UNIVERSIDAD COMPLUTENSE DE MADRID

FACULTAD DE PSICOLOGÍA

TESIS DOCTORAL

Análisis factorial multinivel aplicado a las encuestas de evaluación de la actividad docente universitaria

MEMORIA PARA OPTAR AL GRADO DE DOCTOR

PRESENTADA POR

Lady Catheryne Lancheros Florián

DIRIGIDA POR

Jesús María Alvarado Izquierdo

UNIVERSIDAD COMPLUTENSE DE MADRID
FACULTAD DE PSICOLOGÍA

PROGRAMA DE DOCTORADO EN PSICOLOGÍA

TESIS DOCTORAL

ANÁLISIS FACTORIAL MULTINIVEL APLICADO A LAS ENCUESTAS DE EVALUACIÓN DE LA ACTIVIDAD DOCENTE UNIVERSITARIA

MEMORIA PARA OPTAR AL GRADO DE DOCTORA

PRESENTADA POR

Lady Catheryne Lancheros Florián

DIRECTOR
Jesús María Alvarado Izquierdo

Dedicado a mi madre María Clementina Florián,

a mi padre Luis Alfonso Lancheros Páez

a mi hermanita Diana Marcela Lancheros Florián y

a mi compañero de vida Antonio Daniel Gil Lozano,

Ustedes son mi motivación principal para cada uno

de los logros y etapas de mi vida.

Agradecimientos

Al culminar esta gran, desafiante y larga etapa de mi vida quiero agradecer a las personas que participaron y contribuyeron a culminar este proceso.

En primer lugar, las palabras para agradecer a mi gran guía y maravilloso director de Tesis Jesús Alvarado son infinitas y a la vez insuficientes, porque con cada encuentro, correo, charla o aclaración iluminaba este camino que se tornó oscuro muchas veces y que con su luz volvía a encender en mi la esperanza para continuar adelante con este sueño. Mil y mil gracias por acompañarme en este camino y por compartir conmigo tu inmensa sabiduría, ojalá muchos de los docentes fueran como tú que enseñan con pasión, calidad y respeto.

Agradezco a mi madre Clementina Florián, a mi padre Luis Lancheros y a mi hermanita Marcela Lancheros, quienes siempre han sido mi motor de vida y quienes se han sentido orgullosos de cada logro alcanzado y me han dado fuerzas para continuar en cada etapa, son ustedes por quienes he buscado siempre superarme y superar las dificultades de la vida, con la ilusión de tener un futuro mejor para nuestra familia, a pesar de los retos que ninguno esperaba enfrentar, con resiliencia y amor seguimos transitando este camino llamado vida, los amo con todo mi corazón.

A mi compañero de vida Antonio Gil, quien ha sido mi apoyo incondicional, desde que te dije que quería emprender esta aventura y aun sabiendo del sacrificio que implicaba para nuestra relación, siempre has estado presente tanto en la distancia, como en el día a día desde que nos reencontramos, gracias por el amor, la paciencia y la comprensión tanto en mis mejores como en mis peores momentos de esta historia, no habría sido posible sin ti.

A Sebastián Castro y Eduar Ramírez quienes siempre tuvieron disposición para ayudarme con mis inquietudes metodológicas y me apoyaron hombro a hombro cuando fue necesario hasta encontrar las soluciones que muchas veces me fueron esquivas, no tengo cómo agradecer su generosidad con el conocimiento, el tiempo y el trabajo colaborativo.

A mi amiga incondicional en España Pilar Sánchez, que siempre me ha abierto las puertas de su hogar y su generosidad conmigo y mi familia ha sido infinita.

A mi profe Olga Rodríguez, quien siempre ha estado para apoyarme académica y personalmente, a mirar otras alternativas, a persistir, y a ser un ejemplo para seguir sus pasos y recordarme que con perseverancia y esfuerzo todo es posible. Gracias por tu guía, cariño y amistad.

A Paula Senior, Catalina Calderón, Stephany Quesada y Jenny Cárdenas, amigas del alma, quienes además de su consejo académico, me brindaron su escucha, apoyo y acompañamiento durante el proceso, especialmente en el último año que estuvo lleno de muchos retos personales, gracias por rescatarme una y mil veces, son un gran tesoro en mi vida.

"Uno puede devolver un préstamo de oro, pero está en deuda de por vida con aquellos que son amables"

(Proverbio).

Prólogo: Motivación personal para el desarrollo de esta tesis

Esta aventura comenzó cuando cursaba el Máster de Metodología de las Ciencias del Comportamiento y de la Salud, con mi Trabajo Fin de Máster (TFM) realizado bajo la tutela del Catedrático en Psicometría Vicente Ponsoda, en el servicio de evaluación de la calidad docente de la Universidad Autónoma de Madrid (UAM).

Bajo la tutela del Dr. Ponsoda realicé mis primeros análisis de cuestionarios de evaluación docente. Las cuestiones y dudas que surgieron a raíz de los análisis que se aplican a este tipo de cuestionarios me convencieron de la necesidad de profundizar en los modelos psicométricos. Esta idea la valoré junto a mi tutor y con mi profesor de la asignatura de Validez del Máster el profesor Jesús M.ª Alvarado, quién ante la jubilación del Dr. Ponsoda aceptó continuar con la tutela y dirección de mi tesis doctoral. Estoy enormemente agradecida tanto con Vicente como con Jesús, ya que, sin su inestimable ayuda y valiosas orientaciones, esta tesis no habría sido posible.

El tema de la evaluación docente comenzó a apasionarme, no solamente por el interés de los análisis realizados, sino por mi propia experiencia al ser docente universitaria y receptora de las evaluaciones estudiantiles, en las que muchas veces no me sentí representada por lo reportado por mis estudiantes. La revisión del estado del arte me permitió confirmar mucho de lo que se había aprendido durante el desarrollo del máster.

Mi motivación principal para desarrollar esta tesis fue contribuir al mejor entendimiento y manejo de los datos de las evaluaciones docentes, aplicando técnicas que consideren su estructura natural jerárquica. Al hacerlo, espero aportar a la calidad de la evaluación docente y también abrir caminos para mejorar la toma de decisiones fundamentadas en la educación superior (Arreola, 2000), dado que en muchas oportunidades y en algunas universidades que he trabajado, la decisión de la continuidad de un profesor depende de su evaluación docente, que como ya se verá en el documento son evaluaciones influenciadas por varios sesgos, tratadas

inadecuadamente a nivel metodológico y por ende pueden derivar en conclusiones equivocadas (Hornstein, 2017).

Este último elemento es el que me parece más relevante, dado que la medición procura un proceso objetivo para tener información fiable y en la que se pueda estar seguro de realizar las interpretaciones, según los usos propuestos. En ese sentido, este problema apunta a las fuentes de validez, principalmente a la estructural, aunque también a las relacionadas con el constructo y las consecuencias de uso.

A lo largo del camino, en la revisión y el desarrollo de la tesis, se identificaron dificultades asociadas al uso de las diferentes técnicas avanzadas de corte multinivel, dado que, la aplicabilidad en datos reales mostraba una realidad diferente a la simulación.

En la simulación sabemos que la mayoría de los planteamientos funcionan, otra cosa es la que presenta la realidad, que, en este caso, es posible ver afectadas las conclusiones con algo tan esencial como el tamaño de la muestra. Este viaje, a través de la evaluación docente, ha traído aprendizajes académicos, personales y que también me posicionan diferente frente a mi labor docente.

La esperanza principal es que esta investigación pueda apoyar a quienes realizan análisis de los datos de las evaluaciones docentes, a los tomadores de decisión en las instituciones educativas, y esencialmente, que permita reflexionar sobre la evaluación y el análisis de este tipo de información.

En las siguientes páginas se encontrará que, a través de estudios empíricos y simulaciones, se aportan propuestas para una evaluación más precisa y equitativa de la calidad docente o como se discutirá más delante de la satisfacción con el profesorado.

Índice

Índice de Tablas

Índice de Figuras

Resumen

A lo largo de las décadas, las encuestas estudiantiles o SET (Student Evaluations of Teaching, por sus siglas en inglés) han sido la principal herramienta para la evaluación docente, si bien su eficacia ha sido cuestionada debido a limitaciones teóricas, prácticas y posibles sesgos (Kreitzer y Sweet-Cushman, 2022). Esta tesis destaca la falta de claridad en la interpretación de los resultados y la desconexión entre la satisfacción estudiantil y la calidad de la enseñanza (Stark y Freishtat, 2014; Uttl et al., 2017).

Un problema práctico y metodológico radica en el tratamiento de los datos, el cual habitualmente se realiza sin contemplar la naturaleza jerárquica o anidada de estos (Toland y De Ayala, 2005). Si bien existen propuestas de análisis multinivel más apropiadas (Rampichini et al., 2004), la realidad muestra que este tipo de procedimientos no se aplican, entre otras razones debido a dificultades técnicas o metodológicas. Situación que afecta a las estimaciones realizadas, así como a la comprensión de las propiedades psicométricas de los instrumentos utilizados.

En esta tesis se plantean tres estudios para abordar estas problemáticas. El primer estudio, se centró en realizar el análisis clásico con datos empíricos de los SET, sin considerar la estructura jerárquica de los datos, presentar las bondades y limitaciones de la información obtenida con el enfoque clásico de la Teoría Clásica de los Tests (TCT) y el modelo de respuesta graduada de la Teoría de Respuesta al Ítem (TRI) (Samejima, 2010).

El segundo estudio, muestra la importancia de considerar la naturaleza jerárquica de los datos y el uso de técnicas de análisis acordes a esta, distinguiendo la variabilidad de los estudiantes y de los docentes en las puntuaciones globales. Los resultados mostraron que el aporte de la varianza de los estudiantes era mayor que la de los profesores. Al intentar aplicar los procedimientos multinivel, se identificaron problemas propios de los datos reales (Rantanen, 2013), por lo que fue necesario diseñar una solución que permitiera separar la

varianza para una adecuada interpretación de la validez estructural; encontrando que la dimensionalidad y estructura del instrumento se afectaba cuando no se separaba la varianza.

En el tercer estudio, se realizaron simulaciones para validar la solución técnica propuesta y evaluar las consecuencias de ignorar la estructura jerárquica de los datos (Eber et al., 2021), a través de los índices de ajuste de los modelos y la fiabilidad resultante. Se resalta la importancia de establecer nuevos criterios de validez y considerar diversas fuentes de información.

En conclusión, esta investigación destaca la complejidad de la evaluación docente y propone soluciones para el análisis que permitan mejorar las evidencias de validez de estos instrumentos que soportan sus interpretaciones y usos. Esto resalta la necesidad de fomentar la capacitación en métodos estadísticos y psicométricos para quienes trabajan con este tipo de datos. La tesis contribuye al debate sobre la interpretación y el uso de las encuestas estudiantiles y proporciona una base metodológica para investigaciones futuras.

Palabras clave: SET, evaluación docente, multinivel, simulación, solución técnica, evidencias de validez y estructura de datos.

Abstract

Over the decades, student surveys or SETs (Student Evaluations of Teaching) have been the main tool of teacher evaluation, but their effectiveness has been questioned due to theoretical and practical limitations and biases (Kreitzer y Sweet-Cushman, 2022). This thesis highlights the lack of clarity in the interpretation of results and the disconnection between student satisfaction and the quality of teaching (Stark y Freishtat, 2014; Uttl et al., 2017).

A practical and methodological problem detected refers to the processing of data, which is usually carried out without considering the hierarchical or nested nature of the data (Toland y De Ayala, 2005). Although multilevel methodological solutions have been developed (Rampichini et al., 2004), reality shows that this type of procedure is not used, among other reasons, due to technical or methodological difficulties, situation that affects the estimates made, as well as the understanding of the psychometric properties of the instruments used.

In this thesis three studies are proposed to address these problems. The first study focused on performing the classic analysis with empirical data from the SETs, without considering the hierarchical structure of the data. It also aimed to present the benefits and limitations of the information obtained with the Classical Test Theory (CTT), approach and the Item Response Theory (IRT) with graded response model (Samejima, 2010).

The second study aimed to highlight the importance of considering the hierarchical nature of the data and using procedures consistent with it, in analyzing the variance and the contribution that students and teachers make in the global scores. The results showed that the variance contribution of the students was greater than that of the teachers. When trying to apply the multilevel procedures, problems specific to the real data were identified (Rantanen, 2013), so it was necessary to design a solution that allowed separating the variance for proper structural construct validity, finding that the dimensional structure of the instrument was affected when the variance was not separated.

In the third study, simulations were carried out to validate the proposed technical solution and evaluate the consequences of ignoring the hierarchical structure of the data (Eber et al., 2021), through the model fit indices and the resulting reliability. The importance of establishing new validity criteria and considering various sources of information is highlighted.

In conclusion, this research focus on the complexity of teacher evaluation and proposes statistical solutions to improve the evidence of validity of these instruments that support their interpretations and uses. This highlights the need to promote training in statistical and psychometric methods for those working with this type of data. The thesis contributes to the debate on the interpretation and use of student surveys and provides a methodological basis for future research.

Keywords: SET, teaching evaluation, multilevel, simulation, technical solution, validity evidence and data structure.

1. Introducción

La evaluación docente universitaria, a través de encuestas, se usa principalmente para medir la calidad de la enseñanza y la satisfacción estudiantil (Marsh, 2011). Durante casi un siglo, se ha empleado, investigado y discutido el uso de las evaluaciones de la enseñanza por parte de los estudiantes. Existe una percepción extendida de que estas evaluaciones se centran en la popularidad de los docentes (Sproule, 2000), lo que puede afectar las calificaciones.

Se argumenta que, profesores excelentes pueden recibir calificaciones bajas y viceversa, y que la preocupación por las calificaciones genera una atmósfera que limita la innovación pedagógica y motiva a los profesores a suavizar el contenido de sus cursos (Stark y Freishtat, 2014).

Estas encuestas, realizadas por los estudiantes para evaluar la eficacia o la calidad de los docentes, tienen popularidad en las instituciones, en parte, debido a su facilidad de medición, demandando poco tiempo tanto de la clase como de los profesores. Sin embargo, un problema que afecta a muchos de los cuestionarios es haber sido diseñados sin una comprensión clara, acerca de lo que implica la enseñanza efectiva (Marsh y Hattie, 2002).

Lo anterior deriva en una falta de evidencia sobre la validez de su contenido, que ha llevado a criticar estos instrumentos por la ausencia de elementos que respalden las interpretaciones realizadas y que permitan determinar la eficacia de un profesor (Onwuegbuzie et al., 2009). En este sentido, las respuestas de los estudiantes muchas veces son una mezcla entre su percepción del aprendizaje y sus creencias acerca de las enseñanzas impartidas por el profesor (Kember y Wong, 2000; Spooren, 2013).

Por otra parte, los encargados de analizar los resultados de las evaluaciones docentes consideran que, obtener los promedios numéricos de las calificaciones de estas evaluaciones

es un procedimiento sencillo y que le da un carácter de objetividad, debido a su naturaleza numérica, además permite generar rápidamente una medida de comparación entre profesores (Pounder, 2007).

Sin embargo, estas comparaciones carecen de fundamento estadístico, dado que implícitamente se asume que la distancia que existe entre números como 3 y 4, es igual de significativa que la diferencia entre 6 y 7, al momento de realizar una calificación, generalmente con escalas tipo Likert. Adicionalmente, estas afirmaciones carecen de validez (Stark y Freishtat, 2014) porque esas puntuaciones son variables categóricas ordinales, lo que implica que las calificaciones están dispuestas en categorías con un orden específico, que va desde la peor (1) hasta la mejor (5 o 7). Así, estos números actúan como etiquetas y no como valores numéricos.

En lugar de los números, podríamos emplear descripciones, como "nada efectivas" o "muy efectivas", por lo que calcular promedios de estas etiquetas carece de sentido (McCullough y Radson, 2011). Ahora bien, estas interpretaciones también son incorrectas porque: a) los valores son sensibles si hay datos perdidos (baja tasa de respuesta); b) efecto de las muestras pequeñas puede producir sesgos; c) las respuestas de los estudiantes pueden no ser honestas (falta de anonimato en grupos pequeños) y, d) las respuestas pueden variar por condiciones contextuales (importancia del curso).

Como se señalaba anteriormente, los estudiantes a veces optan por no completar las encuestas o hacerlo de forma parcial, lo que resulta en tasas de respuesta bajas. Conforme disminuye esta tasa, las respuestas pueden volverse menos representativas, es decir, aquellos que optan por no participar pueden diferir significativamente de aquellos que sí lo hacen. Claro está que, las tasas de respuesta por sí solas no ofrecen una indicación clara sobre la calidad de la enseñanza.

Cuando la tasa de respuesta es baja, los datos recolectados no reflejan adecuadamente la totalidad de la calificación de los estudiantes. En realidad, las muestras pequeñas son más susceptibles a fluctuaciones aleatorias producto de los datos extremos, lo que puede resultar en evaluaciones más extremas en clases con pocos estudiantes, incluso si la tasa de respuesta es del 100%. Además, los estudiantes en aulas con pocos alumnos pueden percibir que su anonimato está menos protegido, lo que podría afectar su disposición para responder con honestidad.

Por otra parte, el interés de los estudiantes en los cursos puede variar dependiendo del tipo de curso, si es optativo o principal. Además, la dinámica entre profesores y estudiantes se ve influenciada por la naturaleza y el tamaño de los cursos, así como por cursar el primer o el último año. Estas diferencias suelen ser significativas y pueden afectar los resultados de las evaluaciones del profesorado (Cashin, 1995). Estas variables generan incertidumbre sobre cómo realizar comparaciones equitativas entre distintos tipos de cursos, lo cual complica la interpretación de las evaluaciones docentes (Worthington, 2002).

Sumado a lo anterior, y dada la trascendencia de estos cuestionarios y las decisiones cruciales que conllevan, como la permanencia, promoción y beneficios para los profesores (Spooren et al., 2013), es esencial reflexionar sobre su validez, el análisis de datos y las interpretaciones que se derivan de estos.

Es importante señalar que los datos resultantes de la evaluación de la calidad docente y que, generalmente, provienen de cuestionarios de satisfacción de los estudiantes en entornos universitarios tienen una naturaleza jerárquica o agrupada, es decir, estudiantes que pertenecen a una clase, con un profesor, a una facultad específica y de una universidad, lo que significa que los individuos están agrupados en unidades de un nivel superior, que a su vez también pueden estar agrupados en otras unidades.

A pesar de esta evidencia, cuando se realiza el análisis de los cuestionarios se ignora esta realidad. Esto se debe a veces a falta de conocimiento y en otras, a los problemas metodológicos para realizar análisis que incluyan estas jerarquías, lo que lleva a dos alternativas para investigadores o analistas de datos, 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 lo anterior, a pesar de que el manejo, análisis e interpretación de estos datos son cruciales para la trayectoria profesional de los docentes y la mejora del proceso educativo, existe una escasez notable de investigación que examine el impacto que tiene el limitado conocimiento estadístico y psicométrico de quienes analizan estos datos, así como la tendencia a realizar inferencias basadas únicamente en análisis descriptivos.

Aunque existen directrices para recopilar e interpretar los datos de las evaluaciones de satisfacción estudiantil, muchos usuarios carecen de capacitación especializada en el manejo, análisis e interpretación de estos (Penny, 2003) o desconocen los modelos estadísticos necesarios para llevar a cabo los análisis pertinentes (Franklin, 2001).

Esto implica que, aunque en muchas oportunidades se utilicen las técnicas específicas para este tipo de datos, el desconocer otras características de la información puede llevar a una interpretación errónea de los resultados. Por ejemplo, una preferencia por el uso de medidas agregadas y generales de la satisfacción estudiantil, como los promedios, deja de lado otras herramientas que podrían proporcionar una comprensión más completa del fenómeno o evaluar la calidad de los instrumentos utilizados para recopilar datos (Spooren et al., 2013).

Este aspecto señala directamente la importancia de las evidencias de validez, dado que es crucial reunir pruebas que respalden las interpretaciones propuestas para el uso específico de los tests, es decir, las implicaciones de su aplicación (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).

El considerar la estructura anidada de estos datos representa un desafío significativo en la interpretación de resultados y la toma de decisiones fundamentadas (Spooren et al., 2013). La presente tesis doctoral se justifica por la necesidad de abordar esta complejidad, a través de un enfoque analítico más sofisticado, que el cálculo de los promedios y los procedimientos clásicos. Se busca así, proporcionar una comprensión más profunda y precisa de las dimensiones subyacentes de la evaluación docente, lo que puede tener un impacto directo en la mejora continua de la enseñanza universitaria y la toma de decisiones informadas (Turull y Buxarrais, 2018).

La aplicación de técnicas que permitan comprender los datos anidados, tales como estrategias de carácter multinivel, se alinea con la necesidad de avanzar en la metodología del análisis de la evaluación docente, aportando herramientas más adecuadas para manejar la complejidad inherente a estos datos (Bacci y Caviezel, 2011). Este enfoque metodológico no solo beneficiará el ámbito universitario, sino que también podría tener implicaciones en otros campos que enfrentan desafíos similares en la interpretación de datos jerárquicos y complejos.

Además, se espera que esta investigación contribuya a la identificación de elementos clave que afectan las mediciones de la calidad docente y la satisfacción estudiantil, permitiendo así una mejor comprensión de las necesidades y áreas de mejora en la educación superior (Espinosa et al., 2017).

Con el fin de revisar estas cuestiones acerca de la evaluación docente universitaria, este estudio plantea una serie de objetivos estratégicos. En primer lugar, hacer una revisión detallada de los cuestionarios de evaluación docente, destacando sus características y posibles sesgos asociados, con el fin de entender mejor la medida y sus implicaciones metodológicas.

En segundo lugar, se pretende abordar el problema metodológico relacionado con el constructo evaluado en las encuestas de evaluación docente. Esto implica, analizar las consecuencias de pasar por alto la estructura multinivel al intentar modelar teóricamente los datos provenientes de estas encuestas.

Un tercer objetivo consiste en evaluar las repercusiones de no considerar la estructura de los datos al realizar un análisis tradicional de los mismos. Esto permitirá observar y comprender cómo esta omisión puede influir en los resultados y en la interpretación de los datos reales.

Además, se propone ofrecer una solución viable para realizar un análisis adecuado de los datos obtenidos de las encuestas de evaluación docente. Esta solución busca resolver los problemas metodológicos que puedan surgir al tratar con este tipo de información.

Finalmente, se propone validar y simular las posibles soluciones en diferentes escenarios y con distintas condiciones para comprender mejor su eficacia y limitaciones.

El desarrollo de estos objetivos se realiza a través de un capítulo de introducción, dos capítulos del estado del arte en cuestión, tres capítulos de estudios empíricos y un capítulo de discusión y conclusiones. Los capítulos 2 y 3 se centran en presentar el estado del arte de los cuestionarios de evaluación docente y los análisis alrededor de la evaluación docente. Los capítulos 4 y 5 son estudios realizados con datos reales y el capítulo 6 presenta un estudio de simulación.

Cada capítulo se enfoca en aspectos cruciales para el análisis y la comprensión de la evaluación docente, desde una revisión histórica hasta estudios de simulación y análisis de datos reales. Se espera que, a partir de los resultados obtenidos en esta investigación, sea posible contribuir a la calidad y la interpretación adecuada de las encuestas de evaluación docente, influyendo así en la toma de decisiones basada en estos instrumentos. A continuación, se describe el contenido de cada capítulo.

El capítulo 2, denominado Los Cuestionarios de Evaluación docente, presenta el escenario histórico, el significado y características de los SET (Student Evaluations of Teaching, por sus siglas en inglés), los usos previstos para los cuestionarios de evaluación docente, las principales dificultades y barreras de la implementación de estos cuestionarios, los sesgos que influyen sobre la evaluación docente, tanto respecto a las características del docente, como del estudiante y la clase. También se realiza una breve exposición sobre otras estrategias de evaluación docente, que deberían considerarse, diferentes a las encuestas realizadas por los estudiantes. En este capítulo, se realiza el planteamiento de una de las preguntas más importantes que orientan esta investigación, respecto al constructo medido y la distinción entre la medición de la calidad de la docencia y la satisfacción de los estudiantes. Finalmente, se enuncian aspectos referidos a lo mostrado por la investigación, respecto a las evidencias de validez y cuestiones sobre las dimensiones que suelen evaluarse en estos instrumentos.

En el capítulo 3, se expone teóricamente las cuestiones metodológicas que ayudan con la identificación del problema, abarcando conceptos fundamentales, tales como, el error de medición y la varianza, para luego exponer en detalle el aspecto de la varianza en la evaluación docente, dado que, de acuerdo con lo observado en los diferentes estudios presentados en el capítulo 2, se generan inquietudes respecto a ¿qué es lo que influye en la evaluación docente? ¿se está indagando por los profesores o por los estudiantes? Entonces, se señalan los problemas respecto a la identificación correcta de la varianza generando la inquietud de ¿qué porcentaje

de varianza proporcionan los estudiantes a las evaluaciones y qué porcentaje los docentes? y se expone como más pertinentes el modelo multinivel y el análisis factorial multinivel para el tratamiento de estos datos de naturaleza jerárquica.

Con el fin de abordar los elementos presentados por la revisión teórica se propusieron tres estudios. En el capítulo 4, llamado análisis tradicionales en los cuestionarios de evaluación docente: un estudio empírico; se realiza un análisis clásico utilizando datos reales provenientes de una universidad española, la Universidad Autónoma de Madrid, cuyos datos se obtuvieron provenientes de una práctica de máster en la que se solicitó realizar el análisis al instrumento de evaluación docente, los resultados preliminares de este estudio fueron publicados para evidenciar los procedimientos que comúnmente se realizan en las universidades, las bondades de cada perspectiva y las interpretaciones a las que es posible llegar, sin tener en cuenta ninguna característica de la naturaleza de los datos o de la muestra. En este estudio se refleja cómo, a partir de los resultados del análisis clásico y sus estimaciones, es posible utilizar un procedimiento TRI y obtener más información sobre la calidad del instrumento para la evaluación del constructo. Sin embargo, no se resuelven las inquietudes respecto a la naturaleza de los datos jerárquicos, las variables que pueden afectar la medición y la identificación de la verdadera medición del constructo.

El capítulo 5 evalúa el problema respecto a cuáles son las fuentes de varianza que explican la puntuación total de las evaluaciones docentes. Tal como se presenta en el marco teórico, analizar estas mediciones tiene un carácter importante de complejidad, 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 jerárquica se suele ignorar, por lo que es posible que se incurra en sesgos, al no diferenciar las distintas fuentes de variabilidad, que afectan a las puntuaciones. Este estudio empírico pretendía utilizar un modelo de análisis factorial multinivel, una vez los datos estuvieran limpios y se controlara el sesgo, sin embargo,

dado las características de los datos y la dificultad para tener muestras adecuadas para su aplicación, se vio la necesidad de optar por un procedimiento más sencillo, pero que fuera aplicable a la realidad de los datos. Puesto que, tal como es posible observar en los datos descriptivos, los tamaños de muestra por clase no son representativos y si se hacía uso de aquellos cursos que cumplían los criterios para utilizar procedimientos más sofisticados, la muestra se reducía a un 30%. Esta situación evidenció que la falta de uso de modelos que permiten tener mejores estimaciones está dada por su inaplicabilidad en el contexto real. Por lo anterior, en este estudio se propone un modelo de análisis de cuatro pasos, o, dicho de otra manera, se ofrece una solución técnica sencilla, que se aproxima al análisis factorial multinivel y que permite detectar las fuentes de variabilidad para su control, y así, realizar una estimación más precisa del constructo, controlando el sesgo, a pesar de no ser viable la aplicación de técnicas más avanzadas. Se muestra, además, cómo una vez corregidas las puntuaciones, es posible observar la verdadera estructura factorial del cuestionario, así como hacer una comparación más precisa en términos de satisfacción de los estudiantes, entre las titulaciones y asignaturas.

El capítulo 6, evalúa las consecuencias observadas de ignorar la estructura multinivel en la recuperación del modelo teórico, a través de un estudio de simulación centrado en incluir niveles progresivos de diferencias entre profesores, que fueron controladas en el estudio empírico, en esta simulación se pretendía validar la solución propuesta empíricamente y evidenciar si las diferencias obtenidas en las puntuaciones de un test pueden dar lugar a especificaciones incorrectas del modelo final, lo cual introduce sesgo y es posible que, esta situación comprometa la validez de la medida, específicamente la referida a la fuente de validez estructural. En este estudio, se simularon las condiciones de muestra que no fue posible obtener de los datos reales, para analizar esas consecuencias en el ajuste de cuatro modelos diferentes: un modelo de factores no correlacionados, un modelo unidimensional, un modelo bifactor y un

modelo multinivel. El análisis permitió evidenciar qué tanto influyen las diferencias de calificaciones entre los profesores, respecto a las estimaciones de los modelos factoriales. Encontrando que, aumentar las diferencias de calificación entre los profesores, generó un incremento paradójico en la calidad de los índices de ajuste y por tanto, errores en la identificación de la estructura factorial; destacando la necesidad de, en primer lugar, trabajar los datos provenientes de las encuestas docentes con su estructura jerárquica y en segundo lugar, identificar el aporte de la varianza de las variables que acompañan el proceso de la evaluación, para así controlar su sesgo y poder realizar estimaciones e interpretaciones precisas del constructo. Esto último sería viable identificarlo con la solución propuesta, al corregir la fuente de variabilidad y así recuperar el modelo correcto de los datos.

Los análisis sobre evaluación docente mediante la TCT y la contribución del análisis TRI han sido publicados (Lancheros-Florián, 2022), los estudios de simulación parcialmente están recogidos en la publicación (Lancheros- Florián et al., 2022a) y finalmente la solución para un análisis más adecuado de la validez estructural de los cuestionarios fue publicada en (Lancheros- Florián et al., 2022b).

2. Los Cuestionarios De Evaluación Docente

2.1 Contexto Histórico

La evaluación docente en la educación superior ha sido un tema de importancia mundial desde la década de 1920, cuando se originó en Estados Unidos. Según Banta y Blaich (2011), esta se ha utilizado para medir la calidad de la enseñanza, así como la satisfacción con el aprendizaje, con el fin de mejorar la toma de decisiones en las universidades.

La evaluación de la docencia ha sido caracterizada principalmente a través de encuestas administradas directamente a estudiantes matriculados en circunstancias controladas, generalmente cerca del final de un período académico. Estas encuestas también se conocen como evaluaciones de la enseñanza de los estudiantes, evaluaciones de la docencia por parte

de los estudiantes, calificaciones de la instrucción de los estudiantes, evaluaciones de final de curso, entre otros.

A lo largo de la historia, los procedimientos de evaluación han ido cambiando. En el principio de los años 20's las universidades utilizaban métodos improvisados para evaluar la enseñanza, como las puntuaciones de exámenes de los estudiantes, el número de admitidos y los datos de empleo (Centra, 1993).

En los años 50, los líderes académicos se interesaron por proporcionar evidencia más precisa de la calidad de la enseñanza en sus instituciones (Bowen,1979; Marsh, 1987). Sin embargo, existían muchas dudas para invertir capital para investigación por parte de las Instituciones, debido a la probabilidad, de que las iniciativas de evaluación docente propuestas se vieran obstaculizadas por disputas metodológicas, escepticismo y debates sobre las prioridades institucionales. Además, los expertos en evaluación no lograban llegar a un consenso sobre qué medir y cómo medir la efectividad del profesorado (Bowen,1979).

En los años setenta, surgió una mayor conciencia sobre la relevancia del aprendizaje enfocado en el estudiante y la importancia de incluir su percepción sobre la calidad de la enseñanza, mientras que, a principio de los 80's, los instrumentos de evaluación docente en la educación superior eran impulsados principalmente por fundaciones, organizaciones profesionales y editoriales, cuyo enfoque se centraba en cumplir con las demandas de rendición de cuentas institucional (Gelber,2020).

A finales de los 80s, la mayoría de las universidades en Estados Unidos consultaban los cuestionarios de los cursos de los estudiantes durante las audiencias de promoción y titularidad, y la mitad de estas incluía como evidencia también las observaciones de pares, los trabajos de los estudiantes u otros métodos para evaluar el desempeño docente. Por lo que, la relación entre la evaluación docente y la valoración externa comenzaba a ser tema de interés y de suma importancia para algunos profesores. Por su parte en España, las universidades realizaron

evaluaciones iniciales basadas en encuestas a estudiantes, un proceso voluntario que se llevó a cabo en instituciones como Cantabria, Barcelona y Valencia (Tejedor y Jornet, 2008). Posteriormente, en 1989, se introdujo el Decreto de Retribuciones, estableciendo un sistema de revisión salarial vinculado a la productividad y centrado en evaluar tanto la docencia como la investigación, con evaluaciones periódicas en intervalos regulares (Tejedor y Jornet, 2008).

En los 90s, los defensores de nuevos métodos de evaluación docente seguían siendo optimistas, dado que había apoyo por parte de la literatura académica y los especialistas en evaluación hacia las prácticas de enseñanza centradas en el estudiante (Hutchings et al., 2011; Ory, 1991; Seldin, 1999). La investigación señalaba que en los cursos que se incluían estrategias de aprendizaje "activas", como discusiones o tareas cooperativas, mejoraba la capacidad de los estudiantes para aplicar sus conocimientos, al momento de resolver problemas en nuevos contextos (Kilgo et al., 2015; Pascarella y Terenzini, 1991) y, en consecuencia, los resultados de las encuestas de satisfacción de los estudiantes serían mejores.

En España, en 1992, se creó el Plan Nacional de Evaluación de las Universidades, que tenía como objetivo principal controlar el uso eficiente de los fondos públicos y mejorar la calidad de las instituciones académicas. Este plan abarcaba la evaluación institucional y del profesorado, considerando aspectos como la docencia, la investigación y la gestión universitaria. Esta fase marcó el inicio de una auténtica cultura de evaluación en las universidades españolas, orientada hacia la mejora continua (Tejedor y Jornet, 2008).

La década del 2000 fue muy importante para la historia de la evaluación docente en España, dado que, en el 2001, se implementó un sistema de acreditación para profesores contratados y la habilitación nacional de funcionarios, a través de la Ley Orgánica de Universidades 6/2001. Este sistema se basaba en la valoración realizada por especialistas en diversas áreas, evaluando la formación, la docencia, la investigación y la participación en actividades de gestión por parte de los docentes. Las evaluaciones, principalmente

administrativas, permitieron que solo las personas acreditadas pudieran optar a las plazas universitarias disponibles.

En 2002, se estableció un cambio significativo con la creación de la Agencia Nacional para la Evaluación de la Calidad y la Acreditación (ANECA). Esta agencia se formó en el marco del Espacio Europeo de Educación Superior y tenía como misión contribuir a mejorar la calidad del sistema de Educación Superior en España. La ANECA se encargó de evaluar y acreditar no solo a los profesores, sino también a las enseñanzas y las instituciones universitarias, consolidando así un sistema integral de evaluación y acreditación en el ámbito académico del país (Tejedor y Jornet, 2008).

En 2013, debido en gran parte a políticas estatales y estándares de acreditación, casi la mitad de las universidades y colegios en Estados Unidos utilizaban los resultados de al menos una prueba estandarizada como parte de sus procesos de evaluación docente (Koretz, 2008).

Alrededor del año 2016, la mayoría de los profesores consideraban que la evaluación de la calidad de la enseñanza y la satisfacción con el aprendizaje tenía como objetivo principal responder a los acreditadores, los políticos y otro personal externo (Straumshein, 2016); y no al objetivo de las evaluaciones, a la mejora de la educación, a la retroalimentación de la enseñanza, entre otras. Es por esto, que muchos profesores se distanciaron de estos procedimientos, debido a la aversión a las críticas y por una preocupación por posibles infracciones a la libertad académica (Straumshein, 2016).

En las últimas décadas, se ha observado un cambio significativo a nivel mundial en la mentalidad de los estudiantes de finales del siglo XX y principios del XXI. Estos estudiantes han demostrado una mayor atención a las perspectivas laborales y una conciencia más aguda sobre los costos y beneficios económicos que proporciona la educación superior. Los estudiantes son más exigentes con los beneficios reales de hacer una carrera, por lo que son más estrictos en las evaluaciones docentes (Gelber, 2020).

Este cambio en la mentalidad de los estudiantes ha coincidido con transformaciones importantes en las dinámicas de la educación superior. Uno de estos cambios se ha reflejado en los recortes en los fondos de los gobiernos para la educación superior, lo que ha llevado a que las instituciones educativas dependan, cada vez más, de sus propios recursos para funcionar de manera efectiva (Mitchell et al., 2016).

En este contexto, la diferenciación de las instituciones se ha convertido en un factor crucial para que puedan destacarse y competir en un entorno educativo cada vez más desafiante (Staley y Trinkle, 2011). Esta competencia se hace evidente en el proceso de reclutamiento de nuevos estudiantes, en el que las instituciones educativas se han visto obligadas a adaptarse y diferenciarse en la calidad de su enseñanza, reflejada en las evaluaciones docentes, para atraer a los estudiantes más prometedores y competir en un entorno educativo desafiante (Chou et al., 2012; Staley y Trinkle, 2011).

La evaluación docente ha sido ampliamente empleada a nivel mundial, y al mismo tiempo, ha sido objeto de numerosos estudios (Theall y Franklin, 2000). Esto ha generado que, las investigaciones sobre la evaluación docente realizada por estudiantes superan en número a cualquier otro tema de investigación en educación superior (Marsh et al., 2009). Estos estudios han proporcionado valiosos conocimientos sobre cómo medir la calidad de la enseñanza y mejorar la experiencia educativa en la educación superior.

2.2 Evaluaciones de la enseñanza por parte de los estudiantes (SET)

Los instrumentos para la valoración del desempeño del profesorado también son ampliamente denominadas o conocidas como SET (Student evaluations of teaching, por sus siglas en inglés), la cuales, de acuerdo con por la Organización de las Naciones Unidas para la Educación, la Ciencia y la Cultura, se describen como el proceso de identificar las percepciones de los estudiantes en relación con la labor y la actitud de los docentes. Estas observaciones

permiten a los evaluadores, determinar el grado de conformidad entre las expectativas de los estudiantes y los enfoques de enseñanza utilizados por los maestros (Stroebe, 2016).

El objetivo principal de la construcción y uso de las SET se centra en la recolección de información que permita verificar la calidad de la enseñanza (Muñoz et al., 2011; Stake et al., 2017). Adicionalmente, se le atribuyen tres propósitos principales: (a) mejorar la calidad de la enseñanza; (b) proporcionar información para los ejercicios de evaluación (por ejemplo, decisiones de promoción o ascenso); (c) proporcionar evidencia para la responsabilidad institucional, por ejemplo, demostrar la presencia de procedimientos adecuados para garantizar la calidad de la enseñanza (Kember y Wong, 2000).

Una de las particularidades en este tipo de evaluaciones es que cada universidad suele establecer sus propios criterios acerca de la calidad de la docencia (Oermann et al., 2018). Esta situación ha llevado a la construcción y aplicación de miles de cuestionarios a nivel mundial (Spooren et al., 2013), así como al desarrollo de estudios que indagan por sus propiedades psicométricas (Turull y Buxarrais, 2018).

En educación superior, estas encuestas siguen siendo el único método, el más utilizado o el que tiene mayor relevancia en las universidades del mundo para recolectar datos de evaluación docente (Miller y Seldin, 2014; Muñoz et al., 2011).

Aunque, no se puede desconocer que algunas universidades han ampliado su modelo de evaluación, enfocándose en las directrices establecidas por el Espacio Europeo de Educación Superior (EEES), en el que las evaluaciones docentes se incluyen como una parte esencial de un modelo, complementado por otras medidas, que incluye autoevaluaciones de los docentes, evaluaciones de pares o directores de departamentos, como el programa *Docentia* que pretende unificar el significado de calidad docente en todas las universidades de España (Muñoz et al., 2011).

En el ámbito universitario, las SET son herramientas comunes utilizadas para recopilar la percepción de los estudiantes sobre la enseñanza de sus profesores, suelen responderse a través de una escala Likert, con una serie de categorías ordenadas (de acuerdo - en desacuerdo) y que permiten a los estudiantes expresar su opinión de manera anónima sobre su percepción de la calidad de la enseñanza. Estas suelen abordar aspectos como la facilitación, regulación, interacción grupal, relación individual, entusiasmo del profesor, organización, claridad del curso, carga de trabajo, dificultad, la actitud en clase y/o las habilidades de enseñanza de un maestro, entre otras (Marsh, 2007; Vlăsceanu et al., 2004).

La evaluación del desempeño docente se puede llevar a cabo mediante una variedad de enfoques teóricos y prácticos, adaptados a las circunstancias específicas de la evaluación (Marczely, 1992). Uno de estos enfoques, implica la identificación y valoración de los rasgos y comportamientos de los profesores, utilizando escalas cuantitativas y cualitativas. Otro enfoque, se centra en los objetivos de la enseñanza, aunque su aplicación es limitada, debido a que el rendimiento académico de los estudiantes escapa al control total de los docentes y no es ampliamente aceptado como criterio de evaluación.

Además, existe un modelo que evalúa el proceso de enseñanza, considerando aspectos como la preparación, el diseño, los objetivos, los métodos y el control del aprendizaje (Marczely, 1992). A pesar de las potencialidades de este enfoque, su utilización no está extendida. También existe un modelo centrado en las preocupaciones del profesor, donde los docentes evalúan su propio rendimiento y establecen objetivos basados en esta autoevaluación (Boyan y Copeland, 1978).

A menudo, al hablar sobre la evaluación de profesores, se pone énfasis en la evaluación de la práctica profesional. Sin embargo, es crucial comprender que la evaluación de profesores no se limita solo a su desempeño en el campo profesional, también abarca su formación previa, así como su desarrollo y aprendizaje continuo (Escudero, 2019). Es por eso que, una evaluación

integral debe considerar tanto las habilidades demostradas en el aula, como el crecimiento profesional y educativo a lo largo del tiempo.

Así, la evaluación docente se puede centrar en tres ámbitos distintos: lo que el profesor sabe, lo que hace en el aula y los resultados que obtiene. El conocimiento y las habilidades del profesor, así como sus acciones profesionales y logros, se refieren a su competencia, práctica profesional y productividad, respectivamente. También es importante aclarar, que la productividad no se limita al rendimiento del estudiante, sino que se define como la contribución del profesor al aprendizaje del alumno, una medida más holística y completa (Schalock et al., 1993).

En este escenario, es importante señalar que existen tres tipos de herramientas SET de uso común: aquellas que se desarrollan internamente por una institución, que en este grupo entran la mayoría de SET; aquellos que se obtienen de forma gratuita, como el SEEQ (Marsh, 1980, 1991); y aquellos que se desarrollan comercialmente para su compra, como el ETS SIR-II vendido por el Educational Testing Service, el IDEA SRI vendido por IDEA Center; y el CIEQ vendido por C.O.D.E.S Inc.

A pesar de las diversas variaciones de los SET, las comparaciones entre ellos sugieren que, siempre que el objetivo general sea evaluar la enseñanza efectiva. Las dimensiones dentro de los SET están interrelacionadas y pueden superponerse (Marsh, 1987, 2007; Marsh y Bailey, 1993; Marsh y Dunkin, 1997). Un ejemplo de estas dimensiones SET es el propuesto por Feldman (1997), que incluía tres categorías: presentación, facilitación y regulación. Otro estudio relevante consideró cinco dimensiones para medir la efectividad de la enseñanza: habilidades analíticas/sintéticas, organización/claridad, interacción grupal del profesor, interacción individual entre el profesor y el estudiante, y dinamismo/entusiasmo (Ching, 2018).

Otro ejemplo es el SEEQ que mide nueve dimensiones: tareas y lecturas, amplitud de cobertura, exámenes y calificaciones, interacción grupal, relación individual, entusiasmo del

instructor, aprendizaje y valor académico, organización y claridad, y, por último, la carga de trabajo y dificultad del contenido (Marsh, 2007).

El diseño de los cuestionarios SET depende de la perspectiva desde la cual se realice la evaluación, ya que existen diferencias significativas entre las perspectivas de los miembros de una facultad (quienes enseñan) y las perspectivas de los estudiantes (quienes aprenden) (Rosen, 2018). Estudios realizados por Ching (2018), en la Universidad de Hawái, analizaron los SET desde ambas perspectivas, generando un conjunto de criterios compuestos por elementos como el conocimiento de la materia, la capacidad para estimular el interés y motivar a los estudiantes, la organización del curso, la preparación para el curso, la preocupación por los estudiantes, la calidad de los materiales del curso y una evaluación global sumativa del maestro.

En resumen, la evaluación profesional de los docentes se despliega en un complejo entramado de habilidades, acciones y resultados. Desde la competencia y la práctica profesional del profesor, hasta la contribución concreta al aprendizaje del alumno, cada aspecto es esencial para una evaluación completa y significativa. A través de herramientas como las SET, se busca capturar la percepción de los estudiantes sobre la enseñanza de sus profesores.

2.3 Usos de los cuestionarios de evaluación docente

Los cuestionarios de evaluación docente, que evidencian la satisfacción expresada por parte de los estudiantes, se han constituido en una de las principales fuentes de información sobre la labor realizada por el docente (Calderón y Escalera, 2008; Harvey, 2003).

En España, esta información se recolecta principalmente a través de encuestas de opinión, cuyos datos suelen ser utilizados con diversos fines, evaluación del proceso enseñanza y aprendizaje, la acreditación de los docentes, la acreditación de las titulaciones, el insumo para planes de mejora, entre otros (Denson et al., 2010; Espinosa et al., 2017).

En Latinoamérica, los procesos de evaluación docente se suelen centrar en la regulación de la carrera del docente, que pueden implicar ascensos, incrementos salariales, promociones

o destituciones, la acreditación de los programas de formación y en algunos casos, como insumo para orientar la mejora de la calidad en el trabajo docente (OCDE, 2013a, 2013b).

Respecto a estos usos, algunos profesores consideran las SET como un medio adecuado para evaluar la calidad de las instituciones. Un estudio realizado por Beran y Rokosh (2009) encuestó a 262 docentes universitarios, encontrando que el 84% de ellos respalda el uso de SET y el 62% confía en la forma en que los jefes de departamento y decanos interpretan los resultados de estas evaluaciones. Además, investigaciones como la de Yao y Grady (2005) han revelado que los profesores están interesados en las opiniones de los estudiantes, a pesar de experimentar ansiedad y tensión debido a los propósitos sumativos y administrativos de las SET.

A pesar de su amplio uso, sus limitaciones de tipo teórico y práctico, el uso que se hace de los resultados, el modelo idóneo de profesor y la baja tasa de respuesta de los estudiantes, han hecho que sean ampliamente cuestionados como instrumentos eficaces para los usos propuestos (Turull y Buxarrais, 2018).

2.4 Dificultades, preocupaciones y barreras de las SET

A pesar de su amplio uso en diferentes países del mundo, la evaluación docente es un proceso tan importante como delicado (Escudero et al., 2010), lo que lo ha llevado a ser objeto de análisis y críticas. Algunas de las dificultades identificadas de la evaluación docente se relacionan con: a) la definición conceptual de criterios para la evaluación docente, b) la calidad técnica de los instrumentos usados en la medición, c) la inserción de la evaluación entre los procesos de los sistemas educativos y sus políticas, d) la definición de un marco legal que legitime los procesos, los oficialice y garantice el cumplimiento de los derechos, (e) la instalación de culturas de evaluación para la mejora y (f) la protección de la información y el honor de los evaluados (Mateo, 2000).

Por su parte, Escudero (2019) señala que existen tres tipos de dificultades en el proceso de evaluación docente. En primer lugar, están los problemas técnicos relacionados con las estrategias y recursos de la evaluación docente, que generan incertidumbre entre los diferentes usuarios del sistema de evaluación. En segundo lugar, surgen los problemas de gestión política, vinculados a la falta de integración adecuada de la evaluación docente en el sistema educativo. Por último, se presentan problemas de legitimación de la evaluación, debido a la falta de una cultura adecuada entre los profesores sobre este tema.

2.4.1 Percepción de los docentes sobre la evaluación docente

Kezar (2014), señala que algunas de las preocupaciones manifestadas por los docentes se refieren a que las evaluaciones docentes puedan afectar su capacidad para seleccionar el contenido del curso y/o los métodos de enseñanza; manifestaciones sobre que la calidad de la educación superior no se puede medir de manera efectiva mediante pruebas estandarizadas y que la evaluación docente es un producto de las presiones del mercado. El dilema asociado con las evaluaciones docentes se complica, aún más, debido a la percepción generalizada entre profesores y estudiantes de que las calificaciones otorgadas por los estudiantes pueden influir en las evaluaciones institucionales de los docentes.

Esta percepción, en muchos casos, afecta la legitimidad de las mediciones como herramientas efectivas para mejorar la calidad educativa. No es de extrañar, que muchos profesores no respalden las encuestas realizadas por los estudiantes sobre sus cursos, ya que creen que estas evaluaciones aumentan la presión sobre sus interacciones con los estudiantes, con el fin de obtener calificaciones más altas. Esta presión se vuelve especialmente intensa para los profesores adjuntos, de cátedra o con contratos temporales, ya que a menudo son evaluados exclusivamente basándose en la retroalimentación de los estudiantes (Gelber, 2020).

Las evaluaciones realizadas por los estudiantes continúan siendo una fuente constante de descontento entre los profesores, en gran parte, debido a la incomodidad que surge al leer

críticas detalladas o al ver el desempeño docente reducido a simples números. En la misma línea, dado que las instituciones educativas enfrentan crecientes presiones económicas y a las expectativas cambiantes de los estudiantes, se ha exacerbado la preocupación de los profesores sobre el posible mal uso de estas evaluaciones (Schwab et al., 2018). El estudio de Sánchez-Escobedo et al. (2009) señala que algunos docentes manifestaron resistencia hacia la evaluación debido a experiencias negativas, como ser evaluados con sus grupos más difíciles, la percepción de que solo se considera la opinión de los alumnos y la falta de alineación con los objetivos de sus materias. Además, expresaron temores en relación con la forma en que se usa la evaluación, que se difunda que tienen poca capacidad para enseñar, a ser perjudicados en su trabajo y la preocupación por la falta de claridad en los criterios de evaluación.

Otro factor que contribuye a la incertidumbre por parte de los profesores es la falta de contexto histórico en torno a las evaluaciones de los cursos. La ausencia de este contexto impide una comprensión completa del papel que desempeñan en el desarrollo y la adopción de estas evaluaciones. A pesar de su compleja historia, los cuestionarios de evaluación docente se han convertido en elementos propios de la evaluación de su desempeño y se perciben como un síntoma de la disminución del poder de los profesores (Valsan y Sproule, 2008).

Antes de la década de 1980, cuando las instituciones educativas tenían menos presión para presentar datos cuantitativos a audiencias externas, algunos profesores apoyaban los esfuerzos para examinar la eficacia de la enseñanza y el aprendizaje en sus campus. Estos defensores creían que la evaluación podría mejorar la calidad de su trabajo, una idea difícil de lograr pero que podría señalar una posible solución para el actual estancamiento en este tema. Sin embargo, se ha identificado que parte de la renuencia del profesorado a ser evaluado, se constituye en muchas oportunidades como la barrera más importante para una evaluación significativa en las instituciones (Kuh y Ikenberry, 2009).

Es por esto que, a medida que la evaluación docente se ha vinculado más estrechamente con la rendición de cuentas externa, ha resultado difícil involucrar a los profesores en estos esfuerzos (Ewell, 2002). Muchos profesores y administradores dudan de la sinceridad de los responsables políticos cuando expresan preocupación por la efectividad de la educación superior. Así, la evaluación docente a menudo se percibe como una justificación para otros objetivos, como recortes presupuestarios, mayor necesidad de profesores adjuntos o temporales, expansión de programas en línea o reducción de la autonomía curricular, en lugar de un intento genuino por mejorar la enseñanza (Bolívar, 2008).

Incluso sin estas preocupaciones, para los docentes, la evaluación puede parecer una tarea sin sentido que no impacta de manera constructiva en el aula (Gelber,2020). Se ha encontrado que, los profesores de mayor edad solicitan menos retroalimentación de sus colegas (Kunst et al., 2018; Runhaar et al., 2010). Respecto a la retroalimentación proveniente de los estudiantes, los profesores con mayor experiencia profesional muestran mayor escepticismo sobre su utilidad y los profesores de mayor edad la utilizan con menor frecuencia (Dretzke et al., 2015). Algunos descubrimientos sobre los efectos de género en relación con la retroalimentación indican que las profesoras buscan más retroalimentación de sus colegas (Runhaar et al., 2010) y tienden a mejorar más su enseñanza tras una intervención de retroalimentación por parte de los estudiantes (Buurman et al., 2018).

En este punto, es importante señalar que los diversos enfoques para valorar la calidad de la enseñanza han sido imperfectos desde sus orígenes, pero sus defectos se volvieron más difíciles de aceptar a medida que la evaluación parecía originarse fuera del mundo académico (Gelber, 2020). Lo que se ha identificado en diferentes lugares del mundo, es que la mayoría de los profesores todavía creen que la esencia de la educación universitaria no se cuantifica fácilmente y les preocupa que el acto mismo de medirles afecte la forma de impartir la enseñanza (Birnbaum y Snowdon, 2003). Esto se acentúa, cuando los profesores notan que la

evaluación suele ser promovida por aquellos que desean operar las instituciones de una manera más empresarial.

2.4.2 Percepción de los estudiantes sobre la evaluación docente

Aunque la percepción de los estudiantes es la que más se usa como insumo para hablar de la evaluación docente, se encuentran muy pocas referencias respecto a la percepción de los estudiantes sobre el uso, aplicación y efectos de las evaluaciones docentes, sin embargo, se presentan algunas de las investigaciones encontradas.

La investigación realizada por Spencer y Schmelkin (2002) revela que los estudiantes, en su mayoría, muestran disposición para participar en las evaluaciones del profesorado sin temor a las posibles consecuencias negativas. Sin embargo, a pesar de su participación, existe una falta de confianza generalizada entre los estudiantes en que sus opiniones sean tomadas en cuenta por los docentes y administrativos de las instituciones educativas.

Por su parte, Pardo (2017) señala que los estudiantes consideran que la evaluación docente se basa en criterios generales, pero poco específicos, lo que resulta insuficiente para medir con precisión la calidad del trabajo del profesor. Sugieren que la evaluación debería adaptarse a cada proyecto curricular y a las particularidades de cada comunidad, con una participación de otros profesores en el proceso de calificación. Sin embargo, detectan una falta de coherencia en el proceso, donde se percibe que las observaciones de los estudiantes son ignoradas y carecen de valor en los procesos de mejora, dado que, a pesar de las críticas, sugerencias, bajas calificaciones o recomendaciones dirigidas a algunos docentes, no se evidencia un cambio significativo o mejoramiento en la labor del docente, lo que convierte la evaluación en un trámite obligatorio sin un impacto real en la mejora educativa.

Finalmente, el estudio de Greimel-Fuhrmann (2014) reveló que tanto el interés por la temática de la materia, como el grado de atención hacia los estudiantes por parte de los profesores, pueden predecir la calificación general de calidad de la enseñanza otorgada por los

estudiantes en los SET. También en esa misma línea, los estudiantes señalan que, si el profesor explica con claridad, se preocupa por averiguar si los conceptos explicados han sido entendidos y prepara sus clases, la valoración sobre la calidad de su docencia será muy satisfactoria (Ruiz Esteban y Santos del Cerro, 2020).

2.4.3 Algunas críticas

La evaluación de la enseñanza y de la docencia suelen ser tratadas como sinónimos, sin embargo, en la docencia universitaria aparecen tres elementos diferenciadores: la investigación, la gestión y la enseñanza, cada uno de los cuales puede implicar un objeto distinto de evaluación (Rodríguez, 2000). Las valoraciones realizadas por los estudiantes se enmarcan únicamente en la satisfacción del estudiante con la enseñanza, sin embargo, esta valoración puede estar mediada por otras características, que suelen no estar incluidas en las encuestas de valoración (García, 2000).

En cuanto a la relación entre las SET y la mejora de la enseñanza, existen opiniones encontradas en la literatura. Estudios como el de Davidovitch y Soen (2006) sugieren que las SET mejoran con el tiempo, a medida que los profesores adquieren experiencia y antigüedad, lo que se traduce en una mejora en la calidad de la enseñanza. Sin embargo, otras investigaciones como la de Kember et al. (2002), que utilizaron datos de SET recopilados durante varios años, no encontraron evidencia de que las puntuaciones de SET mejoren con el tiempo. Esta discrepancia, destaca la necesidad de investigaciones adicionales para comprender mejor la relación entre las SET y la mejora de la enseñanza a lo largo del tiempo.

Otra crítica se refiere a que, debido a la naturaleza anónima de las evaluaciones de los cursos, los resultados están disponibles solo para el subconjunto de estudiantes que eligen completar las evaluaciones, situación que dificulta contrastar su validez y por la cual ha recibido amplias críticas (García, 2014). También se ha identificado que las evaluaciones docentes desarrolladas en línea suelen tener una menor participación que las realizadas

presencialmente, lo que puede afectar la información disponible sobre la valoración del profesor (Anderson et al., 2005; Avery et al., 2006).

Para llevar a cabo una evaluación efectiva, es fundamental que las fuentes, procedimientos, agentes y estrategias de evaluación estén alineados con los contenidos y objetivos educativos. Es importante tener en cuenta que, muchos instrumentos de evaluación están diseñados con propósitos específicos y no son adecuados para evaluaciones múltiples (Denham y Almeida, 1987). Además, la evaluación de los estudiantes, especialmente cuando se utiliza para tomar decisiones cruciales como la promoción o el despido de profesores, ha sido objeto de críticas por varios autores (Wilson y Wood, 1996). Por lo tanto, es esencial abordar estas críticas y considerar cuidadosamente las implicaciones de las evaluaciones estudiantiles en la toma de decisiones académicas importantes.

A pesar de estas percepciones, algunos siguen considerando que la evaluación docente puede ser una herramienta valiosa para mejorar la enseñanza y el aprendizaje, cuando se realiza de manera reflexiva y significativa (Kezar, 2014).

2.5 Sesgos identificados en la evaluación docente

Se ha encontrado que las evaluaciones de la enseñanza tienen una correlación débil o totalmente nula con la eficacia de la enseñanza (Stark y Freishtat, 2014; Uttl et al., 2017) y suele estar mediada por diferentes variables. En este sentido, las evaluaciones están influenciadas por factores tales como la disciplina, el interés de los estudiantes, el nivel de la clase, la dificultad de la clase, el tiempo de reunión de la clase y otras características específicas del curso, como la hora del día en que se imparte la clase y el tamaño de la clase (Wachtel, 1998), pero no necesariamente por la calidad real del profesor (Miles y House, 2015; Spooren et al., 2013; Uttl et al., 2017). De la misma manera, factores como el entusiasmo del docente, el atractivo físico y la simpatía pueden generar sesgos en la valoración estudiantil (Feistauer y Richter, 2018; Wolbring y Riordan, 2016; Gruber et al., 2012).

Benton et al. (2012), señalan que en las evaluaciones docentes ha sido posible identificar dos tipos de sesgo: el sesgo de medición y el sesgo de equidad. El primero ocurre cuando variables no relacionadas con la eficacia de la enseñanza afectan sistemáticamente los resultados, tales como las características del curso, el tamaño de la clase, la dificultad del material, y también características individuales de los estudiantes, como su nivel de interés en el tema o sus experiencias previas en cursos similares. Por otro lado, el sesgo de equidad se presenta cuando factores externos al control del docente afectan de manera sistemática los resultados de las evaluaciones. Esto puede incluir sesgos relacionados con el género, la raza, el origen étnico, el acento, la orientación sexual o la discapacidad del profesor (Benton et al., 2012).

2.5.1 Características personales del docente

2.5.1.1. Género.

Solo en algunos estudios se han identificado pocas diferencias de género (Wallisch y Cachia, 2019; Wright y Jenkins-Guarnieri, 2012). La mayor parte de la literatura indica que los hombres reciben puntuaciones evaluativas más altas en comparación con las mujeres (MacNell et al., 2015; Mengel et al., 2018). Por ejemplo, es más probable que los hombres sean vistos como instructores expertos, brillantes u organizados, en una variedad de entornos institucionales y metodologías de investigación, en comparación con las mujeres (Abel y Meltzer, 2007; Arbuckle y Williams, 2003; Boring et al., 2016; MacNell et al., 2015; McPherson et al., 2009; Mengel et al., 2018; Ridgeway, 2011; Wagner et al., 2016). Los estudios con diseños cuasiexperimentales, tal como se definen en Ato et al. (2013), que han estudiado la influencia del género del docente en procesos de enseñanza en línea, han demostrado que los estudiantes otorgaron puntuaciones más bajas cuando creían que el tutor de un curso era una mujer, a pesar de que el curso impartido por hombres fuera exactamente igual (Boring et al., 2016; MacNell et al., 2015).

También se ha identificado un efecto en relación con la afinidad de género, es decir, los estudiantes califican más alto a los profesores que tienen su mismo género (Bachen et al., 1999; Young et al., 2009). Los estudiantes de género masculino califican más bajo a las docentes mujeres (Basow, 1995; Fan et al., 2019; Mengel et al., 2018), y en general, las estudiantes mujeres suelen asignar puntuaciones más bajas en sus valoraciones, mientras que los hombres suelen otorgar calificaciones más altas (Centra, 2000; Rowden y Carlson, 1996).

2.5.1.2 Origen y edad.

Reid (2010) encontró que los profesores de raza negra tienen evaluaciones más bajas que los profesores de raza blanca, así como que, a los profesores con acento diferente al nativo les va peor, que a los profesores nativos de habla inglesa. Se ha identificado también, que los profesores con acento y apellidos asiáticos suelen recibir calificaciones más bajas en las SET (Fan et al., 2019; Subtirelu, 2015). Ante un estilo de enseñanza igualmente estricto, las mujeres latinas se valoran como menos cálidas que las anglosajonas (Anderson y Smith, 2005) y las mujeres de raza negra tienen calificaciones más bajas y mayores comentarios negativos que los hombres de raza blanca (Chávez y Mitchell, 2020). No obstante, Ching (2018) señala que aspectos como el sentido del humor, la calidez y la claridad en la comunicación pueden mejorar la percepción del estudiante y evitar estereotipos basados en apariencia o el país de origen.

Mengel et al. (2018) señala que no hay acuerdo frente a la antigüedad del profesor, algunas investigaciones evidencian que esta característica contribuye a disminuir el sesgo con el que valoran los estudiantes, mientras que otras investigaciones evidencian que los docentes jóvenes son más populares y reciben evaluaciones más altas (Arbuckle y Williams, 2003; McPherson et al., 2009), y en consecuencia la edad del docente también ha mostrado tener un efecto negativo significativo (Arnold y Versluis, 2019).

2.5.2 Características de la clase

Las evaluaciones de la enseñanza también son más bajas para los cursos no optativos y cuantitativos (Benton et al., 2012; Boring et al., 2016; Mengel et al., 2018; Uttl et al., 2017, 2021). Los estudiantes brindan evaluaciones más elevadas en clases basadas en debates o que tienen mayor posibilidad de interacción que en cursos introductorios más grandes (Miles y House, 2015; Spooren et al., 2013); así como, en clases que tienen poca carga de trabajo o que presentan distribuciones de calificaciones más altas (Miles y House, 2015; Rosen, 2018). Así, los docentes tienen más probabilidades de recibir calificaciones más altas en algunos tipos de cursos (optativos, seminarios y cursos de especialización), que en otros (cursos obligatorios, cursos de conferencias extensas y cursos STEM) (Aleamoni y Hexner, 1980; Stark y Freishtat, 2014).

Finalmente, existen diferencias según la disciplina objeto de evaluación, los cursos de ciencias naturales reciben calificaciones más bajas, mientras que las clases de humanidades tienen mayores valoraciones (Basow y Montgomery, 2005; Basow y Silberg, 1987).

2.5.3 Interacción con el docente

Las evaluaciones de los estudiantes pueden verse influenciadas por el relacionamiento con los docentes, es decir, aquellos que tienen una relación positiva con su docente pueden estar inclinados a otorgar una calificación elevada, a pesar de que la calidad de la enseñanza sea baja (Fernández-García, 2022). Incluso, que el docente lleve galletas o chocolate a clase, puede influir en las evaluaciones de los cursos (Hessler et al., 2018; Youmans y Jee, 2007). De igual forma, las clases compuestas por estudiantes internacionales suelen tener calificaciones más bajas debido a diferencias o poca comprensión de las interacciones relativas a la cultura (Arnold y Versluis, 2019).

Puesto que son diversas las variables que afectan a la evaluación de la actividad docente (personales, institucionales, organizacionales, etc.) es necesario analizar diferentes

perspectivas que permitan comprender el fenómeno de la evaluación docente, es por esto que se recomienda introducir en su análisis y estudio componentes contextuales (Pascual, 2007).

2.6 Estrategias usadas en la evaluación docente

Los expertos coinciden en que, si bien todas las aproximaciones metodológicas son útiles en ciertos contextos y para aspectos específicos de la evaluación, cada una también tiene sus alcances y limitaciones inherentes. Por esta razón, debería ser común en la práctica combinar múltiples enfoques metodológicos de manera complementaria. Un desafío crucial para los evaluadores radica en proponer la selección adecuada de fuentes, procedimientos y métodos para cada caso específico, una tarea fundamental según Denham y Almeida (1987).

Peterson (1997) destaca que, algunas fuentes valiosas para evaluar a los profesores incluyen las opiniones de los estudiantes, valoraciones de directivos, valoraciones de colegas y expertos, rendimiento de los estudiantes, autoevaluaciones, pruebas específicas para profesores, observaciones en el aula y portafolios.

2.6.1 Opiniones de los estudiantes

La valoración de los estudiantes es un componente importante al evaluar al profesorado universitario (De la Orden, 1990). Además, las encuestas dirigidas a los estudiantes suelen ser la única forma o la más utilizada en la evaluación de los docentes en España y Estados Unidos (De Juan Herrero et al., 2007).

Las escalas de evaluación docente emplean tanto valoraciones cuantitativas como cualitativas, evaluando dimensiones como la dirección en el aula, las habilidades de enseñanza, la personalidad del docente y la colaboración con otros (Spooren et al., 2013). A lo largo de la historia, el uso de escalas y cuestionarios de evaluación por parte de los participantes directos del proceso educativo, como estudiantes y exalumnos, ha sido predominante y sigue siendo relevante, especialmente en la educación superior. Estos métodos proporcionan información

principalmente sobre el desarrollo en clase, la comunicación profesor-estudiante y la equidad en el aula (Marsh y Hocevar, 1991).

2.6.2 Valoraciones de directivos

Los responsables de los programas y departamentos tienen la autoridad para evaluar a los profesores basándose en sus conocimientos sobre los métodos de enseñanza, técnicas de evaluación en el aula y el contenido de las disciplinas. Estos líderes pueden evaluar la eficacia de la enseñanza y revisar documentos provenientes de diversas fuentes de información.

Los informes sobre el rendimiento en todas las áreas de enseñanza proporcionan datos que los directores utilizan para evaluar al cuerpo docente (Berk, 2005). Los administradores pueden realizar una evaluación sobre la carga laboral de los profesores, la matrícula de los estudiantes en los cursos, los esfuerzos para mejorar la calidad de la enseñanza y el servicio brindado a la institución, por lo que su valoración es mucho más integral.

2.6.3 Valoraciones de colegas y expertos

Quienes se han formado y ejercido la docencia son quienes tienen la capacidad de examinar los materiales del curso, como el plan de estudios, las pruebas y los recursos en línea, entre otros. También pueden asistir a las clases y documentar sus observaciones en un informe que luego comparten con el profesor del curso (Palmer, 1998). Esta evaluación puede llevarse a cabo a través de comités encargados de revisar diversos indicadores del desempeño del profesor, como los exámenes creados por el evaluado y las opiniones de los alumnos.

Además, pueden realizar observaciones directas de las clases del profesor y entrevistas, utilizando herramientas que aseguren la imparcialidad en sus evaluaciones, como rúbricas o escalas que incluyan descripciones numéricas.

Los compañeros tienen la competencia para evaluar diferentes aspectos de la calidad de la enseñanza, tales como los objetivos, el contenido, el diseño y la organización del curso, así como los métodos y materiales educativos utilizados, que evidencian la pertinencia de las

estrategias, respecto al contenido (Elizalde y Reyes, 2008). De igual forma, los expertos pueden realizar las mismas observaciones que los colegas, dado su saber, estos pueden ser internos o externos a la institución, aunque es recomendable que sean externos para asegurar la imparcialidad del proceso de evaluación.

2.6.4 Rendimiento de los estudiantes

El rendimiento de los estudiantes es a menudo considerado como una medida de la efectividad del profesorado. Si un grupo de estudiantes mejora entre el inicio y el final del curso, se asume que el profesor ha tenido un impacto positivo. Sin embargo, este enfoque es polémico, debido a las preguntas sobre la verdadera validez de los exámenes y la variabilidad en la capacidad de los estudiantes. A pesar de las críticas, la evaluación del desempeño docente, a través del rendimiento estudiantil, ha ganado atención (Jornet et al., 2012). Aunque, se ha evidenciado que, la relación entre los resultados de aprendizaje y las opiniones de los estudiantes sobre los profesores es moderada, por lo que no se aplica en todos los casos, profesores, materias o niveles de enseñanza (Clayson, 2009).

2.6.5 Autoevaluaciones

La autoevaluación de los profesores se fundamenta en la creencia de que la reflexión personal sobre su práctica docente puede conducir a una mejora en su desempeño. Esta reflexión está orientada a analizar tanto errores como logros, lo que les permite hacer correcciones para mejorar su enseñanza en el aula. Esta metodología de autoevaluación se suele realizar a través de tres estrategias: informes escritos, cuestionarios y listas de verificación (Elizalde y Reyes, 2008).

2.6.6 Portafolios

Esta es una herramienta que guarda ejemplos de trabajos y prácticas junto con reflexiones del profesor, la cual surgió como una opción evaluativa y formativa en la educación, desde los años noventa en Norteamérica (Cerbin y Hutchings, 1993). Los portafolios pueden

incluir tanto el progreso de los estudiantes, como los logros obtenidos (Arbesú, 2009; Cisneros, 2008). A pesar de que implica una elevada inversión de tiempo, los portafolios se utilizan en algunas oportunidades con fines evaluativos y de mejora pedagógica, aunque existen debates sobre su efectividad y valor para mejorar la enseñanza (Burns, 2000). Un aspecto para considerar es que los portafolios muy rara vez se utilizan en decisiones de contratación o promoción (Seldin, 2000).

2.6.7 Observaciones en el aula

Estas observaciones de la actividad del docente en el aula pueden ser realizadas por expertos o colegas. Estas se basan en el uso de herramientas como rúbricas o escalas gráficas, que procuran asegurar la objetividad en los juicios emitidos sobre las observaciones realizadas, o apoyarse también mediante entrevistas (Elizalde y Reyes, 2008).

2.7 Dificultades de la medida del constructo: calidad o satisfacción

Aunque algunas investigaciones continúan respaldando la premisa de que los estudiantes pueden calificar la efectividad de sus docentes, la mayoría de los estudios más recientes respaldan la conclusión contraria (Beleche et al., 2012), dado que las evaluaciones de los estudiantes no evalúan realmente la enseñanza, sino que representan la percepción de los estudiantes o experiencia en un curso. Otra razón es que, cómo se presentó anteriormente, los estudiantes pueden estar influenciados por prejuicios y sesgos, lo que puede llevar a evaluaciones injustas y poco precisas. Y finalmente, los estudiantes pueden tener diferentes expectativas y preferencias en cuanto a la enseñanza, lo que puede afectar la evaluación de su profesor (Kreitzer y Sweet-Cushman, 2022).

La dificultad de los estudiantes para poder evaluar a sus docentes radica en que no tienen el conocimiento o la experiencia necesarios, así, por ejemplo, para un estudiante puede ser difícil valorar la profundidad de comprensión del docente sobre un tema, o su capacidad para adaptar su enseñanza a las necesidades de los estudiantes (Luengo, 2019). Otra razón para

dudar de la idoneidad de los estudiantes en la valoración a sus docentes está relacionada con un bajo conocimiento global de las complejidades del proceso educativo, incluyendo los objetivos del curso, el contenido del plan de estudios y las metas institucionales a largo plazo (Feistauer y Richter, 2016).

De forma complementaria, los estudiantes, suelen centrarse más en la satisfacción con el curso, que en la efectividad del proceso de aprendizaje. Es por esto, que Kreitzer y Sweet-Cushman (2022) señalan que las evaluaciones docentes son métricas deficientes de la satisfacción del aprendizaje de los estudiantes, y a su vez, son medidas imperfectas del desempeño de los docentes.

Se cuestionan también algunos intentos por validar las evaluaciones de cursos, mediante comparaciones con el rendimiento académico de los estudiantes. Dado que, las calificaciones altas en las evaluaciones docentes no garantizan un aprendizaje significativo en el aula (Clayson, 2009). Esto muestra que podría ser necesario llevar a cabo más estudios para comprender mejor la conexión entre las calificaciones de los estudiantes, los enfoques de enseñanza y los resultados de los estudiantes, antes de llegar a conclusiones definitivas sobre la validez de las evaluaciones docentes.

Por otra parte, Beleche et al., (2012) señalan que las calificaciones son propensas a ser mal interpretadas y utilizadas incorrectamente por la comunidad académica, a menos que los administradores y los miembros superiores del profesorado posean suficiente conocimiento en análisis cuantitativos.

A pesar de que las SET ofrecen una forma rápida y fácil de evaluar el desempeño docente, el manejo de los administradores con estas evaluaciones puede ser insuficiente. Franklin (2001) informó que, aproximadamente la mitad de los administradores de SET involucrados en su estudio, no pudieron responder correctamente preguntas estadísticas básicas.

Esto sugiere que algunos administradores pueden carecer de comprensión sobre cuestiones estadísticas y metodológicas fundamentales para interpretar adecuadamente las SET, como los tamaños de muestra y las tasas de respuesta de los estudiantes (Spooren et al., 2012). En este sentido, a pesar de que las calificaciones parecen precisas, a menudo pueden carecer de significado estadístico, debido a que se ignoran esas consideraciones (Stark y Freishtat, 2014).

Los expertos coinciden en que las evaluaciones docentes sólo son significativas cuando muestran diferencias notables con las normas del campus o tendencias a lo largo de varios años, es decir, las puntuaciones tienden a agruparse en los extremos superior e inferior (Flaherty, 2015).

Entre los elementos que pueden afectar la interpretación de los resultados de las evaluaciones, se han identificado problemáticas relacionadas con: a) omitir el contexto de la evaluación, como el tamaño de la clase, el nivel de los estudiantes y el contenido del curso; b) no considerar múltiples fuentes de información para evaluar la enseñanza, como observaciones en el aula, muestras de trabajo de los estudiantes y valoraciones de los colegas; c) desconocer la variabilidad natural en las calificaciones de los estudiantes; d) ignorar la posibilidad de que los resultados sean influenciados por factores externos, como la dificultad del curso o la personalidad del profesor; e) proporcionar una comunicación poco clara, acerca de los resultados de la evaluación a los profesores; f) no proporcionar oportunidades para que los profesores respondan a los resultados y mejoren su enseñanza; f) utilizar los resultados sin verificar la calidad de los datos para la toma de decisiones sobre la contratación, promoción y tenencia de los profesores, entre otros (Simpson y Siguaw, 2000) .

2.8 Dimensionalidad

La multidimensionalidad de la enseñanza ha generado debates sobre la naturaleza del SET como un instrumento unidimensional o multidimensional. Dado que la enseñanza está

compuesta por muchos aspectos, se estima que las SET deben ser instrumentos multidimensionales (Suárez et al., 2022). Sin embargo, existen autores que defienden la posibilidad de tener puntuaciones únicas y globales en esta evaluación (Apodaca y Grad, 2005). Ante esto, se cuestiona el número y las dimensiones que se puede distinguir en una enseñanza efectiva y por la posibilidad de compilar una puntuación general basada en estas dimensiones (Spooren et al., 2013).

Frente a este aspecto, la literatura carece de consenso en cuanto al número y la naturaleza de las dimensiones que deberían evaluarse (Jackson et al., 1999). La medición de estas dimensiones se basa en datos y técnicas de análisis post hoc, lo que cuestiona la validez de estos enfoques (Onwuegbuzie et al., 2009). Algunos estudios han explorado la dimensionalidad de las SET, utilizando modelos de ecuaciones estructurales y revelando relaciones complejas entre los factores, que deben considerarse al interpretar los resultados (Paswan y Young, 2002).

2.9 Evidencias de validez

En cuanto a la validez de las interpretaciones que se pueden hacer con las puntuaciones de los SET, existen preocupaciones significativas. Las diferencias en las percepciones de los docentes y estudiantes sobre una enseñanza efectiva han llevado a divergencias en las evaluaciones (Richardson, 2005). La validez del ítem y del muestreo en estos cuestionarios se ve afectada por la variabilidad en la construcción y contenido de los instrumentos, lo que a menudo resulta en una falta de validez relacionada con el contenido de la prueba (Spooren et al., 2013).

Algunos autores como Onwuegbuzie et al. (2009) indicaron a los diseñadores de cuestionarios, que el uso de puntos medios o categorías neutrales en escalas SET, puede reducir la consistencia interna de las puntuaciones. Asimismo, Sedlmeier (2006) afirmó que, la construcción de las escalas de calificación también puede influir en las puntuaciones y

Robertson (2004) concluyó que, las puntuaciones SET pueden verse afectadas por la importancia del ítem y la posición de las preguntas en el cuestionario.

Por otra parte, la validez convergente de los SET se ha evaluado contrastando las puntuaciones con el rendimiento objetivo de los estudiantes o con la percepción subjetiva de su propio aprendizaje, estos estudios llamados multisección han observado correlaciones que varían entre 0.10 y 0.47 (Cohen, 1981; Feldman, 1997), mostrando correlaciones positivas y moderadas, así como también se han evidenciado muestras de validez concurrente en estas investigaciones mencionadas.

A pesar de los esfuerzos, la generalización de los resultados se ve perjudicada por la diversidad de métodos, medidas y poblaciones utilizadas en estos estudios. Algunos autores sugieren ajustar las puntuaciones para eliminar efectos de sesgo conocidos, especialmente cuando se usan para la clasificación (McPherson, 2006; McPherson et al., 2009; Santhanam y Hicks, 2001). Además de estas preocupaciones metodológicas, las limitaciones tecnológicas, como la falta de capacitación para manejar datos electrónicos y las bajas tasas de respuesta, plantean desafíos adicionales en la interpretación y aplicación de los resultados del SET (Gamliel y Davidovitz, 2005).

Rahmatpour et al. (2019) concluye que a pesar de que las escalas de satisfacción estudiantil tienen más de 48 años de desarrollo, cada escala tiene al menos una propiedad psicométrica “deficiente” según la lista de verificación COSMIN.

Esta lista de verificación propuesta por los Estándares basados en el consenso para la selección de instrumentos de medición en salud (COSMIN), se centra en valorar la calidad de los artículos publicados en este tema. Mediante su uso, es posible valorar diferentes propiedades psicométricas (A = consistencia interna, B = confiabilidad, C = error de medición, D = validez de contenido, E = validez estructural, F = prueba de hipótesis, G = validez

transcultural, H = validez de criterio y I = capacidad de respuesta). En su estudio encontraron que ninguno de estos artículos tuvo una calidad "Excelente" en sus propiedades psicométricas.

Teniendo en cuenta las limitaciones mencionadas, existe una línea de investigación para el desarrollo de instrumentos que se centra en identificar sus evidencias de validez. Por ejemplo, la Evaluación del Desarrollo de la Instrucción y la Efectividad (IDEA) (Cashin y Perrin, 1978); la Evaluación de los Estudiantes de la Calidad de la Educación – SEEQ (Marsh, 1987; Marsh et al., 2009); la Escala de calificación de la efectividad de la enseñanza (SETERS) (Toland y De Ayala, 2005), el Informe de instrucción del estudiante (SIR II) (Centra, 1998) y el Cuestionario del curso de maestros ejemplares- ECTQ (Kember y Leung, 2008), entre otros.

A lo largo de la tesis, se analizarán las posibles consecuencias sobre los parámetros estimados y las propiedades psicométricas de los instrumentos, al utilizar los análisis tradicionales en las evaluaciones de los estudiantes, se considera al profesor como la unidad principal de análisis en una clase específica.

Los estudiantes evalúan al profesor durante una clase particular y luego se promedian los resultados para obtener una calificación general de los estudiantes en esa clase. Las preguntas clave en esta área de investigación incluyen determinar en qué medida las evaluaciones pueden diferenciar de manera precisa entre los profesores, si las calificaciones dadas por los estudiantes a un profesor en una clase son consistentes y aplicables a otras clases impartidas por el mismo profesor, y si las diferencias entre los profesores tienen validez en relación con otros indicadores de efectividad docente (Marsh et al., 2009).

3. Identificación del problema metodológico y los análisis existentes

La psicometría se ocupa de la evaluación de aspectos psicológicos y se centra en las cualidades que deben tener las mediciones, sin importar el área de aplicación ni los instrumentos utilizados. El proceso de medición puede simplificarse en una secuencia de etapas. En primer lugar, se crea un modelo probabilístico que vincula el desempeño con una variable latente. Luego, se recolectan respuestas a preguntas que se supone están relacionadas con esa variable latente, y a partir de esos datos se calculan los parámetros del modelo. Finalmente, se asigna a cada evaluado un valor probable en la escala vinculada a esa variable latente utilizando estos parámetros estimados. Este último paso se conoce comúnmente como la puntuación (Muñiz, 2018).

Actualmente, no se ha identificado un límite para el número de constructos psicológicos que pueden medirse y, pareciera que es posible construir un instrumento para cada atributo humano (Clark y Watson, 2019). Aunque, cuanto mayor es la dimensionalidad del constructo, más importante es explicar la naturaleza de su multidimensionalidad.

El surgimiento de modelos jerárquicos ha permitido entender cómo las escalas, incluso aquellas que son muy homogéneas, incorporan diversas fuentes de variación que representan distintos niveles jerárquicos. Estos modelos permiten representar, cómo en una escala de orden inferior, coexisten componentes únicos (faceta de orden inferior) y compartidos (rasgo de orden superior) (Clark y Watson, 2019). En estos casos, es posible emplear técnicas multivariadas como la regresión múltiple (Watson et al., 2013), así como el modelo bifactor para discernir las influencias específicas de estos diversos elementos (Mansolf y Reise, 2016), entre otros.

3.1 Error de medición vs puntuación verdadera

La puntuación de una prueba representa un resumen de la evidencia derivada de las respuestas de un examinado a los ítems, que están vinculados al constructo o constructos que

se están midiendo. Los modelos reflectivos consideran que hay una variable latente que se refleja en los ítems (son los modelos de variable latente, como la TRI, Factorial, etc.), mientras que, en los modelos formativos, se considera que las respuestas son una muestra de preguntas o ítems a conveniencia del investigador con la que intenta predecir una variable (ej. regresión).

Los modelos psicométricos son fundamentalmente de tipo reflectivo, es decir, el grado en que se pretenden generalizar las respuestas, dependen de la orientación teórica de la prueba. En este escenario se encuentran dos enfoques: uno empírico que considera la puntuación de la prueba como un resumen de las respuestas sin generalización, y otro que ve las respuestas como indicadores de la capacidad del examinado en niveles subyacentes de algún rasgo. Este último enfoque, se alinea con la tradición de escalamiento psicológico y la teoría de respuesta al ítem (TRI). Si bien existen diversos algoritmos para calcular las puntuaciones y los procedimientos pueden variar, todos comparten el objetivo común de utilizar, de manera óptima, la evidencia proporcionada por las respuestas a los ítems, para inferir el nivel de aptitud del examinado (Wainer y Thissen, 2001).

Existen al menos tres concepciones diferentes de puntuación verdadera (Lord y Novick, 1968; Santisteban y Alvarado, 2001). La perspectiva Platónica se asemeja al uso convencional del término, postulando que un individuo tiene un único valor "verdadero" en una medida específica, mientras que la medida observable difiere de este valor por diversas razones.

Aunque en ciencias como la física, el concepto de medida verdadera tiene sentido, algunos críticos lo consideran "místico" y carente de relevancia teórica. Por otro lado, la definición de puntuación verdadera como Valor Límite; y que se recoge en la mayoría de los manuales clásicos, implica tomar suficientes medidas para calcular el valor medio de las observaciones, esperando que este converja a la puntuación verdadera.

Finalmente, Lord y Novick (1968) en su teoría axiomática del modelo clásico definen la puntuación verdadera de un individuo *a* sobre una medida *g*, denotada como τ_{ga}, como el valor

esperado de la puntuación observada X_{ga}. De este modo la puntuación verdadera es un concepto teóricamente útil con implicaciones verificables en la obtención de resultados prácticos (Santisteban y Alvarado, 2001).

$$\tau_{ga} = E(X_{ga}) \quad (1)$$

En el modelo axiomático de Lord y Novick (1968), la variable realmente relevante es el error de medida, en cuanto a que, este tiene una distribución conocida (distribución normal), lo que hace posible conocer el rango de puntuaciones en el que se encuentra la puntuación verdadera y si, la diferencia de puntuaciones observadas entre dos personas expresa una diferencia verdadera. La naturaleza del error y cómo se maneja es la clave para una adecuada estimación de la puntuación verdadera de los sujetos.

3.2 La varianza y sus componentes

La variabilidad de las puntuaciones reales se relaciona con las auténticas diferencias entre individuos en una medición, reflejando las diferencias reales en las habilidades, rasgos o características evaluadas. Por otro lado, la variabilidad debida a errores comprende imprecisiones en la medición, incluyendo factores aleatorios o de error que no representan la verdadera habilidad o característica evaluada. La combinación de la variabilidad de las puntuaciones reales y la variabilidad debida a errores; da lugar a la variabilidad observada en las puntuaciones empíricas de un test.

Es fundamental cuantificar correctamente la variabilidad en las mediciones respecto a la variabilidad de la puntuación verdadera, lo que proporcionará información esencial sobre la precisión y la fiabilidad de la medición.

3.2.1 Varianza en la evaluación docente

Una de las medidas de fiabilidad entre evaluadores para calificaciones en escala de intervalo, es el Coeficiente de Correlación Intraclase (CCI). El CCI para las evaluaciones

docentes estaría determinado por la correlación entre evaluaciones de pares de estudiantes, determinados aleatoriamente que evalúan el mismo curso o maestro. También puede entenderse, como la proporción de la varianza total de las evaluaciones de los estudiantes que puede explicarse por cursos y profesores.

En términos del modelo de relaciones sociales de Kenny (1994), los componentes de la varianza de cursos y profesores puede verse como varianza objetivo, es decir, el grado en que diferentes evaluadores (los estudiantes) califican al objetivo (el curso o el profesor) de la misma manera.

La fiabilidad se aumenta al maximizar la proporción de la varianza objetivo. Al examinar distintas correlaciones entre profesores y cursos, específicamente al contrastar un docente que enseñaba múltiples cursos, con un solo curso impartido por distintos profesores; se descubrió que, la evaluación mediante cuestionarios resultaba más precisa, cuando se observaba un mayor impacto de los profesores en cursos similares con contenido idéntico, en contraposición a cursos con temarios diferentes (Rindermann y Schofield, 2001).

La estimación de los componentes de la varianza, mediante un análisis de varianza, con factores aleatorios, permitió concluir que la variabilidad asociada a los cursos (6%), era significativamente inferior a la relacionada con los profesores (40%). Asimismo, identificaron un impacto considerable debido a la interacción entre los profesores y los cursos (Gillmore et al., 1978).

En la mayoría de los estudios, el CCI para las evaluaciones de los estudiantes de cursos o profesores fue de aproximadamente 0.20 (Marsh, 1987; Solomon et al., 1997). Un CCI de esa magnitud indica una fiabilidad modesta de las evaluaciones individuales, pero corresponde a una precisión aceptable o incluso alta, cuando se consideran las medias de las evaluaciones, de un número considerable de estudiantes; por ejemplo, 0.90 para 25 estudiantes (Marsh y Roche, 1997).

A partir de estos estudios, surgen inquietudes sobre la influencia de los estudiantes en las valoraciones de los cursos. El enfoque de los modelos multinivel puede abordar esta cuestión de manera más completa (Raudenbush y Bryk, 2006; Richter, 2006). Los análisis multinivel permiten considerar la compleja estructura de estos datos de evaluación. En este tipo de estructura, es común que los estudiantes evalúen múltiples cursos y profesores, que los profesores enseñen en varios cursos y que, un mismo curso sea dictado por diferentes profesores, lo que genera una jerarquía de datos cruzada.

Los anteriores modelos permiten realizar una estimación de los tamaños del efecto de profesores, cursos y estudiantes, y también sus interacciones, siempre que se obtenga un número suficiente de combinaciones diferentes de unidades en los posibles niveles. Utilizando las estimaciones de los componentes de la varianza, se pueden calcular los CCI (proporciones de varianza) para las evaluaciones de los estudiantes de cursos o profesores, y para todas las demás fuentes de varianza incluidas en el modelo (como estudiantes o interacciones entre estudiantes y profesores).

Algunos autores como Spooren (2010), Rantanen (2013) y Staufenbiel et al. (2016) quienes, aplicando diferentes cuestionarios, realizaron análisis multinivel jerárquicos o de clasificación cruzada, en los cuales se obtuvieron estimaciones de los componentes de la varianza para estudiantes, cursos y profesores de magnitud similar. Ellos encontraron que, la varianza podría explicarse entre un 16% (Rantanen, 2013) y un 28% (Spooren, 2010) debido al estudiante; debido al docente en un 21% (Staufenbiel et al., 2016) y un 28% (Rantanen, 2013); debido a la interacción entre docente y curso un 25% (Spooren, 2010); así como una varianza sin explicarse entre el 47% (Spooren, 2010) y el 69% (Staufenbiel et al., 2016).

A su vez, se logró identificar que la proporción de varianza explicada por los profesores fue del 27% en las clases magistrales, pero sólo del 6% en los seminarios (Feistauer y Richter, 2016), lo que refleja la diferencia en el rol de los profesores en los seminarios y las clases

magistrales. Todos estos estudios usaron como variable dependiente una puntuación global promedio de los instrumentos utilizados.

Un aspecto relevante de la investigación de Feistauer y Richter (2016), señaló que, aunque el desempeño docente es el foco de las evaluaciones docentes, los cursos fueron una fuente de variación igualmente fuerte (16% en seminarios y 14% en conferencias), que los docentes (17% en seminarios y 14% en conferencias). Por lo tanto, los estudiantes también fueron influenciados por las características de los cursos, al calificar el desempeño docente. Estas calificaciones individuales que difieren considerablemente se ven afectadas por las características individuales de los estudiantes y, por lo tanto, están sujetas a numerosas fuentes de error.

Feistauer y Richter (2016) realizaron análisis utilizando modelos lineales de efectos mixtos, con clasificación cruzada (Baayen et al., 2008), que permitieron separar los componentes de la varianza de profesores, cursos y estudiantes. Estas tres fuentes de varianzas se incluyeron como efectos aleatorios en el análisis. Con base en los modelos de efectos aleatorios, se calcularon CCI que reflejan las proporciones de varianza debida a estudiantes, docentes y cursos. El CCI representa el cociente entre la varianza de los sujetos y la varianza total, así:

$$CCI = \frac{\sigma^2}{\sigma^2 s + \sigma^2 e} \text{ ó en sus expresiones } CCI = \frac{\sigma^2_{profesores}}{\sigma^2_{total}} \quad (2)$$

Donde σ^2 se refiere a la varianza del sujeto/ objeto/ centro o profesor y $\sigma^2 e$ se refiere a la varianza debida al error.

Los valores altos de CCI, es decir, valores cercanos a 1, indican que los estudiantes estuvieron muy de acuerdo o fueron muy homogéneos en sus valoraciones hacia los profesores. Así mismo, que distintos cursos impartidos por el mismo profesor fueron juzgados de manera

similar. Valores bajos de este CCI, es decir, valores cercanos a 0, indican que los estudiantes, las clases o ambos difieren en sus evaluaciones, lo que implica que las evaluaciones no evalúan con precisión a los docentes.

El CCI oscila entre 0 y 1, por lo que mayores proporciones de varianza entre niveles están indicadas por valores más altos del CCI. Si no se tiene en cuenta la estructura de datos multinivel, la dependencia entre los grupos violaría la supuesta independencia de las observaciones y, por lo tanto, constituyen un sesgo (Pardo y Ruíz, 2015). Que, en términos prácticos, implicaría estar considerando como factor explicativo de las puntuaciones (las habilidades docentes del profesor), parte de lo que en realidad depende de otras variables.

El componente de varianza de los estudiantes fue un poco menor a las proporciones promedio de varianza explicadas por cursos y profesores (27%). Este hallazgo tiene implicaciones que van más allá de la fiabilidad y afectan la validez relacionada con la estructura interna de la prueba, de las evaluaciones realizadas por los estudiantes. Aparentemente, las influencias sistemáticas de las características de los estudiantes fueron casi tan fuertes, como los efectos del curso o las características de los docentes. Esto sugiere que, las evaluaciones de los estudiantes no pueden considerarse estimaciones precisas de la calidad de la enseñanza, al incluir otras fuentes de error, como las características de los estudiantes.

Algunos estudiantes, evalúan a determinados profesores de manera consistentemente positiva, mientras que otros estudiantes, evalúan a los mismos maestros de manera consistentemente negativa. Por lo tanto, la forma en que un estudiante evalúa la calidad de la enseñanza parece depender, en gran medida, de la adecuación de un estudiante con su profesor. La práctica común de promediar las evaluaciones para obtener una puntuación que describa al docente esencialmente ignora la interacción estudiante-profesor y, por lo tanto, puede ser engañosa hasta cierto punto (Feistauer y Richter, 2016).

La heterogeneidad en las evaluaciones de los estudiantes puede deberse a tres factores: (1) se utiliza información diferente para hacer la evaluación, (2) los estudiantes hacen referencia a las mismas conductas en su evaluación, pero las interpretan de manera diferente (3) o se utilizan diferentes tipos de información, no conductual, como la simpatía u otras características del docente para su valoración.

Esto también puede considerarse como un problema general para las evidencias de validez de estas evaluaciones, porque los datos son casi tan informativos con respecto a los estudiantes que proporcionan las evaluaciones, como lo son, con respecto a los cursos o profesores que son el foco real de las evaluaciones (Kenny, 1994).

3.3 Satisfacción del estudiante vs calidad del docente

Generalmente, en las encuestas para evaluar al docente, se suele confundir la satisfacción que se tiene con el profesor con la calidad de su docencia; aunque son dos aspectos claramente diferentes. Por esta razón, a continuación, se amplía la diferencia entre estos dos términos, que en muchos casos se usa de la misma manera.

3.3.1 Satisfacción estudiantil

La satisfacción de los estudiantes puede entenderse como, el nivel de bienestar que experimentan estos al ver cumplidas sus expectativas académicas, producto de las acciones que realiza la institución para abordar sus necesidades educativas. También, suele entenderse como una actitud de corto plazo que refleja las percepciones de los estudiantes, acerca de hasta qué punto sus expectativas sobre diversas experiencias educativas han sido alcanzadas o superadas (Elliot y Healy, 2001). Dada la complejidad de estas expectativas, varios investigadores definen la satisfacción estudiantil como un concepto multidimensional (Hanssen y Solvoll, 2015; Jereb et al., 2018; Nastasić et al., 2019 y Weerasinghe et al., 2017).

Los estudios que consideran la satisfacción estudiantil en las universidades como un medidor de la calidad educativa son de gran relevancia, dado que dicha satisfacción muestra

relaciones significativas, tales como: mejoras en el desempeño académico (Garbanzo, 2007), disminución de las tasas de abandono (Caballero et al., 2007; Osorio y Pérez, 2010), menor tasa de cambios de carrera y, mayor éxito en el proceso de aprendizaje (Sinclaire, 2014).

Además, una calidad educativa comprobada contribuye a consolidar la imagen y el prestigio de una institución. De igual forma, la satisfacción de los estudiantes puede abordarse desde tres enfoques: el relacionado con el bienestar emocional, el ámbito laboral y la perspectiva del consumidor (Melo et al., 2015). En consecuencia, la alteración de alguna de estas áreas provoca una sensación de insatisfacción en el estudiante (Flores-Mamani y Arce-Ortiz, 2019).

Dentro de los aspectos académicos, el clima educativo, tiene gran relevancia en los logros y la satisfacción del estudiante, representando la combinación de todas las experiencias dentro del aula universitaria (Palomer et al., 2018). Otros factores importantes que afectan la satisfacción implican la calidad percibida de la enseñanza, la retroalimentación proporcionada por los profesores, los estilos de enseñanza, la calidad de las experiencias de aprendizaje y el tamaño de la clase (Nastasić et al., 2019; Paul y Pradhan, 2019; Weerasinghe et al., 2017).

Wong y Chapman (2023) destacan que las interacciones informales entre estudiantes son un indicador predictivo de su nivel de satisfacción en la educación superior; y que factores como la edad y el género tienen una asociación significativa con las interacciones formales entre estudiantes y profesores. Esto se alinea con las observaciones de Burnett et al. (2007), quienes encontraron que tanto la frecuencia, como la intensidad de las interacciones entre estudiantes, docentes y compañeros son cruciales para determinar la satisfacción estudiantil.

En síntesis, lo que se pretende lograr con la aplicación de encuestas de satisfacción estudiantil, es identificar qué tan satisfechos se encuentran los estudiantes con la educación que reciben (Chuyma-Huilca et al., 2021), y de esta manera, identificar las áreas de oportunidad en las cuales enfatizar la mejora. Dado que la satisfacción estudiantil está íntimamente ligada a

las actitudes y expectativas, es razonable esperar que la satisfacción estudiantil evolucione a lo largo del tiempo, lo que sugiere que debe ser monitoreada de manera constante (Wong y Chapman, 2023).

3.3.2 Calidad o excelencia docente

La excelencia docente es un concepto controvertido y las definiciones pueden variar según los contextos sociales, económicos y políticos (Skelton, 2004), lo que aumenta la complejidad de la evaluación de los profesores y la identificación de instrumentos que evalúen la excelencia.

Los docentes son un punto clave en el proceso de aprendizaje (Hanushek, 2016) porque determinan con sus comportamientos, que las actividades de enseñanza-aprendizaje seleccionadas, orienten la construcción del aprendizaje de los estudiantes en una determinada dirección.

El análisis comparativo de la enseñanza efectiva, realizado por varias regiones europeas, utilizando el instrumento de observación desarrollado por van de Grift y Lam (1998); un instrumento fiable y válido para comparar la calidad de la educación, independientemente de las diferencias culturales entre países. Fue integrado como parte del Proyecto de Análisis del Aprendizaje y la Enseñanza (ICALT), y dio como resultado un test, con una estructura final de seis dominios, que aparecen repetidamente en varias de las evaluaciones: clima de aprendizaje seguro y estimulante, gestión eficiente del aula, claridad de la enseñanza, enseñanza activadora, estrategias de enseñanza-aprendizaje y diferenciación (Fernández-García et al., 2022).

Lo anterior basado en que, un *clima de aprendizaje seguro y estimulante* es relevante por su influencia en los resultados del aprendizaje y la participación de los estudiantes en el proceso (Reyes et al., 2012). *La gestión eficiente en el aula* permite que el docente alcance sus objetivos más fácilmente, ya que incluye la gestión del tiempo, la presentación ordenada del

contenido, garantizar el respeto por el inicio y fin de la sesión, establecer equilibrio entre las actividades individuales y grupales y abordar con eficacia la conducta de los estudiantes (Danielson, 1996; Oliver y Reschly, 2007; van de Grift, 2007).

La *claridad de la enseñanza* implica aspectos de la calidad, como dar instrucciones por etapas, dejar claro si una respuesta es correcta o incorrecta y comprobar de forma continua si los estudiantes han entendido de qué se trata la lección. Por lo tanto, los estudiantes no aprenderán tanto, si las instrucciones son poco claras (Maulana et al., 2017). Esta categoría incluye que los objetivos sean explícitos, se realice retroalimentación sobre los errores y, finalmente, se expliquen los procedimientos requeridos en cada clase (Blaich et al., 2016; Maulana et al., 2015; van de Grift et al., 2014).

La *enseñanza activadora* implica que los estudiantes sean conscientes sobre su aprendizaje, conecten sus conocimientos previos y utilicen sus procesos mentales complejos (Bonwell y Eison, 1991). Implica también, adecuadas relaciones entre estudiantes y profesores (Maulana et al., 2015). *Las estrategias de enseñanza-aprendizaje* requieren que se usen técnicas de enseñanza variadas, para adaptarse a los diferentes estilos de aprendizaje de los estudiantes.El uso de estrategias metacognitivas, fomentando la agencia en el estudiante (Jensen et al., 2018), para que pueda ejercer opciones, asumir responsabilidades, asumir diferentes roles e internalizar expectativas de aprendizaje y lograr habilidades de aprendizaje más avanzadas (Carriedo, 1995; Maulana et al., 2015).

Finalmente, *la diferenciación* se centra en la inclusión de cualquier estudiante, independientemente de su capacidad (De Jager, 2011), y también, la introducción de dimensiones culturales particulares, tales como, el conocimiento local de los estudiantes y las experiencias fuera del aula (Jensen et al., 2018).

3.4 Equidad en la evaluación y especificación del constructo

Sulis et al. (2019) evidenciaron cómo las SET basadas en medidas no ajustadas de las calificaciones de los estudiantes, puede conducir a resultados sin significado, cuando se utilizan para hacer comparaciones entre profesores.

Después de 50 años de investigación, todavía existe confusión sobre qué factores específicos contribuyen a las puntuaciones producidas por estas evaluaciones, y cómo interpretarlas (Gravestock y Gregor-Greenleaf, 2008). Determinar qué representan las puntuaciones de evaluación de los estudiantes, es importante porque estos datos se utilizan con frecuencia en decisiones muy relevantes para los profesores universitarios, como aumentos, retención, promoción y titularidad (Boysen, 2015; Emery et al., 2003; Gravestock y Gregor-Greenleaf, 2008).

Las respuestas aleatorias, descuidadas, o realizadas con un esfuerzo insuficiente, podrían ser una fuente de error de medición que limite la calidad de los conjuntos de datos obtenidos (Barnes, 2016; Furnham et al., 2015). El uso de evaluaciones de estudiantes en el proceso de revisión del profesorado supone que el profesor del curso es el principal responsable de las diferencias en las calificaciones; y que los estudiantes aprenden más de profesores altamente calificados. Sin embargo, muchos ven estas calificaciones con escepticismo, señalando que las calificaciones de los estudiantes pueden no ser indicadores válidos de la calidad de los instructores o de la eficacia de la enseñanza (Nasser y Fresko, 2002; Spooren et al., 2013; Uttl et al., 2017).

Las evaluaciones que se basan en una instancia única o una única fuente, y que conducen a posibles consecuencias negativas para las asignaciones de los profesores, no ofrecen oportunidades para que los profesores mejoren (Curby et al., 2019).

En un sistema de evaluación perfecto, toda la variación en las calificaciones se debería al profesor y específicamente a la calidad de la enseñanza. La variación atribuible al profesor

representa el grado en el que diferentes instructores, de los mismos cursos, difieren en sus calificaciones; independientemente de la influencia de las cualidades individuales que influyen en esas calificaciones (por ejemplo, edad, género, estilo de enseñanza).

Cuando las calificaciones del curso se utilizan para evaluar a los profesores, se supone que la variación respecto al docente debe representar la mayoría de las calificaciones en la evaluación del curso. Es preocupante que, incluso si una gran parte de la variación se debe al profesor, existen atributos que influyen en las calificaciones de los cursos y que pueden no tener que ver con la calidad de la enseñanza, tal como se explicó ampliamente en el capítulo 2, relacionados con la edad (Wilson et al., 2015), el género (Boring et al., 2016; MacNell, et al., 2015), el atractivo físico (Hamermesh y Parker, 2005), la experiencia (Zabaleta, 2007) y la personalidad (Delucchi, 2000), entre otras.

Es importante considerar que, ninguna de estas características se refiere a la eficacia docente o calidad de la enseñanza. Adicionalmente, un conjunto de estudiantes que toman un curso puede variar en motivación (Chen y Hoshower; 2003), expectativas de calificaciones (Zabaleta, 2007), distribución de género (Bonitz, 2011), promedio de calificaciones colectivo (Griffon et al., 2013), y otros factores. También, la ocasión, referida al momento del día o de la semana en que se ofrece el curso, el aula física y los eventos externos que podrían hacer que un curso sea más o menos destacado (Gravestock y Gregor-Greenleaf, 2008).

En el estudio de Curby et al. (2019), para estimar las fuentes de varianza, emplearon el enfoque de la teoría de generalizabilidad (Cronbach et al., 1963). Cada díada se analizó utilizando un procedimiento de análisis de varianza estándar (ANOVA), pero en lugar de pruebas de significancia, utilizaron las estimaciones de varianza (cuadrados medios o MS) para cada "faceta" (es decir, profesor, curso y ocasión) para un análisis más detallado. El error fue el componente más importante de las calificaciones, representando casi una cuarta parte de la

variación, lo que sugiere que había una falta sustancial de fiabilidad en las percepciones de los estudiantes sobre la enseñanza en general.

Estos resultados indican que es probable que múltiples factores, diferentes al profesor, influyen en la calificación de los estudiantes. Estas puntuaciones de evaluación pueden interpretarse como una estimación aproximada de la satisfacción del estudiante y no deben tomarse como únicos indicadores de la calidad de la enseñanza. Además, estas calificaciones dicen algo sobre el docente, pero la mayoría de la varianza es atribuible a interacciones bidireccionales, no a un efecto principal.

Se encontraron diferencias consistentes entre algunos profesores, pero esas diferencias eran pequeñas, puede ser mejor pensar en los profesores como efectivos o ineficaces para enseñar un curso en particular, pero no como un atributo global general de este (Hildebrand, 1973). Finalmente, los cursos y las ocasiones tienen mayores interacciones que los efectos principales, esto quiere decir que enseñar un curso difícil o a un grupo de estudiantes difíciles, puede no alejarse tanto de la calificación promedio, porque es probable que esos factores interactúen con otros factores de una manera un tanto impredecible (Curby et al., 2019).

Las investigaciones que se centran exclusivamente en estas evaluaciones podrían pasar por alto las interacciones que influyen en las evaluaciones finales de un curso. Las calificaciones de los cursos podrían aportar una perspectiva más valiosa, al analizar qué combinaciones pueden resultar exitosas. Por ejemplo, ofrecer ciertos cursos por la mañana podría generar entornos de aprendizaje más efectivos que en horarios nocturnos. Evaluar la capacidad de un profesor basándose solo en un factor principal, cuando las interacciones son responsables de la mayoría de la variabilidad en los resultados, proporciona una visión limitada, e incluso potencialmente errónea, de la verdadera habilidad del profesor.

Por otra parte, autores como Braga et al. (2014), señalan que las altas calificaciones en las evaluaciones de cursos no son sinónimo de enseñanza de alta calidad, sino en cierta medida

un índice de satisfacción del consumidor, en este caso, el estudiante. Si bien algunas características de los profesores influyen más que otras en la experiencia de los estudiantes (Pepe y Wang, 2012), examinar la variación general de los profesores, parece ser un camino útil para determinar, en qué medida, una diferencia particular en la calidad de los profesores influye realmente en las evaluaciones de los estudiantes.

Los hallazgos de Curby et al. (2019) indican que, si bien el profesor contribuye a parte de la variabilidad en las calificaciones de los cursos de los estudiantes, la única influencia (es decir, el efecto principal) del docente, no explica suficiente variación para atribuir la puntuación general únicamente a la calidad de la enseñanza. Por lo tanto, la interacción entre múltiples facetas que influyen en las calificaciones de los cursos de los estudiantes debe tenerse en cuenta al interpretar las calificaciones como representativas de la calidad de la enseñanza.

Aunque se obtiene cierta información sobre el profesor en las calificaciones de final de curso (Stroebe, 2016), la cantidad de errores y las interacciones bidireccionales, deberían hacer reflexionar a cualquiera que tome decisiones importantes, sobre la carrera de un profesor, basado únicamente en la evaluación realizada por los estudiantes, esto puede entenderse como el sesgo de fuente única, el cual, se suele pasar por alto y, si no se reconoce, puede confundir la interpretación de los resultados (Baugh et al., 2006).

Los SET generalmente se utilizan para identificar áreas tanto del profesor, como de aspectos administrativos, reflejando si el profesor cumple o no cumple con los estándares considerados importantes por los autores de un instrumento de evaluación. Sin embargo, con los efectos del sesgo de fuente única, múltiples constructos medidos con un instrumento de evaluación no son necesariamente identificables estadísticamente, y en ausencia de pruebas concluyentes; se asume su existencia sin fundamentos sólidos. Esta suposición tiene consecuencias en las que una covariación significativa, debido al sesgo de fuente única, indica

una valoración para el profesor que habría sido diferente con otros métodos de recopilación de información (Garger et al., 2019).

Además, como se ha señalado anteriormente, está presente la dificultad de los estudiantes para diferenciar los rasgos de enseñanza, de la simpatía emocional generada por un profesor. Por lo que, centrarse únicamente en el instrumento descuida o aborda parcialmente una cuestión básica de la recopilación de datos: la fuente de datos (Garger et al., 2019).

Sproule (2000) describió posibles falacias estadísticas asociadas con las evaluaciones de los estudiantes respecto a la enseñanza. El sesgo de fuente única podría evitarse mediante el uso de varios métodos de recopilación de información referente a la enseñanza. Teóricamente, un buen instrumento de medición debería estar compuesto por ítems que produzcan medidas fiables de la ubicación de los estudiantes, en términos de su satisfacción, y de los docentes, en términos de la calidad de la enseñanza, a lo largo de los rasgos latentes, reflejados en las dimensiones de la calidad de la enseñanza evaluadas con el cuestionario (Sulis et al., 2019).

3.5 Importancia de la claridad de la unidad de análisis de interés

Para construir indicadores significativos de las respuestas de los estudiantes a los ítems de un cuestionario de evaluación docente, se deben ensamblar cumpliendo algunas restricciones: (1) los ítems deben definir el mismo rasgo latente (unidimensionalidad); de lo contrario, se debe considerar la dimensionalidad de los ítems para elaborar el modelo multidimensional adecuado; (2) debe evaluarse la estructura de dependencia entre las respuestas a los ítems y las características de las unidades relevantes (a nivel de estudiante, curso, clase, profesor, etc.); y (3) el efecto de posibles factores de confusión (por ejemplo, tipo de escuela secundaria a la que asistió, actitud negativa hacia un tema específico, tamaño de la clase, etc.), debe considerarse, siempre que el objetivo del análisis sea evaluar la contribución

del profesor, al hacer comparaciones entre profesores, sobre la base de indicadores de calidad (Draper y Gittoes, 2004; Leckie y Goldstein, 2009).

La detección de una relación significativa del rasgo latente, es decir, la calidad de la enseñanza, con características de los estudiantes (o profesores) que son externas al proceso de evaluación, es una señal de que factores exógenos pueden haber influido potencialmente en las calificaciones observadas (Boring et al., 2016; Fayers y Hand, 2002).

3.6 El modelo multinivel

Muchos tipos de datos tienen una estructura jerárquica, anidada o agrupada, por ejemplo, hijos de los mismos padres tienden a ser más parecidos en sus características físicas y psicológicas, que aquellos elegidos aleatoriamente de la población en general (Goldstein, 2011). El reconocimiento adecuado de estas jerarquías naturales nos permite obtener resultados más satisfactorios al interpretar datos. Dado que, una jerarquía se entiende como aquella formada por unidades agrupadas en diferentes niveles.

Así, por ejemplo, los estudiantes pueden ser unidades de nivel 1 agrupadas o anidadas dentro de escuelas, que están en unidades de nivel 2. La existencia de tales jerarquías de datos no es accidental, ni puede ignorarse, dado que ello genera el riesgo de pasar por alto la importancia de los efectos de grupo, y también puede invalidar los resultados obtenidos mediante las técnicas tradicionales de análisis estadístico, utilizadas para estudiar relaciones de datos. El análisis multinivel plantea cuestiones conceptuales y estadísticas, qué son importantes para una consideración cuidadosa, teniendo en cuenta que un modelo multinivel es un modelo de covarianza (Mehta y Neale, 2005; De Leeuw et al., 2008).

Los modelos multinivel o modelos lineales jerárquicos se caracterizan por una estructura de datos anidada en la que hay múltiples niveles o unidades de análisis (por ejemplo, estudiantes anidados en aulas, aulas anidadas en escuelas, etc.). Muthén (1991) presentó un modelo de análisis factorial multinivel de variable latente (Hox, 2002; Kamata et al., 2008);

mediante este análisis es factible analizar numerosos modelos, que pueden incluir situaciones en las que a) hay igual cantidad de factores en cada nivel con cargas iguales entre ellos; b) se replican las mismas estructuras factoriales, pero con diferentes cargas en los niveles; o c) un número diferente de factores aparecen en los dos niveles. La evaluación de estos modelos factoriales, a través del análisis factorial multinivel, se convierte en una manera de recolectar evidencia empírica y teórica, en el proceso de validación de constructos de naturaleza multinivel (Kim et al., 2016).

Como señaló Sirotnik (1980) "el problema de la unidad de análisis no es un problema estadístico" (p. 246, citado en Kim et al., 2016); es por esto que autores como Zumbo y Forer (2010) han establecido un compilado sobre la validación de constructos multinivel, que se centra en el significado sustantivo de los constructos. Especialmente, en los casos en que los datos se recopilan a un nivel inferior de análisis (p. ej., estudiantes), pero se hacen inferencias sobre una unidad de análisis de nivel superior (p. ej., escuela); situación, que también ocurre, cuando se recolectan datos de los estudiantes para realizar análisis sobre la calidad de la docencia.

Stapleton et al. (2016) han identificado cuatro tipos de constructos multinivel, en función del nivel en el que los constructos son conceptualmente significativos (a) constructo con el individuo como unidad de interés, (b) constructo de clúster compartido, (c) constructo de clúster configural y (d) constructo de clúster configural y compartido simultáneamente.

Los constructos de clúster compartido y configural también se denominan constructos reflexivos (o de composición) y formativos (o de compilación), respectivamente (Bliese, 2000; Lüdtkeet al., 2008). Cuando los individuos son las unidades de interés y el constructo está conceptualmente en el nivel 1, Stapleton et al. (2016) argumentan que no puede existir ninguna configuración verdadera en el segundo nivel, y, por lo tanto, especificar una estructura factorial en el nivel 2 puede no ser apropiado.

A partir de datos multinivel, Stapleton et al. (2016) recomiendan un enfoque basado en un diseño en el que los errores estándar se ajustan para tener en cuenta la dependencia de los datos o un Análisis Factorial Confirmatorio Multinivel, con un modelo saturado en el Nivel 2.

Un constructo de conglomerado compartido es fundamentalmente un constructo de Nivel 2, pero se mide por las respuestas de los individuos a los ítems (por ejemplo, la calidad de la enseñanza docente percibida por los estudiantes). En este contexto, los individuos actúan como informantes del constructo y, por ende, representan fuentes intercambiables de información sobre dicho constructo de grupo. Las respuestas de los individuos, dentro de los grupos, que reflejan un constructo compartido, deben demostrar un alto grado de acuerdo dentro de ese grupo (Kim et al., 2016), por lo que diferentes tipos de constructos multinivel requieren diferentes especificaciones de modelo.

La importancia de interpretar las estimaciones de los parámetros y las estadísticas asociadas, como la fiabilidad de la escala, qué varía según el nivel específico en el contexto multinivel. Por tanto, es crucial que se realice una validación del constructo empleando métodos multinivel, adoptando modelos coherentes con la conceptualización multinivel del constructo, antes de realizar un Análisis Factorial Multinivel (MFA por sus siglas en inglés). Es necesario identificar en qué nivel está el constructo de interés, su relevancia, el tipo específico de constructo en cuestión y, posteriormente, cómo estimar y especificar el modelo para ese constructo.

Hox (2010), distinguió dos enfoques para el análisis factorial multinivel. El primero implica la división de la matriz de covarianza en dos: una para el nivel individual (covarianza interna) y otra para el nivel grupal (covarianza entre grupos). En este sentido hay dos modelos factoriales, uno para el nivel individual (dentro) y otro para el nivel grupal (entre), los cuales pueden ser modelados de manera independiente o simultánea. Por otro lado, el segundo enfoque consiste en modelar directamente los datos multinivel observados. Este modelo

incluye variables en cada nivel de análisis, y las pendientes e intersecciones se estiman considerando la variación a nivel grupal.

3.6.1 Dentro del grupo y entre los grupos

Un supuesto del modelado estructural multinivel es que la población de individuos está anidada dentro de grupos. La puntuación del individuo se puede describir como un vector Y_{ij}, donde i es el subíndice del individuo y, j es el subíndice del grupo. Este vector Y_{ij} comprende que el grupo desagregado significa $\overline{Y}_j$, que representa el nivel del grupo, y la desviación individual de la media del grupo $Y_{ij} - \overline{Y}_j$, que representa el nivel individual.

La puntuación individual total $Y_T = Y_{ij}$ se divide en un componente entre grupos Y_B y un componente dentro del grupo Y_W:

$$Y_T = Y_B + Y_W \quad (3)$$

En esta ecuación, el componente entre grupos es igual al grupo desagregado, significa ($Y_B = \overline{Y}_j$) y el componente dentro del grupo es igual a la desviación individual de la media del grupo ($Y_W = Yij - \overline{Y}_j$).

El cálculo matricial utilizado en SEM se basa en este principio de descomposición. Las medias del grupo desagregado $\overline{Y}_j$ se utilizan para calcular la matriz de covarianza de grupos (Σ_B), y la desviación individual de las medias entre el grupo.

$Y_{ij} - \overline{Y}_j$ se utilizan para calcular la matriz de covarianza dentro de grupos Σ_W.

La siguiente ecuación formula la descomposición de la matriz de covarianza poblacional:

$$\Sigma_T = \Sigma_B + \Sigma_W \quad (4)$$

Los datos de muestra se pueden descomponer utilizando la misma lógica:

$$S_T = S_B + S_O \quad (5)$$

En el modelado de ecuaciones estructurales multinivel, las matrices de covarianza empíricas S_B y los S_W describen dos modelos separados: Un modelo para la estructura entre grupos y otro para la estructura dentro de los grupos.

Se necesitan matrices de covarianza poblacional Σ_B y Σ_W para estimar el parámetro del modelo, coeficientes de trayectoria, cargas factoriales y varianzas residuales. Sin embargo, los datos disponibles son los de la matriz empírica entre grupos S_B y la matriz empírica dentro de grupos S_W. Las matrices no se pueden utilizar simplemente para estimar las matrices de covarianza de la población, es decir que, se hace una inferencia desde lo muestral a lo poblacional, lo que implica problemas de representatividad de las muestras utilizadas.

3.7 Análisis Factorial Confirmatorio Multinivel

Muthén (1994) formuló un protocolo para el desarrollo de un análisis factorial confirmatorio multinivel (AFCM). En este, desglosó el proceso en cinco pasos: los primeros cuatro pasos exploran la estructura factorial y justifican el enfoque multinivel, mientras que el quinto paso realiza el análisis real.

Así que, el proceso comienza con una prueba convencional, utilizando la matriz de covarianza de la muestra total y luego se procede a estimar modelos utilizando matrices separadas para dentro del grupo y entre los grupos. Evalúa la idoneidad del enfoque multinivel y utiliza la proporción de variación entre grupos para detectar su naturaleza multinivel.

Posteriormente, lleva a cabo dos estimaciones separadas: una para la matriz dentro del grupo y otra para la matriz entre grupos, antes de realizar un análisis factorial simultáneo, que debe ser especificado como un modelo multigrupo (Dyer, 2005). En el último paso, realiza el análisis factorial confirmatorio multinivel simultáneamente. Por lo tanto, el análisis factorial confirmatorio multinivel debe especificarse como un modelo multigrupo.

Hox (2010) también utilizó un procedimiento gradual para estimar un análisis factorial confirmatorio multinivel. El proceso comienza con la estimación de un modelo a nivel individual utilizando la matriz de covarianza S_{PW}. Inicialmente, se realiza un análisis factorial exploratorio para determinar el número de factores latentes. Luego, se procede con un análisis factorial confirmatorio para validar la estructura encontrada, comparándola con otras

estructuras similares. Finalmente, se emplea la matriz de covarianza S_B y realiza una estimación de máxima verosimilitud para la matriz de covarianza poblacional entre grupos Σ_B, para realizar análisis separados.

3.7.1 Modelos

Para determinar la presencia de estructuras entre grupos, se inicia un análisis de nivel intermedio estimando modelos de referencia. El primer modelo de referencia, conocido como el modelo nulo, examina si existe alguna estructura multinivel en los datos. Si el modelo nulo no es rechazado, indica que no hay una estructura significativa a nivel de grupo, y, por ende, un análisis a un solo nivel sería apropiado.

El segundo modelo, el modelo de independencia, no especifica covarianzas a nivel de grupo, únicamente considera variaciones. Si el modelo de independencia no es rechazado, indica que la variación observada a nivel de grupo no puede ser explicada mediante un modelo estructural. En caso de que el modelo de independencia sea rechazado, sugiere la presencia de algún modelo estructural a nivel de grupo. El modelo no restrictivo para el nivel de grupo es el modelo saturado. Este modelo ajusta una matriz de covarianza completa a las observaciones a nivel de grupo, manteniendo el modelo interno a nivel individual (Muthén, 1994).

En comparación con el AFC de un solo nivel, el AFCM permite a los investigadores considerar ambos niveles de datos simultáneamente. Más específicamente, usar el AFCM implica dividir la matriz de la covarianza de la población total, Σ_T, en una matriz dentro de covarianza, Σ_W, y una matriz entre covarianzas, Σ_B, para estimar los efectos, tanto dentro como entre grupos.

Los dos componentes de la varianza son ortogonales y aditivos, lo que significa que la relación entre variables entre grupos no tiene que ser la misma, que la relación que existe dentro de los grupos.

Utilizando datos de muestra, la matriz de covarianza total (o general), S_T, también se puede descomponer en Matrices S_B y S_W. Sin embargo, ejecutar un AFCM utilizando las matrices S_B y S_W para estimar tanto Σ_W como Σ_B no es tan sencillo (Hox 2002). En cambio, para dos covarianzas de muestra, es necesario definir matrices: S_{PW}, la matriz de covarianza interna agrupada y S_B, la matriz entre matriz de covarianza grupal.

La matriz S_{PW} es una estimación insesgada de la población dentro de la matriz de covarianza de los grupos (Σ_W) (Muthén, 1994). La matriz de covarianza agrupada dentro se calcula mediante:

$$S_{PW} = (n - G)^{-1} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (y_{ig} - \bar{y}_g)(y_{ig} - \bar{y}_g)' \quad (6)$$

Donde *n* es el tamaño total de la muestra, *G* es el número de grupos, y_{ig} es la puntuación de la observación i anidado en el grupo g; y, y_g es la media específica del grupo en el grupo *g*. S_{PW} también es equivalente, a la matriz de covarianza de las puntuaciones de desviación individual de las medias del grupo, con la excepción de que el denominador es n−G en lugar de n−1.

El análisis factorial de la matriz S_{PW} es sencillo y no presenta ningún desafío de modelado. Una forma sencilla de generar el S_{PW} puede ser por media grupal, centrando todas las variables de interés, generando una matriz de covarianza, usando las variables centradas, multiplicando la matriz de covarianza por n − 1 y luego dividiendo el producto por n - G.

La matriz de covarianza muestral entre grupos S_B se puede calcular usando:

$$S_B = (G - 1)^{-1} \sum_{g=1}^{G} n_g (y_{ig} - \bar{y}_g)(y_{ig} - \bar{y}_g)' \quad (7)$$

donde *y* representa la gran media general. S_B se puede calcular generando una matriz de covarianza utilizando las puntuaciones de desviación de las medias del grupo, repetidas del total de la gran media, multiplicando la matriz por n − 1 para calcular las sumas de cuadrados, y luego dividiendo nuevamente por G−1. Desafortunadamente, S_B es un estimador sesgado de Σ_B, y en realidad estima una combinación de Σ_W y Σ_B, así:

$S_B = \Sigma_W + \zeta\Sigma_B$ (8)

donde ζ representa el tamaño promedio del grupo (Muthén, 1994).

Para casos desequilibrados (que es el caso más frecuente), y en muchos casos, ζ será aproximadamente n/G. Como resultado, Σ_B puede ser aproximadamente estimado por ζ-1 (S_B-S_O). El valor esperado entonces de Σ_B está compuesto por una unidad de varianza dentro del grupo y ζ unidades de varianza entre grupos.

Z se calcula como:

$$Z = \zeta\left[n^2 - \sum_{g=1}^{G} n_g^2\right][n(G-1)]^{-1} \quad (9)$$

Según Muthén (1994), los Modelos de Análisis Factorial Confirmatorio (MCFA) podrían no siempre converger, por lo que los expertos han sugerido pasos detallados para construir y depurar los modelos.

Muthén (1994) propuso dos métodos similares, mientras que Hox (2002) también ofreció un enfoque simplificado. Aunque el método de Muthén es el más utilizado, el de Hox (2002) se considera más directo (Selig et al., 2008). Ambos métodos requieren las matrices S_{PW} y S_B, así como el factor de escala estimado ζ.

A continuación, se describen los pasos propuestos por Hox (2002), aunque esta misma configuración funciona para la metodología propuesta por Muthén (1994):

Paso 1: el modelo de nivel uno.

El primer paso propuesto por Hox (2002) es realizar un análisis factorial utilizando únicamente la matriz S_{PW}, ignorando S_B. El tamaño de muestra efectivo para el análisis es n – G. Si no se encuentra un ajuste adecuado, se continúa y sería necesario revisar la teoría detrás del CFA. Se realiza una prueba utilizando un modelo básico de uno y dos factores, siguiendo tres pasos: 1) especificar el modelo, 2) ajustar el modelo y 3) ver el modelo. Cuando se centran las variables en la media del grupo, se eliminan los efectos a nivel de grupo. Esta técnica no

difiere mucho de un modelo CFA estándar, solo que se utiliza S_{PW} en lugar de S_T y con un número de observaciones reducido (n - G).

Paso 2: el modelo nulo.

Para el paso 2, se especifica un modelo nulo donde se utilizan las matrices S_{PW} y S_B en una configuración multigrupo utilizando la estructura factorial definida en el paso 1, en ambas matrices, con igualdad de restricciones. Realmente no hay dos grupos, pero la configuración multigrupo se puede utilizar para analizar simultáneamente las matrices "dentro del grupo" y "entre grupos". Además de garantizar las cargas factoriales, es necesario que la varianza y la covarianza de cada variable y factor latente sean consistentes en ambos grupos. Un ajuste deficiente del modelo sugiere la presencia de variación entre grupos. Si el modelo nulo muestra un ajuste aceptable, se podría concluir que no hay una variación estadísticamente significativa a nivel de grupo (Stapleton, 2006).

Paso 3: El modelo de independencia.

En este paso, se propuso estimar un modelo de independencia a nivel de grupo. Como $S_B = \Sigma_W + \zeta\Sigma_B$, se estimar la parte intermedia del modelo creando nuevos “factores” a nivel de grupo. El paso 3 ahora requiere el uso de ζ factor de escala y cada varianza de la variable manifiesta se compone de una unidad para S_{PW} y una porción de varianza $S_{B.}$ Cada nuevo factor de nivel de grupo tiene una carga específica de $\sqrt{\zeta}$ a su correspondiente variable manifiesta (Muthén, 1994).

Si el modelo de independencia se ajusta adecuadamente, se concluye que hay una variación significativa entre grupos, pero no existe un modelo estructural relevante (Hox, 2002). Por otro lado, si el modelo de independencia no se ajusta bien, sugiere la presencia de algún modelo estructural a nivel de grupo que requiere ser desarrollado.

Paso 4: el modelo saturado.

Para probar el modelo saturado, ahora se permiten covariar entre sí las variables latentes definidas en el paso 3. Es importante destacar que, en este paso, dado que las variables solo existen en el nivel 2, la varianza debe estimarse sólo para el nivel 2. El ajuste del modelo en el paso 4 debe ser similar al ajuste en el paso 1, ya que todos los grados de libertad en el nivel intermedio se utilizan en un modelo completamente saturado. Si el ajuste resulta ser pobre en el paso 4, esto podría indicar un error, o que, el modelo en el paso 1 tenía deficiencias iniciales (Stapleton, 2006). Si el investigador busca modelar relaciones entre las variables del nivel 2, estas hipotéticas relaciones se pueden evaluar en el siguiente paso.

Es importante señalar que en el enfoque de Muthén (1994) para los MCFA, no se estiman los modelos nulos, de independencia y saturado.

Paso 5: El modelo hipotético.

En esta etapa, se especifica y prueba el modelo teórico de medición a nivel 2. Aquí, la estructura de covarianza definida en el paso 4, que representa el modelo saturado entre todas las variables de nivel 2, será reemplazada por el modelo hipotético con un factor general. No obstante, se busca una solución más concisa y parsimoniosa del modelo por lo que se evalúa si es este o un modelo anterior el mejor para ser seleccionado.

3.7.2 Invarianza factorial

La invariancia entre niveles o invarianza de factores entre niveles es esencial para la validez del conglomerado configural porque el constructo en el Nivel 2 “simplemente refleja el agregado del conglomerado del constructo individual en el nivel 1” (Stapleton et al., 2016). También se espera invariancia entre niveles cuando un constructo se representa a nivel individual, pero tiene algunos efectos de agrupamiento espurios que deben modelarse en el Nivel 2, puesto que, las relaciones del Nivel 2 solo representan las relaciones del nivel 1 (Stapleton et al., 2016).

Así, si se mantiene la invarianza métrica o las cargas factoriales del Nivel 1 son iguales en todos los conglomerados, entonces las cargas factoriales del Nivel 2 deberían ser idénticas a las del nivel 1.

La violación de la invarianza entre niveles se puede interpretar como un sesgo de conglomerado, es decir, cargas factoriales no invariantes entre conglomerados (Jak et al., 2013). Además, el cálculo del CCI requiere la igualdad de cargas factoriales en todos los niveles porque la métrica utilizada en el Nivel 1, y el Nivel 2 debe ser consistente para tomar una relación entre la varianza factorial del Nivel 2 y la varianza factorial total (es decir, la suma de las varianzas de los factores de nivel 1 y 2).

Es importante informar, de ser posible, la invarianza entre niveles y cómo se prueba esta. Debe tenerse en cuenta que cuando los factores no son invariantes entre niveles, el CCI no puede estimarse. De igual forma, el significado de los factores especificados en un AFC o extraídos de un AFE debe revisarse para cada nivel.

3.7.3 Fiabilidad

Dado que la fiabilidad se define como el cociente de la varianza de la puntuación verdadera sobre la varianza total observada (Lord y Novick, 2008), estimar la fiabilidad con datos multinivel es complejo porque las variaciones se descomponen dentro y entre grupos. Geldhof et al., (2014) han señalado la importancia de estimar la fiabilidad específica del nivel: estimación de la fiabilidad en cada nivel, de datos multinivel, porque la fiabilidad de un solo nivel es “difícil de interpretar cuando la fiabilidad no es idéntica en todos los niveles, promediando, entre niveles de medición” (p. 77).

En particular, la fiabilidad compuesta de nivel específico se puede estimar fácilmente utilizando las estimaciones de parámetros de un AFC multinivel (Geldhof et al., 2014). Los modelos AFCM ajustados normalmente se evalúan con el ajuste de los índices comúnmente utilizados para el análisis factorial de un solo nivel, tales como la raíz del error cuadrático

medio de aproximación (RMSEA) y el índice de ajuste comparativo (CFI). Se ha identificado que la raíz cuadrática media estandarizada residual (SRMR) podría detectar la especificación errónea del modelo de nivel 2, razonablemente.

Evaluar el ajuste del modelo a un nivel específico es importante, particularmente, cuando el constructo de interés está en el segundo nivel. Los CCI elevados indican una variabilidad considerable a nivel 2 y, por tanto, la necesidad de modelar factores latentes en este nivel, especialmente constructos de conglomerados compartidos en el AF multinivel (Kim et al., 2016).

4. Estudio 1: Análisis tradicionales en los cuestionarios de evaluación docente: Una aplicación en datos reales

4.1 Introducción

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), este es el modelo por defecto utilizado en la evaluación de la calidad de los test y a pesar de la aparición de otros modelos continúa siendo el de mayor uso; sin embargo, presenta algunas limitaciones en relación con la dependencia de los índices que ofrece, en función de la muestra que se use en el análisis.

Es decir, si la muestra cambia, la información sobre la calidad del instrumento también cambia; esta situación afecta la interpretación de los datos, especialmente cuando las muestras carecen de homogeneidad (Martínez, 2014). En este sentido, algunos profesionales que trabajan con estos datos consideran que las propiedades psicométricas de un test pueden tener un estudio más detallado al utilizar la Teoría de Respuesta al Ítem (TRI) (Abad et al., 2011), en el que ha prevalecido el Modelo de Respuesta Graduada (MRG) (Samejima, 2010) para el análisis de ítems polítomos o politómicos, como son las escalas tipo Likert.

De acuerdo con lo anterior, este estudio tiene como objetivo evidenciar los aportes que realizan dos modelos de análisis: la TCT y la TRI, y así observar la información que es posible obtener de estos cuestionarios haciendo uso de cada método.

4.1.1 Teoría Clásica de los Test

La TCT asume entre sus supuestos que la puntuación de una persona en un test (X), está compuesta por la puntuación verdadera (V) y el error de medida (e). Las deducciones que se hacen de este modelo permiten llegar a fórmulas que estiman las propiedades de los ítems y los test (Muñiz, 2010). El análisis global del test implica revisar elementos de la fiabilidad y las

evidencias de validez, pero también cuando se quieren analizar los ítems que componen un *test* es necesario conocer si los parámetros de los ítems guardan relación con los parámetros del *test* y para ello, se observan sus índices de dificultad, discriminación y flujo de opciones.

La *dificultad* se refiere a la proporción de personas que aciertan un ítem, del total de personas que lo han intentado resolver. La *discriminación* se centra en el poder para diferenciar a las personas que tienen o no la característica medida y de forma más precisa es la correlación entre las puntuaciones de las personas en el ítem y sus puntuaciones en el test. Finalmente, el *flujo de opciones*, es decir la proporción de personas que responde cada alternativa de respuesta suele brindar información sobre la distribución de las respuestas de las personas a cada opción, indicador de calidad de la elaboración del ítem (Muñiz, 2018).

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 es una fortaleza su simplicidad y a la vez una limitación al no poder profundizar con detalle en las conclusiones que permite extraer de los análisis (Abal et al., 2014; Muñiz, 2010). Algunas de las críticas más relevantes de la TCT coinciden en que, las mediciones de los constructos no son invariantes de los instrumentos, es decir, que diferentes instrumentos diseñados para medir el mismo constructo no son comparables entre sí y además las propiedades psicométricas que se obtienen de un *test*, como la dificultad, discriminación, fiabilidad y evidencias de validez dependen de las personas con las que se estimen; quiere decir que, si la muestra cambia, las propiedades del *test* también (Muñiz, 2010).

A pesar de estas limitaciones, la TCT y sus extensiones factoriales siguen vigentes y generalmente se utilizan de forma complementaria con la Teoría de Respuesta al Ítem (TRI), para hacer un análisis más exhaustivo de la calidad del test. No obstante, los distintos modelos no deben verse como opciones confrontadas, ya que existen correspondencias entre los

indicadores de la TCT, el análisis factorial (AF) y sus equivalentes en la TRI (Barbero et al., 2001; Kramp, 2006; Santisteban y Alvarado, 2001).

4.1.2 Teoría de Respuesta al Ítem

Desde la perspectiva de la TRI, se revisa qué ítems se ajustan al modelo y si las opciones de respuesta planteadas logran diferenciar entre niveles de habilidad. De esta forma, se puede indicar si un instrumento cuenta con adecuadas características psicométricas que determinen la calidad de este, como vía de evaluación válida de un atributo.

Entre los modelos de la TRI, el MRG pretende establecer la localización de cada umbral en el continuo del rasgo latente (Attorresi et al., 2011; Samejima, 2010). Existen tres elementos importantes en el MRG.El primer elemento, corresponde a las CCO (Curvas Características Operantes) que representan la probabilidad de elegir una categoría igual o superior a k, que aumenta con el nivel de rasgo. Estas CCO son un paso intermedio para obtener las CCR (Curva de la Categoría de Respuesta) que indican la probabilidad de escoger cada opción k en cada nivel de rasgo. El segundo elemento, se refiere a los parámetros de posición, el parámetro que indica la relación entre el nivel de θ y la categoría que tiene la máxima probabilidad de ser elegida, la media de dos parámetros b_k consecutivos indica el nivel de rasgo en el que la probabilidad de elegir la opción k es máxima (Abad et al., 2006) y el parámetro de discriminación (*a)* que señala la relación entre el ítem y el rasgo medido, se denomina también cómo la discriminación y sus valores se suelen encontrar entre 0.3 y 2.5. El tercer elemento corresponde a las medidas locales de precisión, que se evidencian a través de la Función de Información tanto del ítem como del test (Abad et al., 2011).

Sin embargo, para el uso de la TRI se deben considerar las dificultades respecto a las especificaciones del tamaño de la muestra, la comprensión de los resultados (Asún y Zúñiga,

2008), además del cumplimiento de supuestos o tener ítems que midan apropiadamente en toda la escala del constructo (Abal et al., 2010).

4.1.3 Puntos en común

Uno de los principales aspectos a evaluar cuando se analizan las propiedades de un *test*, es su fiabilidad, que se refiere principalmente, a la precisión de la medida que se realiza con un instrumento y se reporta mediante un indicador llamado coeficiente de fiabilidad. De los coeficientes existentes, el más utilizado en ciencias sociales es el alfa de Cronbach (Zumbo et al., 2007), el cual, brinda información acerca de la consistencia interna del test.

La investigación ha mostrado que este coeficiente asume tau-equivalencia y le afectan factores como la longitud de la prueba, la covarianza entre los ítems y la varianza de la prueba, así como, que suele aumentar su valor, si la muestra o el número de ítems aumenta (Abad et al., 2011; Cortina, 1993). A su vez, existen modelos de fiabilidad basados en el análisis factorial como el coeficiente Omega, en el que para datos unidimensionales se redefine como la proporción de varianza del test que explica el factor común, como ventajas se reporta que no asume tau-equivalencia, trabaja con las cargas factoriales y no depende del número de ítems (Ventura-León y Caycho-Rodríguez, 2017).

El coeficiente Omega reporta mejores indicadores de fiabilidad cuando se trabaja con escalas tipo Likert (Elosua y Zumbo, 2008). En la TRI, también se realiza un análisis de la fiabilidad mediante la función de información, que permite observar qué tan informativo o preciso es el instrumento en los diferentes niveles de rasgo que se miden (Muñiz,2010).

Con el objetivo de presentar los anteriores elementos se realiza el análisis de un cuestionario acerca de la actividad docente, desde las dos perspectivas: la TCT y la TRI, para observar el funcionamiento global del cuestionario y de los ítems bajo cada escenario, rescatar

las bondades e información que brinda cada una y los aportes que realizan a la valoración de la calidad de este tipo de cuestionarios, sin realizar revisiones sobre los datos.

Es importante señalar que, en este estudio, los datos serán tratados conforme suelen analizarse tradicionalmente, es decir, sin considerar su estructura multinivel, ni realizar análisis de posibles sesgos identificados en los procedimientos de simulación. Esto con el fin de observar, qué tipo de información brindan los datos y cómo se puede asumir de forma falaz, a pesar de utilizar las técnicas adecuadamente, llegar a conclusiones equivocadas sobre la estructura factorial del instrumento y sus calidades psicométricas.

4.2 Método

4.2.1 Base de datos

La base de datos utilizada en este primer estudio con datos reales corresponde a los datos obtenidos de la aplicación del cuestionario (ver Apéndice A) que realizaron los estudiantes universitarios y facilitaron la información de forma anónima; se incluyeron ocho facultades y se consolidaron las respuestas de los estudiantes que presentaron el cuestionario durante los años 2012 - 2017. Se analizaron 16.950 respuestas de estudiantes, 61.93% fueron mujeres y 38.07% fueron hombres, se tuvo en cuenta un único cuestionario por estudiante y sin valores perdidos.

4.2.2 Instrumento

Este instrumento se creó con base en los siguientes propósitos: a) servir al profesorado para identificar sus puntos fuertes y aspectos por mejorar; b)servir a los centros para detectar y corregir las malas prácticas; c) servir al profesorado como evidencia para acreditarse, ser contratado o promocionarse; d) formar parte de la información analizada en el programa que evalúa al profesorado; e) cumplir con los programas de seguimiento y renovación de la

acreditación de los estudios que exigen el análisis de sus resultados y la puesta en marcha de medidas correctoras; f) estar disponibles para el análisis que requieran los responsables de los centros, departamentos y titulaciones; g) es uno de los principales indicadores de rendimiento que incide en el reparto presupuestal de los centros; h) estar disponibles para los estudiantes que tengan interés por conocer los resultados de las encuestas; i) estar disponibles para los delegados en caso de requerirlos a través de los representantes en los órganos de gobierno.

El cuestionario de satisfacción de la actividad docente denominado: Cuestionario de estudiantes: opinión sobre la actividad docente (se incluye como Apéndice A) fue desarrollado y aplicado semestralmente en todos los estudios oficiales de la Universidad, de forma telemática a través de una herramienta web, de forma presencial en una universidad pública española.

La encuesta original estaba conformada por 28 ítems, que incluía aspectos como organización y planificación, desarrollo de la docencia, sistema de evaluación, motivación y aprendizaje, interacción con los estudiantes, opinión global y datos sobre el estudiante.

Posteriormente fue dividida en cuatro nuevas encuestas: actividad docente, asignatura, plan académico y prácticas externas. La encuesta de actividad docente se centró en la valoración de la satisfacción con el docente, compuesta por ocho ítems. Después de 4 años de aplicación, tras análisis estadísticos y psicométricos, se eliminó el ítem 3 y se creó una nueva encuesta de 7 ítems.

El cuestionario analizado 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, considerando que los 7 ítems medían un único factor.

La escala de calificación se estableció como respuesta a cada afirmación 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 0.90, según la información reportada por la universidad, dado que no se cuentan con estudios psicométricos publicados.

4.2.3 Procedimiento

El análisis psicométrico se divide en dos momentos: en el primero se realiza el análisis utilizando la TCT y en el segundo, de acuerdo con los resultados de la dimensionalidad, se realiza utilizando el MRG como modelo de la TRI. Los análisis estadísticos y psicométricos se realizaron por medio del lenguaje de programación R versión 3.5.2 (R Core Team, 2022). En este estudio, los análisis se realizaron ignorando la estructura jerárquica de los datos, es decir se asume que las observaciones son independientes, con el fin de observar los resultados que tradicionalmente se obtienen en los análisis realizados en las universidades.

Los indicadores que se estimaron, a partir de la TCT, incluyeron el porcentaje de respuestas por opción en cada ítem, la dificultad del ítem con la media de cada uno, la discriminación del ítem a partir de la correlación ítem - resto del test (considerando el criterio superior a 0.2 para indicar que el ítem discrimina) y la fiabilidad si se elimina el ítem. Se realizó también la evaluación del índice de fiabilidad, por medio del coeficiente alfa de Cronbach.

Antes de iniciar con el procedimiento de TRI, se realizó la evaluación de la pertinencia de aplicar los métodos de la TRI, mediante la comprobación de los supuestos que señala esta

teoría. En primer lugar, se cumplió con el criterio del tamaño de la muestra que sugiere que la cantidad de personas sea superior a mil, siendo en este caso 16.950 estudiantes. Además, se realizó un análisis paralelo para poner a prueba el cumplimiento del supuesto de unidimensionalidad. La estimación de parámetros se realizó con el MRG R:ltm[gmr] (Rizopoulos, 2006) y consistió en que, para cada ítem se estimó un parámetro de discriminación (a), y cuatro parámetros de localización (b1, b2 y b3), se construyeron las CCR y la Función de Información para cada ítem, así como la Función de Información para el Test.

4.3 Resultados

4.3.1 Teoría clásica de los test

En la Tabla 1 se pueden observar los indicadores correspondientes al análisis de ítems para cada uno de los que conforman el cuestionario de actividad docente. Se presenta para cada ítem: la puntuación media (dificultad), las distribuciones en los diferentes niveles de respuesta, el nivel de discriminación (Correlación corregida ítem - *test*) y el coeficiente Alfa de Cronbach si se elimina el ítem.

Tabla 1.

Descriptivos del análisis de ítems utilizando la TCT

ITEM	Media	Desv. típica	Correlación Corregida Ítem-resto del test	Alfa si se elimina el ítem
1	**4.15**	1.11	0.8	0.96
2	3.98	1.26	0.86	0.96
3	3.88	1.35	0.88	0.96
4	3.93	1.32	0.87	0.96
5	3.75	1.32	0.82	0.96
6	**3.7**	1.43	0.88	0.96
7	3.93	1.32	**0.93**	0.96

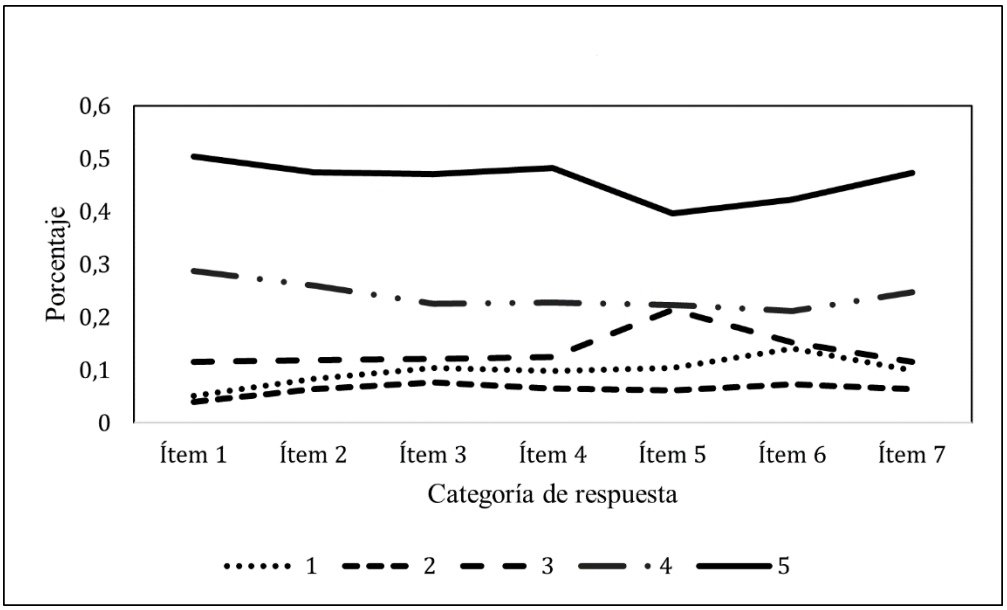
Se evidencia que en general todos los ítems presentan una media cercana al valor 4 de la escala (“Más bien de acuerdo”). El *Ítem 1* presenta la media más alta, mientras que el *Ítem 6* la más baja, el cual también evidencia una mayor desviación típica. La correlación ítem - resto del test es positiva y con valores superiores a 0.80 en todos los ítems, siendo el ítem 7 el que evidencia mayor poder discriminativo. No se identifica que, con la eliminación de alguno de los ítems se mejore el coeficiente alfa, ya que el valor global del cuestionario es de a = 0.96.

4.3.1.1. Elección categoría de respuesta.

En relación con el porcentaje de respuesta, entre el 40% y el 50% de los estudiantes selecciona la categoría 5, seguido de la categoría 4. La tendencia es similar en todos los ítems, por ejemplo, en el ítem 7, el 47.3% de los estudiantes seleccionaron la categoría 5; el 21.2% la categoría 4; el 15.12% la categoría 3; el 7.3% la categoría 2 y el 14.1% la categoría 1, tal como puede observarse en la Figura 1.

Figura 1.

Porcentaje de respuesta de los ítems



Nota: El gráfico representa el porcentaje de elección de respuesta en escala Likert de 1 a 5, por parte de los estudiantes.

Al observar los seis ítems restantes se evidencia un comportamiento similar, en el que en todos los ítems la categoría de respuesta más elegida es la cinco (5); es decir, *Totalmente de acuerdo*, frente al criterio de satisfacción en cada ítem, tal como es posible observar en la Figura 1.

4.3.1.2. Evaluación del supuesto de unidimensionalidad.

Para poder aplicar los modelos TRI clásicos, el primer paso consiste en comprobar el supuesto de unidimensionalidad. Para esto, se realizó el análisis paralelo inicialmente propuesto por Horn en 1965 (*R:psych*, Revelle, 2017).

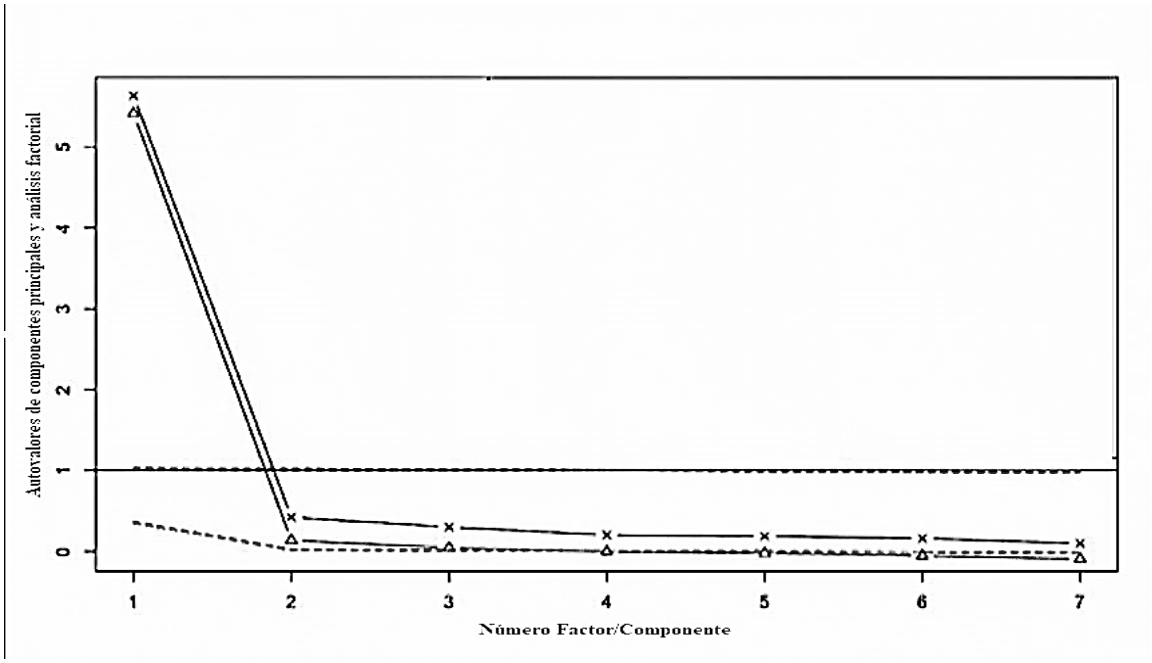
Tabla 2.

Autovalores de análisis paralelo

	Autovalores de la Muestra	Media de los valores aleatorios	Percentil 95 de autovalores aleatorios
C1	5.635	1.028	1.034
C2	0.422	1.018	1.024
C3	0.299	1.008	1.015
C4	0.201	1.000	1.006
C5	0.186	0.992	0.997
C6	0.159	0.982	0.995
C7	0.090	0.971	0.979

Figura 2.

Gráfico análisis paralelo



Nota: Resultados del Análisis Paralelo de Horn

Los resultados de la Tabla 2, en conjunto con los gráficos de la Figura 2, sugieren que por el método del análisis paralelo es recomendable retener un único factor, tanto al comparar con el promedio de las matrices aleatorias, como con el percentil 95.

Se observa que el único autovalor empírico superior al obtenido aleatoriamente es el primer autovalor (5.635), lo cual también se evidencia en el gráfico, en el que se representa el cambio de pendiente en el paso del autovalor 1 al autovalor 2, ya que a partir del segundo se estabilizan las cantidades de los autovalores. Obsérvese como en la Figura 2, en el análisis de componentes principales (ver en la figura PC) como del análisis factorial (ver en la figura FA para el método ejes principales) indican claramente una estructura unidimensional para el cuestionario de actividad docente.

4.3.1.3. Análisis de Invarianza.

Con el fin de comprobar si la estructura factorial se mantiene en diversos grupos de la misma variable, se realizó un Análisis Factorial multigrupo para observar si existe invarianza en los grupos y de qué tipo. Las variables analizadas bajo este procedimiento fueron Género del Estudiante (Femenino y Masculino) y Facultad (8 centros).

Respecto al género y la facultad, se evidenció que al revisar el indicador RMSA ($0.02<0.05$) se indica un buen ajuste de cada uno de los modelos contrastados. Estos resultados permiten asumir invarianza factorial en todos los niveles (configuracional, métrica, escalar y estricta), lo que implica que es posible señalar que los parámetros de los grupos son los mismos.

Por lo tanto, es posible asumir, que la encuesta mide lo mismo para los hombres y las mujeres, así como en las ocho facultades analizadas, lo cual se refleja en que los pesos factoriales son los mismos, las medias tanto de los indicadores (ítems) como de los factores (variables latentes) son iguales en los dos grupos y además de todo lo anterior, los parámetros de unicidad son iguales en los grupos (Ver Apéndice B).

4.3.2 Teoría de Respuesta al Ítem: MRG

4.3.2.1. Estimación de Parámetros.

Una vez comprobado el supuesto de unidimensionalidad, se estimaron los parámetros para el modelo de respuesta graduada (MRG) acorde con el ajuste a este modelo. Los ítems del cuestionario contenían cinco categorías de respuesta, por lo que el parámetro b1 se entiende como el mínimo valor del nivel de rasgo necesario para tener una probabilidad mayor de 0.5 de responder “Totalmente en desacuerdo”, así́ que ese valor es el umbral que separa las categorías “Totalmente en desacuerdo” y “Más bien en desacuerdo”.

De la misma forma b2 separa las categorías en “Más bien en desacuerdo” y “Ni de acuerdo ni en desacuerdo”; b3 diferencia las categorías “Ni de acuerdo ni en desacuerdo” y “Más bien de acuerdo”; y b4 las categorías “Más bien de acuerdo” y “Totalmente de acuerdo”. La estimación de los parámetros se realizó con el método de máxima verosimilitud y se observan en la tabla 3.

Tabla 3.

Estimación de parámetros con el MRG

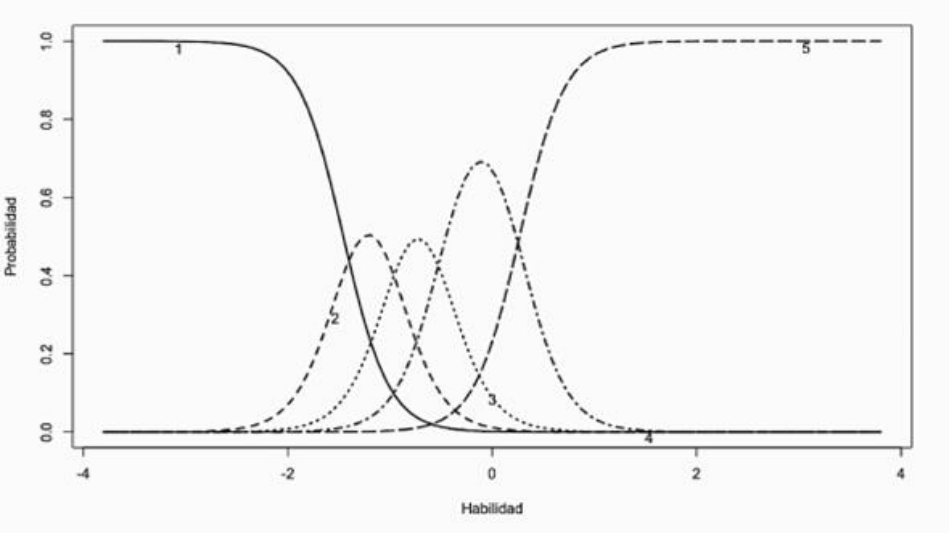
	Estimación parámetros de los ítems				
ITEM	b1	b2	b3	b4	A
1	-2.63	-1.72	-0.82	0.127	3.025
2	-1.69	-1.16	-0.584	0.273	3.348
3	-1.44	-0.94	-0.423	0.279	3.256
4	-1.54	-1.059	-0.498	0.243	3.145
5	-1.59	-1.121	-0.216	0.505	2.582
6	-1.27	-0.801	-0.2	0.409	3.446
7	-1.46	-0.965	-0.486	0.267	4.526

En los parámetros de los ítems se identifica que el parámetro de discriminación (*a),* tiene los valores más altos para el ítem 7, que el ítem denominado de satisfacción global *“En general, el trabajo llevado a cabo por el/la profesor/a ha sido satisfactorio”* esto es posible de evidenciar en la Curva Característica de Respuesta (CCR), en la que cada línea corresponde a la probabilidad de responder a una de las cinco categorías de respuesta en función del nivel de

q. (ver Figura 3). En todos los ítems, el parámetro de discriminación (*a),* es superior al criterio usual utilizado (a>2.5).

Figura 3.

Curva característica de respuesta, ítem 7



Nota: El gráfico representa las cinco categorías de respuesta para el ítem 7.

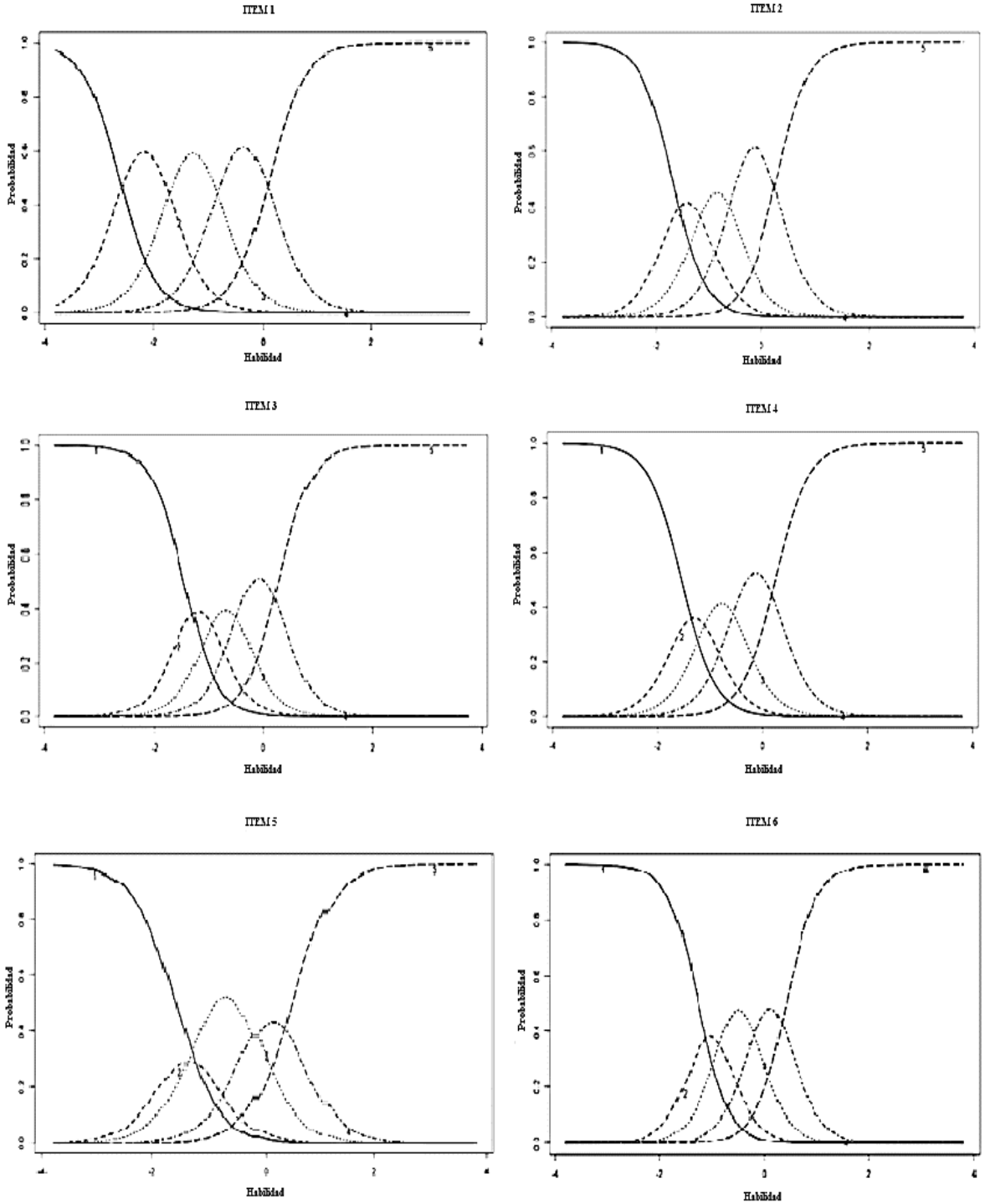
4.3.2.2. Curva Característica de Respuesta (CCR).

En el análisis de las CCR, la línea que se encuentra más a la izquierda corresponde a la probabilidad de responder la categoría que implica menor nivel de rasgo, así mismo la función de la máxima categoría de respuesta tiene mayor probabilidad de ser elegida, entre mayor sea el nivel de rasgo.

Se evidencia que el ítem 1 *"El/La profesor/a ha cumplido con lo explicitado en la guía docente"* es más difícil encontrar estudiantes que señalen la categoría "Totalmente en desacuerdo", dado que se requiere muy poco nivel de rasgo para puntuar bajo en ese ítem. Por otra parte, en el ítem 5 "*Las tutorías académicas con este/a profesor/a han resultado útiles"* es el ítem más difícil puntuar en la categoría "Totalmente de acuerdo", dado que solo las personas con un nivel de rasgo muy alto de satisfacción eligen la máxima categoría. (Ver Figura 4).

Figura 4.

Curva característica de respuesta para 6 ítems.



Nota: Se representa la tendencia de respuesta para las cinco categorías de los 6 ítems del instrumento.

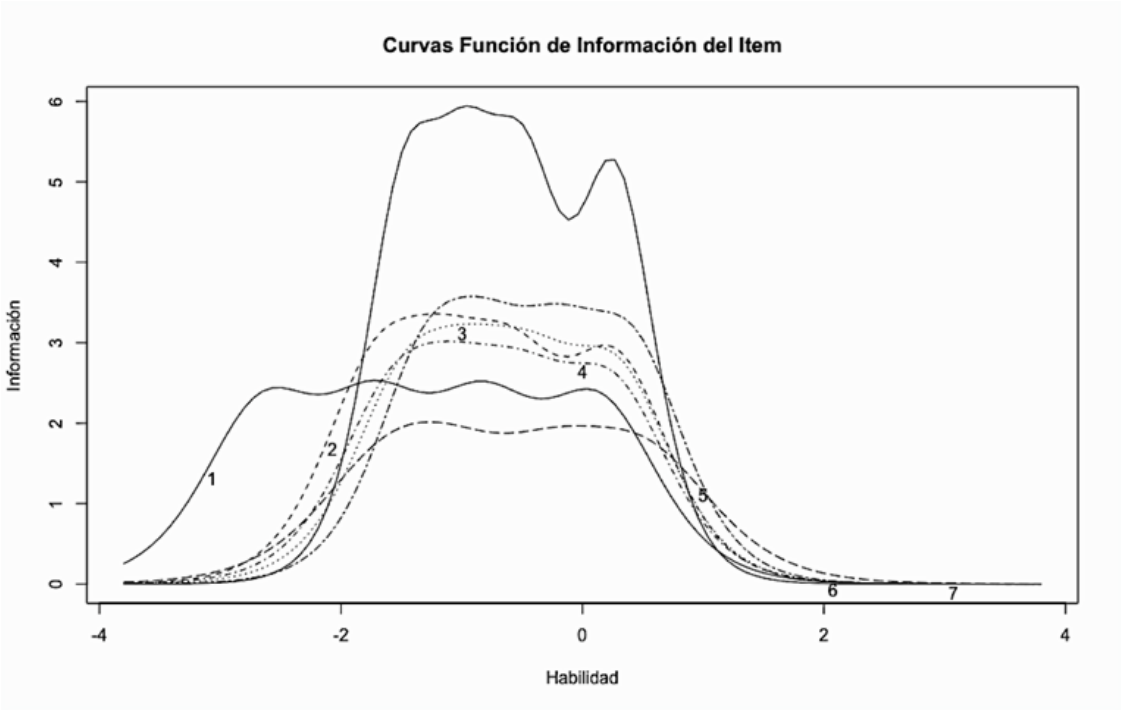
Al observar los parámetros de los ítems se evidencia que no se requiere un alto nivel de satisfacción para elegir la máxima categoría; es decir, brindar una calificación de 5 a los ítems.

4.3.2.3. Función de información del ítem y del test.

Las gráficas de precisión están principalmente representadas en dos elementos: la Función de Información de los Ítems (Figura 5), que como su nombre indica señala la precisión con la que el ítem mide el rasgo latente a lo largo de todo el rango de valores; mientras que la Función de Información del Test (Figura 6), es la sumatoria de las diferentes funciones de información del ítem y muestra en general la precisión de la escala total para discriminar entre los niveles de rasgo de los individuos (Abal et al., 2014).

Figura 5.

Función de Información de los 7 Ítems



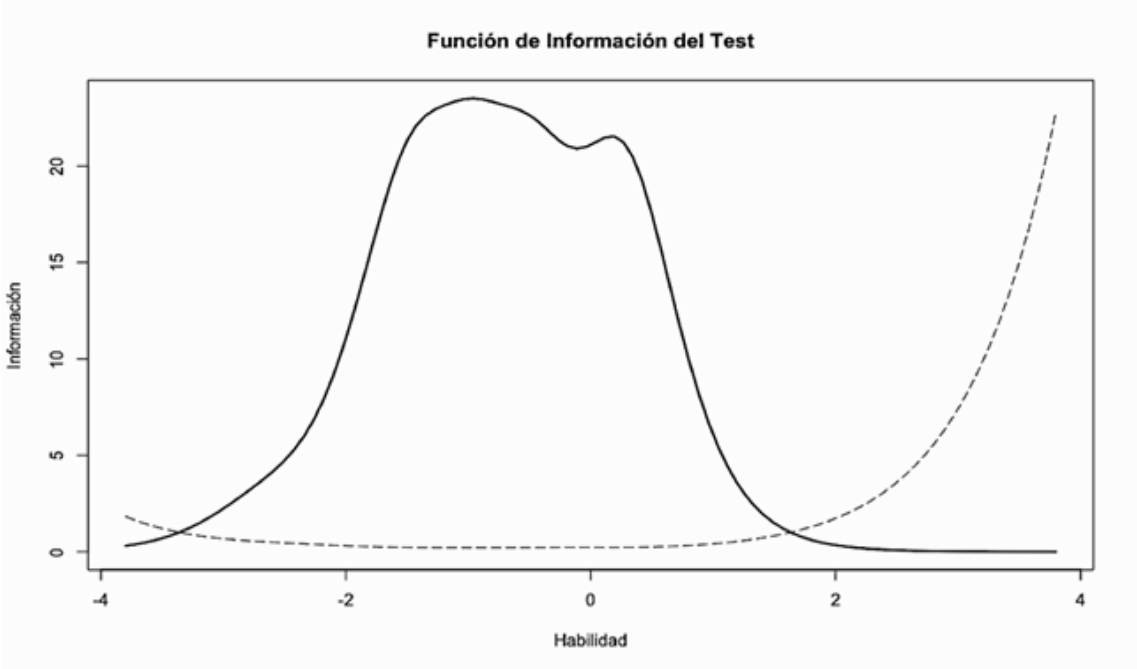
Nota: En el gráfico se observa la curva de función de información para los 7 ítems del test.

En la figura 5 es posible observar la función de información para todos los ítems; la mayoría miden mejor en los niveles de rasgo medios y bajos, el ítem que más información brinda sobre la satisfacción y para todos los niveles de rasgo es el 7. Los que menos información aportan son el 5 y el 1, ofreciendo el mismo nivel de información, en todo el continuo del rasgo. También se evidencia que el ítem 1 brinda más información para los niveles bajos de satisfacción.

En la figura 6 de la Función de Información del test se observa que en general este proporciona un alto nivel de información, en especial entre -1.5 y 0.5.

Figura 6.

Función de Información de la escala completa de 7 Ítems



Nota: En el gráfico se observa la función de información que representa al test completo.

Se evidencia, a partir de la Función de Información del test, que este cuestionario de actividad docente proporciona mayor información en los niveles de rasgo medio - bajo.

4.4 Discusión y conclusiones

La evaluación de la calidad del instrumento, desde las dos perspectivas TCT y TRI, permitió encontrar adecuados indicadores que convergen al momento de evidenciar la calidad del instrumento y que proporcionan evidencias de validez y fiabilidad del cuestionario de evaluación docente y de las interpretaciones que se pueden realizar de sus puntuaciones en cada uno de los ítems que la componen.

Desde la perspectiva clásica (TCT), el análisis de los ítems muestra que estos tienen una media en torno al valor 4 de la escala, lo que sugiere que los estudiantes están "Más bien de acuerdo" con las afirmaciones del cuestionario de actividad docente y que el mayor porcentaje de las respuestas se encuentra en la categoría 5 "Totalmente de acuerdo", también se evidenció un alto nivel de correlación entre cada uno de los ítems y el resto del test (>0.8).

En términos generales estos indicadores reflejan un alto nivel de satisfacción con la labor del docente. Por su parte, el análisis paralelo y los pesos factoriales superiores a 0.8, dan muestra de una estructura unidimensional, siempre es recomendable apoyarse en diversas fuentes y compararlas para tener evidencia de la decisión sobre la estructura factorial, ya que se conoce, que los criterios de ajuste estadístico no siempre apoyan las mejores decisiones (Abad et al., 2011), por ello una buena forma de complementar los análisis fue observar el análisis paralelo y los datos referentes a los autovalores. Así como, la confirmación de la invarianza factorial, en todos sus niveles, para los grupos de la variable género y facultad.

El análisis realizado desde la TRI con el MRG ofrece otros indicadores sobre la calidad del instrumento. Además de identificar el nivel de atributo necesario para seleccionar una categoría, que se considera equivalente al nivel de dificultad de la TCT en el que el ítem más fácil es el 1 y el más difícil es el 5; este modelo proporciona el indicador de discriminación, que, en este caso concreto, es superior para todos los ítems (>2.5). También, permite identificar

los ítems más informativos que, para este caso, fueron los ítems 7, 6, 3, 2 y 4, mientras que aquellos que aportan menos información sobre la satisfacción de los estudiantes son los ítems 1 "*El/La profesor/a ha cumplido con lo explicitado en la guía docente"* y 5 "*Las tutorías académicas con este/a profesor/a han resultado útiles".* Finalmente, con la función de información del test se determinó que el cuestionario de actividad docente proporciona mayor información sobre la satisfacción de los estudiantes que se encuentran en los niveles de rasgo medio - bajo del rasgo medido.

Un punto en común a tener en cuenta en la aplicación de los dos modelos corresponde al análisis de la fiabilidad. En la TCT utilizando el Coeficiente Alfa de Cronbach arrojó un indicador elevado de 0.96 que está asociado directamente con la alta correlación entre los elementos de la escala. En algún momento de la investigación, y dados los valores elevados del Alfa, se consideró que podría existir redundancia entre los ítems (Abad et al., 2011). Sin embargo, al hacer lectura de estos se evidencia que cada uno de los ítems indaga por diferentes elementos que integran la actividad docente. Esta cuestión será abordada más adelante en esta tesis en los siguientes capítulos.

El coeficiente Omega total por basarse en la estructura factorial que es más precisa, cuando la escala es unidimensional (Ventura-León y Caycho-Rodríguez, 2017), como es el caso que demuestra el cuestionario de actividad docente. En este análisis, se obtuvo un indicador también elevado de 0.97, y que debido al incumplimiento del supuesto de tau-equivalencia, ya que, el poder discriminante de los ítems no era exactamente igual, es ligeramente más elevado que si se calcula el coeficiente alfa (Santisteban y Alvarado, 2001; Zumbo et al., 2007).

En el caso de la TRI, esta medida de precisión se traslada a la Función de Información, que, a diferencia de la TCT, se estima para cada nivel de rasgo y no como un único indicador.

En el caso del cuestionario de actividad docente se detectó que, si bien es elevada la precisión de la medida, esta se da principalmente para los niveles bajo - medio del rasgo. Adicionalmente, en los procedimientos basados en la TRI se estima para cada ítem, un parámetro de discriminación y tantos parámetros b como el número de alternativas menos uno, además de las funciones de información del ítem y el test, que permiten realizar un análisis más detallado del ítem.

Una relación que no fue abordada y que podría ser objeto de otros estudios es la relación entre la TCT y la Teoría de la Generalizabilidad (TG). Como complemento a los valores dados por la TCT se asemeja a la conexión entre ANOVA simple y factorial. Así como el ANOVA factorial permite explorar múltiples factores que contribuyen a la variabilidad de los datos, descomponiendo la varianza de la puntuación observada en efectos atribuibles a cada factor, interacciones y "error aleatorio", la TG amplía la TCT al dividir la varianza en varias fuentes. Mientras que la TCT se limita a dos fuentes (varianza sistemática y error), la TG considera la varianza sistemática de los objetos de medida, diversas fuentes de error y sus interacciones (Santisteban y Alvarado, 2001), lo que podría complementar de forma valiosa el análisis de estos cuestionarios.

Es importante resaltar lo señalado por Penny (2003), en relación con la formación que reciben las personas que analizan y manejan la información resultante de los cuestionarios de actividad docente, así como la escasa capacitación que reciben sobre las posibilidades de análisis, por lo que es más frecuente encontrar análisis de tipo clásico, y en menor medida análisis TRI.

Se requiere comprender y desarrollar modelos matemáticos de mayor complejidad y con supuestos difíciles de cumplir, si bien en esta oportunidad los supuestos aparentemente se cumplen para proceder con análisis de corte TRI, un análisis exhaustivo sobre la naturaleza de

los datos hace falta para identificar que los indicadores tan ajustados, altos y casi perfectos, pueden estar presentando una falacia, que con el tratamiento adecuado conduciría a resultados diferentes.

Los indicadores obtenidos a partir del MRG evidencian que, con su uso es posible realizar un estudio más detallado de las propiedades de los ítems y que como ventaja adicional a la TCT permite seleccionar los ítems que se ajusten mejor a la función de información objetivo de la medición realizada. Ahora bien, funcionará mejor en la medida que el instrumento sea unidimensional, dado que no es necesario analizar las relaciones entre los factores, y esto es lo que se debe garantizar en primer lugar.

Otra ventaja del MRG es que informa sobre el comportamiento de cada ítem, por medio de las funciones de respuesta para las categorías; con esta información y el análisis de su comportamiento en función del nivel de habilidad requerido para elegir una opción, es posible decidir sobre la pertinencia de tener en cuenta o no una categoría de respuesta. Estas características hacen que se recomiende el uso del MRG para el análisis de instrumentos de evaluación que usen una escala de respuesta graduada, tipo Likert, como la utilizada por los cuestionarios de evaluación docente.

Entre las herramientas gráficas se resaltan las que se obtienen con el MRG, dado que, en comparación con el modelo clásico, que es más descriptivo; la función de información para cada ítem y para el test total facilitan la identificación de la precisión con la que se mide un atributo y en qué niveles de rasgo.

Finalmente, como criterio práctico, el uso de modelos de TRI también permite proponer nuevos ítems que permitan realizar una medición precisa de los niveles de rasgo que se observan infraestimados. Por ejemplo, en el cuestionario analizado para esta universidad, los

resultados señalan la necesidad de diseñar ítems que midan y proporcionen información para los niveles altos de satisfacción en la evaluación docente.

Los análisis acá expuestos permiten evidenciar diferentes aspectos de los ítems, que, utilizados de forma complementaria, permiten obtener más información de la calidad de los instrumentos utilizados en la evaluación docente. Como señalan Barbero et al. (2001) y Kramp (2006), es posible establecer correspondencia entre las estimaciones realizadas con TCT e IRT y al analizar la información de cada técnica, se obtiene una mayor comprensión tanto de los sujetos como de los ítems.

Estos dos modelos se pueden adecuar con facilidad para las diferentes instituciones universitarias que utilizan cuestionarios de evaluación docente con el fin de recolectar información sobre la satisfacción de los estudiantes y que se convierten en los principales indicadores de acreditación y como insumo principal en la toma de decisiones acerca de la labor del docente (Holland, 2019). Esto cobra relevancia, dado que, si el instrumento es utilizado para tomar decisiones sobre los docentes, es fundamental conocer las características, ventajas y desventajas de dichos instrumentos. En ese sentido, un instrumento de valoración confiable permitirá obtener información más precisa.

Un aspecto para reflexionar, que va más allá de la potencialidad de las técnicas para presentar los resultados, es que los modelos aplicados parten de un supuesto básico y es que existe una distribución de sujetos hacia un mismo objeto; sin embargo, la satisfacción de los estudiantes puede estar condicionada por los diferentes tipos de profesores que pueden existir en una universidad. Por lo que, a pesar de estos excelentes resultados, surgen algunas inquietudes, dado que, si es tan fiable el instrumento en los niveles bajos de rasgo, quiere decir que la mayoría de las evaluaciones de los estudiantes son favorables y si son todas tan favorables, no parece que exista mucha variabilidad en los docentes.

Sin embargo, sería interesante analizar qué pasaría con los resultados y con la información que brinda el test si existiera una alta variabilidad en el profesorado que imparte los cursos. Hasta aquí, aparentemente se encuentra un instrumento que funciona adecuadamente, que es unidimensional y cuyos resultados son bastante satisfactorios, y que permite aplicar modelos de TRI como el de Samejima (2010), basados en el cumplimiento de los supuestos.

A pesar de estos resultados, es importante recordar que estos datos se analizaron sin ningún tipo de tratamiento, ni consideración de la naturaleza multinivel de los datos, ni tampoco la revisión respecto a la varianza que se explica por el profesor o por el estudiante, lo cual puede estar generando una información inadecuada. Especialmente, para las personas que manejan esta información, puede ser aparentemente correcta, dado el rigor metodológico, la estimación de los parámetros y el uso de las técnicas estadísticas adecuadas, pero aún no se ha solucionado la inquietud sobre ¿Qué miden en realidad estos cuestionarios? ¿Cuál es el constructo que está detrás? ¿Se está midiendo la satisfacción de los estudiantes o la calidad de los docentes? Todos los análisis presentados en este capítulo se han realizado considerando que el alumnado es homogéneo y los profesores también ¿Qué ocurre si hay dos fuentes de variabilidad mezclada? Será necesario realizar un siguiente estudio para abordar estas inquietudes.

5. Estudio 2: Potenciales sesgos y propuestas de análisis para su evaluación

5.1 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 los profesores que las imparten. Esta información es de importante trascendencia, dado 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 evaluad; 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 implicadas, esto puede dar lugar a estimaciones sesgadas por incumplimiento de la necesaria propiedad de invarianza.

Este estudio se centra en la fuente de evidencia estructural (AERA, APA y NCME, 2018), ya que se muestra como la estructura factorial y como consecuencia la fiabilidad de la medida puede estimarse de forma 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.

5.1.1 Modelos clásicos de análisis

La información derivada de los datos, los análisis estadísticos efectuados y los informes entregados a los profesores, que provienen de las Evaluaciones de Satisfacción Estudiantil o SET, comúnmente se elaboran utilizando la TCT, fundamentadas principalmente en la media de puntuación de la clase (Toland y De Ayala, 2005), tal como se presentó en el capítulo 3. 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 (Marsh y Hattie, 2002; Toland y 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).

Aunque se reconocen sus limitaciones, la TCT continúa siendo relevante y, en ocasiones, se emplea de manera complementaria con la TRI. Esto se hace con el fin de realizar un análisis más detallado de la calidad del test o para obtener una evaluación más precisa de

las características psicométricas de los elementos que componen 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.

5.1.2 Modelos para el análisis de datos anidados

Los tests 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. (2017), 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; Vermunt y Magidson, 2002).

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 medición de 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.

Entonces, 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; Vermunt y Magidson, 2002). 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 y 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 y Cohen, 2010) que permiten estimar la pertenencia de los sujetos a alguna de las clases latentes y además de terminar 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 y 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 y Carstensen, 2013) y a la dimensionalidad de los test (Hong y 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 y 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 una contribución importante 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 TRI, también se ha demostrado que no proporcionan información sobre los miembros de cada grupo, más allá de los predictores incluidos en el modelo (Cho y 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 y Yamamoto, 2007).

5.1.3 El presente estudio

Como se ha revisado brevemente, existen diferentes estrategias de análisis, sin embargo, la mayoría de estas no están implementadas aún 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. 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, pueden existir otras fuentes de variabilidad como la titulación, la asignatura, el tipo de asignatura, entre otras, que afectan la satisfacción. Esto lleva a que, a pesar de la diversidad de métodos existentes, en la aplicación de los datos reales, las condiciones no están dadas para su uso, lo que mantiene las mismas prácticas sobre la evaluación docente.

Para obtener una medida de satisfacción válida, la propuesta de esta tesis se basa en plantear una solución sencilla, que, en términos generales, consiste 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 que permita hacer una estimación más precisa, controlando las otras fuentes de variabilidad.

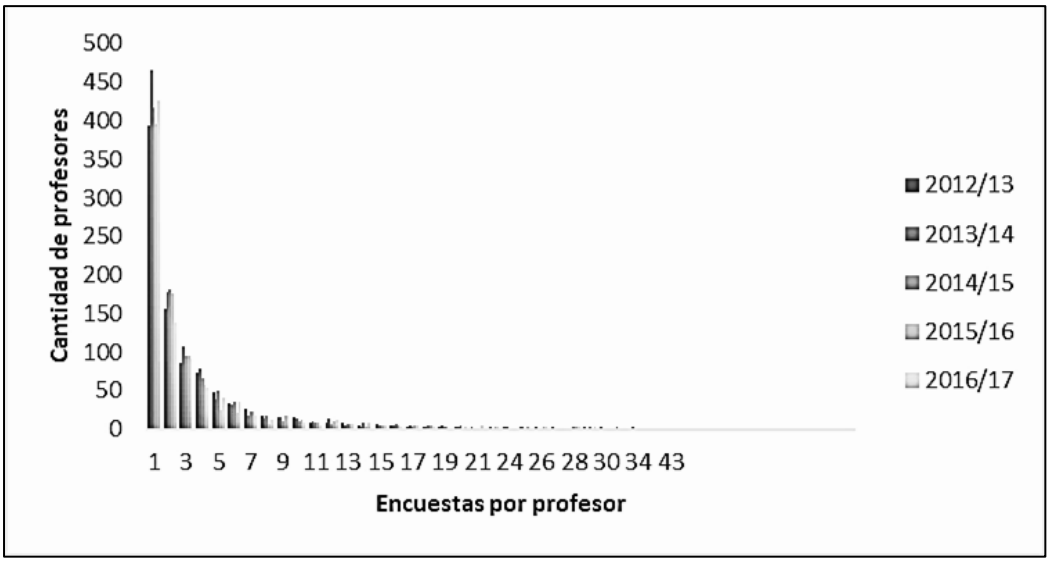
5.2 Método

5.2.1 Datos

Los datos analizados, son los mismos que se utilizaron en el primer estudio. A continuación, se presenta el detalle descriptivo de estos, con el fin de resaltar las dificultades de la muestra en el desarrollo de análisis más complejos.

Figura 7.

Número de encuestas recibidas por profesor

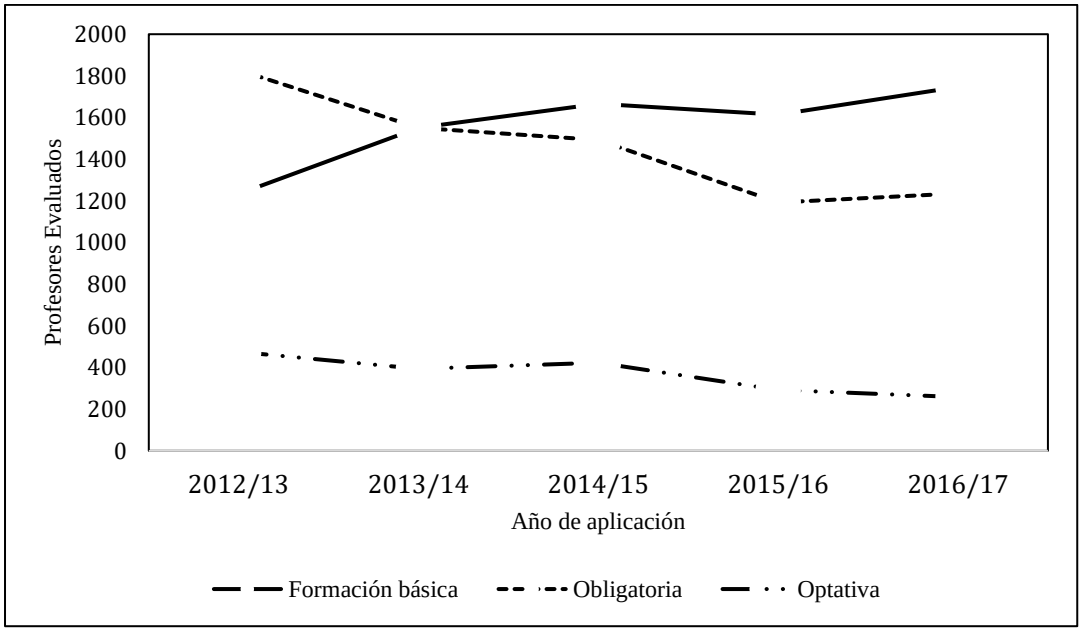


Nota: El gráfico evidencia la cantidad de profesores que reciben una cantidad específica de encuestas.

Cada año en esta universidad, respondieron las evaluaciones docentes entre 3.097 y 3.581 estudiantes, de aproximadamente 30.000 matriculados en cada uno de estos años. Esto, señala en primer lugar, que aproximadamente, sólo el 12% de los estudiantes de la universidad realizaron la evaluación, situación que ya de por sí indica que la muestra no es representativa de la población. Sumado a esto, la cantidad de evaluaciones docentes por profesor es baja, tal como se puede observar en la Figura 7. En los datos analizados, se encontró que la mayoría de los profesores (93%), cada año reciben 10 encuestas o menos, y que un porcentaje muy pequeño de profesores tiene muestras representativas para el análisis de sus evaluaciones. Bajo este escenario resulta difícil cumplir con los requisitos necesarios para el uso de procedimientos de gran complejidad como los modelos TRI, modelos multinivel, entre otros.

Figura 8.

Frecuencia de evaluaciones por tipo de asignatura

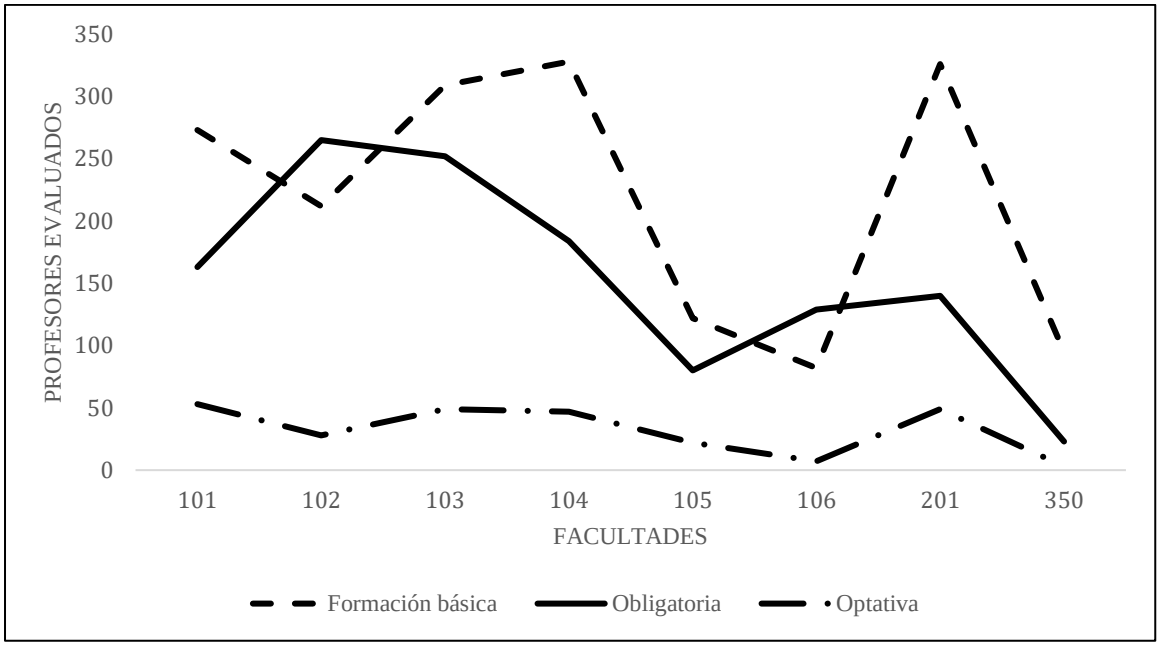


Nota: La Figura muestra la cantidad de profesores evaluados por año en cada tipología de asignatura.

En la figura 8 es posible evidenciar que la mayoría de las evaluaciones se realizan para las asignaturas de formación básica y obligatoria, que a su vez tienen menores puntuaciones, en comparación con las asignaturas optativas que reciben menos valoraciones y tienen mayores puntuaciones.

Figura 9

Distribución de frecuencias de las encuestas por facultad y tipo de asignatura



Nota: La figura muestra los profesores evaluados en cada tipo de asignatura para cada una de las facultades de la universidad.

En la figura 9 es posible identificar, desagregando más los datos, cómo en un año de evaluaciones, si se revisa en detalle la cantidad de evaluaciones realizadas por profesor, por tipo de asignatura para cada facultad, el mínimo es 3 y el máximo 328, cuya tendencia se mantiene cada año. Esto lleva a relacionar que, si se desagrega aún más la medida por programa y por asignatura específica, se consideran entre 10 y menos evaluaciones por asignatura, cómo se evidencia en la figura 7.

5.2.2 Instrumento

El cuestionario de satisfacción de la actividad docente es el mismo utilizado para el primer estudio.

5.2.3 Análisis de datos

5.2.3.1. Análisis de Regresión Lineal Múltiple.

Con el fin de especificar un modelo predictivo para la variable puntuación total, a partir de análisis multivariado, se llevó a cabo un análisis de regresión, en el que la variable dependiente es la puntuación en el instrumento y las variables independientes fueron edad, tipo de asignatura, sexo alumno y sexo docente, variables ampliamente señaladas en la revisión teórica (Abel y Meltzer, 2007; Arbuckle y Williams, 2003; Boring et al., 2016; MacNell et al., 2015; McPherson et al., 2009; Mengel et al., 2018; Ridgeway, 2011; Wagner et al., 2016).

5.2.3.2. Aplicación de modelos multinivel.

Con el objetivo de identificar las fuentes de varianza que aportan a la puntuación total, se pretendía utilizar un análisis factorial confirmatorio multinivel, sin embargo, dadas las características de la muestra, la mayoría de los profesores evaluados tenían menos de 10 valoraciones. Cuando se intentó analizar cursos que tuvieran un número de encuestas representativo, la muestra se reducía al 30%, por tanto, se planteó la solución descrita a continuación para poder separar las fuentes de varianza y proceder con los análisis.

5.2.3.3. Solución propuesta para el análisis de datos.

Con el fin de realizar el filtrado de las variables que introducen sesgo en la prueba, se usó un procedimiento de regresión por pasos. Este procedimiento incluye el desarrollo de cuatro pasos: (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 o facultad, plan de estudio, 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 (2022) 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 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 los 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 una puntuación 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 transformación 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 (2022).

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 0.10 representan un buen ajuste, y los valores inferiores a 0.05 representan un muy buen ajuste a los datos. Para el SRMR, los valores por debajo de 0.08 representan un ajuste razonable y los valores por debajo de 0.05 indican un buen ajuste. Con respecto al CFI, los valores por encima de 0.90 indican que el modelo se ajusta bien, y los valores por encima de 0.95 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 puntuaciones 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 la 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.

5.3 Resultados

En la tabla 4 se muestran los coeficientes para el modelo de regresión.

Tabla 4.

Coeficientes y niveles de significación para las variables incluidas en el modelo

Variable	*B*	Error estándar	*t*	*p*
(Intercepto)	24.709	0.284	86.875	0.000
Edad	0.122	0.012	9.480	0.000
TipoasignaturaObligatoria	-0.325	0.135	-2.399	0.016
TipoasignaturaOptativa	1.664	0.217	7.647	0.000
SexoalumnoM	-0.220	0.129	-1.709	0.087
SexodocenteM	0.084	0.126	0.668	0.504

Se observa que la variable que tiene mayor efecto en el modelo es tipo asignatura (formación básica, obligatoria y optativa). Según esta información cuando la asignatura es obligatoria tiene un efecto negativo sobre la puntuación en la satisfacción, por lo tanto, disminuye la puntuación, mientras que cuando la asignatura es optativa y, a mayor edad del estudiante hay un efecto positivo, por lo que, la puntuación de satisfacción tiende a incrementar, siendo resultados estadísticamente significativos,

En la Tabla 5, se puede observar que la varianza total más alta es la correspondiente a la adición entre profesores y alumnos con una puntuación de 72.52. Por lo tanto, la varianza debida a los profesores sería 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 5.

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	0.44	66.86	66.90	0.66 %
Sexo Alumno	0.14	66.83	66.97	2.10 %
Edad	1.05	65.94	67.00	1.57 %
Centro	0.70	66.38	67.08	1.04 %
Tipo de Asignatura	3.19	65.86	69.08	4.62 %

Respecto a los demás componentes, en la Tabla 6 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 o insignificante nula, por lo tanto, el procedimiento máximo de la conservación de varianza correspondía a los alumnos.

Tabla 6.

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 %

La evaluación de los datos corregidos con el test de Mardia (Joanes y Gill, 1998; Mardia, 1970) mostró unas puntuaciones de 3.22 para el contraste b1 y 88.72 para el contraste b2, en los dos se observaron valores estadísticamente significativos a $p<0.001$, con lo que se rechaza la normalidad multivariante en los datos corregidos. En este caso, se recomienda Omega total como estimador de la fiabilidad (Trizano-Hermosilla y Alvarado, 2016; Zinbarg et al., 2006) estimándose un valor de 0.87, inferior al obtenido en el primer estudio con datos completos.

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 7, 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 (χ^2), cuyo desajuste se justifica por el gran tamaño muestral utilizado en la investigación.

Tabla 7

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]	0.136 [0.133 – 0.139]	0.190 [0.187 – 0.194]	0.056 [0.052 – 0.060]
CFI	0.995	0.958	0.997
TLI	0.993	0.936	0.995
SRMR	0.023	0.046	0.012
χ^2 (gl)	4395.26*(14)	8611.21*(14)	590.12*(11)
BIC	6323.23	5307.28	21.41

* p<0.001

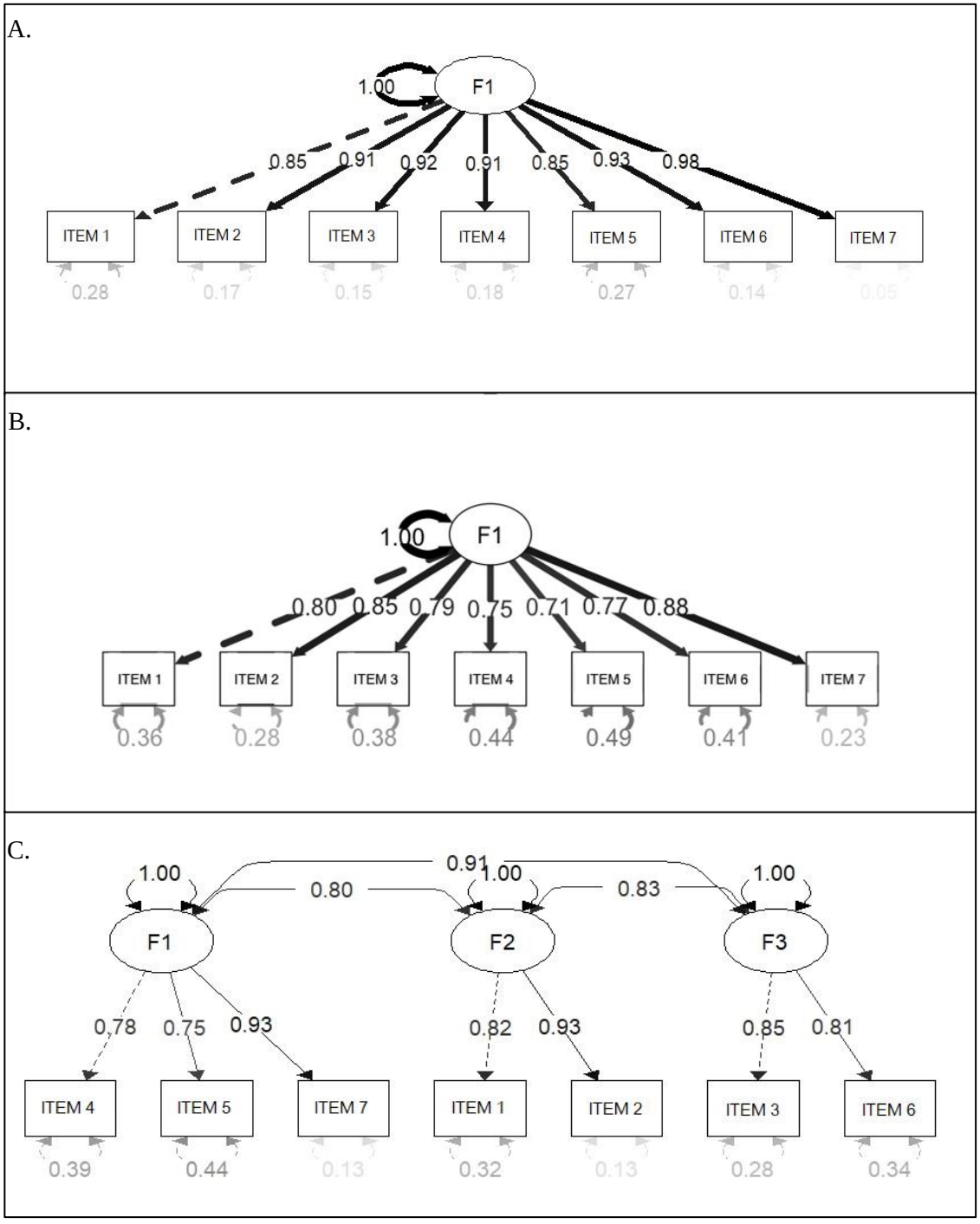
Nota. RMSEA = Error de aproximación de la media cuadrática (*Root Mean Square Error of Approximation*); CFI = índice de ajuste comparativo (*Comparative Fit Index*); TLI = índice de Tucker Lewis (*Tucker Lewis Index*); SRMR = Residuo estandarizado de la media cuadrática (*Standardized Root Mean Square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

En la Figura 10, 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 factores son altas, no se consideró que el modelo presentara redundancia factorial, al revisar los residuos entre las matrices de 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 0.05.

Figura 10.

Pesos estandarizados de los modelos



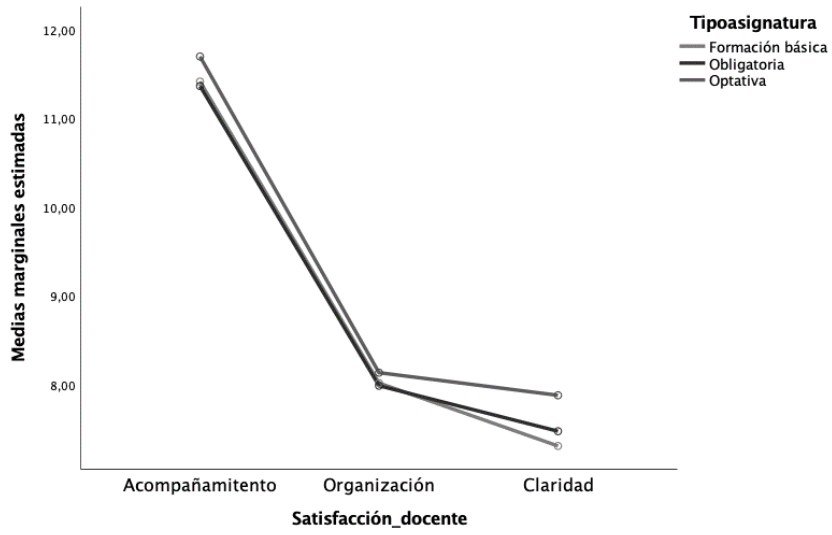
Nota: La figura muestra los Pesos estandarizados del modelo de un factor original sin corrección (A), modelo de un factor con datos corregidos (B) y del modelo de tres factores con datos corregidos (C).

Finalmente, en las comparaciones de medias de la variable seleccionada se evidenció que el nivel crítico asociado al estadístico *W* de Mauchly fue de *sig.< 0.000*, por lo que no se asumió esfericidad en las matrices de varianza-covarianza, por ello, se eligieron estadísticos multivariados para realizar los contrastes de hipótesis 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=0.001* con un coeficiente eta cuadrado de $\hat{\eta}^2$ = *0.005* 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 11 muestra que las asignaturas optativas tienen puntuaciones más elevadas en los tres factores, con respecto a las medias de asignaturas obligatorias y de formación básica.

Figura 11.

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



Nota: La Figura compara según el tipo de asignatura las medias entre los factores.

En la Figura 11, se observa también 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 calidad 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.

5.4 Discusión

El propósito de este estudio 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 considerando su naturaleza multinivel.

En primer lugar, se observó que la varianza explicada por el centro o facultad, asociado a la titulación que se imparte, era insignificante, lo mismo se concluyó 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 y 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 del 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 (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 9).

En este estudio, 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 (Cho y Cohen, 2010; Fox y Glas, 2001; Goodman, 2002; Vermunt y Magidson, 2002; 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 (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.

Sin embargo, 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, se consideraría 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 cual puede llevar a diferentes reflexiones, según la precisión con el que se utilice el modelo.

6. El Problema del Análisis de la Evaluación Docente: Un Estudio de Simulación

6.1 Introducción

Tal como se presentó en los capítulos 2 y 3, se han identificado múltiples variables que influyen en las evaluaciones de los estudiantes hacia sus profesores. Algunas se centran en atributos específicos del docente, como su entusiasmo (Gruber et al., 2012) y atractivo físico (Wolbring y Riordan, 2016), mientras que otros aspectos como la simpatía, así como el interés previo en el tema, pueden introducir posibles sesgos en estas valoraciones por parte de los estudiantes (Feistauer y Richter, 2018).

Asimismo, se han examinado otras variables que se relacionan con la inclinación del estudiante a responder de manera favorable o con aquiescencia (Spooren et al., 2013), el género y la disciplina del estudiante, entre otros factores (Boring et al., 2016). Además, se han identificado efectos asociados con variables como el género del docente, el tamaño de la clase y el tipo de asignatura (Arámburo y Luna, 2013; Luna et al., 2010), todos los cuales se convierten en factores que pueden influir en estas evaluaciones (DeFrain, 2016) y que requieren un enfoque específico para su abordaje (Acevedo y Olivares, 2010).

Estos sesgos tienen un impacto significativo en las valoraciones, llegando al punto en el que profesores altamente competentes pueden obtener calificaciones más bajas que profesores menos efectivos (Boring et al., 2016).

Por lo general, para analizar los datos de estas evaluaciones se emplean modelos univariados, los cuales asumen la independencia de las variables (Garduño, 2000; Marsh, 2007). Esta metodología trata los datos sin considerar la jerarquía entre las variables, pasando por alto su estructura multinivel. Este enfoque común en los métodos de estimación tradicionales puede ser impreciso, ya que conlleva uno de los errores más frecuentes: el uso e

interpretación de puntuaciones promedio brutas o naturales directamente derivados de las respuestas de los estudiantes (Franklin, 2001).

Este enfoque se refleja en la mayoría de los estudios revisados sobre evaluación docente (Alsarhan, 2017; Lang y Kersting, 2007), los cuales han descuidado la naturaleza anidada y jerárquica de los datos. Esta omisión resulta fundamental para que las técnicas estadísticas y las interpretaciones de los datos tengan sentido en relación con el objeto evaluado (Park y Yu, 2016).

Para abordar este fenómeno de una manera más adecuada, se han desarrollado diversos Modelos Multinivel, también conocidos como modelos jerárquicos, los cuales consideran la variación de parámetros estadísticos en múltiples niveles (Hox et al., 2017). En estos modelos, la variable dependiente se mide en el nivel más bajo, mientras que las variables independientes se evalúan en todos los niveles disponibles, como la facultad, las clases, el profesor, la edad, entre otros (Pascual, 2007). Estos modelos son especialmente útiles para abordar estructuras jerárquicas al enfocarse en la covarianza entre los resultados de los evaluados (Pardo et al., 2007). Han sido reconocidos como los más apropiados para representar estos conjuntos de datos, ya que buscan identificar grupos o combinaciones que expliquen la variabilidad observada.

La correcta especificación de estos modelos tiene un impacto significativo en el análisis de las estructuras factoriales de los instrumentos utilizados para medir la satisfacción de los estudiantes. Sin embargo, se ha identificado también, que la variabilidad de los datos afecta a los indicadores de ajuste en el Análisis Factorial Confirmatorio (AFC), según los hallazgos, cuando la variabilidad de los datos aumentaba, los indicadores de ajuste incremental tendían a disminuir. Además, este efecto resultaba más notable en muestras de menor tamaño en comparación con aquellas de mayor tamaño (Hu y Bentler, 1999; Ruíz et al., 2010).

6.1.1 Índices de ajuste

Algunos de los índices de ajuste más utilizados son la prueba de chi-cuadrado y su significancia (p-value), que rechaza la hipótesis nula a partir de un valor $\chi 2$ significativo ($p < .05$), lo cual implica que el modelo teórico propuesto es inadecuado, por lo que es necesario volver a especificarlo (Batista y Coenders, 2000; Cea, 2004). Sin embargo, dadas las limitaciones del índice por su sensibilidad al tamaño muestral y por fundamentarse en la distribución central de $\chi 2$ (Bollen, 1989; Byrne, 1998), se recomienda complementar sus resultados con otros índices de bondad de ajuste.

En este contexto, el índice RMSEA, conocido como el Error Cuadrático Medio de Aproximación por Grado de Libertad (Root Mean Square Error of Approximation), se destaca como uno de los más informativos para evaluar modelos en ecuaciones estructurales. Este índice considera los grados de libertad, lo que lo hace sensible al número de parámetros que el modelo estima (Barbero et al., 2011; Byrne, 1998; Cea, 2004). Los valores de RMSEA tienden a disminuir a medida que aumenta el número de grados de libertad o el tamaño de la muestra (McCallum et al., 1996; Kline, 2011). El índice de ajuste RMSEA constituye una medida absoluta de ajuste que cuantifica la discrepancia entre las predicciones del modelo y los datos observados.

A diferencia de los índices incrementales de ajuste, como el TLI y el CFI, el RMSEA muestra una menor sensibilidad a la variabilidad de los datos, sin importar el tamaño de la muestra (Cho et al., 2017). Hallazgos similares provienen de investigaciones como la de Lai (2021), que señala una ligera reducción en el RMSEA ante el aumento de la variabilidad, aunque dicho efecto resulta ser pequeño.

Así mismo, la raíz cuadrada media residual estandarizada, conocido como el índice de ajuste SRMR (Standardized Root Mean Square Residual) es un índice absoluto de ajuste que mide la discrepancia entre los residuos estandarizados del modelo y los datos observados. Este

índice es menos sensible a la varianza de los datos que los índices de ajuste incrementales, como el CFI y el TLI. Un estudio realizado por Cho et al., (2017) examinó el efecto de la varianza de los datos en el SRMR. Los autores encontraron que el SRMR no se vio afectado significativamente por la varianza de los datos, independientemente del tamaño de la muestra.

Otro estudio realizado por Lai (2021) examinó el efecto de la varianza de los datos en el SRMR. Los autores encontraron que el SRMR se redujo ligeramente cuando la varianza de los datos aumentó, pero que este efecto era pequeño.

Los índices de ajuste TLI (Tucker-Lewis Index) y CFI (Comparative Fit Index) los índices de ajuste TLI (Tucker-Lewis Index) y CFI (Comparative Fit Index) son medidas incrementales que evalúan cómo se ajusta un modelo propuesto en comparación con un modelo más simple, generalmente uno donde cada ítem se relaciona solo con un factor único. En general, estos índices disminuyen cuando los datos tienen una gran variabilidad. Esto sucede porque comparan el ajuste del modelo propuesto con un modelo más básico.

Cuando los datos son muy variables, es posible que los ítems estén correlacionados entre sí, aun si no pertenecen al mismo factor. Esto puede llevar a que el modelo propuesto se ajuste mejor a los datos que el modelo más simple, aunque no sea el modelo correcto. Investigaciones como la de Hu y Bentler (1999) han explorado cómo la variabilidad de los datos afecta estos índices incrementales. Descubrieron que estos índices disminuyen cuando la variabilidad de los datos aumenta, especialmente en muestras pequeñas.

Otros estudios, como el de Kline (2011), también han abordado este tema y concluyen que, si bien la variabilidad afecta estos índices, su impacto suele ser pequeño, más notable en modelos con muchos factores y difícil de detectar en muestras grandes.

Los estudios de Marsh et al. (2009), Hu y Bentler (1999), y Byrne y van de Vijver (2010) resaltan la limitada capacidad de los índices de bondad de ajuste para detectar problemas de sesgo en modelos. Marsh et al. (2009) indican que los índices incrementales como el CFI y

el TLI son más sensibles a la falta de normalidad que los índices absolutos como el SRMR y el RMSEA.

Sin embargo, incluso los índices incrementales tienen dificultades para identificar con precisión problemas de no normalidad. Tanto Hu y Bentler (1999) como Byrne y van de Vijver (2010) coinciden en que, si bien los índices de bondad de ajuste son útiles para evaluar el ajuste en Análisis Factorial Confirmatorios (AFC), su capacidad para detectar problemas de sesgo es limitada. Byrne y van de Vijver (2010), destaca la relativa robustez de estos índices frente a la falta de normalidad en muestras grandes, aunque advierte que en muestras pequeñas pueden ser más sensibles a estas desviaciones, incluyendo la asimetría y la curtosis de los datos.

6.1.2 Estudio de la dimensionalidad

Cuando se aborda el tema de la fiabilidad, se encuentran diversos métodos para su estimación; sin embargo, de manera inmediata, destaca el coeficiente Alfa de Cronbach, ya que es el más utilizado en estudios psicométricos (Ventura-León y Caycho-Rodríguez, 2017).

El alfa de Cronbach se considera una medida de consistencia interna (Cronbach, 1951), que refleja la magnitud de la covarianza entre los ítems (Morales, 1988) y en qué medida el constructo está presente en dichos ítems (Oviedo y Campo-Arias, 2005). Sin embargo, también ha existido una amplia investigación sobre las falencias que presenta este coeficiente como que la reducción de opciones de respuesta disminuye la variabilidad de la escala, impactando negativamente en el coeficiente alfa (Lozano et al., 2008); trabaja con variables continuas, lo cual no es común en ciencias sociales, desvalorizando la fiabilidad (Elosua y Zumbo, 2008) y se ve influenciado por el error muestral (Ledesma, 2004).

Además, cuando los datos son multidimensionales, como en un modelo bifactor, el coeficiente alfa está influenciado por todas las fuentes de varianza y pierde su idoneidad como

indicador de información (Rodríguez et al., 2016). Sin embargo, una de sus principales ventajas radica en que, solo requiere una aplicación de la prueba para su cálculo (Schmidt y Ilies, 2003).

A diferencia del coeficiente alfa, el coeficiente omega se basa en las cargas factoriales, que son la suma ponderada de las variables estandarizadas, una transformación que estabiliza los cálculos (Gerbing y Anderson, 1988), y que refleja de manera más precisa el nivel real de fiabilidad. Otra distinción es que es independiente del número de ítems (McDonald, 1999), el coeficiente omega se considera que es una medida confiable adecuada, en situaciones donde no se cumple el principio de tau-equivalencia, el cual puede ser violado cuando los coeficientes de los ítems que componen una matriz de solución factorial muestran valores significativamente distintos.

Omega estima la proporción de varianza en la puntuación total observada, atribuible a todas las fuentes "modeladas" de varianza común (Reise et al., 2013; Revelle y Zinbarg, 2009). Omega utiliza las cargas factoriales de un modelo específico como base, a diferencia de alfa, que generalmente se calcula según las observaciones de varianzas y covarianzas. Alfa asume cargas iguales (tau-equivalencia esencial), mientras que omega es más adecuado cuando las cargas varían. OmegaH estima la proporción de variabilidad en las puntuaciones totales que se puede atribuir a un único factor general, tratando así la variabilidad en las puntuaciones, debido a factores grupales como error de medición (McDonald, 1999; Reise et al., 2013; Zinbarg et al., 2005; Zinbarg, et al. 2006).

La investigación psicométrica se ha centrado en realizar evaluaciones del grado de unidimensionalidad o multidimensionalidad (Reise et al., 2013b). Si OmegaH tiene un valor elevado mayor a 0.7, es posible interpretar que las puntuaciones totales, ponderadas por unidad, son "esencialmente unidimensionales", en el sentido de que su variabilidad confiable está mayormente influenciada por una sola fuente. No obstante, esto no refleja la "dimensionalidad"

real de los datos. El modelo bifactor proporciona una manera muy simple y elegante de evaluar la fuerza dimensional relativa.

Una medida sencilla y más clara del grado de unidimensionalidad esencial es la VCE – Varianza Común Explicada (Berge y Socan, 2004; Reise et al., 2013a; Reise et al., 2010; Sijtsma, 2009), que es la varianza específica de los índices de un factor general, tomando la relación de la varianza explicada por un factor general y dividiéndolo por la varianza explicada por un factor general y un grupo de factores, donde se supone que los factores no están correlacionados (Sijtsma, 2009).

También puede usarse para juzgar la unidimensionalidad esencial de la varianza común en un conjunto de ítems y ayudar a decidir, si tratar los datos multidimensionales con una estructura bifactorial, como esencialmente unidimensional, en un modelo de medición SEM. Si la varianza común explicada es 0.80, significa que el factor general explica el 80% de la varianza común extraída, con el 20% de la varianza común repartida entre los factores de los grupos.

Para resumir, el Omega Jerárquico o el OmegaH es el porcentaje de la variación total de la puntuación atribuible a un único factor general. VCE es el porcentaje de varianza común explicada por el factor general, es decir, es un grado del índice de unidimensionalidad y está directamente relacionado con la fuerza de la relación con el factor general.

6.1.3 Problemas del tratamiento de los datos

Las calificaciones de los estudiantes tienen una estructura jerárquica, donde las calificaciones se encuentran dentro de cursos que, a su vez, están vinculados con profesores (Pekka, 2013; Rampichini et al., 2004). En general, se ha observado que en los grupos de

estudiantes que comparten un mismo entorno, no hay independencia entre ellos, lo que desafía el supuesto fundamental del modelo lineal general: la independencia entre observaciones.

Por ejemplo, en el estudio de Bacci y Caviezel (2011), se evidenciaron varianzas significativas en los efectos aleatorios de segundo y tercer nivel para cada factor analizado. Esta significancia señala que la jerarquía en la estructura de los datos impacta significativamente en la medición de la satisfacción de los estudiantes. Por esta razón se resalta, el hecho de ser cautelosos con las comparaciones entre enseñanzas, basadas en residuos de tercer nivel. Este estudio evidencia cómo al ignorar la estructura de los datos y analizarlos bajo la suposición de independencia, sin considerar esta estructura multinivel, se pierde información valiosa sobre cómo contribuyen las variables en la variabilidad de los datos.

Además, una problemática interesante que se ha venido estudiando se refiere a un problema de validez, no solamente del instrumento utilizado, sino también de cómo fuentes de variabilidad no controladas pueden conducir a estimaciones incorrectas en la medición (Bacci y 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, como ya se ha señalado anteriormente, 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 (Marsh y Hattie, 2002; Toland y DeAyala, 2005).

En el capítulo 5, se aplicó 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.

En ese segundo 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. Se pretendió entonces mostrar un procedimiento en el que quedaran expuestas consideraciones que no se habían tenido en cuenta en estudios previos analizados, con el análisis clásico y sin considerar la naturaleza anidada de los datos. De acuerdo con los resultados, se identificó que antes de medir la eficiencia del profesor, primero se requiere centrar las puntuaciones, controlando la variabilidad generada 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 verificar esta situación, este 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 puntuaciones, de acuerdo con la valoración de sus estudiantes.

Dado lo anterior, se consideró relevante observar el comportamiento de los índices de ajuste absolutos, cuando se estiman modelos que podrían explicar estas diferencias, principalmente para evaluar hasta qué punto una estructura en la que no hay un factor común, puede pasar por una estructura esencialmente unidimensional. Así, 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, situación que, introduce sesgo y es posible que, esto comprometa la validez de medida, específicamente la referida a la fuente de validez estructural.

Esta 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 y la dimensionalidad de cuatro modelos: (a) modelo original, (b) modelo multinivel, (c) modelo bifactor y (d) modelo unidimensional. Centrándose 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 y en tercer lugar el estudio de la dimensionalidad.

6.2 Método

El principal interés consistió en observar cómo afecta la diferencia entre profesores a los índices de ajuste absolutos y evaluar la afectación a la identificación de la dimensionalidad, pues las observaciones empíricas muestran que, sin la debida corrección de las diferencias, las estructuras factoriales pueden ser erróneamente especificada, tal como se evidenció en el capítulo anterior. Para esto, en la simulación se consideraron las características presentadas a continuación.

6.2.1 Variables

1. VI: Rangos de diferencias de calificaciones entre los profesores, correspondientes a desviaciones típicas (Z=0; Z= [-1.1]; Z= [-2.2]; Z= [-3.3])
2. VD: Índices de ajuste del modelo simulado (unidimensional, tres factores no correlacionados, Bi factor y Multinivel).
3. VD: OmegaH y VCE

6.2.2 Generación de los datos

Los parámetros de los datos se muestrearon, creando un modelo de 9 ítems en una estructura de tres factores no correlacionados (tres ítems por factor), en un modelo tau-equivalente con pesos de 0.7.

El objetivo de la simulación fue evaluar hasta qué punto, una estructura en la que no existe un factor común, y por lo tanto, se corresponde con un OmegaH=0, puede llegar a confundirse con una estructura esencialmente unidimensional, por efecto de una variabilidad no controlada (esto es, la variabilidad del profesorado).

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 (2022), 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 del tipo de asignatura que imparten o de sus habilidades.

Se simuló una estructura de tres factores no correlacionados para mostrar cómo incluso en el caso más extremo de no correlación entre factores, estos pueden pasar por factores correlacionados hasta el punto de confundirse con una estructura unidimensional, cuando existe una fuente de varianza superpuesta (estructura multinivel).

Respondiendo a la pregunta de interés principal, las muestras se dividieron en partes. La primera incluía datos donde no se asumen rangos de diferencias entre los profesores (Z=0); en la segunda muestra, se incluyó un rango de una desviación típica de diferencias en las calificaciones de los profesores (Z=1). Finalmente, las demás muestras incluyeron rangos de diferencias de 2 (Z=2) y 3 (Z=3) desviaciones típicas de calificaciones para los profesores.

Se usó una muestra de 1000 sujetos, por réplica y se realizaron 100 réplicas para cada condición de diferencia de profesores. 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.

6.2.3 Análisis de datos

Una vez generados los datos, la estructura original un modelo de tres factores no correlacionados, se comparó 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.

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 y la significancia estadística para el análisis de los datos, los índices de ajuste absoluto el RMSEA y SRMR, y los índices de ajuste incremental TLI y CFI.

En términos generales, valores de RMSEA inferiores a 0.05 sugieren un buen ajuste, mientras que aquellos comprendidos entre 0.05 y 0.08 indican un ajuste razonable (Browne y Cudeck, 1993; Hu y Bentler,1999; Kelloway, 1998), los valores de RMSEA menores de 0.08 representan un buen ajuste, y los valores inferiores a 0.05 representan un muy buen ajuste a los datos. Para el SRMR, los valores por debajo de 0.08 representan un ajuste razonable y los valores por debajo de 0.05 indican un buen ajuste. Con respecto al CFI y TLI, los valores por encima de 0.95 representan un ajuste muy bueno a los datos.

Finalmente, se estimó el Omega Jerárquico (OmegaH) y la Varianza Común Explicada (VCE), con el fin de, encontrar mayor evidencia respecto a la identificación del modelo y verificación de la estructura factorial en las condiciones simuladas; específicamente este análisis se realizó evaluando el modelo bifactor. Se utilizaron los siguientes puntos de corte para el análisis de la unidimensionalidad esencial: valores de VCE > 0.60 y OmegaH > 0.70 (Reise et al., 2012; Rodríguez et al., 2015).

6.3 Resultados

Los resultados se dividieron en cuatro apartados, cada uno representando un modelo. Las comparaciones de los índices de ajuste se hicieron en relación con los diferentes niveles de puntuaciones típicas al simular rangos de 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. Los valores de cada tabla representan el promedio de los índices de ajuste obtenidos de las 100 réplicas para cada condición de Z.

6.3.1 Modelo de tres factores no correlacionados - modelo teórico

Este modelo, fue el modelo recuperado que cumple con las características base propuestas para la simulación, es decir, el caso de un modelo de tres factores no correlacionados, cada factor con un peso factorial de 0.7. En la Tabla 8 se pueden observar los índices de ajuste de este modelo.

Tabla 8.

Índices de ajuste para el modelo de tres factores no correlacionados

	Diferencias en puntuaciones típicas			
Índice	**Z=0**	**Z= [-1.1]**	**Z= [-2.2]**	**Z= [-3.3]**
χ^2 (p-value)	0.526	0.525	0.536	0.542
SRMR	0.016	0.013	0.007	0.004
RMSEA	0.006	0.006	0.005	0.005
CFI	0.998	0.999	0.999	0.999
TLI	1.000	1.000	1.000	1.000

Los resultados del análisis revelan una consistente y sólida adecuación del modelo propuesto a los datos observados. Los valores de Chi-cuadrado y sus correspondientes p-values, aunque ligeramente superiores al estándar convencional, sugieren un ajuste aceptable

entre el modelo y los datos para todas las diferencias de puntuaciones típicas evaluadas. Más aún, los índices de ajuste, como el SRMR, el RMSEA, el CFI y el TLI, muestran consistentemente valores excepcionalmente favorables en todas las comparaciones de diferencias de puntuaciones. Estos valores extremadamente bajos de SRMR y RMSEA, junto con valores perfectos en CFI y TLI, que indican bondad de ajuste, observándose una mejora de los índices a medida que aumentan las diferencias en las puntuaciones.

6.3.2 Modelo unidimensional

El primer modelo por comparar fue el modelo unidimensional, este modelo suele evidenciarse comúnmente en los análisis básicos factoriales de las encuestas de evaluación docente. En la Tabla 9 se pueden observar los índices de ajuste de este modelo.

Tabla 9.

Índices de ajuste para el modelo unidimensional

Diferencias de calificaciones en puntuaciones típicas				
Índice	**Z= 0**	**Z= [-1.1]**	**Z= [-2.2]**	**Z= [-3.3]**
χ^2 (p-value)	<0.001	<0.001	<0.001	<0.001
SRMR	0.163	0.131	0.075	0.043
RMSEA	0.221	0.245	0.245	0.245
CFI	0.336	0.555	0.786	0.860
TLI	0.115	0.407	0.714	0.814

En este modelo, al revisar el ajuste de los datos es posible observar que, en primer lugar, los valores de Chi-cuadrado y sus respectivos p-values fueron notablemente bajos ($p < 0.001$), sugiriendo que el modelo teórico propuesto es inadecuado. Por otro lado, se observa una tendencia interesante en los índices de ajuste incremental, como el CFI y el TLI, los cuales

mejoran gradualmente a medida que aumentan las diferencias de puntuaciones típicas, sin embargo, no alcanzan los valores óptimos de ajuste. Por su parte, el RMSEA se mantiene constante en 0.2 independientemente del nivel de diferencia, sugiriendo que no existe un ajuste óptimo en todos los casos. De forma particular, se identifica que el SRMR empieza a ajustar cuando hay más diferencia entre los profesores, en los casos de las condiciones Z= [-2.2] (0.075) y Z= [-3.3] (0.043).

6.3.3 Modelo bifactor

En este caso, el modelo bifactor fue definido por un factor general y tres factores específicos. En la Tabla 10 se pueden observar los índices de ajuste de este modelo.

Tabla 10.

Índices de ajuste para el modelo bifactor

	Diferencias en puntuaciones típicas			
Índice	**Z= 0**	**Z= [-1.1]**	**Z= [-2.2]**	**Z= [-3.3]**
χ^2 (p-value)	0.653	0.644	0.649	0.639
SRMR	0.009	0.007	0.003	0.002
RMSEA	0.003	0.003	0.004	0.005
CFI	0.999	0.999	0.999	0.999
TLI	1.000	1.000	1.000	1.000

Estos resultados muestran indicadores que apuntan a un buen ajuste entre el modelo propuesto y los datos observados. Los valores de p> 0.05 para Chi-cuadrado indican una bondad de ajuste. El SRMR disminuye progresivamente a medida que aumentan las diferencias

en las puntuaciones típicas, alcanzando valores muy bajos (0.002 a 0.009), solamente para la condición Z= 0, el modelo no ajustaría que es donde no se encuentra variabilidad en las calificaciones de los profesores, indicando un ajuste muy preciso. El RMSEA se mantiene en valores bajos (0.003 a 0.005), lo que sugiere un ajuste razonablemente bueno, aunque ligeramente más elevado para las mayores diferencias. Por otro lado, tanto el CFI como el TLI muestran valores muy altos próximos a 1, lo que indica un excelente ajuste del modelo en todas las condiciones evaluadas.

6.3.4 Modelo multinivel

En este modelo, se representa la estructura real que debe analizarse, de acuerdo con las características del fenómeno estudiado, en el que primero se toma una muestra de unidades del más alto nivel (profesores), y luego se muestrea las subunidades de las unidades disponibles (estudiantes). En la Tabla 11 se pueden observar los índices de ajuste de este modelo.

Tabla 11.

Índices de ajuste para el modelo multinivel

	Diferencias en puntuaciones típicas			
Índice	**Z= 0**	**Z= [-1.1]**	**Z= [-2.2]**	**Z= [-3.3]**
χ^2 (p-value)	0.988	0.964	0.964	0.964
SRMR	0.634	0.036	0.022	0.020
RMSEA	<0.001	<0.001	<0.001	<0.001
CFI	1.000	0.999	0.999	0.999
TLI	1.000	1.000	1.000	1.000

Los resultados del análisis del modelo multinivel revelan un patrón interesante de ajuste en relación con las diferencias en las puntuaciones típicas. Si bien los valores de $p>0.05$ de Chi-cuadrado indican que no hay diferencia estadísticamente significativa entre la matriz de entrada y la matriz reproducida, otros indicadores muestran algunas variaciones.

El SRMR exhibe una marcada disminución, a medida que aumentan las diferencias en las puntuaciones típicas, con valores iniciales altos, cuando no hay diferencias entre los profesores (0.634), indicando que para la condición Z= 0, es decir cuando no hay diferencias entre profesores, el modelo genera algunos residuos altos y se observa como estos se reducen cuanto mayores son las discrepancias (hasta 0.020). En cualquier caso, los valores de RMSEA son consistentemente bajos (<0.001), con valores que indicarían lo que podría indicar un ajuste prácticamente perfecto en todas las circunstancias. Además, tanto el CFI como el TLI reflejan valores muy altos (0.999 a 1), lo que sugiere un buen ajuste global del modelo en todas las condiciones evaluadas.

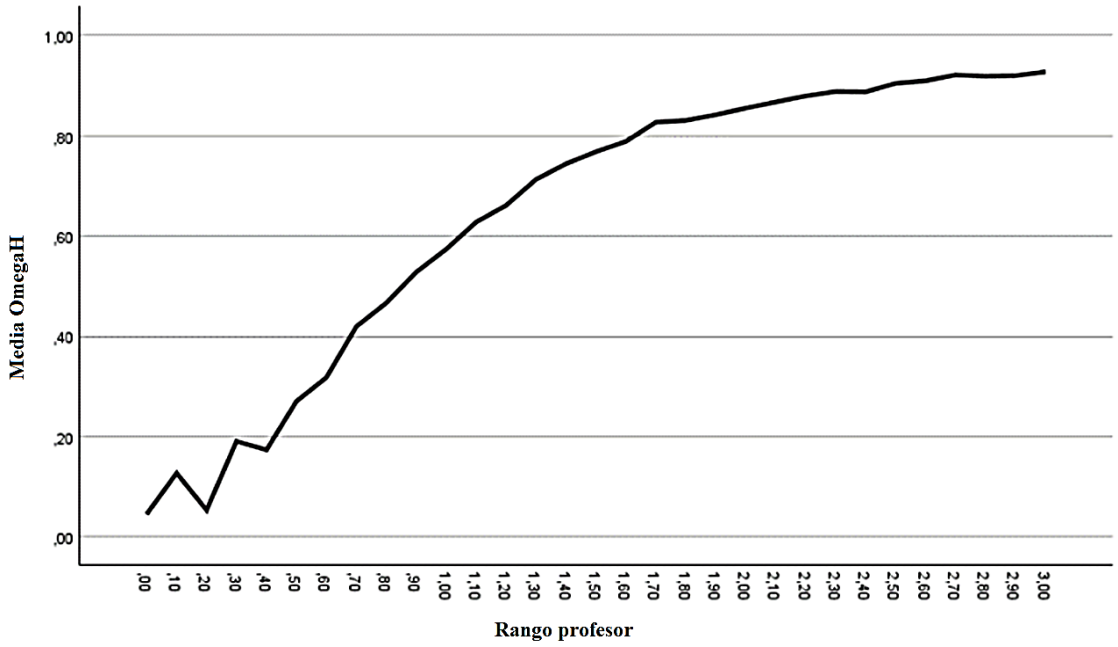
En resumen, en relación con el ajuste de los datos, se identifica que el modelo recuperado de tres factores no correlacionados, el modelo multinivel y el modelo bifactor alcanzaron los criterios de calidad en todos los índices estimados. El modelo de un factor no alcanzó un adecuado nivel de ajuste, sin embargo, los datos de la condición Z= [-2.2], 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.

6.3.5 Estudio de la dimensionalidad

Para un análisis más detallado de cómo una estructura tridimensional de factores no correlacionados puede confundirse con una estructura unidimensional o esencialmente unidimensional, en lugar de agrupar las puntuaciones de los profesores en rango; se realizó una simulación con 200 réplicas, tomando como variable independiente el rango de dispersión de la variable profesor. En la Figura 12 se presenta el valor del Omega Jerárquico para las diferencias simuladas para los profesores desde Z= 0 hasta Z=3.

Figura 12.
Omega Jerárquico en relación con la diferencia de puntuaciones de los profesores

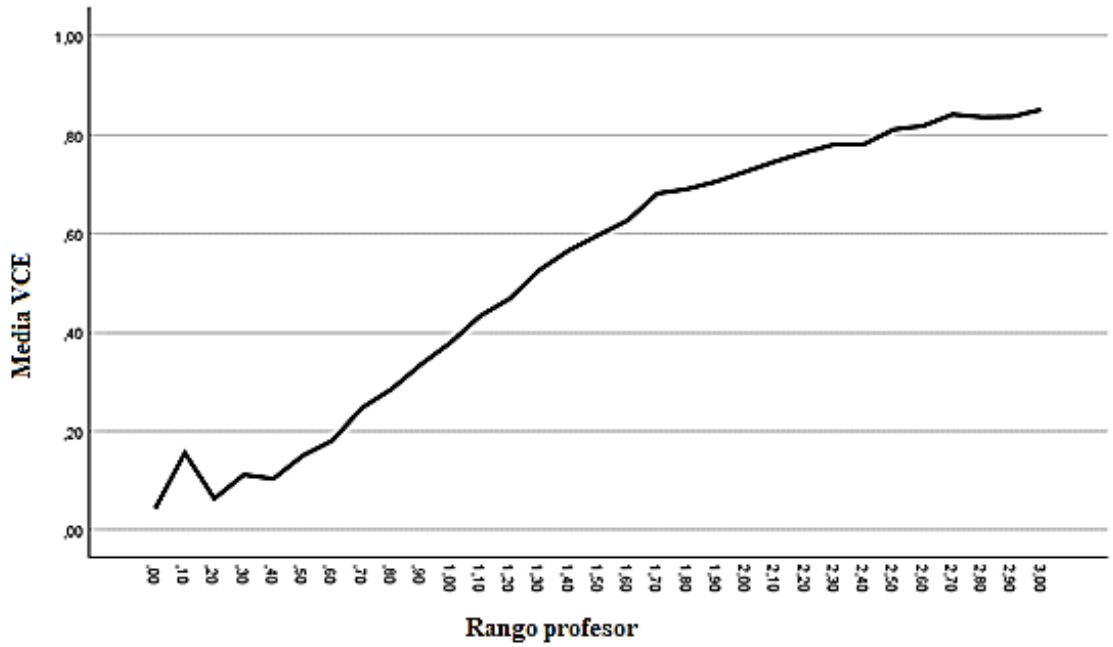


Nota: La figura representa la relación entre el valor medio del OmegaH y su variación en función de la diferencia de calificación entre profesores.

En la figura 12 se observa que el omega jerárquico a partir de una dispersión o diferencia en Z=1, alcanza un valor de omegaH= 0.6 que sería sugerente de una unidimensionalidad esencial, y a partir de Z= 1.5 el omega jerárquico aumenta a valores de 0.80.

Figura 13.

Varianza común explicada (VCE) en relación con la diferencia de puntuaciones de los profesores



Nota: La figura representa la relación entre el valor medio de la varianza común explicada y su variación en función de la diferencia de calificación entre profesores.

En la figura 13 se observa una varianza común explicada superior a 0.60, a partir de una diferencia de Z= 1.5 en las puntuaciones de los profesores, lo cual al combinarlo con el OmegaH se concluiría, erróneamente, que el instrumento tiene una estructura si no unidimensional, al menos esencialmente unidimensional.

6.4 Discusión

El objetivo de este estudio fue evaluar las consecuencias de ignorar la estructura multinivel en la recuperación del modelo teórico. De igual forma, probar mediante simulación la solución técnica propuesta, a través del estudio de la sensibilidad de los índices de bondad de ajuste, habitualmente utilizados en el contexto del análisis factorial confirmatorio. Esto 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 (Marsh y Hattie, 2002; Toland y 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, especialmente por el tamaño de la muestra.

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, se presentan problemas de validez (Messick, 1998), especialmente de la fuente estructural.

En el segundo estudio, se observó adicionalmente que, cuanto mayor eran las diferencias de las puntuaciones entre los profesores, las puntuaciones en el modelo tendían a la unidimensionalidad y en el presente estudio, se identificó que los índices de bondad de ajuste son insensibles a este problema en la evaluación de los modelos unidimensional y bifactor.

En consecuencia, modelos mal especificados con un fuerte factor general, mostraron bondad de ajuste, cuando el modelo correcto era un modelo de tres factores no correlacionados.

En el modelo multinivel los índices de bondad de ajuste mostraron los criterios esperados y así, cuando no había diferencias entre profesores (Z= 0), mostraron ausencia de bondad de ajuste lo que es coherente, si no hay variabilidad por parte del segundo nivel, entonces no es posible hablar de un modelo multinivel.

De igual forma, cuando se generaron diferencias en los diferentes rangos Z= [-1.1], Z= [-2.2] y Z= [-3.3] los índices mostraron un ajuste adecuado. Es importante señalar, que

solamente el índice SRMR mostró cierta sensibilidad en el escenario sin diferencias, tanto el modelo multinivel como en el bifactor.

Sin embargo, en el modelo unidimensional, el índice SRMR tendía a mejorar indebidamente cuando se adiciona variabilidad en las puntuaciones, tal como lo señalaba Lai (2021). 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 del segundo estudio, 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 alcanza un criterio aceptable, lo que sugiere que la varianza explicada por las variables de nivel inferior podría modificar la estructura original de los datos, lo cual confirma lo señalado por Bacci y Caviezel (2011) respecto al hecho de que fuentes de variabilidad no controladas pueden conducir a estimaciones incorrectas en la medición.

Además, específicamente se evidencia como el modelo unidimensional ajusta, cuando hay una mayor dispersión o variabilidad en la calificación de los profesores; también que se podría suponer, erróneamente, que existe una estructura bifactorial con un factor general relevante, dado que ajusta el modelo. Esto indica que una fuente de variación no controlada puede parecer o hacer que se genere una correlación, que no se deba específicamente a los factores, sino a los profesores u otras variables.

A su vez, es posible observar que cuando todos los profesores tienen calificaciones iguales Z= 0, es decir cuando no hay varianza o diferencias en las calificaciones, se observa solo la variación de los sujetos, lo que ayuda a explicar los resultados encontrados.

De igual manera, estos resultados confirman lo señalado por varios autores, respecto a que, los índices de bondad de ajuste tienen una pobre capacidad para detectar problemas de sesgo (Byrne y Van de Vijver, 2010); Hu y Bentler, 1999;), por lo que deberían examinarse

con precaución cuando se sabe que existe variabilidad en los datos, cómo lo es el caso de las evaluaciones docentes.

En este estudio, se identificaron hallazgos similares a los reportados por Cho et al., (2017), dado que los valores de RMSEA mostraron una menor sensibilidad a la variabilidad de los datos, presentando valores similares a pesar de las diferencias. Por su parte, el SRMR no ajustó en los modelos unidimensional, bifactor y multinivel, cuando no había variabilidad, lo cual es coherente con lo que representa cada modelo.

Respecto al estudio de la dimensionalidad, se confirman los resultados obtenidos al observar los índices de ajuste, es importante recordar que se simularon tres factores no correlacionados, por lo tanto, el omega jerárquico verdadero debía ser cero, así como la varianza común explicada. Sin embargo, lo que se observa es que el efecto de la variabilidad de los profesores genera el sesgo reflejado en las figuras, en el que a partir de una diferencia de Z=1.5 se obtienen valores en el OmegaH y la VCE que reflejan una estructura esencialmente unidimensional, lo cual sería erróneo frente al modelo simulado, en el que la varianza del factor general es nula. En situaciones en las que haya correlación entre factores, el error en la interpretación de la dimensionalidad no requeriría de diferencias tan grandes entre la varianza debida al profesor para concluir erróneamente que el modelo es unidimensional.

Este estudio 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., 2022) u otras valoraciones del desempeño del estudiante, que no se centre únicamente en las valoraciones de una encuesta. 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 se conoce que estas evaluaciones no están

desprovistas de limitaciones, estos primeros pasos pueden ser un buen inicio para tener mediciones de calidad.

7. Discusión y Conclusiones Generales de la Tesis

El objetivo principal de esta investigación fue identificar la forma adecuada de tratar los datos provenientes de las encuestas de evaluación docente, profundizar en la comprensión del constructo, tanto por la estructura anidada natural de sus datos, como por la forma en que se recoge la información. Así como, proponer una solución técnica sencilla para que investigadores y otros profesionales que utilizan estos datos puedan obtener estimaciones fiables y válidas.

Desde los años 20 hasta la actualidad, la estrategia principal, o la única en muchas universidades, para la recolección de datos sobre la docencia, sigue siendo las encuestas que se aplican a los estudiantes para evaluar la calidad de ésta, a partir de la satisfacción reportada por los estudiantes (Banta y Blaich, 2011). Dada su relevancia y consecuencias para los profesores, se ha identificado que las investigaciones sobre este tema superan en número a cualquier otro tema de investigación en educación superior (Marsh, 2011).

Además, la revisión teórica muestra la existencia de un gran número de instrumentos de evaluación docente, ya que cada universidad establece sus propios criterios sobre la calidad de la docencia (Oermann et al., 2018). En España, los usos principales que se identifican están relacionados con la evaluación del proceso de enseñanza y aprendizaje, la acreditación de los docentes y las titulaciones, y la entrega insumos para planes de mejora (Denson et al., 2010; Espinosa et al., 2017); y en Latinoamérica su uso se centra en la regulación de la carrera del docente, como ascensos, incrementos salariales, promociones o destituciones, así como la acreditación de programas de formación (OCDE, 2013a, 2013b).

Sin embargo, aunque estas evaluaciones se utilizan ampliamente, se ha cuestionado su eficacia para los propósitos previstos, debido a limitaciones teóricas y prácticas, al uso que se les da a los resultados, al ideal de docente que representan y a la baja tasa de respuesta por parte

de los estudiantes (Turull y Buxarrais, 2018), esto último se vio reflejado en la base de datos reales utilizada en esta investigación.

Esto ha llevado a problemas de diversa índole, tales como las estrategias y recursos empleados en la evaluación docente; la insuficiente integración de la evaluación docente dentro del sistema educativo; y la ilegitimidad de los procesos debido a la carencia de una cultura de evaluación (Escudero, 2019).

Respecto a este último aspecto, que afecta tanto a la medición como al uso, la revisión teórica permitió identificar que gran parte del problema de legitimidad, se debe a que los profesores ven esta evaluación como un elemento aislado de la calidad docente, cuyos datos son principalmente utilizados para reportes externos desvinculados de la academia, es decir, más como una serie de índices, que un elemento de mejora con algún impacto en su trabajo (Gelber, 2020). Por su parte, los estudiantes consideran que sus aportes no son tenidos en cuenta, que son instrumentos subjetivos, largos y de poca utilidad para la mejora de la enseñanza (López, 2019; Silva, 2007).

En este sentido, fue posible identificar varias de las críticas que se han realizado a estos instrumentos, dadas las implicaciones en decisiones académicas y administrativas en las que influyen. Entre las críticas, destacan la presencia de sesgos que pueden abarcar desde las características del docente, del estudiante, la capacidad de los estudiantes como evaluadores y las consideraciones al momento de evaluar, permeado por diversas variables extrañas que no evidencian una relación directa con la calidad de la enseñanza (Miles y House, 2015; Spooren et al., 2013; Uttl et al., 2017, 2021).

De todo esto surge un aspecto relevante, que da sustento al desarrollo de esta tesis, es lo afirmado por Kreitzer y Sweet-Cushman (2022) quienes indican que las evaluaciones docentes son medidas deficientes de la satisfacción del aprendizaje de los estudiantes, y a su vez, son medidas imperfectas del desempeño de los docentes. Esto genera una serie de

inquietudes: ¿Qué miden realmente las encuestas de evaluación docente? ¿Se mide la calidad de la docencia o la satisfacción de los estudiantes? ¿Ambos aspectos en distintas proporciones? ¿Cómo se determina la especificación del constructo? ¿Son útiles estas evaluaciones? ¿Es posible confiar en estas medidas para tomar decisiones?

Los sesgos identificados en la evaluación docente se relacionan con inquietudes acerca de la dimensionalidad, que generan otras preguntas relacionadas: ¿Cuántas y cuáles dimensiones permiten especificar la medición de la enseñanza efectiva? y ¿Es posible compilar una puntuación general basada en estas dimensiones?

De lo identificado en los capítulos 2 y 3, que corresponden al estado del arte sobre la historia y las investigaciones realizadas alrededor de la evaluación docente; se rescatan los elementos teóricos más relevantes y surgieron algunas preguntas que apoyaban y orientaban el desarrollo de esta investigación. ¿En qué medida las evaluaciones pueden diferenciar de manera confiable el desempeño entre los profesores? ¿Las calificaciones dadas por los estudiantes a un profesor en una clase son consistentes y aplicables a otras clases impartidas por el mismo profesor? y ¿Las diferencias entre los profesores tienen validez en relación con otros indicadores de desempeño docente? Como se puede evidenciar, la mayoría de las preguntas que dejó la revisión teórica de este capítulo, apuntan a elementos sobre las evidencias de validez de los instrumentos y la forma en que han sido cuestionados.

En el capítulo 3, dirigido a revisar conceptualmente el problema metodológico, se identificó que en la literatura se pone de manifiesto aspectos relacionados con la naturaleza de los datos obtenidos de la evaluación docente, al señalar que su estructura es anidada, jerárquica o multinivel y que no tenerla en cuenta constituye en sí mismo un sesgo (Pardo y Ruíz, 2015), la importancia de conocer la varianza y los componentes de la varianza de las puntuaciones que se desprenden de este tipo de evaluaciones.

Por ejemplo, se identificó que existía varianza diferenciada para los estudiantes, los profesores y los tipos de curso (Feistauer y Richter, 2016), se dejó de manifiesto que se suele confundir la satisfacción con el profesor y la calidad de la docencia, cuando son dos aspectos claramente diferentes.

La satisfacción estudiantil centrada en qué tan conformes se encuentran los estudiantes con la educación que reciben (Chuyma-Huilca et al., 2021) incluyendo aspectos como el bienestar emocional, la aplicación laboral percibida y los beneficios de la educación (Inzunza-Melo et al., 2015); mientras que la calidad de la docencia se asocia con el clima de aprendizaje seguro y estimulante, gestión eficiente del aula, claridad de la enseñanza, enseñanza activadora, estrategias de enseñanza-aprendizaje y diferenciación (Fernández-García et al., 2022).

Sin embargo, también se evidenció que a pesar de toda la investigación que existe alrededor de este tema, persiste la incertidumbre acerca de los factores que influyen en las puntuaciones generadas por estas evaluaciones y la forma adecuada de interpretarlas (Gravestock y Gregor-Greenleaf, 2008). Por lo tanto, es crucial esclarecer el significado de las puntuaciones de evaluación de los estudiantes, ya que, como se ha mencionado en diferentes oportunidades, estos datos suelen ser determinantes en decisiones importantes que afectan directamente la carrera de los profesores.

Con la medición, toda la variación de las puntuaciones debería depender del profesor y específicamente de la calidad de la enseñanza. En este sentido, se espera que la variabilidad asociada al docente refleje el nivel de diferencia entre profesores que imparten los mismos cursos, sin considerar las particularidades individuales que puedan influir en esas diferencias de calificación, como la edad, el género, el atractivo físico, la experiencia o la personalidad del docente (Boring et al., 2016; Delucchi, 2000; Hamermesh y Parker, 2005; MacNell et al., 2015; Wilson et al., 2015; Zabaleta, 2007).

Por esto, es necesaria la claridad en la unidad de análisis, dado que, si el profesor aporta gran parte de la varianza, pero no explica suficiente o no es el efecto principal, los análisis pueden estar sesgados, apareciendo falacias estadísticas (Sproule, 2000) y llegar a confusiones con los resultados, tanto de la medición como de las propiedades psicométricas y la estructura factorial de los instrumentos (Curby et al., 2019). Es fundamental, utilizar métodos multinivel, adoptando modelos coherentes con la conceptualización jerárquica del constructo, identificar en qué nivel está el constructo de interés, su relevancia, el tipo específico de constructo en cuestión y, posteriormente, cómo estimar y especificar el modelo para ese constructo. Para el tratamiento de estos datos, se recomiendan los modelos de análisis factorial multinivel, resaltando el uso del modelo de Hox (2002).

En el capítulo 4, se realizó un análisis clásico con datos reales, tal como suele hacerse en las instituciones, utilizando la TCT e incluso se presentaron las ventajas del uso de la TRI en función del cumplimiento de los supuestos, realizando los análisis considerando los datos como independientes e ignorando su estructura jerárquica. En este estudio, fue posible identificar excelentes propiedades psicométricas del instrumento y la indicación de una estructura unidimensional, lo cual, ante los ojos de una persona con poca experiencia en investigación o análisis de datos, puede parecer coherente, dado que las técnicas estadísticas se utilizaron de forma correcta y conforme a la evidencia presentada.

Por esto, la importancia de la formación que reciben las personas que analizan y manejan la información resultante de los cuestionarios de actividad docente, así como de la necesidad de capacitación y actualización (Penny, 2003).

Los análisis realizados, ponen en evidencia que si bien los resultados estadísticos y la información que es posible obtener a través de las técnicas básicas de TCT y la complementariedad que ofrecen modelos TRI como el de Samejima (2010), son interesantes e informativos, siempre y cuando los modelos en realidad sean unidimensionales, situación poco

probable en el escenario de la evaluación docente, donde se ha mencionado ampliamente su naturaleza multidimensional (Bacci y Caviezel, 2011; Bacci y Gnaldi, 2015) y según Spooren et al. (2013), aún se sigue cuestionando el número y las dimensiones de los instrumentos para evaluar la enseñanza efectiva, y la posibilidad real de estimar una puntuación general basada en estas dimensiones, dado que se tratan como datos independientes, que desconocen su naturaleza anidada.

Además, la evaluación de la adecuación de un modelo de Teoría de Respuesta al Ítem (TRI) en poblaciones diversas es un desafío complejo que demanda una investigación detallada (Hambleton et al.,1991). Esto implica el uso de muestras grandes y representativas para cada subgrupo relevante, lo que facilita la verificación empírica de la invarianza métrica en cada subpoblación y permite hacer inferencias a nivel poblacional. Sin embargo, dado el carácter exigente de los modelos TRI y la complejidad inherente a la realidad, es probable que se encuentren diferentes niveles de discrepancia y falta de uniformidad en los datos reales.

En consecuencia, es poco probable hallar parámetros completamente invariantes (DeMars, 2010). Esto además, hace énfasis en que los procedimientos estadísticos pueden calcular la información que se le brinde, pero garantizar que esa información tenga un valor teórico y represente la realidad, dependerá de los cuestionamientos y la orientación que un investigador entrenado otorgue, porque aunque las propiedades psicométricas que se reflejan sean excelentes, aún no se responde la pregunta sobre la naturaleza del constructo y lo que realmente se está midiendo ¿calidad docente o satisfacción estudiantil?

Estos resultados evidencian que si bien, el modelo de Samejima (2010) puede tener ventajas sobre la información reportada en la TCT, sigue siendo una información limitada porque no contempla la estructura jerárquica. En ese sentido, modelos más pertinentes de análisis serán aquellos que contemplen la estructura anidada de los datos y a partir del cumplimiento de los supuestos se generen las mediciones, modelos tales como el Análisis

Factorial Confirmatorio Multinivel -MFCA (Hox, 2002; Muthén, 1994), Modelos de Clases Latentes Multinivel (Vermunt, 2003), Modelo Multinivel TRI - MIRT (Fox y Glas, 2001) o los Modelos de Mixturas de la Teoría de Respuesta al Ítem- MMixIRT (Cho y Cohen, 2010).

Ante la imposibilidad de utilizar estos modelos más robustos en los datos aplicados, especialmente por el tamaño de la muestra, así como por la falta de cumplimiento de los supuestos necesarios para su uso; en el segundo estudio presentado en el capítulo 5, fue necesario formular una solución para el análisis de este tipo de datos, de carácter sencillo, que respetase la naturaleza jerárquica de los datos equivalente al análisis factorial multinivel y considerara las limitaciones de muestra.

En primer lugar, se identificaron las fuentes de variabilidad, las dos fuentes que aparecen reflejadas corresponden a los estudiantes y a los docentes, en concordancia con lo señalado por Spooren et al. (2013), frente a la imposibilidad de utilizar una puntuación global para dar cuenta de este fenómeno. Una vez comprendidas estas dos fuentes de varianza fue necesario compensar la diferencia entre los profesores y fijar la varianza del docente para determinar la estructura factorial. Esta solución consta de cuatro (4) pasos, que implica: a) la identificación de la variabilidad de los componentes; b) la alineación de la media de la variable con mayor varianza explicada; c) la estimación del modelo para los datos corregidos; y d) la comparación entre variables.

Esta serie de pasos se ofrece como una opción técnica, al ser inaplicables los modelos más complejos como el MFCA o los modelos multinivel TRI, porque con 10 o menos estudiantes no es viable utilizar estos procedimientos y aunque matemáticamente pueden funcionar, se debe evitar de alguna manera la falacia y el sesgo que se evidencia en los análisis, cuando no se considera la anidación de los datos. En este sentido o se mejora la calidad y cantidad de los datos para que sea viable aplicar las técnicas más robustas, o se aplican

correcciones como la propuesta para mitigar los diferentes sesgos, que siguen determinando en gran medida las mediciones realizadas.

Tal como se reflejó en la investigación de Jibaja (2023) en un estudio en el que se identificaron clases latentes de sujetos, los hallazgos reflejaron que los estudiantes en la clase latente 1, donde se encuentran las puntuaciones más altas en las evaluaciones docentes, en contraste con otras clases latentes, suelen tener más de dos años de estudios universitarios y se matriculan en asignaturas de nivel intermedio o avanzado, que no son necesariamente obligatorias, sugiriendo que los estudiantes más satisfechos con la enseñanza, al tener más experiencia universitaria, poseen un mayor interés en los temas que aquellos que no lo tienen (La Rocca et al., 2017). En contraposición, a los sujetos pertenecientes a la clase latente 4, en la que las respuestas que puntúan más bajo respecto al docente proceden de estudiantes en asignaturas estrictamente obligatorias, lo que refuerza esta relación.

En la propuesta presentada, aparecía un primer nivel en la solución de las limitaciones mencionadas y se expuso en una serie de pasos para el control de sesgo. 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).

Luego de aplicar la solución e identificar una estructura factorial de tres factores, se revisó el contenido del cuestionario y se observaron las diferencias en los ítems, respecto a lo medido. Lo que además permite que el docente identifique fortalezas y debilidades frente a la satisfacción de sus estudiantes y así no asumir como un todo, la puntuación obtenida; cuando

en realidad son aspectos diferenciados los que se están valorando, como en este caso, que se refieren a el acompañamiento directo del docente, la organización y estructuración de las clases y la claridad de las explicaciones del docente.

Este es uno de los aspectos más relevantes a resaltar de esta investigación, dado que, el hecho de que la estructura factorial cambie al controlar las distintas fuentes de varianza puede indicar que la medida no mide lo que pretende medir, lo que afecta directamente a que las evidencias de validez no estarían sustentando ni sus interpretaciones, ni su uso (Borsboom et al., 2004; Messick, 1998). Estos hallazgos, muestran una respuesta más clara sobre qué es lo que se está midiendo, si la satisfacción de los estudiantes o la calidad de los docentes, y los análisis apuntan a la satisfacción de los estudiantes.

Sin embargo, para el análisis de las propiedades psicométricas, así como las interpretaciones que se pueden derivar del instrumento, será necesario revisar si el interés se enfoca en los profesores, será necesario centrar la puntuación para controlar la variabilidad debido a los estudiantes y así evitar confusiones de la que se espera brindar información.

Al aplicar la corrección, se identifica que el constructo es la satisfacción estudiantil. A partir de los planteamientos teóricos y los análisis realizados, se concluye que no es adecuado considerar que estos cuestionarios sean procedimiento adecuados para medir la efectividad de la calidad de la enseñanza, ya que, existen demasiadas variables asociadas a la valoración por parte de los estudiantes (Boring et al., 2016; Delucchi, 2000; MacNell et al., 2015; Hamermesh y Parker, 2005; Wilson et al., 2015; Zabaleta, 2007) y adicionalmente, no se pueden considerar como jueces adecuados los estudiantes, dados los sesgos descritos por las características personales del profesor o las características de la asignatura que se tienen en cuenta, lo que aleja los estudiantes de ser evaluadores idóneos, como señalaba Luengo (2019) por falta de experiencia o conocimiento necesario.

Para que fuera posible medir la calidad de la docencia, sería necesario revisar los aspectos a evaluar de lo que se considera la eficacia de la docencia, el evaluador y el cruce con otras fuentes de información; de lo contrario solo con la encuesta de valoración realizada por los estudiantes solo se mide la satisfacción estudiantil, por lo que los tomadores de decisiones de institutos, centros educativos y universidades, deben ser cuidadosos al momento de basar sus decisiones únicamente en esta fuente de información.

Las actividades realizadas en esta investigación permitieron descontaminar los datos, para poder mejorar la observación del fenómeno respecto a la evaluación de los profesores, dado que se evidencia, que, en la actualidad, es posible realizar la revisión de modelos o estudios parciales, pero se sigue teniendo la dificultad de modelar estos datos completamente por su nivel de complejidad. Cuando no se aplica el modelo adecuado, se está ignorando la naturaleza de los datos que puede llevar como dice Sproule (2000), a falacias estadísticas puesto que se consideran factores correlacionados cuando son independientes, aparecen estructuras factoriales falsas, como el caso del capítulo 4, con estimaciones inadecuadas, que llevan a conclusiones erróneas sobre la fiabilidad y las conclusiones sobre los profesores, tal como lo planteó Curby et al., (2019).

Esto reafirma la necesidad señalada por Bacci y Caviezel (2011) respecto 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). La solución propuesta en la presente investigación permitirá identificar en dónde se debe centrar el foco, si en los estudiantes o en los docentes y así poder proceder con los demás análisis.

En el capítulo 6, se abordó el objetivo de evaluar las consecuencias de ignorar la estructura multinivel en la recuperación del modelo teórico, a partir del desarrollo de una simulación. Además de validar y probar la solución presentada en el capítulo 5, se estudió 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, se investigaron los efectos estructurales al ampliar la proporción de la varianza explicada por las variables de un nivel inferior, considerando la variabilidad de un nivel superior en el que estas variables están incluidas de manera jerárquica.

Se identificó que, de acuerdo con lo señalado por Messick (1998), se presentan problemas de validez, especialmente de la fuente estructural, dado que, una vez realizados los análisis, la estructura factorial podía ajustarse a otros modelos diferentes al previamente simulado. En este sentido, se confirma lo señalado por Bacci y Caviezel (2011), quienes indican que fuentes de variabilidad no controladas pueden conducir a estimaciones incorrectas en la medición, como en el caso de la simulación en la que se identificó que la varianza explicada por las variables de nivel inferior (estudiantes) lograba modificar la estructura original de los datos.

Asimismo, que, si bien una medida relevante para la confirmación de una estructura factorial son los índices de ajuste, estos son poco sensibles para detectar el sesgo, de acuerdo con lo señalado por Hu y Bentler (1999) y Byrne y van de Vijver (2010), por lo que, aunque los procedimientos utilizados sean adecuados, sin la identificación de los efectos de la varianza, pueden aparecer estructuras factoriales incorrectas. En el caso de la simulación, una estructura de tres factores no correlacionados podía asumirse como una estructura bifactorial o una estructura unidimensional, en los casos con mayor dispersión entre las calificaciones de los profesores, lo cual sería una falacia estadística.

Es importante señalar, que, para efectos prácticos, la consecuencia principal es una estimación potencialmente sesgada para el profesor respecto a las dimensiones que mide el cuestionario, así como de sus habilidades docentes. Si esta estimación es tenida en cuenta dentro de la institución, como lo señalaban Denson et al. (2010) y Espinosa et al. (2017) para

evaluación del proceso enseñanza y aprendizaje, la acreditación de los docentes, la acreditación de las titulaciones, el insumo para planes de mejora, entre otros.

Estos resultados podrían generar consecuencias adversas para el profesor con evaluaciones negativas y también, en sentido contrario, obtener beneficios con valoraciones altamente positivas que no son reales. Además, teniendo en cuenta que estas puntuaciones pueden derivar de aspectos que deberían ser controlados o estimados, tales como la titulación (Basow y Montgomery, 2005), la asignatura (Arámburo y Luna, 2013; Luna et al., 2010, la edad (Wilson et al., 2015) el género del profesor (Boring et al., 2016; MacNell, et al., 2015) o el atractivo físico (Hamermesh y Parker, 2005), entre otras tantas variables que han mostrado tener efectos sobre las puntuaciones de satisfacción.

7.1 Limitaciones

Como se observó en la simulación, lo ideal sería contar con datos suficientes de estudiantes y profesores para poder aplicar las técnicas multinivel, sin embargo, en la realidad es difícil tener esas condiciones que permitan usar técnicas más potentes. Aún se están desarrollando los paquetes para analizar estas estructuras de datos y todavía no existen modelos estándar para su representación gráfica.

Generalmente, estos modelos han sido probados con datos simulados, pero es difícil hacerlo con datos empíricos como los utilizados en esta investigación, en la que aparentemente hay una gran cantidad de estudiantes, pero cuando se separan por centro, asignatura, profesor, u otras fuentes relevantes, los grupos representativos disminuyen considerablemente.

Otra limitación que puede afectar la generalización de los resultados es que solamente se contaba con los datos de esta universidad. Es posible que los resultados con otros datos empíricos fueran diferentes, en función del carácter público o privado de la universidad, si la modalidad es presencial o virtual, el tamaño de la universidad y de los grupos por clase, incluso

de las características propias del instrumento. Por ejemplo, mayor cantidad de ítems, dimensiones de evaluación diferenciadas, entre otras.

El problema actual que se está evidenciando en esta investigación, es que existen herramientas metodológicas potentes que permitirían hacer unas estimaciones bastante precisas de datos de estructuras anidadas, pero en la práctica no se pueden utilizar porque no hay suficientes sujetos para controlar las diferentes fuentes de variabilidad, se necesitaría un escenario en el que se tenga suficientes personas para tener diferentes tipologías de profesores, organizados por sus características personales y suficientes variaciones de clases de sujetos.

Por esta razón, es necesario seguir investigando sobre soluciones como las elaboradas para el tratamiento de estos datos y así evitar que se sigan mezclando diferentes tipos de profesores y tipos de alumnos, al momento de realizar estimaciones del constructo y valoración de las propiedades psicométricas de los instrumentos. Y, claro está, mientras se continúa la línea de investigación, se sugiere utilizar la solución técnica propuesta para corregir la fuente de variabilidad de interés, centrando la variable de estudio. Esto haría posible tener una recuperación correcta de la estructura del instrumento aplicado, y, en consecuencia, una estimación más precisa de los parámetros de interés.

La solución propuesta, aporta evidencia a la discusión frente a lo que evalúan estos cuestionarios, es decir, dada la dicotomía entre medir la calidad de la docencia o la satisfacción estudiantil, y a partir de la teoría y los datos, es posible afirmar que estas mediciones realizadas por los estudiantes, están influenciadas por diferentes sesgos y variables intervinientes, que no están relacionadas con la calidad de la docencia, sino esencialmente con la satisfacción subjetiva de los estudiantes. Así, esta solución permite, al separar la varianza, enfocarse en que la variabilidad obtenida se debe a la satisfacción estudiantil, controlando la variabilidad que depende de la calidad de la docencia.

7.2 Futuras líneas de investigación

Con base en la investigación realizada, las futuras líneas de investigación podrían dirigirse a: a) establecer nuevos criterios que permitan asegurar la validez de las mediciones, puesto que desconocer la estructura de datos multinivel hace que la dependencia entre los grupos vulnere la supuesta independencia de las observaciones y, por lo tanto, constituyen un sesgo; b) considerar diversas fuentes de información y compararlas para tener evidencia de la decisión sobre la estructura factorial, ya que esta investigación mostró que los criterios de ajuste estadístico no siempre apoyan las mejores decisiones; c) complementar con otras medidas de los estudiantes o del docente, dado que si bien se habla de evaluación, al usar solamente una fuente de información como son las encuestas, ese está ante un proceso de medición pero no de evaluación (Montero y León, 2002), lo cual está sujeto a la falta de evidencias de validez que sustente su uso e interpretaciones; d) realizar estudios de evidencias adicionales de validez; e) estudiar de modo sistemático el funcionamiento de las técnicas de otras técnicas de detección de dimensionalidad en presencia de estructuras multinivel; f) comparar los métodos de análisis de datos multinivel, explorando la viabilidad de implementar enfoques más eficaces que aborden este sesgo; g) analizar la información ofrecida por diferentes aspectos de la calidad del test al emplear el procedimiento propuesto en esta investigación, en contraposición a no utilizarlo, con el fin de evaluar sus beneficios; h) examinar la influencia del número de ítems en el cuestionario, ya que en instrumentos más extensos se podría lograr un mejor control de las diversas fuentes de variabilidad; i) desarrollar herramientas, programas o paquetes estadísticos en R, Python u otro, que permitan añadir las variables necesarias, que representan la complejidad de fenómenos como la evaluación docente, por ejemplo, si bien se identificaron algunos de los sesgos o la varianza que explican algunas variables, falta desarrollar análisis estadísticos que permitan comprender de manera precisa cómo estas variables influyen específicamente en las variaciones de las puntuaciones en otros niveles jerárquicos, como a

nivel de asignaturas, programas académicos o facultades, entre otros; y j) fortalecer la capacitación en materia estadística, psicométrica y la actualización de modelos a quienes trabajan con este tipo de datos (Beleche et al., 2012); ya que, solo el estudio de estas cuestiones permitirá hacerse nuevas o diferentes preguntas sobre los resultados obtenidos.

Finalmente, una reflexión que va más allá de los métodos se centra en comprender la esencia del fenómeno y la información obtenida mediante estos cuestionarios, la cual puede variar entre instituciones. Esta evaluación puede contribuir a determinar el método más adecuado que brinde una visión más precisa sobre la calidad de los instrumentos. En este sentido, es esencial fomentar la capacitación del personal que trabaja con estos cuestionarios o bases de datos, respecto a soluciones metodológicas que puedan mejorar las decisiones, no sólo en relación con el instrumento, sino también en la retroalimentación proporcionada a los docentes o instituciones.

En conclusión, esta tesis aporta al estado del arte de la evaluación docente, respecto a las evidencias de validez, especialmente en la evidencia de validez de constructo estructural. Se concluye que, a pesar de que existen técnicas robustas y detalladas para realizar el análisis multinivel, éstas suelen ser poco utilizadas en la realidad, dado que no se cumplen los supuestos básicos para su uso.

La solución presentada es una primera aproximación y abre el camino para identificar otras soluciones que permitan el análisis anidado de los datos y se validó su utilidad para evitar dificultades en el análisis de la dimensionalidad. Un hallazgo fundamental y que quedó evidenciado en las pruebas de unidimensionalidad esencial, a través del análisis de varianza común explicada y el omega jerárquico, junto con los índices de bondad de ajuste, es el hecho de que el tratamiento inadecuado de los datos con naturaleza multinivel, su desconocimiento o el hecho de ignorarse, generan una interpretación errónea del instrumento, lo que conduce a

errores en las estimaciones y por ende dificultades en las decisiones basadas que se puedan adoptar respecto a la calidad de la docencia del profesorado.

8. Referencias

Abad, F., Ponsoda, V. & Revuelta, J., (2006). *Modelos politómicos de respuesta al ítem*. La Muralla

Abad, F.; Olea, J.; Ponsoda, V. y García, C. (2011). *Medición en ciencias sociales y de la salud.* Editorial Síntesis.

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. *Summa Psicológica UST*, *11*; 2; 12-2014; 101-113 https://doi.org/10.18774/448x.2014.11.158

Abel, M. H., & Meltzer, A. L. (2007). Student ratings of a male and female professors' lecture on sex discrimination in the workforce. *Sex Roles, 57*(3–4), 173–180 https://doi.org/10.1007/s11199-007-9245-x

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

Aleamoni, L.M. and Hexner, P.Z. (1980). A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. *Instructional Science*. *9*(1), 67-84. https://doi.org/10.1007/BF00118969

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.

https://scholarsarchive.byu.edu/etd/6939/?utm_source=scholarsarchive.byu.edu%2Fetd%2F6939&utm_medium=PDF&utm_campaign=PDFCoverPages

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2018). *Estándares para pruebas educativas y psicológicas* (M. Lieve, Trans.). Washington, DC: American Educational Research Association. (Original work published 2014) and Learning, 87, 85-100. https://doi.org/10.1002/tl.10001

Anderson, H. M., Cain, J., & Bird, E. (2005). Online student course evaluations: Review of literature and a pilot study. *American Journal of Pharmaceutical Education*, *69*(1), 5. https://doi.org/doi:10.5688/aj690105

Anderson, K. J., & Smith, G. (2005). Students' preconceptions of professors: Benefits and barriers according to ethnicity and gender. *Hispanic Journal of Behavioral Sciences*, *27*(2), 184–201 https://doi.org/10.1177/0739986304273707

Apodaca, P., & Grad, H. (2005). The dimensionality of student ratings of teaching: integration of uni and multidimensional models. *Studies in Higher Education*, *30*(6), 723-748. https://doi.org/10.1080/03075070500340101

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

Arbesú, M. I. (2009). Evaluación comprehensiva de la docencia universitaria. *Revista Casa del Tiempo*, *2*(24), 17-21

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

Arnold, I. J. M., & Versluis, I. (2019). The influence of cultural values and nationality on student evaluation of teaching. *International Journal of Educational Research, 98,* 13-24. https://doi.org/10.1016/j.ijer.2019.08.009

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker.

Asún, R., & Zúñiga, C. (2008). Ventajas de los Modelos Politómicos de Teoría de Respuesta al ítem en la Medición de Actitudes Sociales: El Análisis de un Caso. *Psykhe, 17*(2), 103-115. http://doi.org/10.4067/S0718-22282008000200009

Ato, M., López-García, J. J., & Benavente, A. (2013). Un sistema de clasificación de los diseños de investigación en psicología. *Anales de Psicología / Annals of Psychology*, *29*(3), 1038–1059. https://doi.org/10.6018/analesps.29.3.178511

Attorresi, H., Abal, F., Galibert, M., Lozzia, G., & Aguerri, M. (2011). Aplicación del modelo de respuesta graduada a una escala de voluntad de trabajo. *Interdisciplinaria, 28*(2), 231-244. https://explore.openaire.eu/search/result?id=doajarticles::b1267ec660ee2458a7266ac06375e51e

Avery, R.J., Bryant, W.K., Mathios, A., Kang, H. and Bell, D. (2006). Electronic course evaluations: does an online delivery system influence student evaluation? *The Journal of Economic Education*, *36*(1), 21-37 https://doi.org/10.3200/JECE.37.1.21-37

Baayen, R. H., D. J. Davidson, and D. M. Bates. 2008. Mixed Effects Modeling with Crossed Random Effects for Subjects and Items. *Journal of Memory and Language, 59*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

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

Bacci, S., & Gnaldi, M. (2015). A classification of university courses based on students' satisfaction: An application of a two-level mixture item response model. *Quality & Quantity: International Journal of Methodology, 49*(3), 927–940. https://doi.org/10.1007/s11135-014-0101-0

Bachen, C. M., McLoughlin, M. M., & Garcia, S. S. (1999). Assessing the role of gender in college students' valuations of faculty. *Communication Education, 48*(3), 193–210 https://doi.org/10.1080/03634529909379169

Baghaei, P. (2013). Fitting the Mixed Rasch Model to a Reading Comprehension Test: Identifying Reader Types. *Practical Assessment, Research, and Evaluation, 18*(1):5. https://doi.org/10.7275/n191-pt86

Banta, T. W., & Blaich, C. (2011). Closing the assessment loop. *Change: The Magazine of Higher Learning, 43*(1), 22-27. https://doi.org/10.1080/00091383.2011.538642

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*, *13*(2), 324-329. https://www.redalyc.org/articulo.oa?id=72721323

Basow, S. A. (1995). Student evaluations of college professors: When gender matters. *Journal of Educational Psychology, 87*(4), 656 https://doi.org/10.1037/0022-0663.87.4.656

Basow, S. A., & Montgomery, S. (2005). Student ratings and professor self-ratings of college teaching: Effects of gender and divisional affiliation. *Journal of Personnel Evaluation in Education, 18*(2), 91–106. https:10.1007/s11092-006-9001-8

Basow, S. A., & Silberg, N. T. (1987). Student evaluations of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308 https://doi.org/10.1037/0022-0663.79.3.308

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

Batista, J M y Coenders, G. (2000). *Modelos de ecuaciones estructurales*. La Muralla.

Baugh, S.G., Hunt, J.G. and Scandura, T.A. (2006). *Reviewing by the numbers: evaluating quantitative research,* in Baruch, Y., Sullivan, S.E. and Schepmyer, H.N. (Eds), *Winning Reviews: A Guide for Evaluating Scholarly Writing (*156-172), Palgrave Macmillan. https://doi.org/10.1108/HEED-05-2018-0012

Beleche, T., Fairris, D., & Marks, M. (2012). Do course evaluations truly reflect student learning? Evidence from an objectively graded post-test. *Economics of Education Review, 31*(5), 709-719. https://doi.org/10.1016/j.econedurev.2012.05.001

Benton, S. L., Cashin, W. E., & Kansas, E. (2012). Idea paper# 50 student ratings of teaching: A summary of research and literature. *The IDEA Center*. https://www.researchgate.net/publication/308904846_IDEA_Paper_No_50_Student_ratings_of_teaching_A_summary_of_research_and_literature

Beran, T. N., & Rokosh, J. L. (2009). Instructor's perspectives on the utility of student ratings of instruction. *Instructional Science, 37*(2), 171-184. https://doi.org/10.1007/s11251-007-9045-2

Berge, J. M., & Socan, G. (2004). The greatest lower bound to the reliability of a test and the hypothesis of unidimensionality. *Psychometrika, 69,* 613–625 (2004). https://doi.org/10.1007/BF02289858

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education*, *17*(1), 48- 62. http://www.isetl.org/ijtlhe/

Birnbaum, R., & Snowdon, K. (2003). Management fads in higher education. *The Canadian Journal of Higher Education*, *33*(2) https://doi.org/10.1353/rhe.2003.0025

Blaich, C., Wise, K., Pascarella, E. T., & Roksa, J. (2016). Instructional clarity and organization: It's not new or fancy, but it matters. *Change: The Magazine of Higher Learning*, *48*(4), 6-13. https://doi.org/10.1080/00091383.2016.1198142

Bliese, P. D. (2000). *Within-group agreement, nonindependence, and reliability: Implications for data aggregation and analysis*. In: K. J. Klein, & S. W. Kozlowski (Eds.), *Multilevel theory, research, and methods in organization*. Jossey-Bass.

Bollen, K A (1989). *Structural equations with latent variables*. Wiley.

Bolívar, A. (2008). Evaluación de la práctica docente. Una revisión desde España. RIEE. *Revista Iberoamericana de Evaluación Educativa*, *1*(2), 56–74. https://doi.org/10.15366/riee2008.1.2.004

Bonitz, V. S. (2011). Student evaluation of teaching: Individual differences and bias effects. *Graduate Theses and Dissertations. Paper 12211. http://lib.dr.iastate.edu/etd/1221*.

Bonwell, C., y Eison, J. A. (1991). *Active Learning: Creating Excitement in the Classroom*. ASHE-ERIC Higher Education Reports No1. The George Washington University, School of Education and Higher Education

Boring, A., Ottoboni, K., & Stark, P. (2016). Student evaluations of teaching (mostly) do not measure teaching effectiveness. *ScienceOpen Research.* 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

Bowen, H. R. (1979). *Evaluating educational quality*: A conference summary. In A. W. Astin (Ed.), *Evaluating Educational Quality: A Conference Summary* (1-36). Association of American Colleges.

Boyan, N. J., & Copeland, W. D. (1978). *Instructional Supervision Training Program: Training Coordinator's Guide*. Merrill. https://flinders.primo.exlibrisgroup.com/permalink/61FUL_INST/110vnsd/alma9943053501771

Boysen, Guy A. (2015). Preventing the overinterpretation of small mean differences in student evaluations of teaching: An evaluation of warning effectiveness. *Scholarship of Teaching and Learning in Psychology, 1*(4), 269–282. https://doi.org/10.1037/ stl0000042

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

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 disease. *Parkinsonism & related disorders*, *21*(2), 142-146. http://dx.doi.org/10.1016/j.parkreldis.2014.11.016

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

Browne, M. W., & Cudeck, R. (1993). *Alternative ways of assessing model fit*. En K. A. Bollen, & J. S. Long (Eds.), *Testing structural equation models*, 136-162. SAGE.

Burnett, K., Bonnici, L.J., Miksa, S.D., & Kim, J. (2007). Frequency, intensity and topicality in online learning: An exploration of the interaction dimensions that contribute to student

satisfaction in online learning. *Journal of Education for Library and Information Science, 48,* 21-35. https://www.jstor.org/stable/40324318

Burns, C. W. (2000). Another perspective: ¿are teaching portfolios a scam? *Academe, 86* (1), 44-47. https://www.proquest.com/docview/232309162?sourcetype=Trade%20Journals

Buurman, M., Delfgaauw, J. J., Dur, R. A. J., & Zoutenbier, R. (2018). The effects of student feedback to teachers: Evidence from a field experiment (Tinbergen Institute Discussion Paper). Amsterdam & Rotterdam. http://dx.doi.org/10.2139/ssrn.3168466

Byrne, B. M. (1998). *Structural equation modeling with LISREL, PRELIS, and SIMPLIS: Basic Concepts, applications and rogramming*. Lawrence Erlbaum Associates.

Byrne, B. M., & van de Vijver, F. (2010). Testing for measurement and structural equivalence in large-scale cross-cultural studies: Addressing the issue of nonequivalence. *International Journal of Testing*, *10*(2), 107–132. https://doi.org/10.1080/15305051003637306

Caballero, C. C., Abello, R. y Palacios, J. (2007). Relación del *burnout* y el rendimiento académico con la satisfacción frente a los estudios en estudiantes universitarios. *Avances en Psicología Latinoamericana*, *25*(2), 98-111. http://www.redalyc.org/articulo.oa?id=79925207

Calderón, C., & Escalera, G. (2008). La evaluación de la docencia ante el reto del espacio europeo de educación superior (EEES). *Educación XX1, 11*(1), 237-256. https://doi.org/10.5944/educxx1.11.0.316

Carriedo, N. (1995). Hacia la contextualización: La enseñanza de estrategias de comprensión de las ideas principales en el aula. *Comunicación, Lenguaje y Educación*, *28*(4), 123–134. https://doi.org/https://doi.org/10.1174/021470395763771918

Cashin W. E., Perrin P. B. (1978). *IDEA Technical Report No. 4. Description of IDEA Standard Form Data Base*. Center for Faculty Evaluation and Development in Higher Education.

Cashin, W. E. (1995). Student Ratings of Teaching: The Research Revisited. *IDEA* Paper No. 32.

Cea, M A (2004). *Análisis multivariable Teoría y práctica en la investigación social*. Síntesis.

Centra, J. A. (1993). *Reflective Faculty Evaluation: Enhancing Teaching and Determining Faculty Effectiveness. The Jossey-Bass Higher and Adult Education Series*. Jossey-Bass Inc.

Centra J. A. (1998). *Development of The Student Instructional Report II*. Educational Testing Service. http://www.ets.org/Media/Products/283840.pdf

Centra, J. A. (2000). Evaluating the Teaching Portfolio: A Role for Colleagues. *New Directions for Teaching and Learning*, *83*, 87–93. https://doi.org/10.1002/tl.8307

Cerbin, W. y Hutchings, P. (1993). *The teaching portfolio*. Documento presentado en el Bush Summer Institute, Minneapolis, MN.

Chávez, K., & Mitchell, K. M. (2020). Exploring bias in student evaluations: Gender, race, and ethnicity. *PS: Political Science & Politics*, *53*(2), 270-274. https://doi.org/10.1017/S1049096519001744

Chen, Y., & Hoshower, L. B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment & Evaluation in Higher Education*, *28*(1), 71–88. https://doi.org/10.1080/0260293032000033071

Ching, G. (2018). A literature review on the student evaluation of teaching: An examination of the search, experience, and credence qualities of SET. *Higher Education Evaluation and Development*, *12*(2), 63-84. https://doi.org/10.1108/HEED-04-018-0009

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

Cho, G., Hwang, H., Sarstedt, M., & Ringle, C. M. (2020). Cutoff criteria for overall model fit indexes in generalized structured component analysis. *Journal of Marketing Analytics, 8*(4), 189-202. https://doi.org/10.1057/s41270-020-00089-1

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

Chou, C. P., Roberts, A., & Ching, G. S. (2012). A study on the international students' perception and norms in Taiwan. *International Journal of Research Studies in Education, 1*(2), 71-84. https://doi.org/10.5861/ijrse.2012.v1i2.76

Chuyma-Huilca, A., Berrocal-Villegas, S., Mendoza-Hidalgo, M., & Romero-Díaz, A. (2021). Evaluación del clima organizacional y la satisfacción de los estudiantes de la carrera de negocios internacionales de una universidad de Lima, Perú. *Revista Inclusiones*, *8*(11), 56-66. https://revistainclusiones.org/pdf2/20%20Chuyma%20et%20al%20VOL%208%20NUM%20ESPECIAL%20ENERO%20MARZO%202021%20REVINC.pdf

Cisneros, E. J. (2008). *El portafolio como instrumento de evaluación docente: Una experiencia en el sureste de México.* IV Coloquio Iberoamericano sobre la Evaluación de la Docencia. México: IISUE/UNAM/RIED/Conacyt. https://doi.org/10.15366/riee2008.1.3.010

Clark, L. A., & Watson, D. (2019). Constructing validity: New developments in creating objective measuring instruments. *Psychological assessment*, *31*(12), 1412. https://doi.org/10.1037/pas0000626

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

Cohen P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research*, 51, 281–309. https://doi.org/10.2307/1170209

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

Cortina, J. M. (1993). What is coefficient alpha? an examination of theory and applications. *Journal of Applied Psychology, 78*, 98-104. https://doi.org/10.1037/0021-9010.78.1.98

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests.*Psychometrika*, *16*(3), 297-334.

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Mathematical and Statistical Psychology*, *16*(2), 137–163.https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

Curby, T., McKnight, P., Alexander, L., & Erchov, S. (2019). Sources of variance in end-of-course student evaluations. *Assessment & Evaluation in Higher Education, (45)*1, 44-53 https://doi.org/10.1080/02602938.2019.1607249

Danielson, C. (1996). *Enhancing professional practice. A Framework for Teaching*. 2nd Edition. ASCD

Davidovitch, N., & Soen, D. (2006). Using students' assessments to improve instructors' quality of teaching. *Journal of further and higher education*, *30*(4), 351-376. https://doi.org/10.1080/03098770600965375

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

De Jager, T. (2011). Guidelines to assist the implementation of differentiated learning activities in South African secondary schools. *International Journal of Inclusive Education, 17*(1), 80–94. https://doi.org/https://doi.org/10.1080/13603116.2011.580465

De Juan Herrero, J., Pérez Cañaveras, R., Gómez-Torres, M., Vizcaya Moreno, M. y Mora Pascual, J. (2007). Buenas prácticas en la evaluación de la docencia y del profesorado universitario. *Espacio Europeo de Educación Superior, 1*, 155-182. https://core.ac.uk/download/pdf/16365316.pdf

De la Orden, A. (1990). Evaluación, selección y promoción del profesor universitario. *Revista Complutense de Educación, 1*(1), 11-29. https://dialnet.unirioja.es/servlet/articulo?codigo=150060

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

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.) https://repository.arizona.edu/handle/10150/621018

Delucchi, M. (2000). Don't worry, be happy: Instructor likability, student perceptions of learning, and teacher ratings in upper-level sociology courses. *Teaching Sociology*, 220–231. https://doi.org/10.2307/1318991.

DeMars, C. (2010). *Item response theory*. Oxford University Press.

Denham, S. A., & Almeida, M. C. (1987). Children's social problem-solving skills, behavioral adjustment, and interventions: A meta-analysis evaluating theory and practice. *Journal of applied developmental psychology, 8*(4), 391-409. https://doi.org/10.1016/0193-3973(87)90029-3

Denson, N., Loveday, T., & Dalton, H. (2010). Student evaluation of courses: What predicts satisfaction. *Higher Education Research and Development, 29*(4), 339-356. https://doi.org/10.1080/07294360903394466

Draper, D., & Gittoes, M. (2004). Statistical analysis of performance indicators in UK higher education. *Journal of the Royal Statistical Society: Series A, 167*(3), 449–474. https://doi.org/10.1111/j.1467-985X.2004.apm12.x

Dretzke, B. J., Sheldon, T. D., & Lim, A. (2015).What do K-12 teachers think about including student surveys in their performance ratings? *Mid-Western Educational Researcher, 27*(3), 185–206. https://scholarworks.bgsu.edu/mwer/vol27/iss3/2

Dyer, N. G., Hanges, P. J., & Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The leadership quarterly, 16*(1), 149-167. https://doi.org/10.1016/j.leaqua.2004.09.009

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

Elizalde, L. y Reyes, R. (2008). Elementos claves para la evaluación del desempeño de los docentes. *Revista Electrónica de Investigación Educativa, 10,* 1-13. http://redalyc.uaemex.mx/src/inicio/ArtPdfRed.jsp?iCve=15511127004

Elliott, K. M., & Healy, M. A. (2001). Key factors infuencing student satisfaction related to recruitment and retention. *Journal of Marketing for Higher Education, 10*(4), 1–11. https://doi.org/10.1300/J050v 10n04_01

Elosua Oliden, P., & Zumbo, B. D. (2008). Coeficientes de fiabilidad para escalas de respuesta categórica ordenada. *Psicothema, 20*(4), 896–901.

Emery, Charles R., Kramer, Tracy R., & Tian, Robert G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness. *Quality Assurance in Education, 11*(1), 37–46. https://doi.org/10.1108/09684880310462074

Escudero Escorza, T. (2019). Evaluación del profesorado como camino directo hacia la mejora de la calidad educativa. *Revista de Investigación Educativa, 37*(1), 15-37. http://doi.org/10.6018/rie.37.1.342521

Escudero, T., Pino, J. L., & Rodríguez, C. (2010). Evaluación del profesorado universitario para incentivos individuales: revisión metaevaluativa. *Revista de Educación, 351,* 513-537.

Espinosa, S., Sánchez, J. A. F., Tarí, J. J., Sempere, V. S., Conca, J. V., & Fernández, M. G. (2017). Análisis de la calidad de la docencia en la universidad española. *Investigación en Docencia Universitaria: Diseñando el Futuro a partir de la Innovación Educativa*, 145-156. http://hdl.handle.net/10045/71105

Ewell, P. (2002). Perpetual movement: Assessment after twenty years. *Teagle Foundation*. exploratory study of the faculty response. *Journal of Marketing Education, 22*(3), 199-213. https://doi.org/10.1177/0273475300223004

Fan, Y., Shepherd, L. J., Slavich, E., Waters, D., Stone, M., Abel, R., & Johnston, E. L. (2019). Gender and cultural bias in student evaluations: Why representation matters. *PLoS One, 14*(2), e0209749 https://doi.org/10.1371/journal.pone.0209749

Fayers, P. M., & Hand, D. J. (2002). Causal variables, indicator variables and measurement scales: An example from quality of life. *Journal of the Royal Statistical Society: Series B, 165*, 233–261 https://doi.org/10.1111/1467-985X.02020

Feistauer, D., &; Richter, T. (2016). How reliable are students' evaluations of teaching quality? A variance components approach. Assessment & Evaluation in Higher Education, 42(8), 1263-1279 https://doi.org/10.1080/02602938.2016.1261083

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

Feldman K. A. (1997). *Identifying exemplary teachers and teaching. Evidence from student ratings*. In Perry R., Smart J. (Eds.), *Effective teaching in higher education. Research and Practice* (368-395). Agathon. https://doi.org/10.1007/1-4020-5742-3_5

Fernández-García, C. M., Inda-Caro, M., Maulana, R., & Torío-López, S. (2022). Teaching behaviours under observation: an instrument for assessing teaching quality in Spain (La observación del comportamiento del profesorado: un instrumento para evaluar la calidad docente en España). *Culture and Education, 34*(2), 466-513. https://doi.org/10.1080/11356405.2022.2039537

Flaherty, C. (2015). Flawed evaluations. Inside Higher Ed, 10.

Flores-Mamani, E., & Arce Ortiz, N. V. (2019). Tipología de estudiantes según el nivel de satisfacción en su formación profesional en las universidades privadas de Puno, Perú. *Aletheia. Revista de Desarrollo Humano, Educativo y Social Contemporáneo*, *11*(1), 15-36. http://doi.org/10.11600/21450366.11.1aletheia.15.36

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

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

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

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

Gamliel E., & Davidovitz L. (2005). Online versus traditional teaching evaluations: Mode can matter. *Assessment & Evaluation in Higher Education*, *30,* 581–592. https://doi.org/10.1080/02602930500260647

Garbanzo, G. (2007). Factores asociados al rendimiento académico en estudiantes universitarios, una reflexión desde la calidad de la educación superior pública. *Revista Educación*, *31*(1), 43-63. http://www.redalyc.org/articulo.oa?id=44031103

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

García Garduño, J. (2014). ¿Para qué sirve la evaluación de la docencia? Un estudio exploratorio de las creencias de los estudiantes. *Education Policy Analysis Archives/Archivos Analíticos de Políticas Educativas, 22* , 1-20. http://www.redalyc.org/articulo.oa?id=275031898010

Garger, J., Jacques, P. H., Gastle, B. W., & Connolly, C. M. (2019). Threats of common method variance in student assessment of instruction instruments. *Higher Education Evaluation and Development*, *13*(1), 2-17. http://doi.org/10.1108/HEED-05-2018-0012

Gelber, S. M. (2020). *Grading the College: A History of Evaluating Teaching and Learning.* Johns Hopkins University Press. https://doi.org/10.1111/hequ.12317

Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods, 19*, 72–91. https://doi.org/10.1037/a0032138

Gerbing, D. W. & Anderson J. C. (1988). An update paradigm for scale development incorporating unidimentionality and its assessment. *Journal of Marketing Research, 25* (2), pp. 186-192.

Gillmore, G. M., M. T. Kane, and R. W. Naccarato. (1978). The Generalizability of Student Ratings of Instruction: Estimation of the Teacher and Course Components. *Journal of Educational Measurement, 15*(1), 1–13. https://doi.org/10.1111/j.1745-3984.1978.tb00051.x

Glaría L., Rocío, & Carmona SM., Lorena, & Pérez V., Chisthian, & Parra P., Paula (2016). Estructura Factorial y Consistencia Interna de la Escala de Evaluación. https://doi.org/10.15446/revfacmed.v68n2.73025

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* (3-55). Cambridge https://doi.org/10.1017/CBO9780511499531.002

Gravestock, P, & Gregor-Greenleaf, Emily. (2008). *Student course evaluations: Research, models and trends*. Higher Education Quality Council of Ontario.

Greimel-Fuhrmann, B. (2014). Students' perception of teaching behaviour and its effect on evaluation. *International Journal for Cross-Disciplinary Subjects in Education*,5(1), 1557–1563. https://doi.org/10.20533/ijcdse.2042.6364.2014.0218

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

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory* (Vol. 2). Sage.

Hamermesh, D. S., & Parker, A. M. (2005). Beauty in the classroom: Professors' pulchritude and putative pedagogical productivity. *Economics of Education Review, 24*, 369–376. https://doi.org/10.1016/j.econedurev.2004.07.013.

Hanssen, T., & Solvoll, G. (2015). The importance of university facilities for student satisfaction at a Norwegian University. *Facilities, 33*(13/14), 744–759. https://doi.org/10.1108/F-11-2014-0081

Hanushek, E. A. (2016). School human capital and teacher salary policies. *Journal of Professional Capital and Community, 1*(1), 23-40. https://doi.org/10.1108/JPCC-07-2015-0002

Harvey, L. (2003). Student feedback. *Quality in Higher Education, 9*(1), 3-20. https://doi.org/10.1080/13538320308164

Hessler, M., Pöpping, D. M., Hollstein, H., Ohlenburg, H., Arnemann, P. H., Massoth, C. (2018). Availability of cookies during an academic course session affects evaluation of teaching. *Medical Education, 52*(10), 1064–1072 https://doi.org/10.1111/medu.13627

Hildebrand, M. (1973). The character and skills of the effective professor. *The Journal of Higher Education*, 41-50. https://doi.org/10.2307/1980624

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

Holland, E. P. (2019). Making sense of module feedback: Accounting for individual behaviours in student evaluations of teaching. *Assessment and Evaluation in Higher Education, 44*(6), 961-972. https://doi.org/10.1080/02602938.2018.1556777

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/bf02289447

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education, 4*(1), 1304016. https://doi.org/10.1080/2331186X.2017.1304016

Hox, J. (2002). *Multilevel analysis: Techniques and applications*. Lawrence Erlbaum

Hox, J J. 2010. *Multilevel Analysis: Techniques and Applications*. Routledge.

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

Hutchings, P., Huber, M. T., & Ciccone, A. (2011). *The scholarship of teaching and learning reconsidered: Institutional integration and impact* (Vol. 21). John Wiley & Sons. https://doi.org/10.1111/teth.12057

Jackson D. L., Teal C. R., Raines S. J., Nansel T. R., Force R. C., Burdsal C. A. (1999). The dimensions of student's perceptions of teaching effectiveness. *Educational and Psychological Measurement, 59*, 580–596. https://doi.org/10.1177/00131649921970035

Jensen, B., Grajeda, S., & Haertel, E. (2018). Measuring cultural dimensions of classroom interactions. *Educational Assessment, 23*(4), 250 -276. https://doi.org/10.1080/10627197.2018.1515010

Jereb, E., Jerebic, J., & Urh, M. (2018). Revising the importance of factors pertaining to student satisfaction in higher education. *Organizacija, 51*(4), 271–285. https://doi.org/10.2478/orga-2018-0020

Jibaja, C. (2023). *Modelos TRI multidimensionales de clases latentes en encuestas de estudiantes sobre la evaluación de la enseñanza docente*. [Trabajo Final de Máster (Inédito), Universidad Complutense de Madrid]

Joanes, D.N. and 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

Jornet, J. M., González Such, J. y Bakieva, M. (2012). Los resultados de aprendizaje como indicador para la evaluación de la calidad de la docencia universitaria. Reflexiones metodológicas. *Revista Iberoamericana de Evaluación Educativa, 5* (2), 100-115. https://doi.org/10.15366/riee2012.5.2.007

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

Kamata, A., Bauer, D. J., & Miyazaki, Y. (2008). *Multilevel measurement modeling*. In A. A. O Connell & D. B. McCoach (Eds.), *Multilevel modeling of educational data*. Information Age Publishing

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://doi.org/10.1023/A:1004068500314 h

Kember, D., Leung, D. Y. P., & Kwan, K. P. (2002). Does the use of student feedback questionnaires improve the overall quality of teaching? *Assessment and Evaluation in Higher Education, 27*(5), 411-425. http://doi.org/10.1080/0260293022000009294

Kember D., Leung D. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education, 33(1),* 341–353. https://doi.org/10.1080/02602930701563070

Kenny, D. A. 1994. Interpersonal Perception: A Social Relations Analysis. New York: Guilford Press.

Kezar, A. (2014). Higher education change and social networks: A review of research. *The journal of higher education, 85*(1) 91-125. https://doi.org/10.1080/00221546.2014.11777320

Kilgo, C. A., Ezell Sheets, J. K., & Pascarella, E. T. (2015). The link between high-impact practices and student learning: Some longitudinal evidence. *Higher Education, 69,* 509-525. https://doi.org/10.1007/s10734-014-9788-z

Kim, E. S., Dedrick, R. F., Cao, C., & Ferron, J. M. (2016). Multilevel factor analysis: Reporting guidelines and a review of reporting practices. *Multivariate Behavioral Research, 51*(6), 881-898. https://doi.org/10.1080/00273171.2016.1228042

Kline, R. B. (2011). *Principles and practice of structural equation modeling*. Guilford publications.

Koretz, D. (2008). *What Test Scores Tell Us About American Kids.* (74-112) in his *Measuring Up: What Educational Testing Really Tells Us*. Harvard University Press. https://doi.org/10.4159/9780674039728

Kramp, U. (2006). *Efecto del número de opciones de respuesta sobre las propiedades psicométricas de los cuestionarios de personalidad* (Tesis doctoral, Universidad de Barcelona, España) https://www.tdx.cat/handle/10803/2535#page=4

Kreitzer, R. J., & Sweet-Cushman, J. (2022). Evaluating student evaluations of teaching: A review of measurement and equity bias in SETs and recommendations for ethical reform. *Journal of Academic Ethics 20*(3–4). https://doi.org/10.1007/s10805-021-09400-w

Kuh, G., & Ikenberry, S. (2009). More than you think, less than we need. *National Institute for Learning Outcomes Assessment, 1*(2). https://www.learningoutcomesassessment.org/wp-content/uploads/2019/02/2009NILOASurveyReportAbridged.pdf

Kunst, E. M., van Woerkom, M., & Poell, R. F. (2018). Teachers' goal orientation profiles and participation in professional development activities. *Vocations and Learning, 11*, 91–111. https:// doi.org/10.1007/s12186-017-9182-y.

Lai, K. (2021). Fit difference between nonnested models given categorical data: Measures and estimation. Structural Equation Modeling, 28(1), 9920. https://doi.org/10.1080/10705511.2020.1763802

Lancheros- Florián, L. C. (2022). Ventajas de aplicar la TRI en cuestionarios sobre la actividad docente universitaria. *Revista EducaT: Educación Virtual, Innovación y Tecnologías, 3*(1). https://doi.org/10.22490/27452115.5801

Lancheros- Florián, L. C., S. Ramírez, E., & M. Alvarado, J. (2022a). Satisfacción con la calidad docente en el ámbito universitario: Potenciales sesgos y propuestas de análisis para su evaluación. *Revista Iberoamericana De Diagnóstico Y Evaluación Psicológica, 65*(4), 69. https://doi.org/10.21865/RIDEP65.4.06

Lancheros-Florián, L. C., S. Ramírez, E., & M. Alvarado, J. (2022b). El problema del análisis de la evaluación de la satisfacción estudiantil en el ámbito universitario: Un estudio de simulación. *Revista Iberoamericana De Diagnóstico Y Evaluación Psicológica, 66*(5), 81. https://doi.org/10.21865/RIDEP66.5.06

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

La Rocca, M., Parrella, M. L., Primerano, I., Sulis, I. y Vitale, M. P. (2017). An integrated strategy for the analysis of student evaluation of teaching: from descriptive measures to explanatory models. *Quality and Quantity, 51*(2), 675–691. https://doi.org/10.1007/s11135-016-0432-0

Leckie, G., & Goldstein, H. (2009). The limitation of using school league tables to inform school choice. *Journal of the Royal Statistical Society: Series A, 172*(4), 835–851. https://doi.org/10.1111/j.1467-985X.2009.00597.x

Ledesma, R. (2004). AlphaCI: un programa de cálculo de intervalos de confianza para el coeficiente alfa de Cronbach. *Psico-USF, 9* (1), 31-37. https://doi.org/10.1590/S1413-82712004000100005

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

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Lord, F. M., & Novick, M. R. (2008). *Statistical theories of mental test scores*. IAP.

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, 30*(3), 1151-1169. https://doi.org/10.6018/analesps.30.3.19936

Lozano, L., García-Cueto, E. & Muñiz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology, 4* (2), 73- 79. https://doi.org/10.1027/1614-2241.4.2.73

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new more reliable approach to grouplevel effects in contextual studies. *Psychological Methods, 13*, 203–229. https://doi.org/10.1037/a0012869

Luengo, L. M. (2019). La validez de las evaluaciones de los estudiantes sobre sus profesores: una revisión de la literatura en la última década. *Revista de Educación, 390*, 1-24.

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. En Higher Education 2010, 22(3), 337-345 http://www.isetl.org/ijtlhe/

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*(2), 130-149. doi:10.1037/1082-989X.1.2.130

McDonald, R. P. (1999*). Test Theory: A Unified Treatment*. Erlbaum.

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

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. http://doi.org/10.1186/s12955-018-0903-1

Mansolf, M., & Reise, S. P. (2016). Exploratory bifactor analysis: The Schmid-Leiman orthogonalization and Jennrich-Bentler analytic rotations. *Multivariate Behavioral Research, 51*, 698 –717. http://doi.org/10.1080/00273171.2016.1215898

Marczely, B. (1992). Teacher evaluation: Research versus practice. *Journal of Personnel Evaluation in Education, 5*(3), 279-290. https://doi.org/10.1007/BF00125242

Mardia, K.V. (1970). Measures of multivariate skewness and kurtosis with applications. *Biometrika, 57*(3) 519-530. https://doi.org/10.1093/biomet/57.3.519

Marsh, B. (2011). *The evaluation of a university in-school teacher education program in science (INSTEP)*. Unpublished dissertation). Springer.

Marsh H. W., Roche L A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias and utility. *American Psychologist, 52*, 1187–1197. https://doi.org/10.1037/0003-066X.52.11.1187

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

Marsh, H. W. (2007). *Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulnes*s. In R. P. Perry & J. C. Smart (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective* (319-383). Springer. https://doi.org/10.1007/1-4020-5742-3_9

Marsh, H. W., & Bailey, M. (1993). Multidimensional students' evaluations of teaching effectiveness: A profile analysis. *Journal of higher education, 64*(1), 1-18. https://doi.org/10.2307/2959975

Marsh, H.W. & Dunkin, M. J. (1997). *Students' evaluations of university teaching: A multidimensional perspective*. En R. Perry y J. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 241–320). Agathon Press.

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

Marsh, H. W., & Hocevar, D. (1991). Students' evaluations of teaching effectiveness: The stability of mean ratings of the same teachers over a 13-year period. *Teaching and teacher education*, *7*(4), 303-314. https://doi.org/10.1016/0742-051X(91)90001-6

Marsh, H. W., Muthén, B., Asparouhov, T., Lüdtke, O., Robitzsch, A., Morin, A. J., & Trautwein, U. (2009). Exploratory structural equation modeling, integrating CFA and EFA: Application to students' evaluations of university teaching. *Structural equation modeling: A multidisciplinary journal*, *16*(3), 439-476. https://doi.org/10.1080/10705510903008220

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

Martínez, R., y Hernández, V. (2014). *Psicometría*. Difusora Larousse - Alianza Editorial.

Mateo, J. (2000). *La evaluación educativa, su práctica y otras metáforas*. Editorial Horsori.

Maulana, R., Helms-Lorenz, M., y Van De Grift, W. (2015). Pupils' perception of teaching behavior: evaluation of an instrument and importance of academic motivation in Indonesian secondary education. *International Journal of Educational Research,69(1)*, 98-112. https://doi.org/10.1016/j.ijer.2014.11.002

Maulana, R., Helms-Lorenz, M., & Van de Grift, W. (2017). Validating a model of effective teaching behaviour of pre-service teachers. *Teachers and Teaching*, *23*(4), 471-493. https://doi.org/10.1080/13540602.2016.1211102

McCullough, B. D., & Radson, D. (2011). Analysing student evaluations of teaching: Comparing means and proportions. *Evaluation & Research in Education*, *24*(3), 183-202. https://doi.org/10.1080/09500790.2011.603411

McPherson M. A. (2006). Determinants of how students evaluate teachers. *The Journal of Economic Education, 37*(1), 3- 20. https://doi.org/10.3200/JECE.37.1.3-20

McPherson, M. A., Todd Jewell, R., & Kim, M. (2009). What determines student evaluation scores? A random effects analysis of undergraduate economics classes. *Eastern Economic Journal, 35(1), 37–51*. https://doi.org/10.1057/palgrave.eej.9050042

Mehta, P. D., & Neale, M. C. (2005). People are variables too: Multilevel structural equations models. *Psychological Methods, 10*(3), 259–284. https://doi.org/10.1037/1082-989X.10.3.259

Melo, B. I., Moreira, L. O., Villalobos, C. P., Araneda, G. T., Calvo, P. M., Kother, A. M., ... & Durán, C. B. (2015). Estructura factorial y confiabilidad del Cuestionario de Satisfacción Académica en estudiantes de medicina chilenos. *Revista Iberoamericana de Diagnóstico y Evaluación-e Avaliação Psicológica*, *2*(40), 73-82.

Mengel, F., Sauermann, J., & Zölitz, U. (2018). Gender bias in teaching evaluations. *Journal of the European Economic Association, 17*(2), 535–566. https://doi.org/10.1093/jeea/jvx057

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

Miles, P., & House, D. (2015). The tail wagging the dog: An overdue examination of student teaching evaluations. *International Journal of Higher Education*, *4*(2), 116. https://doi.org/10.5430/ijhe.v4n2p116

Miller, J. E., & Seldin, P. (2014). Changing practices in faculty evaluation. *Academe*, *100*(3), 35-38. https://www.jstor.org/stable/24642931

Mitchell, M., Leachman, M. and Masterson, K. (2016). Funding down, tuition up. www.cbpp.org/research/state-budget-and-tax/funding-down-tuition-up

Mittelhaëuser, M., Béguin, A., & Sijtsma, K. (2013). *Modeling Differences in Test-Taking Motivation: Exploring the Usefulness of the Mixture Rasch Model and Modeling.* Springer

Montero, I., &; León, O. G. (2002). Clasificación y descripción de las metodologías de investigación en Psicología. *International journal of clinical and health psychology, 2(3),*503-508.

Morales, P. (1988). *Medición de actitudes en psicología y educación: construcción de escalas y problemas metodológicos*. San Sebastián

Muñiz, 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, 31*, 57-66. https://www.papelesdelpsicologo.es/pdf/1796.pdf

Muñiz, J. (2018). *Introducción a la Psicometría: Teoría clásica y TRI*. Pirámide

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

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*(1), 338–354. https://doi.org/10.1111/j.1745-3984.1991.tb00363.x

Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376–398 https://doi.org/10.1177/0049124194022003006

Nasser, F. and Fresko, B. (2002). Faculty Views of Student Evaluation of College Teaching, *Assessment & Evaluation in Higher Education*, *27*(2), 187-198. https://doi.org/10.1080/02602930220128751

Nastasić, A., Banjević, K., & Gardašević, D. (2019). Student satisfaction as a performance indicator of higher education institution. *Mednarodno Inovativno Poslovanje, 11*(2), 67–76. https://doi.org/10.32015/JIBM/ 2019-11-2-8

Oermann, M. H., Conklin, J. L., Rushton, S., y 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

Oliver, R. M., y Reschly, D. J. (2007). Effective classroom management: teacher preparation and professional development. National Comprehensive center for Teacher Quality.

Pardo Chitiva, J. (2017). Percepciones sobre evaluación docente de los estudiantes de la Facultad de Ciencias y Educación de la Universidad Distrital Francisco José de Caldas. [Tesis de maestría, Repositorio Institucional - Universidad Distrital Francisco José de Caldas]

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

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

Organización para la Cooperación y Desarrollo Económico (OCDE). (2013a). Synergies for Better Learning: An International Perspective on Evaluation and Assessment. París: OCDE.

Organización para la Cooperación y Desarrollo Económico (OCDE). (2013b). Teachers for the 21st Century: Using Evaluation to Improve Teaching. París: OCDE.

Ory, J. C. (1991). Changes in evaluating teaching in higher education. *Theory into Practice, 30*(1), 30-36. https://doi.org/10.1080/00405849109543473

Osorio, J. V. y Pérez, K. M. (2010). *El nivel de satisfacción escolar y su relación con la orientación vocacional en alumnos de psicología educativa* (Tesis profesional, Universidad Pedagógica Nacional, México). http://200.23.113.51/pdf/27385.pdf

Oviedo, H. C. & Campo-Arias, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Revista Colombiana de Psiquiatría, 34* (4), 572-580.

Palmer, P. (1998). *Evaluating: Assessing and enhancing teaching quality*. The Courage to Teach.

Palomer, Leonor, Jana, M Paz, Zuzulich, Soledad, Barriga, M Trinidad, & Heusser, M Ignacia. (2018). Medición del clima educativo y factores que influyen en su resultado. Estudio en una carrera de odontología chilena. *FEM: Revista de la Fundación Educación Médica*, *21*(2), 87-96. https://doi.org/10.33588/fem.212.937

Pardo, A., & Ruiz, M. Á. (2015). *Análisis de datos en ciencias sociales y de la salud. III*. Síntesis.

Pardo, A., Ruiz, M. Á., & San Martín, R. (2007). Cómo ajustar e interpretar modelos multinivel con SPSS. Psicothema, *19*(2), 308-21. 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

Pascarella, E. T., & Terenzini, P. T. (1991). *How College Affects Students: Findings and Insights from Twenty Years of Research*. Jossey-Bass Inc.

Pascual G. 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

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

Paul, R., & Pradhan, S. (2019). Achieving student satisfaction and student loyalty in higher education: A focus on service value dimensions. *Services Marketing Quarterly, 40*(3), 245–268. https://doi.org/10.1080/ 15332969.2019.1630177

Pekka R. (2013) The number of feedback 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

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

Pepe, J., & Wang, M. (2012). What instructor qualities do students reward? *College Student Journal, 46*(3), 603–614. https://www.ingentaconnect.com/content/prin/csj/2012/00000046/00000003/art00014

Peterson, K. D. (1997). *Asesoramiento y evaluación para el profesorado principiante*. En *Manual para la evaluación del profesorado* (147-164). La Muralla.

Pounder, J. (2007) Is student evaluation of teaching worthwhile? An analytical framework for answering the question. *Qual Ass Educ.;15*(2):178–191. https://doi.org/10.1108/09684880710748938

R Core Team (2022). R: *A language and environment for statistical computing*. R

Rahmatpour, P., Sharif Nia, H., & Peyrovi, H. (2019). Evaluation of psychometric properties of scales measuring student academic satisfaction: A systematic review. *Journal of Education and Health Promotion, 8*(1), 256. https://doi.org/10.4103/jehp.jehp_466_19

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

Rantanen, P. (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 Education, 38*(1), 224–239. https://doi.org/10.1080/02602938.2011.625471

Raudenbush, S. W., and A. S. Bryk. 2006. *Hierarchical Linear Models Applications and Data Analysis Methods: Applications and Data Analysis Methods*. 2nd ed. Thousand Oaks, CA: Sage.

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. *Educational and Psychological Measurement, 73*(5), 898-908. https://doi.org/10.1177/0013164413478165

Reid, L. D. (2010). The role of perceived race and gender in the evaluation of college teaching on RateMyProfessors.com. *Journal of Diversity in Higher Education, 3*(3), 137-152. https://doi.org/10.1037/a0019865

Reise, S. P., Moore, T. M., & Haviland, M. G. (2010). Bifactor models and rotations: Exploring the extent to which multidimensional data yield univocal scale scores. *Journal of Personality Assessment, 92*, 544-559. https://doi.org/10.1080/00223891.2010.496477

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013a). Multidimensionality and structural coefficient bias in structural equation modeling. *Educational and Psychological Measurement, 73*(1), 5-26. https://doi.org/10.1177/0013164412449831

Reise, S. P., Moore, T. M., & Haviland, M. G. (2013b). *Applying unidimensional item response theory models to psychological data*. En K. F. Geisinger, B. A. Bracken, J. F. Carlson, J.-I. C. Hansen, N. R. Kuncel, S. P. Reise, & M. C. Rodriguez (Eds.), *APA handbook of testing and assessment in psychology, Vol. 1. Test theory and testing and assessment in industrial and organizational psychology* (101–119). American Psychological Association. https://doi.org/10.1037/14047-006

Revelle, W. (2017). Package 'psych'. *The Comprehensive R Archive Network*. https://cran.r-project.org/web/packages/psych/psych.pdfRevelle, W., & Zinbarg, R. E. (2009). Coefficients alpha, beta, omega, and the glb: Comments on Sijtsma. *Psychometrika, 74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Revelle W (2021). *psych: Procedures for Psychological, Psychometric, and Personality*

Richardson J. T. E. (2005). Instruments for obtaining student feedback: A review of the literature. *Assessment and Evaluation in Higher Education*, *30*(1), 387-415. https://doi.org/10.1080/02602930500099193

Richter, T. 2006. What is Wrong with ANOVA and Multiple Regression? Analyzing Sentence Reading Times with Hierarchical Linear Models. *Discourse Processes 41*(1), 221–250. https://doi.org/10.1207/s15326950dp4103_1

Ridgeway, C. L. (2011). *Framed by gender: How gender inequality persists in the modern world.* Oxford University Press

Rindermann, H., & N. Schofield. 2001. Generalizability of Multidimensional Student Ratings of University Instruction across Courses and Teachers. *Research in Higher Education*, *42*(1) 377-399. https://doi.org/10.1023/A:1011050724796

Rizopoulos D (2006). ltm: An R package for Latent Variable Modelling and Item Response Theory Analyses. *Journal of Statistical Software*, *17*(5), 1-25. https://doi.org/10.18637/jss.v017.i05

Robertson S. I. (2004). Student perceptions of student perception of module questionnaires: Questionnaire completion as problem solving. *Assessment and Evaluation in Higher Education 29*(6), 663-679. https://doi.org/10.1080/0260293042000227218

Rodríguez G. G. (2000). La evaluación de la actividad docente en la universidad: entre el sueño y la realidad. *Revista de Investigación Educativa, 18*(2), 417-432. https://revistas.um.es/rie/article/view/121091/113771

Rodríguez, A., Reise, S. P., & Haviland, M. G. (2016). Applying bifactor statistical indices in the evaluation of psychological measures. *Journal of Personality Assessment, 98*(3), 223. https://doi.org/10.1080/00223891.2015.1089249

Rosen, A. S. (2018). Correlations, trends and potential biases among publicly accessible web-based student evaluations of teaching: a large-scale study of RateMyProfessors. com data. *Assessment & Evaluation in Higher Education*, *43*(1), 31- 44. https://doi.org/10.1080/02602938.2016.1276155

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.

Rowden, G. V., & Carlson, R. E. (1996). Gender issues and students' perceptions of instructors' immediacy and evaluation of teaching and course. *Psychological Reports, 78*(3), 835-839. https://doi.org/10.2466/pr0.1996.78.3.835

Ruiz Esteban, C. M., & Santos del Cerro, J. (2020). Calidad de la docencia: La satisfacción del alumnado universitario con sus profesores. *Anales de Psicología / Annals of Psychology*, *36*(2), 304–312. https://doi.org/10.6018/analesps.335431

Ruíz, 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

Runhaar, P., Sanders, K., & Yang, H. (2010). Stimulating teachers' reflection and feedback asking: An interplay of self-efficacy, learning goal orientation, and transformational

leadership. *Teaching and Teacher Education, 26,* 1154–1161. https://doi.org/10.1016/j.tate.2010.02.011.

Samejima, F. (2010). *The general graded response model*. En Nering, M y Ostini, E. (Eds). *Handbook of polytomous ítem response models*. Taylor & Francis group.

Sánchez Escobedo, P. A., Castillo Blue, E. M., & Valdés Cuervo, A. A. (2009). Percepción de los docentes con respecto a la evaluación de su práctica. *Investigación Educativa Duranguense, (10),* 36-45. http://dialnet.unirioja.es/servlet/oaiart?codigo=2941615

Santhanam E., Hicks O. (2001). Disciplinary, gender and course year influences on student perceptions of teaching: Explorations and implications. *Teaching in Higher Education*, *7*(1), 17–31. https://doi.org/10.1080/13562510120100364

Santisteban, C., & Alvarado, J. M. (2001*). Modelos psicométricos*. UNED

Schalock, M. D., Cowart, B., & Staebler, B. (1993). Teacher productivity revisited: Definition, theory, measurement and application. *Journal of Personnel Evaluation in Education, 7*(2), 179-196. https://doi.org/10.1007/BF00995302

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8*(2), 206-224. https://doi.org/10.1037/1082-989X.8.2.206

Schwab, K., Moseley, B., & Dustin, D. (2018). Grading grades as a measure of student learning. *SCHOLE: A Journal of Leisure Studies and Recreation Education*, *33*(2), 87-95. https://doi.org/10.1080/1937156X.2018.1513276

Sedlmeier P. (2006). The role of scales in student ratings. *Learning and Instruction, 16, (*5), 401-415. https://doi.org/10.1016/j.learninstruc.2006.09.002

Seldin, P. (1999). Changing Practices in Evaluating Teaching. : A Practical Guide to Improved Faculty Performance and Promotion/Tenure Decisions. Peter Seldin and Associations.

Seldin, P. (2000). Teaching portfolios: a positive appraisal *Academe, 86*(1), 36-44. https://doi.org/10.2307/40252334

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

Sinclaire, J. K. (2014). An empirical investigation of student satisfaction with college courses. *Research in Higher Education Journal, 23*. https://doi.org/10.1177/0273475300223004

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*(1), 107-120. https://doi.org/10.1007/s11336-008-9101-0

Skelton, A. (2004). Understanding 'teaching excellence' in higher education: A critical evaluation of the national teaching fellowships scheme. *Studies in Higher Education (Dorchester-on-Thames), 29*(4), 451-468. https://doi.org/10.1080/0307507042000236362

Solomon, D. J., Speer, A. J., Rosebraugh, C. J., & DiPette, D. J. (1997). The reliability of medical student ratings of clinical teaching. *Evaluation & the Health Professions, 20*(3), 343-352. https://doi.org/10.1177/016327879702000306

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

Spooren, P. (2010). On the Credibility of the Judge: A Cross-classified Multilevel Analysis on Students' Evaluation of Teaching. *Studies in Educational Evaluation 36*(1) 121–131. https://doi.org/10.1016/j.stueduc.2011.02.001

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

Spooren, P., Mortelmans, D., & Thijssen, P. (2012). 'Content'versus 'style': acquiescence in student evaluation of teaching? *British Educational Research Journal, 38*(1), 3-21. Springer.

Sproule, R. (2000). Student evaluation of teaching: a methodological critique of evaluation practices. Education Policy Analysis Archives, 8 (50). http://epaa.asu.edu/epaa/v8n50. html

Stake, R. E., García, M. I. A., y 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

Staley, D.J. and Trinkle, D.A. (2011). The changing landscape of higher education. *Educause, (1),* 16-32. http://er.educause.edu/articles/2011/2/the-changing-landscape-of-higher-education

Stapleton, L. M. (2006). Using multilevel structural equation modeling techniques with complex sample data. *Structural equation modeling: A second course,* 345-383.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics, 41*(5), 481-520. https://doi.org/10.3102/1076998616646200

Stark, P. B., & Freishtat, R. (2014). An evaluation of course evaluations. Science Open Research, 0(0), 1-7. https://doi.org/10.14293/S2199-1006.1.SOR-EDU.AOFRQA.v1

Staufenbiel, T., T. Seppelfricke, and J. Rickers. 2016. "Prädiktoren studentischer Lehrveranstaltungsevaluationen." [Predictors of Student Evaluations of Teaching.] *Diagnostica, 62*(1), 44–59. https://doi.org/10.1026/0012-1924/a000142

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

Straumshein, C. (2016). *Doubts about data: 2016 survey of faculty attitudes on technology*. Inside Higher Ed.

Stroebe, W. (2016). Why good teaching evaluations may reward bad teaching: On grade inflation and other unintended consequences of student evaluations. *Perspectives on Psychological Science, 11*, 800–816. https://doi.org/10.1177/1745691616650284.

Suárez Monzón, N., Cáceres Mesa, M. L., Gómez Suárez, V., & Pérez Cruz, I. C. (2022). Evaluación docente y desarrollo profesional universitario: Una revisión basada en los participantes, las dimensiones y los métodos. *PUBLICACIONES*, *52*(3), 139–189. https://doi.org/10.30827/publicaciones.v52i3.22271

Subtirelu, N. C. (2015). She does have an accent but…: Race and language ideology in students' evaluations of mathematics instructors on RateMyProfessors.com. *Language in Society, 44*(1), 35–62. https://doi.org/10.1017/S0047404514000736

Sulis, I., Porcu, M., & Capursi, V. (2019). On the use of student evaluation of teaching: a longitudinal analysis combining measurement issues and implications of the exercise. *Social Indicators Research*, *142*(3), 1305-1331. https://doi.org/10.1007/s11205-018-1946-8

Tabachinik BG y Fidel LS (2001). *Using multivariate statistics* (2ª ed). Pearson Education.

Tejedor. F. J. y Jornet, J. M. (2008). La evaluación del profesorado universitario en España. *Revista Electrónica de Investigación Educativa, número especial.* https://dialnet.unirioja.es/servlet/oaiart?codigo=3051723

Theall, M. y Franklin, J. L. (2000). Creating responsive student ratings systems to improve evaluation practice. *New Directions for Teaching and Learning, 83*(1), 95-107. https://doi.org/10.1002/tl.8308

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

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/REYD2018.17.23484

Uttl, B. (2021). Lessons learned from research on student evaluation of teaching in higher education. *Student Feedback on Teaching in Schools: Using Student Perceptions for the Development of Teaching and Teachers*, 237-256. https://doi.org/10.1007/978-3-030-75150-0_15

Uttl, B., Smibert, D. (2017). Student evaluations of teaching: teaching quantitative courses can be hazardous to one's career, *PeerJ, 5,* e3299, https://doi.org/10.7717/peerj.3299

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

Valsan, C., & Sproule, R. (2008). The invisible hands behind the student evaluation of teaching: The rise of the new managerial elite in the governance of higher education. *Journal of Economic Issues, 42(*4), 939-958. https://doi.org/10.1080/00213624.2008.11507197

van de Grift, W. (2007). Quality of teaching in four European countries: A review of the literature and application of an assessment instrument. *Educational research, 49*(2), 127-152. https://doi.org/10.1080/00131880701369651

van de Grift, W. J., & Lam, J. F. (1998). Het didactisch handelen in het basisonderwijs. *Tijdschrift voor onderwijsresearch, 23*(3), 224-241.

van De Grift, W., Helms-Lorenz, M., y Maulana, R. (2014). Teaching skills of student teachers: calibration of an evaluation instrument and its value in predicting student academic engagement. *Studies in Educational Evaluation, 43*,150-159. https://10.1016/j.stueduc.2014.09.003

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

Ventura-León, J. L., & Caycho-Rodríguez, T. (2017). El coeficiente Omega: un método alternativo para la estimación de la confiabilidad. *Revista Latinoamericana de Ciencias Sociales, Niñez y Juventud, 15*(1), 625-627.

Vermunt, J. K. (2003). Multilevel latent class models. *Sociological methodology, 33*(1), 213-239. https://doi.org/10.1111/j.0081-1750.2003.t01-1-00131.x

Vermunt, J. K., &; Magidson, J. (2002). *Latent class cluster analysis*. In J. Hagenaars, & A. McCutcheon (Eds.), *Applied latent class analysis* (89-106). Cambridge University Press. https://doi.org/10.1017/CBO9780511499531.004

Vlăsceanu, L., Grünberg, L., & Pârlea, D. (2004). Quality Assurance and Accreditation: A Glossary of Basic Terms and Definitions. United Nations Educational, Scientific and Cultural Organization, Bucharest.

von Davier, M., Yamamoto, K. (2007). *Mixture-Distribution and HYBRID Rasch Models*. En: *Multivariate and Mixture Distribution Rasch Models*. Statistics for Social and Behavioral Sciences. Springer. https://doi.org/10.1007/978-0-387-49839-3_6

Wagner, N., Rieger, M., & Voorvelt, K. (2016). Gender, ethnicity and teaching evaluations: Evidence from mixed teaching teams. *Economics of Education Review, 54,* 79–94 https://doi.org/10.1016/j.econedurev.2016.06.004

Wainer, H., & Thissen, D. (2001). *True score theory: The traditional method*. In *Test scoring* (35-84). Routledge.

Wallisch, P., & Cachia, J. (2019). Determinants of perceived teaching quality: the role of divergent interpretations of expectations. *Center for Open Science*. https://doi.org/10.31234/osf.io/dsvgq

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*(2), 191-212. https://doi.org/10.1080/0260293980230207

Watson, D., Clark, L. A., Chmielewski, M., & Kotov, R. (2013). The value of suppressor effects in explicating the construct validity of symptom measures. *Psychological Assessment, 25*, 929 –941. https://doi.org/10.1037/a0032781

Weerasinghe, I.S., Fernando, S., & Lalitha, R. (2017). Students' satisfaction in higher education. American Journal of Educational Research, 5(5), 533 – 539. https://ssrn.com/abstract=2976013

Wilson, B., & Wood, J. A. (1996). Teacher Evaluation: A National Dilemma. *Journal of Personnel Evaluation in Education*, *10*(1), 75-82. https://doi.org/10.1007/BF00139470

Wilson, J. H., Beyer, D., & Monteiro, H. (2015). Professor age and gender affect student perceptions and grades. *Journal of the Scholarship of Teaching and Learning*, *15*(4), 126–138. https://doi.org/10.1080/87567555.2013.825574

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

Worthington, A. (2002). The impact of student perceptions and characteristics on teaching evaluations: a case study in finance education. *Assessment and evaluation in Higher Education, 27*(1), 49-64. https://doi.org/10.1080/02602930120105054

Wong, W.H., Chapman, E.(2023). Student satisfaction and interaction in higher education. *High Education 85*, 957–978 . https://doi.org/10.1007/s10734-022-00874-0

Wright, S. L., & Jenkins-Guarnieri, M. A. (2012). Student evaluations of teaching: combining the meta-analyses and demonstrating further evidence for effective use. *Assessment & Evaluation in Higher Education, 37*(6), 683–699 https://doi.org/10.1080/02602938.2011.563279

Yao, Y., & Grady, M. (2005). How do faculty make formative use of student evaluation feedback? A multiple case study. *Journal of Personnel Evaluation in Education, 18*(1), 107-126. https://doi.org/10.1007/s11092-006-9000-9

Youmans, R. J., & Jee, B. D. (2007). Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course. *Teaching of Psychology, 34*(4), 245–247. https://doi.org/10.1080/00986280701700318

Young, S., Rush, L., & Shaw, D. (2009). Evaluating Gender Bias in Ratings of University Instructors' Teaching Effectiveness. *International Journal for the Scholarship of Teaching and Learning, 3*(2), https://doi.org/10.20429/ijsotl.2009.030219

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.E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and McDonald's ωH : Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika, 70*(1), 123–133.

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.org/10.1177/0146621605278814

Zumbo, B. D., & Forer, B. (2010). *A multilevel view of measurement validity: Some concepts and foundations*. Paper presented at the meeting of the American Educational Research Association, Denver, CO

Zumbo, B. D., Gadermann, A.M., & Zeisser, C. (2007). Ordinal versions of coefficients alpha and theta for likert rating scales. *Journal of Modern Applied Statistical Methods, 6*,(1), 21-29. https://doi.org/10.22237/jmasm/1177992180

Zumbo, B.D., Liu, Y., Wu, A.D., Forer, B., Shear, B.R. (2017). *National and International Educational Achievement Testingeducational achievement testing*: A Case of Multi-level validation framed by the Ecological model of Item responding. *Understanding and investigating response processes in validation research* (341-361). Springer International Publishing AG. https://doi.org/10.1007/978-3-319-56129-5_18

Apéndice A. Ítems del 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.

Escala: NP No procede // 1 Totalmente en desacuerdo // 2 Más bien en desacuerdo // 3 Ni de acuerdo ni en desacuerdo // 4 Más bien de acuerdo // 5 Totalmente de acuerdo

Apéndice B. Resultados Invarianza Factorial

Tabla B1.

Índices de Bondad y Ajuste, Grupo: Género del Estudiante

Modelo	X2	RMSEA	CFI	Comp. Modelos	ΔX	Δgl	p-valor (p-ΔX)	ΔCFI
1. Inv. Configuración	$X^2_{28} = 222.461\ (0.00)$	0.029 (0.025-0.032)	0.999	-	-	-	-	-
2. Inv. Métrica	$X^2_{34} = 2224.461\ (0.00)$	0.026 (0.023-0.029)	0.999	2 vs. 1	2	6	0.9197	0
3. Inv. Escalar	$X^2_{54} = 260.831(0.00)$	0.021 (0.019-0.024)	0.999	3 vs. 2	36	20	0.0139	0
4. Inv. Estricta	$X^2_{55} = 277.888\ (0.00)$	0.022 (0.019-0.024)	0.999	4 vs. 3	17.05	1	0.0000	0

Tabla B2.

Índices de Bondad y Ajuste, Grupo: Facultad o Centro Propietario

Modelo	X2	gl	p-valor	RMSEA	CFI	Comp. Modelos	ΔX	Δgl	p-valor (p-ΔX)	ΔCFI
1. Inv. Configuración	224.501	112	0	0.022 (0.018 -0.026)	1	-	-	-	-	-
2. Inv. Métrica	258.218	154	0	0.018 (0.014 -0.022)	1	2 vs. 1	33.717	42	0.8151	0
3. Inv. Escalar	792.252	294	0	0.028 (0.026 -0.031)	0.998	3 vs. 2	534.034	140	0.000	-0.002
4. Inv. Estricta	1.592.012	301	0	0.045 (0.043 -0.047)	0.995	4 vs. 3	799.760	7	0.000	-0.003