002/tl.10001

Garduño, J. M. G. (2000). ¿Qué factores extra clase o sesgos afectan la evaluación docente en la educación superior? *Revista Mexicana de Investigación Educativa, 5*(10), 303 - 325. https://www.redalyc.org/articulo.oa?id=14001006

Gruber, T., Lowrie, A., Brodowsky, G. H., Reppel, A. E., Voss, R., & Chowdhury, I. N. (2012). Investigating the influence of professor characteristics on student satisfaction and dissatisfaction: A comparative study. *Journal of Marketing Education*, *34*(2), 165-178. https://doi.org/10.1177/0273475312450385

Hox, J. J., Moerbeek, M., & Van de Schoot, R. (2017). *Multilevel analysis: Techniques and applications*. Routledge.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. https://doi.org/10.1080/10705519909540118

Kelloway, E. K. (1998). *Using LISREL for structural equation modeling: A researcher's guide*. Sage.

Lancheros-Florián L., Ramírez E., & Alvarado, J. (2022). Satisfacción con la calidad docente en el ámbito universitario: Potenciales sesgos y propuestas de análisis para su evaluación. *Revista Iberoamericana de Diagnóstico y*

Evaluación – e Avaliação Psicológica, *65*(4), 69-83. https://doi.org/10.21865/RIDEP65.4.06

Lang, J. W., & Kersting, M. (2007). Regular feedback from student ratings of instruction: Do college teachers improve their ratings in the long run?. *Instructional Science, 35*(3), 187-205. https://doi.org/10.1007/s11251-006-9006-1

Lee, S. (2015). Implementing a Simulation Study Using Multiple Software Packages for Structural Equation Modeling. *SAGE Open, 5*(3). https://doi.org/10.1177/2158244015591823

Luna, E., Arámburo, V., & Cordero, G. (2010). Influence of the pedagogical context on students evaluation of teaching. *International Journal of Teaching and Learning in Higher Education, 22*(3), 337-345. http://www.isetl.org/ijtlhe/

Martínez, B. M. T., del Carmen Pérez-Fuentes, M., & Jurado. (2022). Investigación sobre el compromiso o engagement académico de los estudiantes: Una revisión sistemática sobre factores influyentes y instrumentos de evaluación. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, *1*(62), 101-111. https://doi.org/10.21865/RIDEP62.1.08

Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness: Complementary, antagonistic, or independent constructs?. *The Journal of Higher Education*, *73*(5), 603-641. https://doi.org/10.1080/00221546.2002.11777170

Messick, S. (1998). Test validity: A matter of consequence. *Social Indicators Research*, *45*(1), 35-44. https://doi.org/10.1023/A:1006964925094

Muñoz, C. P., Nieto, B. B., Méndez, M. J. M., & Morillejo, E. A. (2011). Evaluación de la actividad docente en el Espacio Europeo de Educación Superior: Un estudio comparativo de indicadores de calidad en universidades europeas. *Revista Española de Pedagogía*, *248*, 145-163. https://dialnet.unirioja.es/servlet/articulo?codigo=3365101

Oermann, M. H., Conklin, J. L., Rushton, S., & Bush, M. A. (2018). Student evaluations of teaching (SET): Guidelines for their use. *Nursing Fórum, 53*(3), 280-285. https://doi.org/10.1111/nuf.12249

Pardo, A., Ruiz, M. Á., & San Martín, R. (2007). Cómo ajustar e interpretar modelos multinivel con SPSS. *Psicothema*, *19*(2), 308-321. https://www.redalyc.org/articulo.oa?id=72719220

Park, J., & Yu, H. T. (2016). The impact of ignoring the level of nesting structure in nonparametric multilevel latent class models. *Educational and Psychological Measurement, 76*(5), 824-847. https://doi.org/10.1177/0013164415618240

Pascual-Gómez, I. (2007). Análisis de la satisfacción del alumno con la docencia recibida: un estudio con modelos jerárquicos lineales. *Revista Electrónica de Investigación y Evaluación Educativa, 13*(1), 127-138 https://doi.org/10.7203/relieve.13.1.4216

Pekka R. (2013) The number of feedbacks needed for reliable evaluation. A multilevel analysis of the reliability, stability and generalisability of students' evaluation of teaching. *Assessment & Evaluation in Higher Educatio*n, *38*(2), 224-239, https://doi.org/10.1080/02602938.2011.625471

Rampichini, C., Grilli, L. & Petrucci, A. (2004). Analysis of university course evaluations: From descriptive measures to multilevel models. *Statistical Methods & Applications 13*, 357-373. https://doi.org/10.1007/s10260-004-0087-1

Rosseel Y. (2012). "lavaan: An R Package for Structural Equation Modeling." *Journal of Statistical Software*, *48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02.

Ruiz, M. A., Pardo, A., & San Martín, R. (2010). Modelos de ecuaciones estructurales. *Papeles del psicólogo, 31*(1), 34-45. https://dialnet.unirioja.es/servlet/articulo?codigo=3150815

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

Stake, R. E., García, M. I. A., & Pérez, G. C. (2017). Evaluando la calidad de la Universidad–Particularmente su enseñanza. REDU: *Revista de Docencia Universitaria, 15*(2), 125-142. https://doi.org/10.4995/redu.2017.6371

Toland, M. D., & De Ayala, R. J. (2005). A multilevel factor analysis of students' evaluations of teaching. *Educational and Psychological Measurement, 65*(2), 272-296. https://doi.org/10.1177/0013164404268667

Turull, M., & Buxarrais, M. R. (2018). La evaluación de la docencia en las universidades públicas catalanas: Análisis comparativo de los diferentes manuales de evaluación. *Revista de Educación y Derecho, 17*, 1-30. https://doi.org/10.1344/re&d.v0i17.21842

Wolbring, T., & Riordan, P. (2016). How beauty works. Theoretical mechanisms and two empirical applications on students' evaluation of teaching. *Social Science Research, 57*, 253-272. https://doi.org/10.1016/j.ssresearch.2015.12.009