Recibido / Recebido: 22.01.2022 - Aceptado / Aceite: 03.06.2022 https://doi.org/10.21865/RIDEP65.4.06

Satisfacción con la Calidad Docente en el Ámbito Universitario: Potenciales Sesgos y Propuestas de Análisis para su Evaluación

Teaching Quality Satisfaction at University: Potential Biases and Proposal Analysis for its Evaluation

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

Resumen

La evaluación de la satisfacción estudiantil es un indicador fundamental para establecer la calidad en los distintos niveles del sistema universitario. Sin embargo, la satisfacción es un constructo complejo, dado que se mide para distintos niveles en una estructura jerárquica anidada, como son la titulación, la asignatura y el profesor que la imparte. Esta estructura multinivel habitualmente se ignora, por lo que es posible que se incurra en sesgos, al no diferenciar las distintas fuentes de variabilidad, que afectan a las puntuaciones. El presente estudio expone un modelo de análisis que permite detectar las fuentes de variabilidad para su control y una estimación más precisa. Se muestra cómo una vez corregidas las puntuaciones en satisfacción, es posible observar la estructura factorial del cuestionario, así como hacer una comparación más precisa en términos de satisfacción, entre las titulaciones y asignaturas.

Palabras clave: evaluación, satisfacción, calidad docente, sesgo, variabilidad

Abstract

The assessment of students' satisfaction is a key indicator for defining the quality of the university system across its various levels. However, satisfaction is a complex construct that is measured in a nested hierarchical structure, such as the degree, the course, and the lecturer who teaches it. This multilevel structure is usually ignored, which might produce biased results, by not differentiating among the distinct sources of variability that influence the scores. This study presents an analysis model that accounts for diverse sources of variability, resulting in more accurate estimates. We show that once the satisfaction scores have been corrected, the factor structure of the questionnaire is unveiled and the comparisons in terms of satisfaction between degrees and courses become more precise.

Keywords: assessment, satisfaction, teaching quality, bias, variability

Agradecimientos: Esta publicación es parte del proyecto de I+D+i PID2019-105177GB-C22 financiado por MCIN/ AEI/10.13039/501100011033/

[1] Magíster en Psicología. Máster en Metodología de las Ciencias del Comportamiento y de la Salud. Candidata a Doctora. Universidad Complutense de Madrid, España. Tel.: +57 3015500973. Correo: ladycala@ucm.es

[2] Psicólogo. Máster en Metodología de la Investigación en Ciencias del Comportamiento y de la Salud. Candidato a Doctor en Psicología. Facultad de Psicología, Universidad Complutense de Madrid, España. Dirección postal: Plaza del Ángel, 13. Piso 5 1 Izq, CP: 28012, Madrid, España. Tel.: +34 611198942. Correo: edrami01@ucm.es

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

ISSN: 1135-3848 print /2183-6051online

Introducción

La evaluación de la satisfacción de los estudiantes en los distintos aspectos de la docencia universitaria es un elemento fundamental en la planificación de las titulaciones y las propuestas de modificación de los planes de estudio (Glaría et al., 2016). Es por esta razón, que interesa conocer la satisfacción de los estudiantes respecto a las distintas titulaciones, asignaturas que forman estas y con los profesores que las imparten. Estas informaciones son de una importante trascendencia en cuanto a que permiten detectar desajustes para la toma de medidas adecuadas.

Ahora bien, para que una evaluación sea válida requiere de que se cumplan un conjunto de requisitos que básicamente se resumen en dos (ver Messick, 1989), la medida tiene que ser completa, es decir, debe muestrear todos los dominios de contenido que forman parte del constructo evaluado, y en segundo lugar debe evitarse la presencia de varianza irrelevante, o dicho de otra manera, debe evitarse que las puntuaciones incluyan como propios elementos que no forman parte del constructo que se intenta medir. Tanto la estructura factorial correcta como la presencia de varianza irrelevante pueden tener graves consecuencias y pueden pasar desapercibidas, si se aplican procedimientos de análisis inadecuados, como podría ser el uso de la puntuación total sin tener en cuenta la naturaleza multinivel de los datos (los datos obtenidos están anidados en titulaciones, asignaturas y profesores). Si se utiliza la medida de satisfacción "global" y no se controla el impacto de las distintas fuentes de varianza implicada, puede dar lugar a estimaciones sesgadas por incumplimiento de la necesaria propiedad de invarianza.

Estructura de los datos

En el área de psicología y educación, así como en otras áreas es común identificar que la estructura de los datos esté organizada jerárquicamente por su naturaleza. Esto es, los individuos están agrupados en unidades de un nivel superior, que a su vez también pueden estar agrupados en otras unidades. Estas estructuras se encuentran, por ejemplo, en los alumnos de las clases o en los pacientes de los hospitales. Sin embargo, en general el tratamiento que se realiza de este tipo de datos se ignora y se realiza el análisis partiendo de la hipótesis de independencia. Esto se debe también, a la ausencia de desarrollos metodológicos que incluyera estas jerarquías en los análisis, lo que generaba dos alternativas, como lo señalaban Bacci y Caviezel (2011): "(a) aplicar un modelo de regresión único a datos individuales, ignorando la presencia de grupos; (b) aplicar un conjunto de modelos de regresión específicos para cada grupo, reconociendo explícitamente los grupos, pero tratándolos como entidades completamente autónomas" (p. 2778). Por esta razón, el presente trabajo intenta mostrar las consecuencias de ignorar la presencia de estructuras anidadas e indaga sobre la forma adecuada de tratar estos datos.

La evaluación de la satisfacción de los estudiantes en el ámbito universitario es un claro ejemplo de estructuras anidadas. Este tipo de evaluaciones se realiza con el objetivo de comprobar la calidad de la educación y ser parte de un proceso de mejora continua (Bolivar, 2008; Stake et al., 2017). Los cuestionarios que se realizan sobre la actividad docente, conocidos también como *Student Evaluations of Teaching* (SET), se utilizan como una medida del desempeño docente en casi todas las instituciones de educación superior del mundo (Zabaleta, 2007). Los SET son cuestionarios diseñados para permitir que los estudiantes valoren la calidad de la enseñanza recibida (Simpson & Siguaw, 2000), mediante diferentes criterios que se enfocan, principalmente, en la forma en que es impartida la docencia de un curso específico. Estas evaluaciones son de carácter obligatorio en la mayoría de los países de la OCDE, en las instituciones de educación superior e incluso son reglamentados por la Ley (Oermann et al., 2018).

Dada la importancia que tienen estos cuestionarios y las decisiones que se toman con ellos, como la permanencia, promoción y estímulos para los docentes (Spooren et al., 2013), es importante hacer reflexiones sobre su validez, el análisis de sus datos y el alcance de sus interpretaciones. Algunos cuestionarios se construyeron sin tener una comprensión clara de lo que es la enseñanza efectiva, por lo que carecen de evidencia de validez de contenido, indicando

que no podrían medir lo que dicen medir (Onwuegbuzie et al., 2009), en este sentido se puede considerar que las respuestas de los estudiantes pueden ser una mezcla entre la concepción de aprendizaje de los estudiantes y sus creencias sobre las enseñanzas del profesor (Kember & Wong, 2000; Spooren, 2013).

Si bien existen un gran número de investigaciones sobre las variables que pueden afectar este fenómeno, también se evidencia que a pesar de las pautas para recopilar e interpretar los datos de SET, muchos usuarios de estas evaluaciones no tienen capacitación para el manejo de estos datos (Penny, 2003), o desconocen los modelos estadísticos que permitirían realizar diagnósticos (Franklin, 2001).

A pesar de que el uso, el análisis y la interpretación de estos datos tienen consecuencias para la carrera de los docentes y la mejora de la enseñanza, existe muy poca investigación sobre cómo afecta el hecho de que los encargados de analizar los datos tengan poco conocimiento estadístico y psicométrico o que realicen inferencias basados en datos descriptivos, por lo que siguen prefiriendo medidas agregadas y generales de la satisfacción de los estudiantes, desconociendo otras herramientas que pueden dar cuenta del fenómeno o de la calidad de los instrumentos que se utilizan para recopilar los datos (Spooren et al., 2013). Este aspecto apunta directamente a las evidencias de validez, dada la necesidad de reunir evidencias que permitan evaluar la solidez de las interpretaciones propuestas para el uso previsto de las pruebas, es decir, las consecuencias de su uso (Messick, 1998). Estas interpretaciones propuestas para un test son indicadores del constructo, el cual se considera como resultado de su variabilidad (Borsboom et al., 2004).

Las evidencias de validez se encuentran detalladas en *The Standards for Educational and Psychological Testing* una publicación conjunta de la *American Educational Research Association* (AERA), *American Psychological Association* (APA) y *National Council on Measurement in Education* (NCME), en estos se identifican cinco grandes fuentes de evidencia, la relativa al contenido, la relativa a los procesos de respuesta, la estructural, relación con otras variables y consecuencial (AERA, APA, & NCME, 2018). El presente artículo se centra en la fuente de evidencia estructural, ya que se muestra como la estructura factorial y como consecuencia la fiabilidad de la medida puede ser incorrecta si no se controlan adecuadamente las fuentes de varianza que afectan a las puntuaciones. Puesto que las evidencias de validez están interconectadas, una estructura factorial incorrecta puede tener consecuencias sobre la estimación correcta del nivel de aptitud de los evaluados lo que afectaría tanto a la red nomológica (relación con otras variables) como a las consecuencias (posibles sesgos), por estas razones es especialmente relevante establecer cuál es la estructura correcta del instrumento.

Modelos clásicos de análisis

La información que se extrae de los datos, los análisis estadísticos realizados y los reportes a los docentes de los SET se suelen realizar haciendo uso de la Teoría Clásica de los Test (TCT), basados principalmente en el puntaje promedio de la clase (Toland & de Ayala, 2005). La aplicación de la TCT funciona adecuadamente con la mayoría de los datos empíricos, situación que puede ser tanto una fortaleza como una debilidad; dado que una fortaleza es su simplicidad y a la vez una limitación al no poder profundizar con detalle en las conclusiones que permite extraer de los análisis (Muñiz, 2010; Abal et al., 2014). En este sentido, las investigaciones serían más precisas si se incorporará, por ejemplo, la variación grupal dentro del análisis, pues al tener una sola media se suprime esta variabilidad (March & Hattie, 2002; Toland & de Ayala, 2005).

Para estos fines se han identificado dos formas que son las más utilizadas, pero también las más criticadas para analizar datos anidados, jerárquicos o multinivel, utilizando un único nivel. La primera implica ignorar la estructura de asociación de los datos, asumiendo que son independientes, lo que constituye una mala especificación que atenta directamente contra la validez de las inferencias; la segunda agrupa los datos en un nivel superior, en los cuales se pierde información dentro de cada grupo, lo que genera que la asociación entre los grupos sea mayor que en la de los datos desagregados, lo cual introduce sesgo en la estimación de los errores estándar (Goldstein, 2011).

A pesar de las limitaciones conocidas, la TCT sigue vigente y en algunos casos se utiliza de forma complementaria con la Teoría de Respuesta al Ítem (TRI), para hacer un análisis más exhaustivo de la calidad del test o un análisis más preciso de las propiedades psicométricas de los ítems que conforman el instrumento (Barbero et al., 2001). Sin embargo, los modelos de rasgo latente de un nivel requieren del supuesto de unidimensionalidad e independencia local (Abal et al., 2010), los cuales se incumplen en mayor o menor medida en los datos anidados, puesto que se confunden las diferentes fuentes de variabilidad que provienen de los distintos niveles de análisis, lo que puede conducir a sobreestimaciones en los poderes discriminantes de los ítems y de su dificultad, así como a estimaciones incorrectas de la fiabilidad del instrumento, con lo que la estimación de la variable latente puede estar fuertemente comprometida o sesgada. Así, por ejemplo, en el caso de un profesor que 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, obviamente no deberían ser comparados, salvo que previamente se controlen estas fuentes de variabilidad.

Modelos para el análisis de datos anidados

Los test de satisfacción de los estudiantes, en relación con la calidad docente, se aplican a grupos de estudiantes diferentes para cada profesor. Pero, además, se comparan las puntuaciones por materia y titulación, dando por hecho que la herramienta es invariante y que por lo tanto estas variables no contribuyen con varianza irrelevante a la puntuación total de satisfacción.

Previo a cualquier análisis en el que se comparen las puntuaciones en satisfacción es preciso conocer la contribución que tienen sobre la varianza total los distintos niveles de análisis. Así, por ejemplo, si la variabilidad de los estudiantes es muy elevada, una vez controladas las demás fuentes de varianza, el informe de los profesores basado en la media de la satisfacción de los estudiantes podría estar seriamente sesgado, dependiendo de las características de los estudiantes que se hayan encuestado. La solución ideal sería que toda o la mayor parte de la variabilidad fuese debida al docente. Este asunto es ampliamente tratado por Zumbo et al. (2010), en relación con el problema de recolectar datos de naturaleza multinivel, en un nivel de análisis inferior, como el de los estudiantes, para hacer inferencias sobre una unidad de análisis de nivel superior, en este caso los docentes. Por esta razón, es crucial que en los estudios de evaluación y como requisito previo a todo procedimiento se identifiquen las fuentes de variabilidad que afectan a las puntuaciones, antes de proceder con cualquier análisis, dado que herramientas valiosas, como la de valoración de la calidad de la docencia, terminan siendo cuestionadas, en términos de validez, por la mezcla de las diferentes fuentes de variabilidad.

A pesar de esto, cuando las fuentes de heterogeneidad se conocen, es posible hacer uso de modelos multigrupos, como el de funcionamiento diferencial del ítem que permiten comparar los parámetros de los grupos, a partir de la varianza irrelevante del constructo (Raykov et al., 2013). Sin embargo, los grupos pueden estar conformados por variables no observadas, que solo pueden inferirse a partir de los datos, generando posibles sub-agrupaciones. En estos casos, los modelos que asumen que la fuente de heterogeneidad no es observable han sido diseñados para identificar conglomerados de sujetos que tienen patrones de respuesta similares en un test (Goodman, 2002; Vermut & Magidson, 2006), 2015).

Estos modelos se constituyen como soluciones adecuadas para datos de naturaleza multinivel y buscan identificar estos grupos o mezclas que expliquen la variabilidad de los datos, dado que en algunos estudios de este tipo de análisis en la medición de la satisfacción docente, como el de Bacci y Caviezel (2011), se ratifica que la estructura de los datos tiene un efecto significativo en la satisfacción de los estudiantes: esto justifica tanto, el uso de un modelo multinivel como la atención que se debe prestar a las comparaciones entre la enseñanza, sobre la base de los residuos de tercer nivel.

En este sentido, la mayoría de estos modelos pueden verse como extensiones del *modelo básico de clases latentes*, el cual, para estimar la

probabilidad de observar un determinado conjunto de respuestas, requiere multiplicar entre sí las probabilidades de respuesta correcta en cada clase latente y luego sumar estos productos (Goodman, 2002; Vermut & Magidson, 2006). Este análisis permite agrupar los patrones de respuesta y por tanto, a los sujetos que tienen esos patrones en un número reducido de clases latentes, de tal forma que los patrones de respuesta de los sujetos de una misma clase latente son más similares entre sí, que con respecto a los patrones de quienes pertenezcan a otra clase latente. De forma complementaria, los Modelos de Clases Latentes Multinivel añaden características al nivel inferior de análisis (Vermunt, 2003).

Por otra parte, desde la TRI, la combinación de un modelo multinivel con una o más variables latentes modeladas mediante un modelo TRI es llamado *modelo multinivel TRI - MIRT* (Fox & Glas, 2001) que involucra la estimación de clases individuales y está compuesto por dos componentes, un modelo TRI de dos parámetros y un modelo que describe la relación entre la variable latente y las variables explicativas del primer y segundo nivel. Por su parte, Kamata (2001) plantea un modelo Rasch multinivel para datos binarios, en el que aparece un modelo de tres niveles, en el que el nivel dos (2) es el nivel de persona y el nivel tres (3) es de predictores, es decir el nivel de grupo. En esta misma línea, aparecen *los modelos de mixturas de la teoría de respuesta al ítem- MMixIRT* (Cho & Cohen, 2010) que permiten estimar la pertenencia de los sujetos a alguna de las clases latentes y además estimar su nivel en el constructo medido, en estos modelos las probabilidades de observar una respuesta en una clase latente, se realiza a través de alguno de los modelos tradicionales de TRI (Formann & Kohlmann, 2002). Esto quiere decir, que proporcionan información tanto a nivel individual (p. ej., examen o estudiante) como a nivel grupal (p. ej., profesor o escuela). En estos modelos, a diferencia de las clases latentes, se señala que existen una o varias variables latentes que explican las diferencias entre los ítems en cada una de las clases y por lo tanto diferencias entre los individuos de una misma clase latente (Sterba, 2013).

Se han explorado una gran diversidad de aplicaciones de estos modelos, como la generación de evidencias de validez referidas a la invarianza de los parámetros de los ítems (Baghaei & Carstensen, 2013) y a la dimensionalidad de los test (Hong & Min, 2007), la identificación de posibles fuentes del funcionamiento diferencial de los ítems (Choi et al., 2014), la identificación de personas que utilizan estrategias distintas de razonamiento (De Boeck & Rijmen, 2003); así como la detección de examinados con bajos niveles de motivación para contestar test de bajas consecuencias (Mittelhaëuser et al., 2013), entre otros.

Si bien los modelos *MMixIRT* suponen un contribución para el tratamiento adecuado cuando se evidencia la existencia de mixturas o poblaciones mezcladas su uso requiere del cumplimiento de supuestos muy restrictivos que limitan su aplicabilidad por ejemplo, el Modelo de Mixtura de Rasch ignora la estructura multinivel básica que está presente más allá del nivel del estudiante en gran parte de los datos de las pruebas educativas; respecto a los modelos multinivel IRT también se ha demostrado que no proporcionan información sobre los miembros de cada grupo más allá de los predictores incluidos en el modelo (Cho & Cohen, 2010). Por su parte, los parámetros de MMixIRT suelen estimarse con algoritmos como la cadena de Markov Monte Carlo (MCMC), lo cual, generalmente requiere un tiempo de cómputo sustancial para obtener resultados utilizables (von Davier & Yamamoto, 2007).

El presente estudio

Como se ha revisado brevemente, existen diferentes estrategias de análisis, sin embargo, la mayoría de estas no están implementadas en paquetes estadísticos y generalmente son de una gran complejidad matemática y de cómputo; además se basan en supuestos que difícilmente se dan en situaciones aplicadas, como por ejemplo, para una estimación precisa de la labor de un docente particular se requieren tamaños muestrales representativos de estudiantes que valoren al docente, del mismo modo puede haber otras fuentes de variabilidad como la titulación, la asignatura, el tipo de asignatura, entre otras, que afectan a la satisfacción.

Para obtener una medida de satisfacción válida en el presente estudio se plantea una

solución sencilla, que se basa en primer lugar en determinar las fuentes de varianza que explican la puntuación total para, a partir de aquí, centrar la medida en la variable de interés para hacer una estimación más precisa controlando las otras fuentes de variabilidad.

Método

Datos

Los datos utilizados en el presente estudio se obtienen de una base de datos facilitada por una universidad española ubicada en Madrid, los datos se encontraban anonimizados y la universidad solicitó mantener el anonimato para las publicaciones realizadas. Se utilizaron 16950 registros de estudiantes pertenecientes a ocho centros de estudio en las distintas titulaciones y asignaturas, se eliminaron los valores perdidos y solo se tuvo en cuenta un cuestionario por estudiante. El 63.8% de la muestra fueron mujeres y la media de edad de los estudiantes fue de 21.18 años.

Instrumento

El cuestionario de satisfacción de la actividad docente denominado: Cuestionario de estudiantes: opinión sobre la actividad docente (ver Anexo 1) fue desarrollado y aplicado por una universidad española como medida global de la satisfacción. El cuestionario está constituido por siete (7) afirmaciones de respuesta graduada tipo Likert y que evaluaron el nivel de acuerdo o desacuerdo frente a cada una de estas, en una escala establecida desde *Totalmente en desacuerdo* hasta *Totalmente de acuerdo*, e incluyendo un No Procede. La calificación es ordinal de 1 a 5, en donde la máxima puntuación (5) refleja mayor satisfacción con el criterio de desempeño docente. El uso que se ha dado al instrumento, en los años en los que se ha aplicado, es el típico de los instrumentos desarrollados desde la Teoría Clásica de los Test, obteniéndose una puntuación suma de la satisfacción de los distintos ítems e indicadores, lo que se apoya en evidencias de unidimensionalidad y una buena fiabilidad del instrumento mediante el coeficiente alfa de Cronbach superior a .90, según la información reportada por la universidad, dado que no se cuentan con estudios psicométricos publicados.

Análisis de Datos

Se usó un procedimiento stepwise para el filtrado de las variables que introducen sesgo en la prueba. Esta vía usó: (1) la identificación de la varianza explicada de cada variable relevante del cuestionario, (2) alineación de la media de la variable con mayor varianza explicada, (3) estimación del modelo para los datos corregidos y (4) comparación entre variables.

Identificación de la variabilidad de los componentes

Para la identificación de la varianza explicada de los componentes se consideraron las diferentes variables de la base de datos usada, como: profesores, Id de alumno, sexo y edad del alumno y datos con respecto al centro, plan de estudio y tipo de asignaturas, puntuación de los ítems y totales. Debido a la gran variabilidad de observaciones por profesor identificada en la matriz de datos, solo se tuvieron en cuenta los docentes que fueron evaluados por al menos diez estudiantes.

Luego de comprobar que las muestras estaban constituidas por observaciones aleatorias e independientes y que se cumplían los supuestos de normalidad, se realizó una estimación de un modelo de efectos mixtos con el objetivo de identificar las características latentes de los datos, bajo el criterio de conservación de la máxima varianza

Esta estimación se hizo variable a variable para rastrear componentes con mayor varianza. El paquete Lme4 (Bates et al., 2015) del software R core Team (2020) se usó para la estimación del modelo sugerido. En las líneas de código mostradas para ilustrar dos ejemplos, se observan las estimaciones de dos componentes:

```
### Varianza explicada por los profesores ###
 > mv_prof <- lmer( TOTAL ~ (1|Profesor), data = datos)
 > coef(mv_prof)
 > summary(mv_prof )

### Varianza explicada por la edad de los alumnos ###
 > mv_edad <- lmer( TOTAL ~ (1|Edad), data = datos)
 > coef(mv_edad)
 > summary(mv_edad)
```

A manera ilustrativa se muestra una parte del código usado (el código completo puede consultarse en https://memopro.weebly.com/ en la pestaña repositorio BBDD) con la que se pueden obtener las estimaciones para los efectos aleatorios y por medio de estos determinar los pesos de las variables. Con esta sintaxis se obtuvo la varianza

explicada por todos distintos componentes que se comentan en el apartado de resultados.

Alineación de la media de la variable con mayor varianza explicada

Una vez fijada la variabilidad de los componentes se determinaron las características de la muestra que podrían contribuir a dicho fenómeno. En la matriz de datos se observó que algunos profesores eran evaluados tan solo por uno o dos alumnos, mientras otros docentes tenían mediciones correspondientes a más de treinta examinados. Esto se controló en primer lugar filtrando los docentes que fueron examinados por al menos diez alumnos.

Debido a estas diferencias, fue necesaria una equiparación que permitiera obtener un puntaje comparable en términos de muestra, partiendo del supuesto de que mediciones de grupos desiguales de profesores podrían interferir en las propiedades psicométricas de la prueba. Para este fin se aplicó una corrección por la media, que consiste en hacer que las puntuaciones de cada uno de los profesores se equiparen en la media de los totales de la prueba, este procedimiento se realizó siguiendo el criterio de Holland y Dorans (2006) y Fuentealba (2010). El resultado del proceso anterior son datos continuos tanto para los ítems, como para los totales de la prueba, por lo que fue necesario aplicar una trasformación a enteros para los análisis posteriores. La funcionalidad del procedimiento se comprobó al repetir el análisis de regresión mixta con los datos corregidos.

Estimación del modelo para los datos corregidos

En esta fase del procedimiento se estimó un modelo factorial confirmatorio, dado su carácter de modelo causal de medida, que permite confirmar las hipótesis sobre las variables latentes que agrupan los ítems (Pérez, 2020). Previo a los análisis se examinaron los datos para verificar la existencia de valores faltantes. La elección del modelo tuvo en cuenta una evaluación de normalidad multivariante con el Test de Mardia y una elección del modelo final. Las estimaciones se realizaron con el paquete Psych (Revelle, 2021) y el paquete Lavaan (Rosseel, 2012) del software R core Team (2020).

Una vez corregidos los datos, se realizó un AFC con estimación robusta de mínimos cuadrados ponderados (DWLS) que utiliza correlaciones policóricas siguiendo el criterio de Lloret et al. (2014). Se estimó la prueba de chi-cuadrado y los siguientes índices de bondad de ajuste para el modelo elegido: (a) la raíz del error cuadrático medio de aproximación (RMSEA), (b) el índice de ajuste comparativo (CFI), y (c) la raíz cuadrada media residual estandarizada (SRMR), así mismo se incluyó el índice de ajuste relativo BIC.

Según Kelloway (1998) y Hu y Bentler (1999), los valores de RMSEA de .10 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, los valores por encima de .90 indican que el modelo se ajusta bien, y los valores por encima de .95 representan un ajuste muy bueno a los datos. Se asume de igual forma que el índice BIC más pequeño refleja mejor ajuste relativo.

Comparación entre variables

Finalmente, se estimó un modelo con medidas repetidas para hacer comparaciones de medias entre las puntaciones factoriales y los tres tipos de asignatura de la base de datos: asignaturas de formación básica, asignaturas obligatorias y optativas. Se usó un modelo lineal general de medidas repetidas, en el que las medidas totales de cada uno de los factores de la prueba de satisfacción docente se estimaron como variables intra sujeto y la variable tipo de asignatura se estimó como variable inter sujeto. Se realizaron pruebas de homogeneidad para determinar esfericidad de las matrices y todos los contrastes fueron simples. Tanto las medias para satisfacción, como para las medias de la interacción entre la satisfacción docente y las variables inter sujeto, fueron ajustadas con el intervalo de confianza de Bonferroni.

Resultados

En la Tabla 1 se puede observar que la varianza total más alta es la correspondiente a la adición entre profesores y alumnos con un puntaje de 72.52. Por lo tanto, la varianza debida a los profesores seria de 40.19% y 59.81 % debida a los alumnos. La asignatura es el componente adicional que más explica la varianza en el modelo especificado.

Tabla 1. Varianza de los componentes de la prueba de satisfacción docente

Nombre del componente	Intersección	Residuos	Varianza total	% de varianza explicada
Profesor	29.15	43.15	72.52	40.19 %
Asignatura	27.09	44.51	71.60	37 .83 %
Año	0.444	66.86	66.90	0.66 %
Sexo Alumno	0.141	66.83	66.97	2.10 %
Edad	1.055	65.94	67.00	1.57 %
Centro	0.704	66.38	67.08	1.04 %
Tipo de Asignatura	3.195	65.86	69.08	4.62 %

Tabla 2. Varianza de los componentes de la prueba de satisfacción docente con datos corregidos

Nombre del componente	Intersección	Residuos	Varianza total	% de varianza explicada
Profesor	0.00	38.06	38.06	0.00 %
Asignatura	0.02	37.85	37.87	0.05 %
Sexo Alumno	0.00	38.06	38.06	0.00 %
Edad	0.00	38.06	38.06	0.00 %
Centro	0.00	38.06	38.08	0.00 %
Tipo de Asignatura	0.18	37.99	38.17	0.47 %

Tabla 3. Índices de bondad ajuste para el modelo de un factor original (sin corrección) y los modelos corregidos de uno y tres factores

Índices de bondad de ajuste	Modelo de un factor Datos no corregidos	Modelo de un factor Datos corregidos	Modelo de tres factores Datos corregidos
RMSEA [90% CI]	.136 [.133–.139]	.190 [.187–.194]	.056 [.052–.060]
CFI	.995	.958	.997
TLI	.993	.936	.995
SRMR	.023	.046	.012
χ^2 (gl)	4395.26*(14)	8611.21*(14)	590.12*(11)
BIC	6323.23	5307.28	21.41

Nota. * p<.001 RMSEA=Error de aproximación de la media cuadrática (*Root Mean Square Error of Approximation*); CFI=Índice de ajuste comparativo (*Comparative Fit Index*); TLI=Índice de Tucker Lewis (*Tucker Lewis Index*); SRMR=Residuo estandarizado de la media cuadrática (*Standardized Root Mean Square Residual*); BIC=Criterio de información bayesiana (*Bayesian Information Criterion*).

Respecto al resto de componentes, en la Tabla 2 se observa que ninguno tiene una mayor incidencia sobre la variable dependiente. Por lo tanto, teniendo en cuenta que la variabilidad introducida por los profesores era mayor al 40% se aplicó una alineación lineal por la media. Para este fin, todos los profesores se igualaron en media ítem a ítem. El modelo de regresión mixta, posterior a la corrección y redondeo de los datos, mostró que la varianza explicada de todos los componentes fue nula, por lo tanto, el procedimiento máximo de la conservación de varianza correspondía a los alumnos.

La evaluación de los datos corregidos con el Test de Mardia (Joanes & Gill, 1998; Mardia, 1970) mostró unos puntajes de 3.22 para el contraste b1 y 88.72 para el contraste b2, en los dos se observaron valores estadísticamente significativos a p<.001, con lo que se rechaza la normalidad multivariante en los datos corregidos. En este caso, se recomienda Omega como estimador de la fiabilidad (Trizano-Hermosilla & Alvarado, 2016; Zinbarg et al., 2006) estimándose un valor de .87.

Se hizo una comparación entre los modelos factoriales de mejor ajuste en los datos no corregidos y posteriormente, se compararon los mismos modelos con datos corregidos. En la Tabla 3 se observa que la estructura trifactorial del modelo alineado es el que presenta un menor valor de BIC y es el que presenta los mejores índices en bondad de ajuste, siendo el único modelo en el que todos los índices están dentro de los valores recomendados, salvo chi cuadrado, cuyo desajuste se justifica por el gran tamaño muestral utilizado en la investigación.

En la Figura 1 se representan los tres modelos, observándose una sobreestimación de los pesos en el modelo original con datos no alineados. El alineamiento planteado permite observar cómo el mejor modelo es el de tres factores que se puede explicar teóricamente cómo: El factor 1 relacionado con la satisfacción frente al acompañamiento directo del docente, el factor 2 referido a la satisfacción con la organización y estructuración de las clases y el factor 3 referido a la satisfacción con la claridad de las explicaciones del docente. Aunque las correlaciones entre los

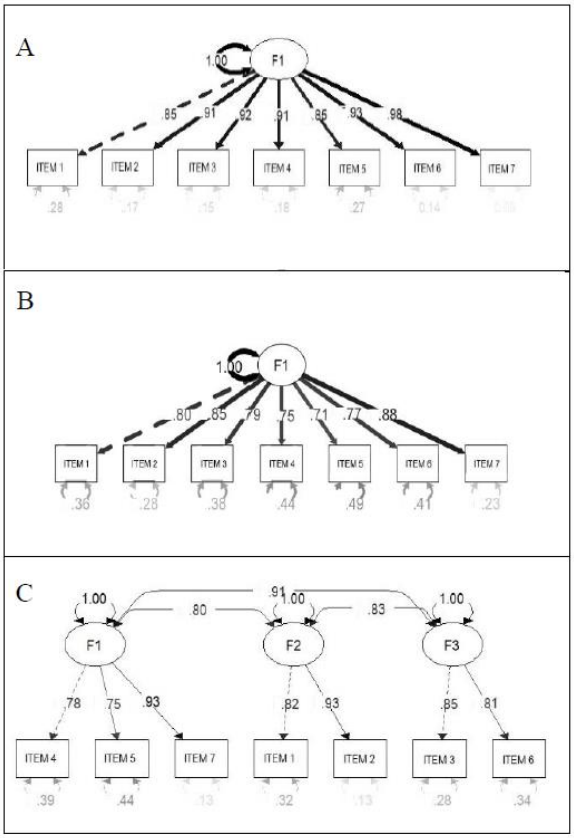

Figura 1. Pesos estandarizados del modelo de un factor original sin corrección (A), modelo de un factor con datos corregidos (B) y del modelo de tres factores con datos corregidos (C)

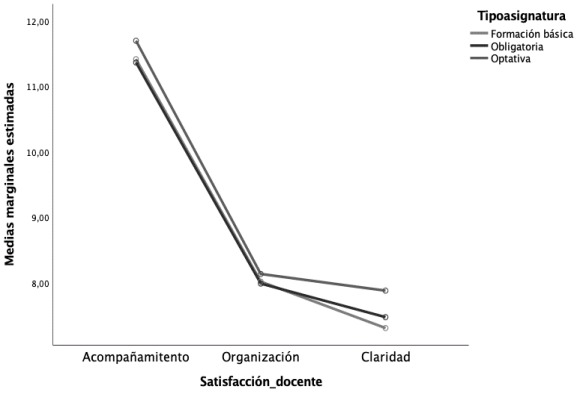

Figura 2. Comparaciones de medias entre los tres factores del cuestionario de satisfacción docente por tipo de asignatura

factores son altas, no se consideró que el modelo presentara redundancia factorial al revisar los residuos entre las matrices la matriz de correlación reproducida y la matriz de correlación. Según Mahmud et al. (2018) y Broen et al. (2015) cuando un modelo ajusta bien tendrá menos del 50% de los residuos no redundantes con valores absolutos superiores a .05.

Finalmente, en las comparaciones de medias de la variable seleccionada evidencio que el nivel crítico asociado al estadístico *W* de Mauchly fue de $p<.001$, por lo que no se asumió esfericidad en las matrices de varianza-covarianza, por ello, se eligieron estadísticos multivariados para realizar los contrastes de hipostasis de igualdad de medias. La Traza de Pillai, la Lambda de Wilks, la T^2 de Hotelling y la Raíz mayor de Roy recomendadas por Tabachinik y Fidel, (2001) muestran un nivel crítico de $p=.001$ con un coeficiente eta cuadrado parcial de $\hat{\eta}^2=.005$ en la interacción del factor *satisfacción docente* y *tipo de asignatura*, por lo que se asume diferencia de medias en los dos factores inter-sujetos estimados en la ANOVA de medidas repetidas.

En términos comparativos de la satisfacción docente, la Figura 2, muestra que las asignaturas optativas tienen puntajes más elevados en los tres factores, con respecto a las medias de asignaturas obligatorias y de formación básica.

En la Figura 2, se observa que el factor 3 es el que mejor indaga por la actividad del profesor, dado que tiene que ver con el acompañamiento directo del docente y además es el que contiene la mayor cantidad de ítems. Se identifica también, que existen diferencias estadísticamente significativas entre los tres tipos de asignatura (obligatoria, formación básica y optativa), con una aparente confusión entre el interés del estudiante por la asignatura y la evaluación de la claridad del docente. El factor 1 está más relacionado con el cumplimiento de la normativa y el factor 2 que es más de tipo organizativo muestran, a su vez, que, en general, los profesores se ajustan al temario y normas de la institución. En cualquier caso, el sesgo positivo hacia las asignaturas optativas es también evidente, así como que los profesores de formación básica tienen las apreciaciones más bajas.

Discusión

El propósito de la investigación era proponer una alternativa de análisis para los cuestionarios de satisfacción de los estudiantes respecto a la docencia universitaria que respetase la naturaleza jerárquica de los datos. De la aplicación de este procedimiento se deducen implicaciones relevantes para el tratamiento adecuado de estos datos que respecta su naturaleza multinivel.

En primer lugar, se observó que la varianza explicada por el centro, asociado a la titulación que se imparte, era despreciable, lo mismo concluimos sobre la asignatura o materia impartida. Estos resultados son relevantes, en cuanto a que permiten prescindir de estas dos fuentes de variabilidad considerando que el instrumento de evaluación fue equilibrado y probablemente invariante respecto a estas dos variables. En cualquier caso, si la institución deseara realizar una comparación respecto a la satisfacción por centro/titulación o asignatura, debería como marcan los estándares realizar previamente un análisis de invarianza (Van De Schoot et al., 2015), para que las diferencias encontradas se deban a la variable latente medida y no a condiciones particulares de los grupos que se comparan (Byrne & van de Vijver, 2010).

En segundo lugar, se identificó que hay dos fuentes de variabilidad que explican la mayor parte de la varianza total de las puntuaciones en satisfacción: estudiantes y docentes. Puesto que hay dos fuentes independientes que conforman la puntuación total del instrumento, es preciso fijar una si se quiere hacer una estimación adecuada de la otra variable, siendo inaceptable el uso de la puntuación total. Dado que la variabilidad de algunos de los componentes, probablemente beneficien a algunos examinados y perjudiquen a otros con estimaciones incorrectas tal como lo señalan Bacci y Caviezel (2011).

Al evaluar la satisfacción, una vez fijada la varianza del docente, se observa que la estructura factorial de cuestionario no era unidimensional, sino que se revela una estructura tridimensional, lo que hace posible obtener una puntuación diferenciada para los tres dominios de contenido que se observan: acompañamiento directo del docente, la organización y estructuración de las clases y la claridad de las explicaciones del docente. Los tres factores observados guardan una estrecha lógica con la prueba porque diferencian constructos de interés que, aunque pueden estar correlacionados obedecen a fenómenos diferentes, por lo tanto, es importante medirlo como constructos diferenciados porque ayudaría a retroalimentar mejor los resultados al docente evaluado con la escala.

El que la estructura factorial cambie al controlar las distintas fuentes de varianza es de especial interés, ya que este aspecto está directamente relacionado con la validez de la medida (Messick, 1998) y con las interpretaciones que se pueden inferir del constructo a través del test, de acuerdo con lo planteado por Borsboom et al. (2004).

Si en lugar de estar interesados en las diferencias en las medidas de satisfacción que estarán relacionadas con características psicológicas o motivacionales de los estudiantes, el interés se centra en los profesores, es necesario centrar la puntuación controlando la variabilidad debida a los estudiantes, puesto que en caso de no hacerlo estaría confundiendo calidad docente con las características concretas del grupo de alumnos evaluadores (Ver Tabla 2). En esta ocasión, el énfasis se ha centrado en la descontaminación de los datos para mejorar el escalamiento de los profesores, sin desconocer la existencia de otros modelos diseñados para este tipo de datos (Goodman, 2002; Vermut & Magidson, 2006; Cho & Cohen, 2010; Fox & Glas, 2001;Vermunt, 2003), lo cual confirma los resultados obtenidos por Bacci y Caviezel (2011) respecto a la necesidad de tratar estos datos como una estructura multinivel y de proporcionar evidencias de la dimensionalidad de los test, tal como lo señalan Hong y Min (2007).

De acuerdo con lo anterior, el procedimiento que se sugiere permite identificar en primer lugar si el foco de interés será la calidad docente o la satisfacción de los estudiantes y una vez realizada la limpieza de los datos e identificado el centro del análisis, se vuelve viable realizar procedimientos que tienen gran potencia para el análisis multinivel, como los presentados en la introducción. No obstante, como una primera aproximación al complejo problema de la estimación de la satisfacción hacia la actividad del docente, esta primera aproximación puede ser fácilmente aplicable para obtener estimaciones más precisas y válidas, ya que aporta avances para la resolución del problema psicométrico de la evaluación docente. Por tanto, no está desprovisto de limitaciones, dado que solo se observa lo que ocurre con los datos respecto a su estructura factorial, luego de realizar este procedimiento de limpieza, que permite separar las fuentes de varianza, sin embargo, se considera necesario continuar investigando, desde el punto de vista

metodológico, en la aplicación correcta de modelos de análisis que respeten la naturaleza multinivel propia de los datos obtenidos en las evaluaciones de la satisfacción de la actividad docente. Por ejemplo, identificar qué ocurre con la estimación de los parámetros de los ítems o el nivel de información que se le puede reportar al docente, a partir de la información que pueda brindar cada cuestionario en particular, si bien este cuestionario se centra en la estimación de la satisfacción como algo global, existen otros cuestionarios que pueden tener una aproximación multidimensional, lo cuál puede llevar a diferentes reflexiones, según la precisión con el que se utilice el modelo.

En ese mismo sentido, futuras líneas de trabajo pueden estar orientadas a comparar los procedimientos de análisis de datos multinivel, identificando si es posible implementar métodos más eficientes que incluyan este control de sesgo, o una comparación de la información proporcionada de los diferentes elementos de calidad del test realizando el procedimiento que se propone en este documento y sin dicho procedimiento con el fin de comparar las bondades del mismo. De igual forma, también es importante revisar la incidencia del número de ítems del cuestionario, ya que en instrumentos más largos es posible hacer un mejor control de las distintas fuentes de varianza.

Una reflexión importante que trasciende los métodos consiste en comprender la naturaleza del fenómeno y la información que se recolecta con estos cuestionarios, que puede variar de una institución a otra. Ese análisis puede ayudar a la comprensión sobre el método más pertinente a usar y que proporcione mejor información sobre la calidad de los instrumentos, en ese sentido es importante promover la formación en las personas que trabajan en el día a día con estos cuestionarios o bases de datos sobre estas sencillas soluciones metodológicas y que pueden favorecer la toma de decisiones, no solo frente al instrumento sino frente a la información que se brinda a los docentes o a las instituciones acerca de la valoración que realizan los estudiantes sobre su satisfacción con la actividad docente.

Referencias

Abal, F. J. P., Lozzia, G. S., Aguerri, M. E., Galibert, M. S., & Attorresi, H. F. (2010). La escasa aplicación de la teoría de respuesta al ítem en tests de ejecución típica. *Revista Colombiana de Psicología*, *19*(1), 111-122. http://www.redalyc.org/articulo.oa?id=80415077010

Abal, F. J. P., Auné, S. E., & Attorresi, H. F. (2014). Comparación del Modelo de Respuesta Graduada y la Teoría Clásica de Tests en una Escala de Confianza para la Matemática. Universidad Santo Tomás. Escuela de Psicología; *Summa Psicológica UST, 11*(2), 12-2014; 101-113 https://doi.org/10.18774/448x.2014.11.158

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). Estándares para pruebas educativas y psicológicas (M. Lieve, Trans.). *American Educational Research Association*. (Original work published 2014) https://doi.org/10.2307/j.ctvr43hg2

Barbero, M. I., Prieto, P., Suárez, J. C., & San Luis, C. (2001). Relaciones empíricas entre los estadísticos de la teoría clásica de los tests y los de la teoría de respuesta a los ítems. *Psicothema [en linea]*. *13*(2), 324-329. https://www.redalyc.org/articulo.oa?id=72721323

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bacci, S., & 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

Baghaei, P., & Carstensen, C. (2013). Fitting the Mixed Rasch Model to a reading comprehension test: Identifying reader types. *Practical Assessment, Research & Evaluation*, *18*(5). https://doi.org/10.7275/n191-pt86

Bolívar, A. (2008). Evaluación de la práctica docente. Una revisión desde España. RIEE.

Revista Iberoamericana de Evaluación Educativa. 1(2), 56-74. http://hdl.handle.net/10486/661519 https://revistas.uam.es/riee/article/view/4666

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

Broen, M. P., Moonen, A. J., Kuijf, M. L., Dujardin, K., Marsh, L., Richard, I. H., ... & Leentjens, A. F. (2015). Factor analysis of the Hamilton Depression Rating Scale in Parkinson's disease. *Parkinsonism & Related Disorders*, 21(2), 142-146. http://dx.doi.org/10.1016/j.parkreldis.2014.11.016

Byrne, B. M., & van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of none-quivalence. *International Journal of Testing*, *10*(2), 107-132. https://doi.org/10.1080/15305051003637306

Cho, S. J., & Cohen, A. S. (2010). A multilevel mixture IRT model with an application to DIF. *Journal of Educational and Behavioral Statistics*, *35*(3), 336-370. https://doi.org/10.3102/1076998609353111

Choi, Y., Alexeev, N., & Cohen, A. (2014). DIF Analysis using a Mixture 3PL Model with a covariate on the TIMSS 2007 Mathematics Test. *KAERA Research Forum*, *1*(1), 4-14. https://doi.org/10.1080/15305058.2015.1007241

Davier, M. V., & Yamamoto, K. (2007). Mixture-distribution and HYBRID Rasch models. En Multivariate and mixture distribution Rasch models (pp. 99-115). Springer.

De Boeck, P., & Rijmen, F. (2003). A latent class model for individual differences in the interpretation of conditionals. *Psychological Research*, 67, 219-231. https://doi.org/10.1007/s00426-002-0092-7

Formann, A., & Kohlmann, T. (2002). Three Parameter Linear Logistic Latent Class Analysis. En J. Hagenaars & A. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 183-210). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531

Fuentealba, R. G. (2010). Equiparación, alineamiento y predicción de puntuaciones en medición educativa. *Revista Iberoamericana de Evaluación Educativa*, *3*(2), 103-126. https://revistas.uam.es/riee/article/view/4493

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

Fox, J. P., & Glas, C. A. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, 66(2), 271-288. https://doi.org/10.1007/BF02294839

Glaría L., Rocío, & Carmona SM., Lorena, & Pérez V., Chisthian, & Parra P., Paula (2016). Estructura factorial y consistencia interna de la Escala de Evaluación del Currículum de Programas Universitarios en estudiantes de fonoaudiología de Chile. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica, 1*(41),80-89. https://www.redalyc.org/articulo.oa?id=459646901008

Goldstein, H. (2011). Multilevel statistical models (Vol. 922). John Wiley & Sons.

Goodman, L. (2002). Latent Class Analysis: The Empirical Study of Latent Types, Latent Variables, and Latent Structures. In J. Hagenaars & A. McCutcheon (Eds.), *Applied Latent Class Analysis* (pp. 3-55). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.002

Holland, P. W. & Dorans, N. J. (2006). Linking and equating. En R. L. Brennan (Ed.) *Educational Measurement* (4th Ed). Praeger Publishers. https://doi.org/10.1111/j.1745-3984.2010.00112.x

Hong, S., & Min, S. (2007). Mixed Rasch Modeling of the Self-Rating Depression Scale: Incorporating latent class and Rasch Rating Scale Models. *Educational and Psychological Measurement*, *67*(2), 280-299. https://doi.org/10.1177/0013164406292072

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A*

Multidisciplinary Journal, 6, 1-55. https://doi.org/10.1080/10705519909540118

Joanes, D. N., & Gill, C. A. (1998). Comparing measures of sample skewness and kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician), 47*(1), 183-189. http://www.jstor.org/stable/2988433

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, 38(1), 79-93. http://www.jstor.org/stable/1435439

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

Kember, D., & Wong, A. (2000). Implications for evaluation from a study of students' perceptions of good and poor teaching. *Higher Education*, *40*(1), 69-97. http://www.jstor.org/stable/3447952

Loret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica, revisada y actualizada. *Anales de Psicología / Annals of Psychology*, *30*(3), 1151-1169. https://dx.doi.org/10.6018/analesps.30.3.199361

Mahmud, I., Clarke, L., Nahar, N., & Ploubidis, G. B. (2018). Factorial structure of the locomotor disability scale in a sample of adults with mobility impairments in Bangladesh. *Health and quality of life outcomes*, *16(*1), 81. https://doi.org/10.1186/s12955-018-0903-1

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

Mardia (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika*, *57*(3), 519-30, 1970. https://doi.org/10.2307/2334770

Messick, S. (1989). Meaning and values in test validation: The science and ethics of assessment. *Educational Researcher*, *18*(2), 5-11. https://doi.org/10.3102/0013189X018002005

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

Mittelhaëuser, M., Béguin, A., & Sijtsma, K. (2013). Modeling differences in test-taking motivation: Exploring the usefulness of the Mixture Rasch Model and Person-Fit Statistics. En R.Millsap, L.vanderArk, D.Bolt & C. Woods (Eds.), New Developments in Quantitative Psychology. Presentations from the 77th Annual Psychometric Society Meeting (pp. 345-355). Springer. https://doi.org/10.1007/978-1-4614-9348-8_23

Muñiz Fernández, J. (2010). Las teorías de los tests: Teoría clásica y teoría de respuesta a los ítems. Papeles del Psicólogo: *Revista del Colegio Oficial de Psicólogos*. https://www.papelesdelpsicologo.es/pdf/1796.pdf

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

Onwuegbuzie, A. J., Slate, J. R., Leech, N. L., & Collins, K. M. (2009). Mixed data analysis: Advanced integration techniques. *International Journal of Multiple Research Approaches*, *3*(1), 13-33. https://doi.org/10.5172/mra.455.3.1.13

Ondé Pérez, D. (2020). Revisión del concepto de causalidad en el marco del análisis factorial confirmatorio. *Revista Iberoamericana de Diagnóstico y Evaluación – e Avaliação Psicológica*, *1*(54), 103-117. https://doi.org/10.21865/RIDEP54.1.09

Raykov, T., Marcoulides, G. A., Lee, C. L., & Chang, C. (2013). Studying differential item functioning via latent variable modeling: A note on a multiple-testing procedure. *Educational and Psychological Measurement*, *73*(5), 898-908. https://doi.org/10.1177/0013164413478165

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Revelle, W. (2021). Psych: Procedures for psychological, psychometric, and personality research. Northwestern University, Evanston, Illinois. R package version 2.1.9. https://CRAN.R-project.org/package=psych

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

Simpson, P. M., & Siguaw, J. A. (2000). Student evaluations of teaching: An exploratory study of the faculty response. *Journal of Marketing Education*, 22(3), 199-213. https://doi.org/10.1177/0273475300223004

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

Sterba, S. (2013). Understanding Linkages Among Mixture Models. *Multivariate Behavioral Research*, *48*, 775-815. https://doi.org/10.1080/00273171.2013.827564

Tabachinik, B. G., & Fidel, L. S. (2001). Using multivariate statistics (2ª ed). Allayn and Bacon/ Pearson Education.

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

Trizano-Hermosilla, I., & Alvarado, J. M. (2016). Best alternatives to Cronbach's Alpha reliability in realistic conditions: Congeneric and Asymmetrical Measurements. *Frontiers in Psychology*, *7*, 769. https://doi.org/10.3389/fpsyg.2016.00769

Van De Schoot, R., Schmidt, P., De Beuckelaer, A., Lek, K., & Zondervan-Zwijnenburg, M. (2015). Measurement invariance. *Frontiers in Psychology*, 6(1064). https://doi.org/10.3389/fpsyg.2015.01064

Vermunt, J. K., & Magidson, J. (2002). Latent class cluster analysis. In J. Hagenaars, & A. McCutcheon (Eds.), Applied latent class analysis (pp. 89-106). Cambridge University Press.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological Methodology*, *33*(1), 213-239. http://www.jstor.org/stable/1519857

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teaching in Higher Education*, *12*, 55-76. https://doi.org/10.1080/13562510601102131

Zinbarg, R., Yovel, I., Revelle, W. & McDonald, R. (2006). Estimating generalizability to a universe of indicators that all have one attribute in common: A comparison of estimators for omega. *Applied Psychological Measurement*, 30, 121-144. https://doi/10.1177/0146621605278814

Zumbo, B. D., Liu, Y., Wu, A. D., Forer, B., Shear, B. R. (2017). National and International Educational Achievement Testing: A case of multi-level validation framed by the Ecological Model of Item Responding. In: Zumbo, B., Hubley, A. (eds) Understanding and Investigating Response Processes in Validation Research. Social Indicators Research Series, vol 69. Springer, Cham. https://doi.org/10.1007/978-3-319-56129-5_18

Anexo 1

Cuestionario de satisfacción de la actividad docente

1. El/La profesor/a ha cumplido con lo explicitado en la guía docente.
2. El/La profesor/a ha organizado y estructurado adecuadamente su actividad docente.
3. El/La profesor/a ha explicado con claridad.
4. El/La profesor/a se ha preocupado por el proceso de aprendizaje de los estudiantes.
5. Las tutorías académicas con este/a profesor/a han resultado útiles.
6. La actividad del/de la profesor/a ha contribuido a aumentar mi interés por esta asignatura.
7. En general, el trabajo llevado a cabo por el/la profesor/a ha sido satisfactorio.