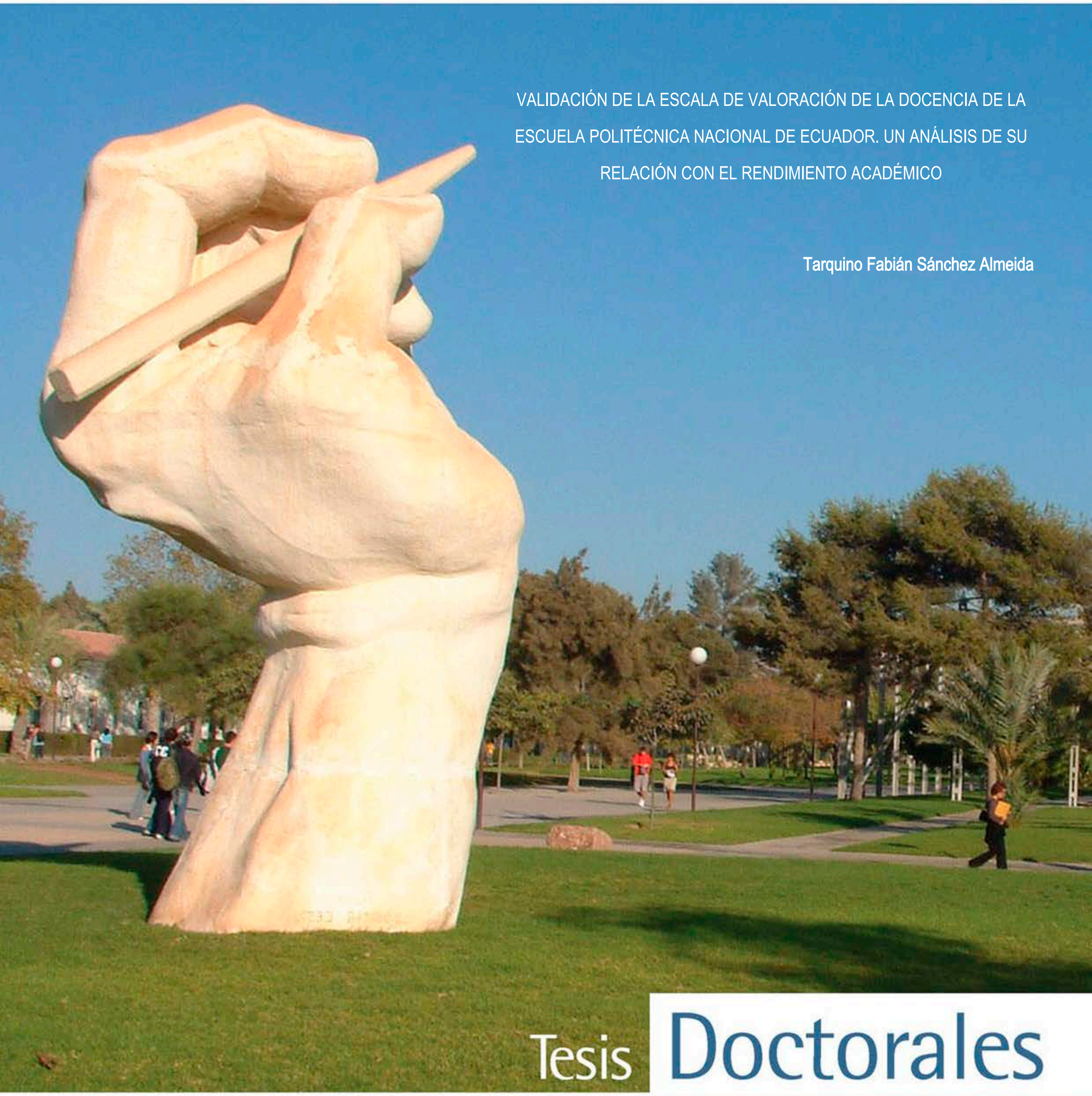


VALIDACIÓN DE LA ESCALA DE VALORACIÓN DE LA DOCENCIA DE LA ESCUELA POLITÉCNICA NACIONAL DE ECUADOR. UN ANÁLISIS DE SU RELACIÓN CON EL RENDIMIENTO ACADÉMICO

Tarquino Fabián Sánchez Almeida

DEPARTAMENTO DE PSICOLOGIA EVOLUTIVA Y DIDÁCTICA

FACULTAD DE EDUCACIÓN

VALIDACIÓN DE LA ESCALA DE VALORACIÓN DE LA DOCENCIA DE LA ESCUELA POLITÉCNICA NACIONAL DE ECUADOR. UN ANÁLISIS DE SU RELACIÓN CON EL RENDIMIENTO ACADÉMICO.

AUTOR:

TARQUINO FABIÁN SÁNCHEZ ALMEIDA

DOCTORADO EN INVESTIGACIÓN EDUCATIVA

Tesis presentada por compendio de publicaciones para aspirar al grado de

DOCTOR POR LA UNIVERSIDAD DE ALICANTE

DIRECTORES:

DR. JUAN LUIS CASTEJÓN COSTA
DR. ALEJANDRO VEAS INIESTA

Con el fin de realizar la presente tesis doctoral, el autor ha recibido ayuda predoctoral de la Escuela Politécnica Nacional

Alicante, Julio de 2021

VALIDACIÓN DE LA ESCALA DE VALORACIÓN DE LA DOCENCIA DE LA ESCUELA POLITÉCNICA NACIONAL DE ECUADOR. UN ANÁLISIS DE SU RELACIÓN CON EL RENDIMIENTO ACADÉMICO

Contenido

1. AGRADECIMIENTOS

Cuando se inicia un gran proyecto de vida, se lo realiza con ilusión y convencido de que esto permitirá ser mejor ser humano como persona y como profesional, esta vez la vara está trazada muy alto, incursionar en los dominios de la investigación no es cosa sencilla, sobre todo, si se ha llegado a la madurez, consolidado una trayectoria profesional, sin embargo, cruzar dicha meta se convirtió en un reto especial sobre todo si se considera que la mayor parte de mi vida lo he dedicado a la docencia universitaria y al estudio de las ciencias, en donde la investigación educativa ocupa un lugar especial por lo que lograr este objetivo con el apoyo de muchas personas se trasformó en un recorrido maravilloso. Es por ello que deseo agradecer a todas las personas que me apoyaron en esta etapa de mi formación.

Especial reconocimiento a mi director de tesis, el doctor Juan Luis Castejón Costa, por ser quien guió mis esfuerzos en conseguir este gran objetivo, sus palabras y acciones llenas de criterio y conocimiento fruto de su gran experiencia como docente-investigador profundizaron en mí la pasión por la investigación y como ser humano la palabra de un amigo carismático y sencillo me deja huella en lo moral y ético.

A Raquel Gilar, con quien compartimos no solo la responsabilidad de la ejecución del proyecto de investigación PIC-18-INE-EPN-002, en conjunto con la Universidad de Alicante, La Universidad Nacional de Educación y la Escuela Politécnica Nacional, sino también la producción científica con los resultados del proyecto, una gran profesional y amiga que aporto con sus conocimientos invaluables a esta investigación.

A los investigadores destacados Alejandro Veas de la Universidad de Alicante, Jaime León de la Universidad de las Palmas de Gran Canaria, por su gran aporte en alcanzar esta meta.

A mis compañeros del proyecto; Jack Vidal; Liliam Molina, Jessica Reina, Marianela Jaramillo, Diego Salazar, Iván Sandoval y Amanda Ordoñez, por su disposición y apoyo constante.

A mis padres, Don Tarquino Sánchez (+) y Doña Hilda Almeida, mi agradecimiento profundo y va para ellos mi dedicatoria especial, a quienes llenaron la esencia de mi ser, con valores éticos y morales. A mis hermanos, cuñados y sobrinas por su apoyo moral y afectivo.

A mi esposa Patricia, compañera noble que todo lo cura y conoce, razón de mi vida. A mis queridos hijos; Sofía, Andrés Tarquino y Emilia, para quienes vivo y son fuente de inspiración a seguir conquistando sueños que se hacen realidad. Muchas gracias por levantarme en mis derrotas y disfrutar conmigo las alegrías y triunfos.

2.a. RESUMEN

La evaluación del profesorado constituye uno de los objetivos más importantes para garantizar la calidad de la educación superior, dicha evaluación responde a la necesidad de mejorar el Sistema de Educación Superior del Ecuador, es así que, la Ley Orgánica de Educación Superior, establece que los profesores se someterán a una evaluación periódica integral según los parámetros de evaluación dados por el Reglamento de Carrera y Escalafón del Profesor e Investigador y las normas de cada institución en la cual se observarán la evaluación que realizan los estudiantes a sus docentes (heteroevaluación).

La evaluación integral del desempeño se aplica periódicamente a todos los docentes de las instituciones de educación superior públicas y particulares, evaluando las actividades de docencia, investigación, vinculación y dirección o gestión académica. Los instrumentos de evaluación y los procedimientos para tan importante objetivo deberán ser fiables y validados de tal manera que la información recopilada permita medir con eficacia el desempeño del profesorado y retroalimentarlo para la mejora de la calidad de la enseñanza, la investigación y la gestión.

Por lo tanto, es una necesidad contar con instrumentos de evaluación de la docencia fiables y validados, cuyos resultados propendan a la excelencia y pertinencia de la educación superior y su relación que pudieran tener con el desempeño académico de los estudiantes, la presente tesis doctoral desarrollada en la modalidad por compendio de publicaciones persigue el cumplimiento de tres objetivos que corresponden a los tres artículos científicos con este fin.

El primer objetivo, valida el cuestionario de evaluación del profesorado de la Escuela Politécnica Nacional aplicando el método de Análisis Factorial (AF) con extracción de Componentes Principales verificando su confiabilidad y validez. Se aplicó inicialmente la prueba de esfericidad de Bartlett y luego el análisis de Kaiser-Meyer-Olkin (KMO) cuyos resultados reflejan que existen correlaciones significativas entre las variables (pertinencia) y que el Análisis Factorial es adecuado para explicar los datos (validez). Se parte de un análisis exploratorio de los datos obtenidos de la aplicación de los cuestionarios de 33 ítems con 5 opciones de respuesta a una base de 6 110 estudiantes de las carreras de ingeniería, ciencias y programas de tecnología superior. Estos estudiantes fueron matriculados en 8 facultades y escuelas distintas, estudiando 24 carreras de grado diferentes, el 68.60% son estudiantes varones y el restante 31.40% mujeres, la edad promedio corresponde a 22.30 años. La muestra de profesores consistió en 670 docentes, los cuales representó una muestra variada en términos de edad, categoría y experiencia en la enseñanza, de los cuales el 62.80% fueron hombres. La aplicación de la escala de 33 ítems fue realizada al final de curso académico 2016-17, antes de que los estudiantes conocieran sus calificaciones finales. El resultado del método descrito fue una escala reducida o corta de 14 ítems, la cual identifica claramente 4 constructos o dimensiones: *Planificación y desarrollo de la docencia (Factor 1)*; *Metodología y recursos (Factor 2); Evaluación (Factor 3) y Relación profesor – alumno (Factor 4).*

El segundo objetivo, consiste en validar la escala de evaluación docente reducida practicada a los estudiantes de grado de la Escuela Politécnica Nacional y la relación entre la versión corta (14 ítems) y larga (32 ítems) de la escala con el rendimiento académico y examinar si los puntajes son invariantes con respecto a variables relevantes como el género del estudiante en el contexto de los estudios científico-tecnológicos. La metodología utilizada incluyó: análisis descriptivo, correlación intraclase, modelación de ecuaciones estructurales, análisis factorial confirmatorio, correlaciones entre la escala corta y larga. La muestra utilizada es la misma que la del primer artículo científico, es decir, estuvo conformada por 6 110 estudiantes de la Escuela Politécnica Nacional del Ecuador que calificaron la enseñanza de 310 de sus profesores, lo que representa una muestra variada en edad, categoría y experiencia docente. Estos estudiantes estaban matriculados en 8 facultades diferentes en 28 programas de grado y asistieron a 358 diferentes clases los cuales calificaron la enseñanza de sus maestros durante el curso académico 2016-17. El mayor porcentaje de estudiantes hombres es representativo de la población de estudiantes politécnicos. Las medidas de rendimiento académico de los estudiantes se obtuvieron para una submuestra de 1 538 estudiantes. Los resultados mostraron un modelo multidimensional con cuatro factores altamente correlacionados que no excluyen un factor general, con un excelente ajuste a los datos, tanto en la escala larga como en la versión corta de la escala. La estructura con el mejor ajuste fue el modelo de ecuaciones estructurales exploratorio (ESEM) de cuatro factores de dos factores; sin embargo, las cargas de los factores en el factor general fueron bajos y, por lo tanto, se mantuvo la estructura ESEM de cuatro factores.

Al analizar la escala corta de 14 ítems, utilizando ESEM de cuatro factores, proporcionó cargas de moderadas a altas y cargas cruzadas bajas específicamente, para la *Planificación y desarrollo de la docencia* (Factor 1), las cargas oscilaron entre 0.508

y 0.857, para *Metodología y recursos* (Factor 2) entre 0.601 y 0.856, para *Evaluación* (Factor 3) entre 0.385 y 0.885, y para *Relación profesor-alumno* (Factor 4) entre 0.629 y 0.958. Por lo tanto, decidimos mantener esta *estructura*. Para la escala de 32 ítems la fiabilidad se evaluó mediante la fórmula de factores correlacionados, para la escala total fue 0.980, para *Planificación y desarrollo de la docencia* 0.949, para *Metodología y recursos* 0.901, para *Evaluación* 0.948, y para *Relación profesor-alumno* 0.947.

Para el análisis de las correlaciones entre la escala de evaluación de la docencia corta y larga con el rendimiento académico, tomando datos individuales y agregados, fueron estadísticamente significativas con valores moderados-bajos. Tomando las subescalas como la escala total mostraron correlaciones significativas con el rendimiento académico.

Además, las correlaciones en los datos agregados en clases o secciones fueron ligeramente más altas que en los datos individuales. Por otro lado, los resultados de este estudio mostraron invariancia de la medición de género configurable, métrica y escalar en el contexto de los estudios científico-tecnológicos.

En conjunto, los resultados demostraron las buenas cualidades psicométricas del Cuestionario de Evaluación del Profesor de la Escuela Politécnica Nacional y su validez de constructo y criterio, así como su alta fiabilidad. Además, los índices psicométricos de la versión corta de esta escala sugieren la posibilidad de desarrollar escalas cortas de tres o cuatro ítems que son igualmente confiables y válidas.

El tercer objetivo, aborda la relación entre la evaluación de la enseñanza por parte de los estudiantes (SET) y rendimiento académico en la educación superior utilizando una gran muestra de estudiantes y profesores de la Escuela Politécnica Nacional, se utilizó diferentes procedimientos metodológicos que consideran como unidades de análisis las clases individuales y grupales, la variabilidad entre los estudiantes dentro de las clases, y la variabilidad entre las medias de las clases grupales y su relación con el rendimiento académico medio de los grupos de clase, mediante correlación y técnicas de regresión múltiple. También se realizó un análisis multisección en aquellas disciplinas del curso en las que había más de un grupo de clase (sección). Los resultados del análisis de clases individual y grupal revelaron que SET fue moderadamente bajo, pero relacionado con el rendimiento académico de manera significativa una vez que se controló el efecto del rendimiento académico previo. Los resultados del análisis realizado en las disciplinas del curso con diferentes secciones, de acuerdo con un diseño multisección, arrojaron resultados similares a los análisis de datos individuales y grupales.

Con base a los resultados de esta investigación, se puede llegar a conclusiones que permiten definir la validez y confiabilidad de una escala de evaluación docente en una universidad y la relación de la evaluación de la enseñanza con el rendimiento académico de los estudiantes, en el marco de un mejoramiento continuo de la calidad del sistema de educación superior. Las principales conclusiones se describen a continuación:

En conjunto, los resultados demostraron las buenas cualidades psicométricas del cuestionario de evaluación del desempeño docente de la Escuela Politécnica Nacional y su validez de constructo, así como su alta confiabilidad. Además, los índices psicométricos de la versión corta de esta escala sugieren la posibilidad de desarrollar escalas cortas de tres o cuatro ítems igualmente fiables y válidos, donde las relaciones obtenidas entre las versiones larga y corta del nuevo instrumento con el rendimiento académico tienen implicaciones prácticas para la enseñanza del docente.

Por otro lado, este instrumento de evaluación puede ayudar a los profesores para adaptar su enseñanza a las necesidades y preferencias de los estudiantes en el contexto de características específicas de los estudios politécnicos, sin perder de vista la controversia entre las percepciones de los estudiantes sobre la calidad de la enseñanza, o percepciones de aprendizaje, y su aprendizaje real en el contexto de las ciencias exactas o de ingeniería.

Respecto a las limitaciones de este estudio, dado que las escalas larga y corto se administraron como parte de la escala completa, y a pesar de la corrección de Levy (1968) y Gower (1971) para el cálculo de la correlación entre las dos versiones, sería necesario administrar las escalas larga y corta a la misma muestra de forma independiente.

2.b. ABSTRACT

The evaluation of the teaching staff constitutes one of the most important objectives to guarantee the quality of higher education, said evaluation responds to the need to improve the Higher Education System of Ecuador thus, the Organic Law of Higher Education establishes that the Professors will undergo a Comprehensive periodic evaluation according to the evaluation parameters given by the Regulation of the Career and Ladder of the Professor and Researcher and the regulations of each institution in which this evaluation that students make of their teachers will be observed (hetero-evaluation). Comprehensive performance evaluation is periodically applied to all teachers of public and private higher education institutions, evaluating the activities of teaching, research, linking, and academic direction or management. The evaluation instruments and the procedures for this important must be reliable and validated in such a way that the information collected allows to effectively mediate the performance of the teaching staff and provide feedback for the improvement of the quality of teaching, research, and management.

Therefore, it is a necessity to have reliable and validated teaching evaluation instruments, whose results tend to the excellence and relevance of higher education and its relationship that they may have with the academic performance of students, the present doctoral thesis developed In the modality by a compendium of publications, it pursues the fulfillment of three objectives that correspond to the three scientific articles for this purpose.

The first objective validates the evaluation questionnaire of the teaching staff of the National Polytechnic School of Factor Analysis applying method with the extraction of Principal Components, verifying its reliability and validity. The Bartlett sphericity test was applied first and then the Kaiser-Meyer-Olkin analysis, the results of which reflect that there are significant correlations between the variables (relevance) and that the factor analysis is adequate to explain the data (validity). It is based on an exploratory analysis of the data obtained from the application of the 33-item questionnaires with 5 response options to a base of 6 110 students from engineering, science, and higher technology programs. These students were enrolled in 8 different faculties and schools, studying 24 different undergraduate careers, 68.60% are male students, and the remaining 31.40% female, the average age corresponds to 22.30 years. The sample of teachers consisted of 670 teachers, who represented a varied sample in terms of age, category, and teaching experience, of which 62.80% were men. The application of the

33-item scale was carried out at the end of the 2016-17 academic period before the students knew their final grades. The result of the described method was a reduced or short scale of 14 items, which identifies 4 constructs or dimensions: Planning and development of teaching (Factor 1); Methodology and resources (Factor 2); Evaluation (Factor 3) and Teacher-student relationship (Factor 4).

The second objective is to validate the reduced teacher evaluation scale practiced on undergraduate students from the National Polytechnic School and the relationship between the short version (14 items) and the version long (32 items) of the scale with academic performance and examine if the scores are invariant concerning relevant variables such as the student's gender in the context of scientific-technological studies. The methodology used included: descriptive analysis, interclass correlation, modeling of structural equations, confirmatory factor analysis, correlations between the short and long scale. The sample used is the same as that of the first scientific article, that is, it was made up of 6 110 students from the National Polytechnic School of Ecuador who rated the teaching of 310 of their professors, which represents a sample varied in age, category, and experienced teacher. These students were enrolled in 8 different faculties in 28-degree programs and attended 358 different classes which rated the teaching of their teachers during the 2016-17 academic year. The highest percentage of male students is representative of the population of polytechnic students. The mean age was 22.6 years (SD = 3.2). Measures of student academic performance were obtained for a subsample of 1 538 students. The results show a multidimensional model with four highly correlated factors that do not exclude a factor generally, with an excellent fit to the data, both in the long scale and in the short scale version. The structure with the best fit was the exploratory structural equation model (ESEM) of four factors of two factors; however, factor loadings in the overall factor were low and therefore the ESEM four-factor structure was maintained.

When analyzing the short scale of 14 items, using ESEM of four factors, it provided loads of moderate to high and low cross loads specifically, for the Planning and development of teaching (Factor 1), the loads ranged between 0.508 and 0.857, for Methodology and resources (Factor 2) between 0.601 and 0.856, for Assessment (Factor 3) between .385 and 0.885, and Teacher-student relationship (Factor 4) between 0.629 and 0.958. Therefore, we decided to keep this structure. For the 32-item scale, reliability was evaluated using the formula of congeneric correlated factors, for the scale total it was 0.980, for Planning and teaching development 0.949, for Methodology and resources 0.901, for Evaluation 0.948, and Evaluation 0.948 Teacher-student relationship 0.947.

For the analysis of the correlations between the evaluation scale of short and long teaching with academic performance, taking individual and aggregated data, they were statistically significant with moderate-low values. Taking the subscales as the scale total, significant correlations with academic performance. Furthermore, the correlations in the data aggregated in classes or sections were slightly higher than in the individual data. On the other hand, the results of this study invariance of the configurable, metric, and scalar gender measurement in the context of scientific-technological studies.

Overall, the results demonstrated the psychometric qualities good of the National Polytechnic School Teacher Assessment Questionnaire and its construct and criterion validity, as well as its high reliability. Furthermore, the psychometric indices of the short version of this scale suggest the possibility of developing short scales of three or four items that are equally reliable and valid.

The third objective addresses the relationship between the evaluation of student teaching (SET) and academic performance in higher education using a large sample of students and teachers from the National Polytechnic School, using different methodological procedures that they consider as units of analysis the individual and group classes, the variability between the students within the classes and the variability between the means of the group classes and their relationship with the average academic performance of the groups of classes, using correlation techniques and multiple regression. A multisection analysis was also carried out in those course disciplines in which there was more than one class group (section). The results of the individual and group class analysis revealed that SET was low moderately, but related to academic performance in a significant way once the effect of previous academic performance was controlled for. The results of the analysis carried out in the disciplines of the course with different sections, according to of the multisection design, yielded similar results to the individual and group data analysis.

Based on the results of this research, conclusions can be reached that would allow defining the validity and reliability of a teacher evaluation scale in a university and the relationship of the evaluation of teaching with the academic performance of students, within the framework of continuous improvement of the quality of the higher education system. The main conclusions are described below:

Overall, the results demonstrated the psychometric qualities good of the Teaching Evaluation questionnaire of the National Polytechnic School and its construct validity, as well as its high reliability. Furthermore, the psychometric indices of the short version of this scale suggest the possibility of developing short scales of three or four items that are equally reliable and valid, where the relationships obtained between the long and short versions of the new instrument with academic performance have practical implications for teacher teaching.

On the other hand, this assessment instrument can help teachers to adapt their teaching to the needs and preferences of students in the context of specific characteristics of polytechnic studies, without losing sight of the controversy between students' perceptions about the teaching quality or learning perceptions, and their actual learning in the context of the exact sciences or engineering.

Regarding the limitations of this study, given that the long and short scales were administered as part of the full scale, and despite the Levy and Gower correction for calculating the correlation between the two versions, it would be necessary to administer the scales long and short to the same sample independently.

3.a. PRESENTACIÓN

Los procesos de enseñanza-aprendizaje en la educación superior, exigen de una actualización y mejoramiento constante de sus técnicas, explorando nuevos métodos de adquirir conocimiento, y técnicas de investigación, que sirven para desarrollar habilidades de comunicación efectivas entre profesor/estudiante; en este contexto, la evaluación del desempeño docente utilizando escalas de valoración viables y confiables, y su relación con el rendimiento académico de los estudiantes, apuntan a garantizan la calidad en el sistema de educación superior. Con este antecedente la presente tesis doctoral, denominada *Validación de la escala de valoración de la docencia de la Escuela Politécnica Nacional de Ecuador, un análisis de su relación con el rendimiento académico*, constituye un trabajo original de investigación con valor científico y configurada como una unidad, elaborado a partir de un conjunto de publicaciones relacionadas con el plan de investigación de la tesis doctoral, está desarrollados a partir de la información recopilada en la Escuela Politécnica Nacional de Quito-Ecuador, como resultado de la SET utilizando una escala de evaluación corta y larga, en los periodos académicos 2017-18, aportando con conocimiento científico en el análisis de esta temática educativa.

Esta tesis se enmarca en el programa de Doctorado en Investigación Educativa de la Universidad de Alicante. La fase de la formación predoctoral se inició en 2018, en paralelo, en noviembre del mismo año, se suscribe el convenio No. 20180143CI, entre la Escuela Politécnica Nacional y la Secretaría de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT), para la ejecución del proyecto de investigación concursable denominado: *Diseño e implementación de un modelo de admisión inclusivo al Sistema de Educación Superior del Ecuador*, tras un proceso de evaluación y selección de 427 postulaciones en el marco de la convocatoria para el financiamiento de proyectos de fomento a la investigación y/o desarrollo tecnológico a través de fondos concursables dirigido a los actores generadores y gestores del conocimiento del sistema nacional de ciencia y tecnología, innovación y saberes ancestrales del Ecuador, denominado “INEDITA” de la SENESCYT. Este proyecto está concebido en la modalidad colaborativa en red, con la Universidad Nacional de Educación y la Universidad de Alicante, con la participación de la Dra. Raquel Gilar, Jefa del Departamento de Psicología Evolutiva y Didáctica en calidad de directora subrogante del proyecto y del investigador principal y director de esta tesis doctoral, el Dr. Juan

Luis Castejón Costa, Catedrático de Psicología de la Educación y profesor del prenombrado Departamento.

Esta tesis doctoral, está elaborada por compendio de publicaciones, lo que permitió difundir los resultados de dicha investigación en revistas de alto impacto. Los artículos científicos que se incluyen en esta tesis doctoral están indexados en Journal Citation Reports (JCR) Science Edition o Social Sciences Edition, en los cuartiles Q2 y Q3. Además, todos los artículos científicos se encuentran redactados en lengua inglesa.

1. Sánchez-Almeida, T. F., Sandoval-Palis, I. P., Gilar-Corbi, R., Castejón-Costa, J. L., & Salazar-Orellana, D. I. (2020). Teaching evaluation questionnaire validation at Escuela Politécnica Nacional, applying the method of Factor Analysis with extraction of principal components. *Ingeniería E Investigación*, *40*(1), 70-77. https://doi.org/10.15446/ing.investig.v40n1.79634
2. Sanchez, T.F., León, J., Gilar-Corbi, R., Castejón, J.L. (2021). Validation of a Short Scale for Student Evaluation of Teaching Ratings in a Polytechnic Higher Education Institution. *Front. Psychol.*, Aceptado para su publicación el 24-5-2021 https://doi.org/10.3389/fpsyg.2021.635543
3. Sánchez T, Gilar-Corbi R, Castejón J-L, Vidal J and León J (2020) Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador. *Front. Psychol*. *11*:233. https://doi.org/10.3389/fpsyg.2020.00233
4. Sánchez, T., Naranjo, D., Vidal, J., Salazar, D., Pérez, C., & Jaramillo, M. (2021). Analysis of academic performance based on sociograms: A case study with students from at-risk groups. *Journal of Technology and Science Education*, *11*(1), 167-179. https://doi.org/10.3926/jotse.1110

La tesis doctoral se estructura en dos partes; *la primera*, analiza una escala de evaluación del desempeño docente (SET) de 32 ítems, aplicada a una gran muestra de estudiantes de las carreras de grado de ingeniería, tecnología superior y ciencias de la Escuela Politécnica Nacional y se reduce a una escala corta de 14 preguntas, utilizando el método del Análisis Factorial exploratorio con extracción de Componentes Principales, reduciendo al número óptimo de factores capaces de explicar el máximo de información contenido en los datos.

Se presenta un segundo artículo científico, aceptado para su publicación en la revista Frontiers in Psychology cuyo objetivo principal de este trabajo fue validar la escala de calificación para la evaluación de la enseñanza por parte de los estudiantes desarrollada en el contexto de la Escuela Politécnica Nacional de Ecuador, utilizando una gran muestra de 6110 estudiantes. El análisis de datos incluyó análisis descriptivo, correlación intraclase, modelado exploratorio de ecuaciones estructurales, análisis factorial confirmatorio (AFC), correlaciones entre la escala corta y larga corregida para la varianza del error compartido, la invarianza de la medición de género, la confiabilidad usando factores correlacionados congenéricos y la correlación con el rendimiento académico , tomando la clase como unidad de análisis siguiendo un diseño multisección.

En cuanto a las implicaciones prácticas de esta primera parte, se debe reconocer la estructura multidimensional y jerárquica de las dimensiones evaluadas en el cuestionario de evaluación de la docencia por parte de los estudiantes, mientras que la retroalimentación brindada a los docentes para la mejora de la práctica docente incluye un perfil de las puntuaciones de las diferentes dimensiones, mostrando las fortalezas y debilidades de los métodos empleados por cada docente. Por otro lado, dada la existencia de sesgo de género en los estudiantes que evalúan la enseñanza, se comprobó la invarianza de género configuracional, métrica y escalar en el contexto de una institución de educación superior de ciencia y tecnología. Los resultados de este trabajo también muestran la validez concurrente de la escala reducida de 14 ítems, que mostró una alta correlación con la escala total de 32 ítems. La correlación corregida de Levy y el índice Gower, reveló una alta concurrencia entre ambas formas, con valores superiores a 0.90.

La *segunda parte* de la tesis doctoral explora la relación entre los resultados de la evaluación del desempeño docente, como una medida de eficacia de los profesores, con el rendimiento académico de los estudiantes de un clase individual o grupal. Para ello se utilizó un diseño multisección cuando las disciplinas del curso tenían más de un grupo de clase; y se consideró el desempeño académico previo, ya que no se aseguró la asignación aleatoria de los estudiantes a los grupos. Se utilizaron métodos estadísticos que consideran la variabilidad individual del estudiante dentro de cada grupo o entre grupos. Los resultados obtenidos con datos agregados, tomando el grupo clase como la unidad de análisis, mostró una correlación moderada pero estadísticamente significativa

(0.28) entre la SET y la calificación final de la disciplina. Sin embargo, la correlación entre el rendimiento previo y la SET no fue estadísticamente significativa, lo que sugiere que la SET no se ve afectado por los logros académicos anteriores. Por otro lado, los resultados también mostraron correlaciones significativas y moderadas entre las versiones larga y corta de la escala con el rendimiento académico, tomando datos individuales y agregados en clases o secciones. También se analizó, el rendimiento académico de estudiantes de grupos vulnerables desde la perspectiva del análisis de redes sociales (ARS). En este sentido, se estudió la información académica y de interacción de 45 estudiantes pertenecientes a grupos vulnerables que asistieron durante un periodo académico a un curso piloto de intervención socio-académica. Con esta información se construyó un sociograma, a partir del cual se determinaron las métricas de centralidad de ARS. Posteriormente, se estudiaron las relaciones entre dichas métricas y las variables académicas a través de análisis de correlación y regresión lineal con regularización LASSO. Finalmente, se determinó que el rendimiento académico de los estudiantes al finalizar el curso piloto estuvo influenciado, por un lado, por los conocimientos académicos previos al ingreso a la universidad, representados por la nota en la sección de Matemática y Geometría de la prueba de diagnóstico y, por otro lado, por la dinámica de la red social en la que se desenvolvieron en el aula, representada por la centralidad de eigenvector. Estos resultados tienen potencial explicativo significativo del rendimiento académico en función de las métricas de ARS y aportan evidencia para sustentar la implementación de prácticas que fomenten un entorno social saludable en el contexto académico.

Como parte del proyecto de investigación, durante la formación predoctoral se han presentado diversas ponencias y artículos en congresos internacionales relacionadas con la tesis doctoral. Estos artículos y ponencias que se encuentran en la página web del Grupo de Investigación *"Investigación Educativa",* https://investigacioneducativa.epn.edu.ec, de la Escuela Politécnica Nacional, dan cuenta de los trabajos de investigación que se han realizó en torno a esta temática educativa:

1. Gilar-Corbi R, Pozo-Rico T, Castejón J-L, Sánchez T, Sandoval-Palis I, Vidal J. Academic Achievement and Failure in University Studies: Motivational and Emotional Factors. Sustainability. 2020; 12(23):9798. https://doi.org/10.3390/su12239798

2. Sánchez, T., Ordoñez, A. Gilar, R., & Castejón J.L.(2020). Confirmatory Factor Analysis of the Assessment Instrument Teacher of the Escuela Politécnica National. (2020) Proceedings of the LACCEI International Multi-conference for Engineering, Education and Technology, 2020-July, https://dx.doi.org/10.18687/LACCEI2020.1.1.110
3. Ramos, V., Sánchez, T., Reina, J., Franco-Crespo, A. (2020). Differences between vulnerable and non-vulnerable students regarding the psychological abilities and self-control skills within the development of learning, INTED2020 Proceedings. pp. 8388-8394. http://dx.doi.org/10.21125/inted.2020.2282
4. Franco-Crespo, T. Sánchez, N. Molina, V. Ramos (2020). The inclusion of students in vulnerable conditions in Ecuador, INTED2020 Proceedings, pp. 8383-8387. http://dx.doi.org/10.21125/inted.2020.2281
5. Sandoval, I., Sánchez, T., Naranjo, D., & Jiménez, A., Proposal of a Mathematics Pilot Program for Engineering Students from Vulnerable Groups of Escuela Politécnica Nacional (2019) Proceedings of the LACCEI International Multi-conference for Engineering, Education and Technology, 2019-July, https://dx.doi.org/10.18687/LACCEI2019.1.1.387
6. Molina Valencia, L. N., Sánchez Almeida, T. F., Vidal Chica, J. I., Guayasamín Pico, R. M., & Reina Trávez, J. L. (2019). Evaluación diagnóstica de conocimientos y propuesta de un curso de intervención a los estudiantes que ingresan al curso de nivelación de la escuela politécnica nacional. Encuentro Internacional de Educación en Ingeniería. https://acofipapers.org/index.php/eiei/article/view/150.

4. INTRODUCCIÓN

4.1 Evaluación del desempeño docente: Caracterización y metodología de análisis.

A lo largo de los años, la innovación y la aparición de nuevas líneas de investigación han permitido incorporar nuevas áreas de conocimiento como instrumentos de formación académica. Es ahí donde se propone la aplicación de la psicología de la instrucción como una nueva herramienta para el profesorado. Esto ha servido de pauta para la aparición de nuevas investigaciones que buscan comprender de manera integral el proceso de enseñanza dentro de la metodología y encontrar la mejor alternativa para transmitir conocimientos en el aula.

Aparicio (2014) indicó que es posible interpretar el aprendizaje como la relación existente entre comunicación e interacción donde la interacción se ve como parte del desarrollo docente y académico. Por tanto, los profesores universitarios requieren de competencias específicas que les permitan potenciar la calidad del proceso de enseñanza-aprendizaje en el aula. Estas competencias pretenden alcanzar la excelencia en términos de resultados, lo que implica una cultura de evaluación y control del proceso de aprendizaje.

Los instrumentos que normalmente se utilizan para medir la evaluación de los profesores por parte de sus estudiantes, los programas de estudio y la satisfacción con su instrucción se conocen como escalas de calificación estándar. Sin embargo, las investigaciones sobre la evaluación de las calificaciones docentes por los estudiantes aún no han brindado respuestas claras a algunas preguntas sobre su validez (Hornstein, 2017; Marsh, 2007 a, b; Spooren, Brockx y Mortelmans, 2013; Uttl, White y González, 2017).

Muchos instrumentos de evaluación se han construido y validado dentro de la propia Institución, y los resultados de dicha validación no siempre se han publicado y, en algunos casos, ni siquiera se ha probado la calidad psicométrica (Richardson, 2005). Además, existe una falta de consenso sobre el número y tipo de dimensiones (Spooren et al., 2013), debido a problemas conceptuales relacionados con la falta de un marco teórico sobre qué es la enseñanza efectiva, y problemas metodológicos en la medición de estas dimensiones como un proceso impulsado por datos (en el que se utilizan diferentes técnicas analíticas post hoc). Parece necesario utilizar las dimensiones más habituales, que se asocian a una mayor eficacia docente.

Desde una perspectiva estadística, existen pocos registros en Ecuador sobre la evaluación del desempeño docente en las universidades, y la limitada evidencia existente es de carácter restringido. En la actualidad, la *Ley Orgánica de Educación Superior*, que rige el sistema educativo ecuatoriano establece en el artículo 151 que los docentes se someterán a una evaluación periódica integral de acuerdo con el *Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior* y las normas estatutarias de cada Institución que lo integra, en ejercicio de su autonomía responsable, la encuesta que realizan los estudiantes sobre sus profesores será considerada como uno de los parámetros de evaluación (Consejo de Educación Superior, 2018).

Los instrumentos de evaluación actuales se diseñaron considerando los componentes establecidos en la normativa citada en el párrafo anterior, tales como, autoevaluación, coevaluación y heteroevaluación. Algunos de los ítems proceden de otras escalas de valoración de la SET, como SEEQ (Marsh, 2007a), STERS (Toland y De Ayala, 2005) y SET37 (Mortelmans y Spooren, 2009), y están adaptados a las características de la Escuela Politécnica Nacional. En general, la validación técnica del instrumento de evaluación no se considera un criterio para garantizar la calidad de la aplicación del instrumento. La evaluación integral del desempeño docente es un componente esencial que permite a un profesor enrolarse como Profesor Asistente o Profesor Asociado. Los requisitos incluyen una calificación de al menos el 75% de la puntuación en la evaluación de desempeño durante sus dos últimos períodos académicos. Adicionalmente, según el artículo 96 del *Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior* (Consejo de Educación Superior, 2017), el personal académico será destituido si ha obtenido:

1) Un desempeño de evaluación integral de menos del 60% por dos veces consecutivas, y
2) Cuatro evaluaciones integrales de desempeño de menos del 60% a lo largo de su carrera.

Además, dicho normativa legal establece que los Profesores Principales titulares serán promovidos al siguiente nivel si cumplen con otros requisitos, como haber obtenido una puntuación de al menos 80% en la evaluación de desempeño de sus dos últimos períodos académicos (Consejo de Educación Superior, 2017).

El presente estudio se realizó en un contexto diferente a la mayoría de los estudios previos (Clayson, 2009), las evaluaciones de los estudiantes de la docencia se efectuaron en una institución de educación superior de ciencia y tecnología como la Escuela Politécnica Nacional del Ecuador, donde los estudiantes cursan materias técnicas, tales como ingeniería, ciencias y tecnología superior. Desafortunadamente, en América del Sur hay una escasez de escalas SET confiables y válidas en las instituciones de educación superior politécnicas, aunque es un procedimiento generalizado desde principios de la década de 1980 (Pareja, 1986).

La metodología propuesta surge como una necesidad para validar los instrumentos de evaluación del profesorado de la Escuela Politécnica Nacional del Ecuador. Esta validación se aplica a profesores de ingeniería, ciencias y programas de nivel tecnológico superior, utilizando el método de Análisis Factorial (AF) con extracción de componentes principales. Esta investigación considera los requisitos de confiabilidad y validez que deben tener los cuestionarios con escalas de calificación de opinión Likert (Alaminos y Castejón Costa, 2006).

El método más utilizado para extraer los factores iniciales de la matriz de variables de correlación observadas es el método del Componente Principal, se caracteriza por un análisis de la varianza total del conjunto de variables observadas. El propósito es descubrir los principales componentes que definen este conjunto. Tanto el análisis factorial como el análisis de componentes principales son técnicas de reducción de datos multivariables.

Las principales características métricas para determinar la precisión de un instrumento de evaluación (cuestionario) son la confiabilidad y la validez. La confiabilidad es la propiedad que designa la constancia y precisión de los resultados obtenidos por un instrumento cuando se aplica en diferentes ocasiones. Por otro lado, la validez se refiere a si el instrumento puede medir lo que se pretende medir (Carvajal, Centeno, Watson, Martínez y Sanz Rubiales, 2011). La confiabilidad se puede estimar por cuatro medios: consistencia interna, estabilidad, equivalencia y armonía entre jueces. El método de elección es la consistencia interna, que utiliza la prueba estadística Alfa (α) de Cronbach. El objetivo de este enfoque es comparar la variabilidad de cada ítem con la variabilidad total del instrumento.

Actualmente, se ha desarrollado una línea de trabajo para reducir el tamaño de la escala o para elaborar nuevas escalas con un número reducido de ítems. La falta de tiempo para su aplicación, el cansancio y las posibles respuestas estereotipadas en escalas demasiado largas o que forman parte de un conjunto que se aplica dentro de un mismo estudio, entre otras razones, ha llevado a propuestas de escalas cortas (Gogol et al., 2014; Lafontaine et al., 2016). Estas escalas deben ser lo suficientemente pequeñas para permitir una evaluación rápida de los constructos propuestos, pero lo suficientemente grandes para garantizar la confiabilidad, validez y estimación de parámetros precisas adecuadas. Los objetivos iniciales del presente trabajo son, por un lado, analizar la validez de constructo del cuestionario de enseñanza-aprendizaje y por otro, proponer una reducción de esa escala, conservando sus propiedades psicométricas. Además, estos métodos permiten identificar los elementos y constructos más relevantes.

El resultado de esta validación es el diseño de un cuestionario cuya aplicación aporta información veraz que mejorará la calidad del Sistema de Educación Superior del Ecuador. En este contexto se parte de una escala original que contiene treinta y tres ítems o variables cuantitativas, por lo que al aplicar la técnica de análisis factorial con el método de extracción de componentes principales se obtiene dos variables ficticias que permitan relacionar y resumir la encuesta al profesorado. Esto permite evaluar los aspectos relevantes del docente, dentro del proceso de enseñanza-aprendizaje.

4.2 Métodos estadísticos para el análisis de validez, confiabilidad y reducción de la escala de evaluación del desempeño docente

4.2.1 Análisis de la información original en cuanto a su relevancia y validez

Se realiza un análisis exploratorio de los datos obtenidos de la aplicación de los instrumentos de evaluación de 33 ítems con 5 opciones de respuesta, los elementos se agrupan teóricamente en los siguientes cuatro factores. 1. *Planificación, dominio y claridad en la explicación de la materia* (es decir, el docente expresa convenientemente los objetivos y contenidos de la clase, indicando su relación con la formación del alumno). 2. *Metodología y 3. Recursos* (es decir, el docente preparó material didáctico aparte del libro de texto y lo dio a conocer). 4. Evaluación (es decir, los eventos de evaluación están relacionados con la enseñanza impartida) 5. *Relación profesor-*

estudiante (es decir, el profesor creó un clima de confianza y productividad en clase). Si bien el número y las dimensiones de la enseñanza eficaz sigue siendo una cuestión abierta (Spooren et al, 2013), estas cuatro dimensiones están presentes en la mayor parte de la literatura sobre SET (Feldman, 1989; Huybers, 2014; Richardson, 2005). Se obtuvo respuestas por 6110 estudiantes de los programas de ingeniería, ciencias y tecnología de nivel superior para los profesores de la Escuela Politécnica Nacional. Estos estudiantes estaban matriculados en 8 facultades y escuelas, cursando 24 titulaciones diferentes. El mayor porcentaje de estudiantes varones es representativo de la población de estudiantes de estudios politécnicos, en la que el 68.60% eran varones y el 31.40% mujeres. La edad media fue de 22.30 años. Estos 6110 alumnos asistieron a 1 380 asignaturas diferentes que se distribuyeron en 1812 grupos de clase. La muestra de maestros consistió en 670 maestros, quienes representaron una muestra variada en términos de edad, categoría y experiencia docente. Más de la mitad de estos profesores eran hombres (62.80%). La aplicación de la escala de 33 ítems se llevó a cabo al final del semestre 2017-A (octubre 2017- marzo 2018), antes de que los estudiantes conocieran sus calificaciones finales. Todos los profesores fueron evaluados por los estudiantes en el mismo período. Todos los alumnos tuvieron que evaluar a los profesores para poder acceder a sus calificaciones finales. La evaluación docente de los estudiantes se realizó a través de una plataforma electrónica, obteniendo 19527 registros (matriz de datos original) en los que se registraron los datos (un mismo estudiante pudo evaluar a varios profesores ya que cursó varias materias).

A partir de la matriz de datos original, se elabora una matriz de correlación entre todas las variables consideradas (ítems). Se realizan diversas pruebas para determinar si es pertinente, desde un punto de vista estadístico, realizar análisis factoriales con la información disponible de la matriz de correlaciones.

4.2.2 Pruebas de validez y confiabilidad

La prueba de *esfericidad de Bartlett*: se basa en la distribución chi-cuadrado, donde valores altos llevan a rechazar la hipótesis nula (H0) que establece que las variables no están correlacionadas dentro de la población. Por tanto, la prueba de esfericidad de Bartlett determina si la matriz de correlación es una matriz de identidad, lo que indicaría que el modelo factorial es inadecuado. Si el valor de significancia (valor p) es menor que 0,050, rechazamos la hipótesis nula (H0) y continuamos con el análisis factorial.

El índice de *Kaiser-Meyer-Olkin (KMO)*: permite la comparación entre la magnitud de los coeficientes de correlación observados y la magnitud de los coeficientes de correlación parcial. La estadística KMO varía entre 0 y 1. Los menores de 0.500 indican que no se requiere análisis factorial para los datos en cuestión.

El *coeficiente de correlación parcial*: describe la relación lineal entre dos variables controlando los efectos de una o más variables adicionales. Estos coeficientes deberían tender a cero cuando se prestan para análisis factorial (Montoya, 2007).

4.2.3 Extracción de componentes principales

La interpretación de los componentes principales suele ser difícil, por lo que la extracción inicial se rota para lograr una solución que la facilite. Varimax con normalización de Kaiser (Kaiser, 1958) es el método de rotación que utiliza la rotación ortogonal de factores previamente normalizados. En otras palabras, mantiene la independencia entre los factores rotados. Este método logra que cada componente rotado presente correlaciones con solo unas pocas variables. Por tanto, este método minimiza el número de variables con cargas elevadas en un factor y es adecuado cuando se reduce el número de componentes.

4.3 Relación entre la evaluación del desempeño docente con el rendimiento académico de los estudiantes.

Uno de los puntos centrales de controversia es la relación de las calificaciones de SET con sus resultados de aprendizaje, como el rendimiento académico (Uttl et al., 2017). La evidencia en apoyo de SET como medida de la efectividad de la instrucción de los docentes proviene de los estudios que muestran una correlación entre las medidas de evaluación del aprendizaje por parte de los estudiantes y el rendimiento de los mismos.

Inicialmente, la validez del comportamiento de los estudiantes podría probarse por la correlación entre SET y el rendimiento académico. Sin embargo, los criterios de evaluación pueden diferir para distintas unidades de curso y las calificaciones de los estudiantes no pueden considerarse como una medida simple de la eficacia de la enseñanza (Richardson, 2005).

La evidencia clave proporcionada a favor de SET como una medida de la efectividad de la instrucción de los docentes son los estudios multisección (Uttl et al., 2017). Leventhal (1975) y Cohen (1981) defienden que la validación más fuerte de SET implica la designación de estudiantes para las diferentes secciones de un curso de múltiples secciones. Si la designación es aleatoria, las diferencias en el desempeño de los estudiantes entre secciones pueden deberse a diferencias a los docentes. Cuando los estudiantes se autoseleccionan en secciones, puede ser difícil inferir la relación calificación/logro. Si este es el caso, Marsh y Overall (1980) consideran que, en estos estudios, se debería proporcionar adecuados controles/medidas para la capacidad inicial o el rendimiento previo.

Algunos investigadores (Cohen, 1981; Clayson, 2009; Uttl et al., 2017) señalan que el rendimiento de los estudiantes depende en gran medida de factores como la inteligencia o el rendimiento previo y que para controlar completamente estos factores, es necesario asignar estudiantes y profesores al azar en las clases o, alternativamente, utilizar otros procedimientos de control de la capacidad o el rendimiento inicial del estudiante, como el análisis de covarianza utilizando medidas de rendimiento académicos o capacidad previos como covariables; se usa el cambio en las calificaciones basado en momentos previos y posteriores a la prueba; o hacer una regresión de los puntajes de desempeño de los estudiantes individuales en las mediciones de los logros anteriores y usar las ganancias residuales en el desempeño, promediadas entre los estudiantes dentro de las secciones, como medidas de aprendizaje. Es aconsejable utilizar un procedimiento estadístico en el que tanto las calificaciones como el rendimiento de los estudiantes se ajustan a la condición inicial del estudiante.

Un estudio de múltiples secciones ideal implica una disciplina de curso o asignatura con muchas clases grupales comparables (secciones) que toman el mismo programa y criterios de evaluación, en las que los estudiantes se asignan aleatoriamente a secciones, con un profesor diferente en cada sección; todos los profesores son evaluados mediante calificaciones antes de un examen final; y el rendimiento académico del estudiante se evalúa empleando el mismo examen final o uno equivalente.

Si un estudiante muestra un mejor rendimiento académico debido a profesores altamente calificados, se debe observar una correlación entre el SET promedio de las secciones y el examen final promedio de las secciones (Uttl et al., 2017).

Esto nos lleva a considerar la unidad de análisis apropiada en este tipo de estudios (Cohen, 1981). Algunos investigadores utilizan al estudiante como unidad de análisis, relacionando el rendimiento académico del estudiante con la calificación de su profesor.

Otros investigadores utilizan la clase grupal como unidad de análisis, correlacionando el logro medio de la clase grupal con el SET medio de la clase. Los investigadores que utilizan datos de estudiantes individuales siguen un diseño que les permite establecer si los estudiantes que se desempeñan mejor, independientemente de la clase a la que asistan, califican mejor a los maestros. Para analizar la asociación entre SET y el rendimiento académico de los estudiantes para los respectivos profesores, la clase grupal (o profesor) debe utilizarse como unidad de análisis en el diseño de validez de la escala (Cohen, 1981; Abrami et al., 1990; Marsh y Roche, 2000 ; Clayson, 2005; Richardson, 2005).

Aunque esta solución es ampliamente aceptada, recientemente han surgido críticas. Se argumenta que la variabilidad entre estudiantes, a pesar de ser promediada, podría confundir la variación entre medias grupales. En consecuencia, se puede encontrar que no existen relaciones entre el SET y el rendimiento para los estudiantes individuales, incluso cuando los datos medios entre clases muestran una relación significativa (Clayson, 2007; Weinberg et al., 2009). Es necesario utilizar métodos estadísticos que consideren tanto la variabilidad individual dentro de la clase de grupo como la variabilidad entre las medias de la clase de grupo. Otro tema metodológico que puede afectar los resultados sobre la relación entre SET y el rendimiento académico de los estudiantes es el número de secciones (Cohen, 1981; Uttl et al., 2017). Kulik y McKeachie (1975) indicaron que a menudo aparecen grandes correlaciones con tamaños de muestra pequeños, lo que sugiere que para encontrar un coeficiente de validez estable, se necesitan al menos 30 secciones en un estudio de múltiples secciones. Más recientemente, Uttl et al. (2017) presentaron resultados específicos sobre este tema en su metaanálisis de la eficacia docente de la facultad.

El presente estudio incluye una medida del rendimiento académico previo y procedimientos estadísticos que ajustan tanto las medidas de SET como el rendimiento para el rendimiento anterior del estudiante. Si bien el logro previo es una de las variables que más influye en el logro final, este estudio examina si el SET hace una contribución significativa al logro final, luego de controlar el efecto del logro previo. Una cuestión metodológica abierta, que busca abordar este estudio, es la unidad de análisis. La mayoría de los investigadores en este campo utilizan el promedio de clase grupal como unidad de análisis, argumentando que se eliminan las diferencias individuales dentro de la clase grupal y las diferencias entre las medias de las clases grupales, secciones o profesores (Cohen, 1981; Abrami et al., 1990; Marsh y Roche, 2 000; Clayson, 2005; Richardson, 2005; Uttl et al., 2017) se reflejan claramente; otros investigadores defienden la necesidad de tener en cuenta la variabilidad individual dentro de las clases grupales (Clayson, 2007; Weinberg et al., 2 009). Algunos estudios en este campo han considerado ambos aspectos por separado (Clayson, 2009), pero hasta donde sabemos, ninguno ha considerado la variabilidad dentro y entre clases grupales o profesores de manera conjunta. En este estudio, utilizaremos métodos que consideran ambas fuentes de variabilidad, los estudiantes y la clase grupal, para el análisis multinivel.

5. OBJETIVOS

Las pruebas psicológicas y educativas, cuestionarios y escalas, son muestras de conducta que permiten llevar a cabo inferencias relevantes sobre la conducta de las personas, si el proceso de aprendizaje en las instituciones de educación superior objeto de estudio, es complejo, no solo porque involucra a diferentes actores como los profesores, alumnos y el currículo académica, sino porque el sistema de educación tiene que ser regulado y retroalimentado por los resultados de dicho proceso, por lo tanto, es necesario contar con un test que evalúen al profesorado con objetividad y justicia, evitando evaluaciones sesgadas por aspectos como la clase social, la raza, el sexo, las creencias, la situación geográfica de los sujetos, y otros aspectos subjetivos pero siempre se buscara evaluar a todos de forma comparable. Por ello, y de acuerdo con los contenidos expuestos en la introducción, los objetivos que se presentan en esta tesis doctoral son los siguientes:

5.1. Objetivos generales

1. Validar un cuestionario de evaluación del desempeño de la docencia universitaria en el marco de la Escuela Politécnica Nacional del Ecuador.
2. Comparar una versión corta de la escala con una larga (original) y examinar su invariancia en relación con variables relevantes como el género y el rendimiento académico de los estudiantes.
3. Analizar la relación de la evaluación del desempeño docente SET con el rendimiento académico de los estudiantes y desde la perspectiva del análisis de redes sociales (ARS).

5.2. Objetivos específicos

1. Analizar los diferentes trabajos realizados en universidades sobre la evaluación de la docencia universitaria a partir de una amplia revisión del estado del arte.
2. Elaborar una escala de valoración de la docencia universitaria, adaptada al contexto de la EPN, teniendo en cuenta otras escalas existentes, y los resultados de la aplicación de una escala anterior en la EPN.
3. Identificar las principales dimensiones o constructos de la docencia universitaria en este contexto.

4. Aplicar la escala a una muestra amplia y representativa de estudiantes de la EPN y llevar a cabo el análisis y selección de ítems, así como el establecimiento de la fiabilidad y validez de la escala, mediante los procedimientos de la Teoría Clásica de los Test (TCT) y un Análisis Factorial Confirmatorio (AFA).
5. Profundizar en el análisis de las correlaciones entre la evaluación de la enseñanza (SET) y el rendimiento académico en la educación superior, considerando la variabilidad entre las clases grupales. Para ello se utilizarán técnicas de ruta multinivel, además técnicas de correlación y regresión jerárquica.
6. Establecer la validez de criterio o predictiva de la escala y/o de sus dimensiones en relación con criterios tales como la calidad de la docencia, el rendimiento académico de los estudiantes, etc.
7. Analizar el poder predictivo de la percepción de la enseñanza recibida sobre el rendimiento académico de los estudiantes, una vez controlado el efecto de variables como sexo, rendimiento previo o calificaciones de ingreso a la EPN y desde la perspectiva del análisis de redes sociales (ARS)

6. MÉTODO

6.1 Participantes

La muestra utilizada para el cumplimiento de los 3 objetivos generales, los cuales corresponden a los 3 artículos científicos que constan en esta tesis doctoral, se basa en una muestra amplia de 6110 estudiantes de los programas de ingeniería, ciencias y tecnología de nivel superior evaluando a sus profesores de la Escuela Politécnica Nacional durante el curso académico 2016-17, con variantes dependiendo del estudio realizado conforme se describe a continuación:

Para el cumplimiento de los objetivos 1 y 2 de la tesis doctoral, los cuales corresponden a los artículos científicos primero y segundo respectivamente, se validó un instrumento de evaluación de la docencia de 33 ítems con 5 constructos y 5 opciones de respuesta: *Didáctica, Recursos, Metodología, Criterio de evaluación y Relación profesor-alumno*, utilizando el método del Análisis Factorial con extracción de componentes principales, para el primer estudio; y para el segundo estudio, se validó la escala de calificación de la enseñanza por parte de los estudiantes tanto corta como larga, se examinó la invarianza en relación con variables relevantes como el género, se utilizó análisis descriptivo, correlación interclase, análisis factorial confirmatorio, modelado de ecuaciones estructural exploratoria y correlaciones con el rendimiento académico de la clase tomada como unidad de análisis siguiendo un diseño multisección; aplicados para los dos estudios a una gran muestra de 6110 estudiantes.

Estos estudiantes estaban matriculados en 8 facultades y escuelas, cursando 24 titulaciones diferentes. El mayor porcentaje de estudiantes varones es representativo de la población de estudiantes de estudios politécnicos, en la que el 68.60% eran varones y el 31.40% mujeres. La edad media fue de 22.30 años. Asistieron a 1380 asignaturas diferentes que se distribuyeron en 1 812 grupos de clase. La muestra de maestros consistió en 670 maestros, quienes representaron una muestra variada en términos de edad, categoría y experiencia docente. Más de la mitad de estos profesores fueron masculino (62,80%). La aplicación de la escala de 33 ítems fue realizada al final del curso académico 2016-17, antes de que los estudiantes conocieran sus calificaciones finales. Todos los profesores fueron evaluados por los estudiantes en el mismo término. Todos los alumnos tuvieron que evaluar a los profesores para poder para acceder a sus calificaciones finales. La evaluación docente del alumno se realizó a través de una plataforma electrónica, obteniendo 19527 registros (matriz de datos original) en los que los datos fueron registrados (el mismo estudiante pudo evaluar varios profesores ya que cursó varias asignaturas).

Finalmente, el artículo 3 estudia el rendimiento académico de estudiantes de grupos vulnerables desde la perspectiva del Análisis de Redes Sociales (ARS), se estudio la información académica y de interacción de 45 estudiantes pertenecientes a grupos vulnerables que asistieron durante un periodo académico a un curso piloto de intervención socio-académica. Para la selección de los participantes del curso piloto de intervención socio-académica, se llevó a cabo una jornada de inducción en la que se informó a los estudiantes acerca de la posibilidad de participar en dicho programa durante un periodo académico previo al curso de nivelación en la Escuela Politécnica Nacional de Ecuador. Dado que el curso piloto de intervención socio-académica no era de carácter obligatorio y que, además, debía contar con un número de estudiantes similar al que conforma un paralelo ordinario del curso de nivelación, se seleccionó una muestra de 45 estudiantes de entre aquellos que aceptaron participar voluntariamente en el programa. En dicha muestra se procuró mantener las proporciones poblacionales concernientes al sexo, el tipo de curso de nivelación (Ingeniería, Ciencias y Ciencias Administrativas o Nivel Tecnológico Superior) y la provincia de procedencia. Asi; el 65% fueron hombres, el 69% se encuentra matriculado en carreras de ingeniería, ciencias y ciencias administrativas y el 31% en carreras de nivel tecnológico superior, el 65% provienen de la proviencia de Pichincha y el 35% corresponde al resto de provincias del país. La nota promedio sobre 10 puntos obtenida en la evaluación diagnóstica de Matemática-Geometría fue 4,52± 1,5842 y en Lenguaje y Comunicación fue de 5,12±1,1162. Asimismo, mediante análisis inferencial, se determinó que el promedio de las notas de los 45 estudiantes en cada una de las secciones de la prueba de diagnóstico no presentaba diferencias estadísticamente significativas con los correspondientes promedios poblacionales. Con esta información se construyó un sociograma, a partir del cual se determinaron las métricas de centralidad de ARS. Posteriormente, se estudiaron las relaciones entre dichas métricas y las variables académicas a través del análisis de correlación y regresión lineal con regularización LASSO. Como avance de resultados, se determinó que el rendimiento académico de los estudiantes del curso piloto estuvo influenciado, por un lado, por los conocimientos académicos previos al ingreso a la universidad, representados por la nota en la sección de Matemáticas y Geometría de la prueba de diagnóstico y, por otro lado, por la dinámica de la red social en la que se desenvolvieron en el aula, representada por la centralidad de eigenvector. Estos resultados tienen potencial explicativo significativo del rendimiento académico en función de las métricas de ARS y aportan evidencia para sustentar la implementación de prácticas que fomenten un entorno social saludable en el contexto académico.

6.2 Medidas

La escala original de evaluación del rendimiento docente se compone de 33 ítems con 5 constructos y 5 opciones de respuesta y solo una opción es correcta: *Didáctica (orientación en el aula), Recursos, Metodología, Criterio de evaluación y Relación profesor-alumno,* las mediciones de los resultados del análisis factorial de la muestra revelan la existencia de dos factores: el factor 1 explica el 70% de la variación en las puntuaciones de la escala y factor 2, 3.20%. El factor 1 está compuesto por los ítems 17 a 33 todos los ítems tienen altas saturaciones entre ellos con un factor (0.780 a 0.680). Dado que todos estos ítems se refieren a la relación profesor-alumno, este factor se puede llamarse *Relación profesor-alumno y establecimiento de un buen ambiente de aprendizaje*. El factor 2 se compone de los elementos 1 a 16. Todos los elementos tienen saturaciones altas o relaciones entre cada uno de ellos con un factor (0.761 a 0.670); estos elementos se refieren a lo que puede llamado *Planificación, dominio y claridad en la explicación del tema*. Para medir la confiabilidad se calculó el coeficiente de consistencia interno α de Cronbach; siendo $\alpha = 0.970$ la confiabilidad del factor 1 y para el factor 2, $\alpha = 0.950$. Una vez que se eliminan los ítems redundantes se redujo a aproximadamente 14 ítems (escala corta) sin pérdida de validez o confiabilidad $\alpha = 0.960$, y con prácticamente el mismo valor informativo que el original instrumento de evaluación.

Para el análisis con el rendimiento académico de los estudiantes se tomaron dos medidas: el rendimiento académico anterior y el rendimiento académico al final del semestre. Los logros acumulados anteriores fueron una medida del rendimiento académico medio alcanzado por los estudiantes en todas las asignaturas anteriores, que estaban matriculados hasta el inicio del semestre académico. Esta medida se obtuvo de registros administrativos computarizados, aunque estrictamente no puede considerarse una medida del desempeño previo en la disciplina en particular, puede verse como un indicativo del conocimiento o habilidad general con la que el estudiante comienza el estudio de la disciplina. La medida del rendimiento académico al final del semestre se obtuvo mediante las calificaciones otorgadas por el docente, en base a un examen final, consistente en exámenes escritos teóricos y prácticos. Estos exámenes finales en algunos casos fueron los mismos en todas las secciones y en otros fueron diferentes. Las diferentes secciones siguen el mismo programa y tienen los mismos criterios de evaluación. Estos criterios se especifican en el programa de estudios de cada curso. También existen reglas generales comunes para todos los exámenes en la Escuela Politécnica. Las medidas de rendimiento académico acumulado anterior y las calificaciones finales variaron de 0 a 40 para todos los cursos. La edad y el sexo de los estudiantes y de los profesores, se recopilaron de los registros administrativos.

6.3 Procedimiento

Los datos se obtuvieron de los registros informáticos existentes en la administración del Escuela Politécnica Nacional, y se otorgó permiso de acceso a los mismos al personal académico de la Institución. Los datos proporcionados por la institución eran anónimos, con un solo código de identificación para cada estudiante.

La aplicación de la escala de evaluación de la docencia por parte de los estudiantes se llevó a cabo al final del semestre, antes de conocer sus calificaciones finales. Todos los profesores fueron evaluados por los estudiantes en un período de tiempo similar. Todos los alumnos tuvieron que evaluar a los profesores para poder acceder a sus calificaciones finales. La evaluación de la docencia de los estudiantes se realizó a través de una plataforma electrónica en la que se grabado.

El impacto que tienen los procedimientos de las evaluaciones de la enseñanza de los estudiantes en una facultad sobre las tasas de respuesta ha sido analizado por varios autores en evaluaciones electrónicas especiales. Así, Young, Joines, Standish y Gallagher (2019) encontró que las evaluaciones realizadas por los estudiantes eran considerablemente más altas cuando el profesorado daba tiempo en clase a los estudiantes para completar la evaluación de la enseñanza, en comparación con un formulario electrónico emitido por la administración. Sin embargo, otros estudios sobre este tema no encontraron diferencias entre las evaluaciones realizadas con cuestionarios electrónicos y cuestionarios en papel y lápiz o cuando respondió una muestra más representativa en lugar de una muestra pequeña y sesgada (Nowell, Gale, & Kerkvliet, 2014).

6.4 Análisis de datos

6.4.1 Validez de la estructura de la escala

Con respecto a la validez de la estructura de la escala, se siguió las recomendaciones de Schmitt, Sass, Chappelle y Thompson (2018). Existen diferentes métodos para retener la "mejor" estructura factorial; por ejemplo, Análisis Factorial Exploratorio (EFA), Análisis Factorial Confirmatorio (CFA) o Modelo Exploratorio de Ecuaciones Estructurales (ESEM). EFA tiene la desventaja de la dificultad de replicar resultados con diferentes muestras, mientras que CFA conduce a cargas sesgadas y correlaciones entre factores porque requiere que las cargas cruzadas sean 0 en los factores no objetivo (Garn, Morin y Lonsdale, 2018). ESEM combina EFA y CFA, proporciona índices de bondad de ajuste y permite probar la invariancia de medición de múltiples grupos (Xiao, Liu y Hau, 2019). Schmitt y col. (2018) recomiendan usar EFA cuando no existe una teoría a priori, usar CFA cuando hay una teoría sólida y evidencia de la estructura de la escala, y usar ESEM cuando la teoría a priori es escasa. Howard, Gagné, Morin y Forest (2018) añaden que ESEM debe mantenerse sobre CFA cuando las correlaciones son diferentes entre estos dos métodos.

Para el análisis factorial, específicamente en estructuras multidimensionales, se utilizó los modelos bifactoriales (Morin, Katrin Arens y Marsh, 2016), los cuales se emplean para dividir la covarianza entre un factor global de segundo orden (es decir, el estilo de los profesores) y factores específicos (es decir, la metodología y los recursos o la relación profesor-alumno). Por lo tanto, con base en la información disponible, se probaron los siguientes modelos: un factor a través de CFA, cuatro factores a través de CFA, cuatro factores a través de ESEM y, cuatro y dos factores a través de ESEM. El método de estimación fue robusto de máxima verosimilitud porque los datos no seguían una distribución normal; además, como las respuestas no eran independientes, corregimos el estadístico χ^2 y los errores estándar mediante un estimador robusto (Muthen & Satorra, 1995; Muthén & Muthén, 2018). Todos los análisis se realizaron con Mplus 8.3.

6.4.2 Versión corta de la escala

Para elegir los ítems para una versión corta, tomamos en cuenta las cargas factoriales, corregidas por correlaciones ítem-test, confiabilidad y la significancia teórica del ítem (Marsh, Martin y Jackson, 2010). Para probar la concordancia de ambas versiones, nos basamos en la corrección de Levy de la correlación entre la escala corta y larga. Esta corrección explica la varianza del error compartido entre ambas formas debido al subconjunto de elementos (Barrett, 2015; Levy, 1967). Además, debido a que la correlación solo tiene en cuenta la monotonicidad entre ambas formas, también nos apoyamos en el índice de Gower (Barrett, 2012; Gower, 1971), cuyos valores oscilan entre 0 y 1, donde valores cercanos a 1 indican concordancia.

6.4.3 Invarianza de medición de género.

Para probar si los estudiantes masculinos y femeninos interpretan la escala de manera similar, realizamos una prueba de invariancia de medición (Vandenberg & Lance, 2000). Específicamente, comparamos tres modelos: configural, métrico y escalar (Muthén & Muthén, 2018). El modelo configuracional tiene cargas factoriales, intersecciones y varianzas residuales libres entre grupos y medias factoriales fijas en cero en todos los grupos. En el modelo métrico, las cargas factoriales se mantienen iguales en todos los grupos, mientras que las intersecciones y las varianzas residuales son libres entre los grupos y las medias de los factores se fijan en cero en todos los grupos. Finalmente, en el modelo escalar, las cargas de los factores y las intersecciones son iguales entre los grupos, mientras que las varianzas residuales son libres entre los grupos y las medias de los factores están restringidas a cero en un grupo y libres en el otro grupo. Para las comparaciones de modelos, se usó la diferencia en χ^2 y los cambios en el índice de ajuste comparativo (CFI) y en el error cuadrático de aproximación (RMSEA).

6.4.4 Fiabilidad

Para probar la confiabilidad de la forma corta y larga, no usamos el alfa de Cronbach porque cada vez hay más evidencia de su falta de precisión y la dificultad de cumplir con sus supuestos: el paralelismo y la equivalencia-tau de los ítems (McNeish, 2018; Zhang y Yuan, 2016). Cho (2016) propone diferentes fórmulas para estimar la confiabilidad cuando los ítems carecen de paralelismo, equivalencia-tau o ambos, no solo para estructuras unidimensionales sino también para estructuras multidimensionales. Se empelna diferentes indices de fiabilida como el indice clásico alpha de Cronbach y el índice omega, asi como la fiabilidad compuesta.

6.4.5 Validez de criterio: Relación con el rendimiento académico

Para analizar las relaciones entre las calificaciones de los estudiantes de la docencia y el rendimiento académico, los datos se tomaron individualmente y se agruparon en secciones. Inicialmente, la validez de las calificaciones de los estudiantes se podría evidenciar por la correlación entre SET y el rendimiento académico. Sin embargo, no se puede suponer que las calificaciones de los estudiantes constituyan una simple medida de la eficacia de la enseñanza porque cada grupo podría tener evaluaciones diferentes (Richardson, 2005). La evidencia clave citada en apoyo de las evaluaciones de la enseñanza por parte de los estudiantes como una medida de la efectividad instruccional de un maestro son los estudios multisección, en los que diferentes profesores enseñan la misma materia siguiendo el mismo esquema, y al final del semestre, todas las secciones tienen la misma examen o equivalentes (Cohen, 1981; Uttl, White, & González, 2017). Para encontrar la correlación entre las puntuaciones de la escala y el rendimiento académico, los datos se tomaron individualmente y se trataron como un estudio típico de varias secciones en el que se utilizó la clase promedio como unidad de análisis.

6.4.6 Relación SET con rendimiento académico de los estudiantes

Para el estudio del análisis de la relación entre la evaluación del desempeño docente SET con el rendimiento académico de los estudiantes, por un lado, se utilizó el grupo de clase promedio como unidad de análisis; por otro, se analizó los datos individuales de los estudiantes. Cuando se utilizó el promedio de clase-grupo como unidad de análisis, se realizó un análisis de correlación y un análisis de regresión jerárquica. El análisis de correlación se calculó con la técnica de correlación producto-momento de Pearson. El análisis de regresión lineal jerárquica múltiple incluyó, en el primer paso, logros académicos previos y, en el segundo paso, SET. Este enfoque metodológico establece la contribución específica de una variable, que entra en último lugar en el análisis, a la predicción de la variable dependiente, en este caso, el rendimiento académico al final del semestre. Además, se puede estimar la cantidad adicional de varianza que se tiene en cuenta en el rendimiento académico final por SET (Cohen y Cohen, 1983). Se realizó un análisis de ruta multinivel sobre los datos individuales, agrupados en secciones. Este análisis da cuenta en conjunto de la variabilidad entre estudiantes individuales dentro de los grupos de clase (nivel 1) y la variabilidad entre grupos, impartida por diferentes profesores (nivel 2). Se establece un análisis de trayectoria en el que se examina la influencia del rendimiento académico previo en el rendimiento académico final y en el SET, también se incluye la relación de SET con el rendimiento académico final. Se observaron todas las variables; no se definió ninguna variable latente.

El programa utilizado fue el modelado de ecuaciones estructurales (EQS) de Bentler (2005). La estimación de parámetros se realizó sobre la base de la máxima verosimilitud (ML); La estimación de ML se basa en las características de normalidad multivariante que se utilizan para producir estimaciones óptimas de los parámetros de la población y, por lo tanto, requiere tamaños de muestra relativamente grandes. Se utilizó una diversidad de índices de ajuste al evaluar el ajuste del modelo, incluidos chi-cuadrado, chi-cuadrado en relación con el grado de libertad, raíz cuadrada media estandarizada residual (SRMR), raíz del error cuadrático medio de aproximación (RMSEA) y la comparación índice de ajuste (CFI) (Hu y Bentler, 1999)

El análisis de datos agrupados, aunque puede considerarse más apropiado que el análisis de datos individuales (Cohen, 1981), plantea algunas cuestiones metodológicas importantes. Un análisis de grupos de clases que mezclan diferentes disciplinas de cursos o materias y secciones de los mismos cursos plantea interrogantes sobre la validez de los coeficientes de correlación estimados a partir de una combinación de datos heterogéneos (Hassler y Thadewald, 2003; Almeida-de-Macedo et al., 2013). El efecto de las varianzas-covarianzas heterogéneas en un conjunto de datos provoca estimaciones menos eficientes de los coeficientes de correlación de Pearson entre grupos comparado con el enfoque de combinar coeficientes de correlación de grupos individuales. Para superar esto, se realiza un análisis de datos agregados siguiendo el procedimiento de un diseño multisección, utilizando los datos de las disciplinas del curso con dos o más secciones. Para considerar el efecto de sesgo de una muestra pequeña, se utilizaron correlaciones ponderadas por tamaño simple.

6.4.7. Relación del rendimiento académico de los estudiantes desde la perspectiva del análisis de redes sociales

Se realizó una encuesta de interacción, con los resultados se construyó un sociograma para representar la red social del curso de intervención piloto. Debido a que se buscó estudiar la potencialidad de las interacciones entre los estudiantes y no su direccionalidad, se consideraron a todas las relaciones como simétricas (no dirigidas) y binarias (Newman, 2003).

A partir del sociograma, se determinó la densidad, el grado medio y las métricas de centralidad de grado (CG), cercanía (CC), intermediación (CI) y eigenvector (CE). Posteriormente, se integraron en una matriz las métricas de ARS con la información de rendimiento académico de los estudiantes. Los estudiantes a quienes les faltaba información de interacción o de rendimiento (por haber abandonado el curso piloto) fueron excluidos del conjunto de datos.

Se llevaron a cabo pruebas t y análisis de correlación entre las variables estudiadas y, finalmente, se aproximó una regresión lineal para determinar las variables que más influyeron en el rendimiento académico de los estudiantes al finalizar el curso piloto.

7. RESULTADOS. TRABAJOS PUBLICADOS O ACEPTADOS

7.1 Trabajos publicados o aceptados

1. Sánchez-Almeida, T. F., Sandoval-Palis, I. P., Gilar-Corbi, R., Castejón-Costa, J. L., & Salazar-Orellana, D. I. (2020). Teaching evaluation questionnaire validation at Escuela Politécnica Nacional, applying the method of Factor Analysis with extraction of principal components. *Ingeniería E Investigación*, *40*(1), 70-77. https://doi.org/10.15446/ing.investig.v40n1.79634
2. Sanchez, T.F., León, J., Gilar-Corbi, R., Castejón, J.L. (2021). Validation of a Short Scale for Student Evaluation of Teaching Ratings in a Polytechnic Higher Education Institution. *Front. Psychol*., Aceptado para su publicación el 24-5-2021. https://doi.org/10.3389/fpsyg.2021.635543
3. Sánchez T, Gilar-Corbi R, Castejón J-L, Vidal J and León J (2020) Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador. *Front. Psychol*. 11:233. https://doi.org/10.3389/fpsyg.2020.00233
4. Sánchez, T., Naranjo, D., Vidal, J., Salazar, D., Pérez, C., & Jaramillo, M. (2021). Analysis of academic performance based on sociograms: A case study with students from at-risk groups. *Journal of Technology and Science Education*, *11*(1), 167-179. https://doi.org/10.3926/jotse.1110

Universidad Nacional de Colombia (Sede Bogotá). Facultad de Ingeniería
ISSN: 0120-5609 (enlace)
ISSN: 2248-8723 (online)
DOI: 10.15446/ing.investig

Sánchez-Almeida, T. F., Sandoval-Palis, I. P., Gilar-Corbi, R., Castejón-Costa, J. L., & Salazar-Orellana, D. I. (2020). Teaching Evaluation Questionnaire Validation at Escuela Politécnica Nacional, Applying the Method of Factor Analysis with Extraction of Principal Components. *Ingeniería E Investigación*, *40*(1), 70-77. https://doi.org/10.15446/ing.investig.v40n1.79634

DOI: http://dx.doi.org/10.15446/ing.investig.v40n1.79634

Teaching Evaluation Questionnaire Validation at Escuela Politécnica Nacional, Applying the Method of Factor Analysis with Extraction of Principal Components

Validación del cuestionario de evaluación docente de la Escuela Politécnica Nacional, aplicando el método de Análisis Factorial con extracción de componentes principales

Tarquino Sánchez-Almeida[1], Iván Sandoval-Palis[2], Raquel Gilar-Corbi[3], Juan Castejón-Costa[4], and Diego Salazar-Orellana[5]

ABSTRACT

This work validates a teaching evaluation instrument applied to professors in engineering, sciences and higher technological level programs of the Escuela Politécnica Nacional, using the method of Factor Analysis with extraction of principal components. The database used for the research was previously examined and refined due to inconsistency, eg. outliers, out of range values, etc. The result of the method described above was a reduced survey of 15 items, which was obtained from an original study of 33 items. This new questionnaire clearly identifies the four main dimensions or aspects required: teaching development and planning, teacher-student relationship, evaluation, and a global assessment question. The reduction of the evaluation scale will allow to improve the process of integral teaching performance evaluation of the faculty at Escuela Politécnica Nacional, and this method could serve as a benchmark for the teaching evaluation process of other universities that belong to the higher education system of Ecuador.

Keywords: factor analysis, teaching evaluation, questionnaire validation, principal component analysis

RESUMEN

Este trabajo valida un instrumento de evaluación docente aplicado a profesores de las carreras de ingeniería, ciencias y de nivel tecnológico superior de la Escuela Politécnica Nacional, utilizando el método de Análisis Factorial con extracción de componentes principales. La base de datos utilizada en la investigación fue examinada previamente y refinada por inconsistencia - por ejemplo, valores atípicos, valores fuera de rango, etc. El resultado del método descrito anteriormente fue una encuesta reducida de 15 ítems, que se obtuvo de un estudio original de 33 ítems. Este nuevo cuestionario identifica claramente las cuatro dimensiones o aspectos: planificación y desarrollo de la docencia; relación profesor-alumno; evaluación; y una pregunta de valoración global. La reducción de la escala de evaluación permitirá mejorar el proceso de la evaluación integral del desempeño docente del personal de la Escuela Politécnica Nacional, y este método podría servir de referencia para el proceso de evaluación de la enseñanza de otras universidades que pertenecen al sistema de educación superior del Ecuador.

Palabras clave: análisis factorial, evaluación docente, validación de cuestionario, análisis de componentes principales

Received: May 10th, 2019
Accepted: March 3rd, 2020

Introduction

Over the years, innovation and the appearance of new lines of research have found use incorporating new areas of knowledge as instruments of academic training. That is where the application of instruction psychology is proposed as a new teaching staff tool. This has served as a guideline for the appearance of new research that seeks to holistically understand the teaching process within the methodology and the best alternative to transmit knowledge in the classroom.

How to cite: Sánchez-Almeida, T., Sandoval-Palis, I., Gilar-Corbi, R., Castejón-Costa, J., Salazar-Orellana, D. (2020). Teaching Evaluation Questionnaire Validation at Escuela Politécnica Nacional, Applying the Method of Factor Analysis with Extraction of Principal Components. *Ingeniería e Investigación*, *40*(1), 70-77. 10.15446/ing.investig.v40n1.79634

[1] Electronics and Telecommunications Engineer, Escuela Politécnica Nacional, Ecuador. MBA. Project Management, Escuela Politécnica Nacional, Ecuador. Affiliation: Full-Professor, Escuela Politécnica Nacional, Ecuador. E-mail: tarquino.sanchez@epn.edu.ec

[2] Electrical Engineer, Escuela Politécnica Nacional, Ecuador. MSc. University Pedagogy, Escuela Politécnica Nacional, Ecuador. Affiliation: Full-Professor, Escuela Politécnica Nacional, Ecuador. E-mail: ivan.sandoval@epn.edu.ec

[3] Degree in Psychopedagogy, Universidad de Alicante, España. Teaching Diploma, Universidad de Alicante, España. PhD. Design, Orientation and Psychopedagogical Intervention, Universidad de Alicante, España. Affiliation: Associate-Professor, Universidad de Alicante, España. E-mail: raquel.gilar@ua.es

[4] Psychology Bachelor, Universidad de Valencia, España. PhD. in Psychology Universidad de Valencia, España. Affiliation: Full-Professor, Universidad de Alicante, España. E-mail: jl.castejon@ua.es

[5] Mathematician, Escuela Politécnica Nacional, Ecuador. Affiliation: Escuela Politécnica Nacional, Ecuador. E-mail: diego.salazar@epn.edu.ec

Aparicio (2014) indicated that it is possible to interpret learning as the existing relationship between communication and interaction where interaction is seen as part of the teaching and academic development. Therefore, university professors require specific skills that allow them to enhance the quality of the teaching-learning process in the classroom. These competences enable them to achieve excellence in terms of results, which involves an evaluation culture and control of the learning process.

The normally used instruments to measure students' evaluation of their teachers, programs, and satisfaction with their instruction are known as standard rating scales. However, research on student evaluation of teaching ratings has not yet provided clear answers to some questions about their validity (Hornstein, 2017; Marsh, 2007 a,b; Spooren, Brockx, and Mortelmans, 2013; Uttl, White, and Gonzalez, 2017).

From a statistical perspective, there exist are records in Ecuador regarding the teaching performance evaluation in universities, and the existing limited evidence is of restricted nature. Nowadays, the "Ley Orgánica de Educación Superior", the law that governs the Ecuadorian educational system, establishes in the article 151 that teachers will submit to an integral periodic evaluation according to the program and teaching scale regulations of the professors and researchers of the Higher Education System and the statutory norms of each institution within it, in exercise of its responsible autonomy. The survey carried out by the students about their teachers will be considered as one of the evaluation parameters (Consejo de Educación Superior, 2018).

The current assessment instruments were designed considering the components established in the program and teaching scale regulations of the professors and researchers of the Higher Education System, such as self-assessment, co-evaluation, and hetero-evaluation. Some of the items are taken from other SET rating scales, like the SEEQ (Marsh, 2007a), STERS (Toland and De Ayala, 2005), and SET37 (Mortelmans and Spooren, 2009), and are adapted to the characteristics of the Escuela Politécnica Nacional. In general, the technical validation of the evaluation instrument is not considered as a criterion to guarantee the quality of the application of the instrument. The integral assessment of teacher performance is an essential component that allows a professor to enroll as Assistant Professor or Associate Professor. The requirements include a qualification of at least 75% of the score in the performance evaluation during his last two academic periods. Additionally, according to article 96 of the regulation (Consejo de Educación Superior, 2017), the academic staff will be dismissed if they have obtained:

1) an integral evaluation performance of less than 60% for two consecutive times, and

2) four integral performance evaluations of less than 60% throughout their career.

In addition, it establishes that the main titular teachers will be promoted to the next level if they comply with other requirements such as having obtained a score of at least 80% in the performance evaluation of their last two academic periods (Consejo de Educación Superior, 2017).

The proposed methodology arises as a necessity to validate the instruments to evaluate the teaching staff at the Escuela Politécnica Nacional of Ecuador. This validation is applied to teachers of engineering, sciences and higher technological level programs, using the method of factor analysis with extraction of major components. This research considers the reliability and validity requirements that questionnaires must have with Likert opinion rating scales (Alaminos and Castejón Costa, 2006).

The most used method to extract the initial factors of the matrix of correlation observed variables is the principal component method. It is characterized by an analysis of the total variance of the set of observed variables. The purpose is to discover the main components that define this set. Both factor analysis and principal component analysis are multivariable data reduction techniques.

The main metric characteristics to determine the accuracy of an evaluation instrument (questionnaire) are reliability and validity. Reliability is the property that designates the constancy and precision of the results obtained by an instrument when applied on different occasions. On the other hand, validity refers to whether the instrument can measure what it is intended to measure (Carvajal, Centeno, Watson, Martínez, and Sanz Rubiales, 2011). Reliability can be estimated by four means: internal consistency, stability, equivalence and inter-judge harmony. The method of choice is internal consistency, which uses the Cronbach Alpha (α) statistical test. The objective of this approach is to compare the variability of each item against the total variability of the instrument.

Currently, a line of work has been developed to reduce the length of scales already used or to elaborate new scales with a reduced number of items. The lack of time for their application , fatigue, and possible stereotyped responses in scales that are too long or that are part of a set that is applied within the same study, among others, has led to proposals of short scales (Gogol et al., 2014; Lafontaine et al., 2016). These scales have to be small enough to allow for a rapid assessment of purposed constructs, but large enough to ensure appropriate reliability, validity, and accurate parameter estimation.

The objectives of the present work are two: on one hand, to analyze the construct validity of the teaching-learning questionnaire, and on the other hand, to propose a reduction of that scale, conserving its psychometric properties.

Finally, the development of this research leads to improvements and appliance of new strategies for the teacher evaluation instrument. Additionally, these methods allow to identify the most relevant items and constructs. The result of this validation is the design of a questionnaire whose application brings accurate information that will improve the quality of the Higher Education System of Ecuador.

Mathematical Model

In factor analysis, a linear model is assumed:

$$X = \mu + LF + \varepsilon$$

where

$X(p \times 1)$ is the observable random vector, with mean vector μ and covariance matrix Σ;

$L(p \times m)$ is the matrix of factor loadings;

$F(m \times 1)$ are common factors, unobserved values of factors which describe major features of members of the population;

$\varepsilon(p \times 1)$ are error specific factors, measurement error and variation not accounted for by the common factors;

μ_i is the mean of variable I;

ε_i is the ith specific factor;

F_j is the jth common factor; and

L_{ij} is the loading of the ith variable on the jth factor.

We assume that the unobservable random vectors F and ε satisfy the following conditions:

F and ε are independent;

$E(\varepsilon) = 0$ and $Cov(\varepsilon) = \Psi$; where Ψ, is a diagonal matrix; and $E(F) = 0$ and $Cov(F) = 1$.

Thus, the factors are assumed to be uncorrelated. This is called the orthogonal factor mode.

Factoring, given the model: $X = \mu + LF + \varepsilon$

The implied covariance structure for X is,

$Cov(X) = LL' + \Psi$

If, $V(X_i) = l_{i1}^2 + \cdots + l_{im}^2 + \Psi_i$ and;

$Cov(X_{(i)}, \ldots, X_{(k)}) = l_{i1}^2 l_{k1}^2 + \cdots + l_{(im)}^2 l_{km}^2$

Furthermore, $Cov(X, F) = L$ so that $Cov(X_{(i)}, \ldots, F_{(j)}) = l_{(ij)}$

The portion of variance of the ith variable that is explained by the m common factors is called the communuality of the ith variable: $\sigma_{ii} = h_i^2 + \Psi_i$ where σ_{ii} is the variance of X_i, i.e., the ith diagonal of Σ; $h_i^2 = (LL')_{ii} = l_{i1}^2 + \cdots + l_{(i2)}^2 + l_{(im)}^2$ is the communality of X_i; and Ψ_i is the specific variance or uniqueness of X_i.

Note that the communality h_i^2 is the sum of squared loadings for X_i (Harman, 1968).

In this case, thirty-three items or quantitative variables are presented, so the factor analysis technique is applied with the extraction method of main components to obtain two dummy variables that allow to relate and summarize the teaching staff survey. This allows to evaluate the relevant aspects of the teacher, within the teaching-learning process.

Methodology

Analysis of the original information regarding its relevance and validity

An exploratory analysis is made of the data obtained from the application of the evaluation instruments of 33 items with 5 answer choices (see Table 1), which were carried out by 6 110 students of the engineering, science and higher level technological programs for the professors of the Escuela Politécnica Nacional. These students were enrolled in 8 faculties and schools, studying 24 different degrees. The higher percentage of male students is representative of the population of students of polytechnic studies, in which 68,60% were male and 31,40% were female. The average age was 22,30 years old. These 6 110 students attended 1 380 different subjects which were distributed into 1 812 class-groups. The teacher sample consisted of 670 teachers, who represented a varied sample in terms of age, category, and teaching experience. More than half of these teachers were male (62,80%). The application of the scale of 33 items was carried out at the end of semester 2017-A (October 2017-March 2018), before the students knew their final grades. All teachers were evaluated by the students in the same term. All students had to evaluate the teachers to be able to access their final grades. The student teaching evaluation was conducted through an electronic platform, obtaining 19 527 records (original data matrix) in which the data were recorded (the same student was able to evaluate several professors since he/she took several subjects).

From the **original data matrix**, a correlation matrix is elaborated between all the considered variables (items). Several tests are carried out to determine if it is pertinent, from a statistical point of view, to carry out factor analysis with the information available from the correlation matrix. The main tests are:

The Bartlett sphericity test: it is based on chi-square distribution, where high values lead to rejecting the null hypothesis (H_0) that states that the variables are not correlated within the population. Thus, Bartlett's test of sphericity determines whether the correlation matrix is an identity matrix, which would indicate that the factorial model is inadequate. If the significance value (p-value) is less than 0,050, we reject the null hypothesis (H_0) and continue with the factor analysis.

The Kaiser-Meyer-Olkin Index (*KMO*): it allows the comparison between the magnitude of the observed correlation coefficients and the magnitude of the partial correlation coefficients. The KMO statistic varies between 0 and 1. Those less than 0,500 indicate that factor analysis is not required for the data in question.

The partial correlation coefficient: it describes the linear relationship between two variables while controlling the effects of one or more additional variables. These coefficients should tend to zero, when they are lent for factor analysis (Montoya O. 2005).

Table 1. Evaluation instrument of 33 items in 5 constructs

N°	QUESTION	N°	ANSWER
I	**DIDACTICS**	**IV**	**EVALUATION CRITERIA**
1	Did the teacher clearly explain the objectives and themes, indicating their interrelation and contribution to professional profile?	17	Has the teacher used objective methods to evaluate students?
2	Did the teacher select class activities appropriately, depending on the objectives?	18	Has the evaluation been used to reorient student learning?
3	Has the teacher been clear in his/her explanations and presentations?	19	Has the teacher considered aspects that are not merely cognitive?
4	Has the teacher related theoretical fundamental concepts and principles with practice?	20	Does the teacher evaluate fairly and impartially?
5	Does the teacher solve the difficulties that arise?	21	Has the minimum grade to approve the course been explained and why?
6	Does the teacher show the mastery of the subject?	22	Were the objectives defined in a clear and concise form?
7	Does the lecturer demonstrate planning his/her lectures before the class presentations?	23	Are the evaluation events related to the teaching taught?
8	Is the teacher creative and dynamic in the classroom?	**V**	**TEACHER-STUDENT RELATIONSHIP**
9	Does the professor show that he/she is up-to-date in the subject that is imparted?	24	Did the teacher ascertain that the students understand what they were being taught?
II	**RESOURCES**	25	Did the teacher encourage the initiatives coming from the students?
10	Does the teacher prepare didactic material additional to the textbook and makes it known?	26	Did the teacher create an environment of participation?
11	Does he/she organize didactic experiences such as visits, excursions, projects, discussions?	27	Did the teacher maintain a cordial relationship with the entire group of students?
12	Has the complementary, recommended or used material been interesting?	28	Did the teacher create an environment of trust and work during class?
13	Does he/she use means that benefit the learning process?	29	Has the teacher motivated students and increased their interest about the subject?
III	**METHODOLOGY**	30	Does the teacher have an attitude of availability outside the class?
14	Did the teacher use different teaching methods properly?	31	Has the teacher openly accepted the suggestions made by students?
15	Has the teacher used a varied methodology?	32	Was the teacher attentive with the evolution of the students?
16	Has the teacher explained the methodologies for evaluating the course?	33	Excluding limitations that are not due to the teacher, could he/she be considered as a good teacher?

Source: Author

Extraction of Main Components

Interpretation of the main components is often difficult, so the initial extraction is rotated to achieve a solution that facilitates it. Varimax with Kaiser Normalization (Kaiser, 1958) is the rotation method that uses the orthogonal rotation of factors previously normalized. In other words, it maintains the independence between the rotated factors. This method achieves that each rotated component presents correlations with only a few variables. Therefore, this method minimizes the number of variables with high loads by one factor and is adequate when the number of components is reduced.

Results

Statistical analysis of teacher evaluation instruments

Bartlett's sphericity test was applied before using the multivariate factor analysis technique in order to verify if the correlation matrix is an identity matrix, which means that the correlations between the variables are zeros. The test consists of an estimation of the chi-square indicator, where high values lead to rejecting the null hypothesis. The test must have a significance value lower than the 0,050 limit, which would indicate that the variables are not correlated within the population. Table 3 shows the result of the Bartlett's Sphericity test that is 0,000. This demonstrates that the null hypothesis is rejected. Therefore, factor analysis is applicable in this case.

The analysis tool that was used was the Kaiser-Meyer-Olkin test (KMO). It is an index that compares the magnitude of the correlation coefficients observed with the magnitude of the partial correlation coefficients, eliminating the effect of the remaining variables included in the analysis. Since the partial correlation between two variables must be small when the factorial model is adequate, the denominator must increase a little compared to the magnitude of the correlation coefficients observed if the data corresponds to a factorial structure, in which case KMO will have a value close to 1.

Table 3 shows the result of the KMO test using the SPSS statistical analysis software, which has a value of 0,990, very close to the unit and therefore fulfills the requirement.

Table 2. KMO and Bartlett Tests

KMO and Bartlett tests		
Kaiser-Meyer-Olkin measure of sampling adequacy		0,990
Bartlett's sphericity test	Chi-square Aprox.	338 959,305
	gl	528
	Sig.	0,000

Source: Authors (data analysis performed by SPSS v.22)

The partial conclusion that can be reached about this first part is that the two types of analysis on the pertinence and validity of the data matrix are satisfactorily verified.

Now, we proceed with the second part, which consists of extracting the principal components by grouping the 33 items or original variables into new variables called "factors". It is based on an exploratory analysis and shows that there is a large number of stereotyped responses, defined as those in which students respond with a single type of score along the whole scale, be it 1, 2, 3, 4 or 5. The data from these students is eliminated and, finally, the number of records on which the analysis is based is 15 771.

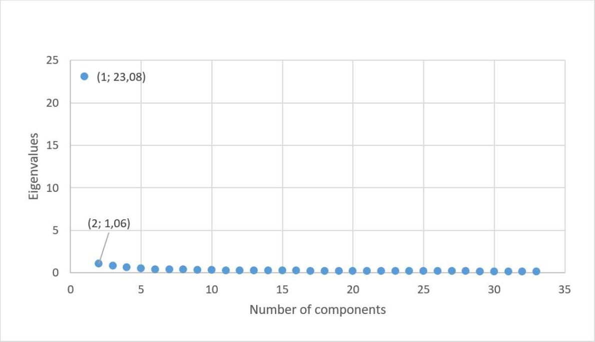


Figure 1. Sedimentation chart
Source: Authors

The results of the factor analysis of the sample reveal the existence of two factors, dimensions, or different constructs, as can be seen in the sedimentation chart in Figure I. These are chosen when the components have eigenvalues greater than 1.

The total variance explained in Table 3 analyzes in detail the selection of the two components, factors, or constructs: factor 1 explains 70% of the variation in the scores of the scale, and factor 2, 3,20%. Only the first two factors have eigenvalues greater than 1 and explain 73,20% of the original problem, resulting in a loss of 26,80% of the original information due to the fact that the survey has a very high number of items, among other aspects.

Factor 1 is composed of items 17 to 33. All the items have high saturations between them and the factor (0,780 to 0,680). This is related to the factor or set of items. The items with higher load or saturations are, in this order, items 32, 28, 31, 30, 25, 26, 24, 29, 27, 33, 22, and 20, to the less representative 23, 18, 19, 21 and 17.

Given that all these items refer to the teacher-student relationship, this factor can be called *Teacher-Student Relationship* and *establishment of a good learning environment*.

Extraction Method: Principal Component Analysis Rotation Method: Varimax with Kaiser Normalization

Factor 2 is composed of items 1 to 16. All the items have saturations or high relationships between each one of them and the factor (0,761 to 0,670); the items with higher loads or saturations are, in that order, the items 3, 4, 5, 2, 6, 7, 1, 9, and 8. In addition, these items refer to what may be called *Planning, mastery, and clarity in the explanation of the subject*.

Given that Factor 1, displays a greater variance percentage than Factor 2, this indicates that the students of the Escuela Politécnica Nacional give the greatest importance to the teacher-student relationship, or, in other words, perform the assessment of the teacher depending on the quality of this relationship, to a greater extent than the aspect of *Planning, mastery and clarity in the explanation of the subject*.

Table 3. Factor loadings and total variance explained

Rotated Component Matrix	Loadings	
	Factor 1	Factor 2
Item32	**0,782**	0,407
Item28	**0,780**	0,398
Item31	**0,789**	0,405
Item30	**0,787**	0,407
Item25	**0,784**	0,414
Item26	**0,770**	0,405
Item24	**0,767**	0,427
Item29	**0,759**	0,442
Item27	**0,759**	0,467
Item33	**0,757**	0,427
Item22	**0,727**	0,494
Item20	**0,723**	0,470
Item23	**0,723**	0,490
Item18	**0,722**	0,494
Item19	**0,712**	0,455
Item21	**0,705**	0,483
Item17	**0,689**	0,507
Item15	0,643	**0,697**
Item14	0,643	**0,695**
Item16	0,622	**0,690**
Item13	0,618	**0,686**
Item12	0,597	**0,680**
Item11	0,525	**0,673**
Item2	0,399	**0,759**
Item4	0,430	**0,760**
Item3	0,415	**0,761**
Item7	0,442	**0,756**
Item5	0,453	**0,760**
Item1	0,369	**0,755**
Item6	0,417	**0,758**
Item9	0,454	**0,747**
Item8	0,506	**0,699**
Item10	0,547	**0,670**
Initial eigenvalues	23,100	1,100
% of Total Variance	70	3,200
Total Variance		73,20%
Sum of charges to the square of the extraction	23,100	1,100
% of Total Variance	70	3,200
Total Variance		73,20%
Sum of charges to the square of the rotation	13,300	10,800
% of Total Variance	40,400	32,800
Total Variance		73,20%

Source: Authors

Another requirement that any questionnaire or rating scale must meet is reliability. If all the items amount to or contribute to measure the same, the reliability will be high. As indicated above, the most used statistical tool to calculate reliability is Cronbach's Alpha (α) internal consistency coefficient. It evinces an adequate reliability when α values range from 0,650 to high values such as 0,800 and above.

To do this, the reliability of each of the factors obtained in the factorial analysis was calculated using Cronbach's α internal consistency coefficient; being α = 0,970 the reliability of Factor 1, and Factor 2, α = 0,950. I was very high in both cases.

Given that it is possible that both factors or aspects are related, the total reliability of the 33 item scale, that amounted to α = 0,980, was obtained. This implies that a total score of the scale can be obtained, as well as scores for each of the previous factors or sub-scales.

Once it is confirmed that the reliability of each of the sub-scales is very high, it is possible to determine which item contributes more to the reliability of the scale and which items are redundant. Moreover, these can be eliminated without decreasing the reliability of the scale.

The sub-scale/*Factor 1*, composed of 17 items, has a reliability α = 0,970. If redundant items are eliminated, this scale can be reduced to 6 items: 26, 28, 30, 31, 32, and 33, with reliability α = 0,950, which is still very high.

The sub-scale/*Factor 2*, composed of 16 items has a reliability α = 0,950. If redundant items are eliminated, this scale can be reduced to 5 items: 3, 4, 5, 6, and 7, with reliability α = 0,930, which remains considerably high.

Thus, the 33-item questionnaire can be reduced to about 11 items without loss of validity or reliability α = 0,960, and with practically the same informative value as the original evaluation instrument . These items would be, as indicated in Table 4:

Factor 1/scale 1: Items 26, 28, 30, 31, 32, and 33. 27 and 29 could also be included in this order, with reliability α = 0,960.

Factor 2/scale 2: Items 3, 4, 5, 6, and 7. We could also include 9, with α = 0,940 and 2, with α = 0,940 in this order.

Table 4. Reduced questionnaire of 11 items

Factor	Items	(∝) Cronbach	Additional Items	(∝) Cronbach
1	26, 28, 30, 31, 32, 33	0,979	27,290	0,961
2	3, 4, 5, 6, 7	0,959	9,200	0,945, 0,944

Source: Authors

Then, the question that arises is: What happens with the other items and with the other theoretical aspects included in the scale as *Resources, Methodology and Evaluation*?

The answer explains that they contribute very little to the assessment of the teaching staff, given what the selected items of the reduced scale do.

However, the dimensions or aspects related to *Resources, Methodology and Evaluation* are important enough to be included in the scale, for which it is necessary to incorporate items that better represent these dimensions than the previous scale of 33 questions.

Therefore, a factor analysis was carried out to determine the extent to which the dimensions or aspects of *Resources, Methodology and Evaluation* influence the results, thus forcing the appearance of four factors, out of which two are new: *Evaluation, Methodology-Resources* and –two factors that had previously appeared– *Teacher-student relationship, and Planning, mastery and clarity in the subject's explanation*. Table 5 summarizes the variance parameters, factor loadings, as well as the Cronbach's Alpha internal consistency coefficients (α) for each aspect.

Table 5. Factorial analysis forced to 4 factors

Factor	Variance	Items	Factor loadings	(∝) Cronbach	Appearance
1	70,98%	24-33	0,720-0,620	0,970	Teacher-student relationship
2	3,72%	1-9	0,730-0,580	0,960	Planning, mastery and clarity in the subject's explanation
3	2,85%	16-23	0,680-0,460	0,950	Evaluation
4	2,28%	10-15	0,770-0,600	0,950	Methodology-resources

Source: Authors

Proposal of a reduced scale

Based on the information displayed in section A, it is considered convenient to better define the items on the *Evaluation* and *Methodology-Resources* aspects. To do this, 15 items are proposed, since they commonly appear in most universities (Alaminos and Castejón, 2006). This scale could include the most effective items of the original questionnaire, along with some new items introduced from other questionnaires, based on the theoretical dimensions of the aspects that are to be measured (Casero, 2008).

The analysis of the data obtained is structured in four aspects or dimensions: *Teaching Development and Planning*, *Teacher-Student Relationship*, *Evaluation* and a *Global Assessment question*, as indicated in Table 6, which shows each question with subscripts that express the following information:

1 = combined items of the aspects in the original questionnaire.

2 = relevant items of the original questionnaire.

3 = New items included

The analysis of the proposed reduced scale observed in Table 6 indicates that two of the included items, related to the *Teacher-student relationship,* were the same ones as in the original questionnaire because they provide relevant information about the evaluation to the professor. In addition to this table, there are items, such as 7, 12, and 15 that

Table 6. Reduced scale proposal

Aspect	N	Question
Planning and Development of Teaching	1	Does the teacher present and explain at the beginning of the period the contents (syllabus), methodologies and teaching activities, evaluation system, presentation of works, etc.?[1,2]
	2	Does the teacher demonstrate that his classes are based on the learning objectives and the syllabus of the subject?[1]
	3	Does the teacher demonstrate that he prepares and plans his classes (activities, methodologies, resources, evaluation, etc.)?[1]
	4	Does the teacher demonstrate mastery of the topics discussed in class?[1,2]
	5	Is the teacher clear in his expositions and explanations? Are the taught topics understood?[1,2]
	6	Does the teacher meet the established schedule?[1]
	7	Does the teacher use different didactic resources (for example, books, posters, maps, photos, slides, articles, videos, software, etc.) as support for the teaching of the subject? The resources and materials used or recommended are useful to take the course (bibliography, slides, virtual campus material, etc.)?[3]
	8	Does the methodology used by the teacher facilitate the learning of the subject and encourage interest in it?[1]
Teacher–Student Relationship	9	Has the teacher created an environment of trust and work in class?[2]
	10	Is the teacher accessible and willing to attend out of class consultations?[1,2]
	11	Has the teacher been made suggestions that he openly accepted? Has the teacher created an atmosphere of class participation?[2]
Evaluation	12	Are the evaluation events related to the topics presented in the course? Is the evaluation adjusted to the contents studied during the course?[3]
	13	Does the professor respect the weighing established by the institution that no evaluation must exceed 40% of the total score?[1]
	14	Does the teacher comply with the review of tests and/or previous exams the record of grades?[1]
	15	Does the teacher make the correction of the evaluations to the students? In the exams and work we the students have the possibility to know the mistakes made and comment on their valuation?[3]
Global Valuation	15/16	Excluding limitations that are not due to the teacher, can he/she be considered a good teacher?[2]

Source: Author

include two alternatives, which coincide in their meaning, but are expressed differently. Therefore, for those questions, it is necessary to choose between the options when applying the new survey.

From the results, two possible options are proposed for item 11, which are items 31 and 26 of the original questionnaires, offering the possibility of choosing between one or the other, since both alternatives have practically the same importance.

Finally, regarding item 16, which is related to the *Global Assessment* is included in the questionnaire to evaluate the general performance of the professor. However, as it not a relevant aspect, it can be considered as a replacement to item 15.

Discussion and conclusions

The first objectives of the present work were to analyze the construct validity of the teaching-learning questionnaire. Factor analysis revealed that the scale was composed of two factors. However, when factor analysis was forced to 4 factors, the theoretical structure of the initial questionnaire was exactly reproduced.

The second objective of the present work was to propose a reduction of the teaching evaluation questionnaire. It is difficult to reduce a questionnaire while maintaining the fundamental aspects of teaching. However, if the objective is to reduce the questionnaire even further to condense it to 13 items, for example, it is recommended to eliminate item 2, which covers the *Planning* aspect, as well as items 14 and 15 that refer to the grading methodology. In addition, it would be optimal to eliminate item 16. These changes are proposed while taking into that the questionnaire would maintain the desired margin of reliability.

For the validation of the reduced questionnaire proposed in Table 6, the data obtained from a large sample would be subjected to the same analysis, along with other techniques such as Confirmatory Factor Analysis and Item Response Theory Analysis (TRI).

The items with the highest saturations are those that best define the factor, while the items with low saturations define the factor less accurately. Based on this, for the original questionnaire of 33 items, Factor 1 has a high level of saturation -within the range of 0,780 to 0,689- and determines a positive teacher-student relationship, as well as a good learning environment. Similarly, Factor 2 has a high saturation level and describes the planning, mastery and clarity in the explanation of the subject, leaving the rest with low saturation levels.

Based on reliability tests with Cronbach internal consistency coefficients (α) and Bartlett's sphericity test, it is concluded that the two types of analysis about the relevance and validity of the matrix data are satisfactorily verified, which means that the original matrix data is reliable. In addition, all the questions have relevant information for the analysis of communalities.

The results obtained satisfy all the objectives established in this research paper and offer a proposal for a tool used for student evaluation of the university's teaching staff, based on the opinions of lecturers and students. The contribution that this work aims is to do is to present an available instrument to be used by universities and polytechnic schools, especially at the Escuela Politécnica Nacional, to validate and reduce the teaching evaluation questionnaires. The positive results of this study confirm it is possible to enter to a new phase for teaching evaluation using a new and well-defined survey.

A limitation of the study is that the assumption of randomness for factor analysis was not followed, because the questions are not arranged in a random order. On the other hand, another limitation is that the construct validity was examined but not the criterion of validity, for example, correlating the questionnaire scores to some external criterion.

In addition to the validation analysis of the teacher evaluation instrument, it is recommended to carry out a multidimensional analysis including aspects of gender, academic record, admission examination score, subjects, degrees, among others, in order to relate the scores in the scales to other variables and their correlations.

Acknowledgements

We would like to thank the Escuela Politécnica Nacional and the National Secretariat for Higher Education, Science, and Technology, who financed the realization of the network research project PIC-18-INE-EPN-002. This paper is part of Tarquino Sánchez's Ph.D. thesis.

References

Alaminos, A. and Castejón, J. L. (2006). *Elaboración, análisis e interpretación de encuestas, cuestionarios y escalas de opinión*. Alicante: Universidad de Alicante. http://hdl.handle.net/10045/20331

Aparicio, E. (2014). *Validación de un cuestionario de evaluación de la docencia universitaria*. (Doctoral thesis, Universidad de Alicante, Alicante, Spain). http://hdl.handle.net/10045/45168

Carvajal, A., Centeno, C., Watson, R., Martínez, M., and Sanz Rubiales, Á. (2011). ¿Cómo validar un instrumento de medida de la salud?. *Anales del Sistema Sanitario de Navarra 34*(1), 63-72. 10.4321/S1137-66272011000100007

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

Consejo de Educación Superior (2017). *Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior*. https://n9.cl/4scm

Consejo de Educación Superior (2018). *Ley Organica De Educacion Superior*, LOES. http://www.ces.gob.ec/documentos/Normativa/LOES.pdf

Gogol, K, Bunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischhach, A., and Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology 39*, 188-205. 10.1016/j.cedpsych.2014.04.002

Harman, H. (1968) *Modern Factor Analysis*, (2nd Ed.) Chicago: The University of Chicago Press, Ltd. https://archive.org/details/ModernFactorAnalysis/mode/2up

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

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187-200. 10.1007/BF02289233

Lafontaine, M.-F., Brassard, A., Lussier, Y., Valois, P., Shaver, P. R., and Johnson, S. M. (2016). Selecting the best items for a short-form of the experiencies in close relationships questionnaire. *European Journal of Psychological Assessment 32*(1), 140-154. 10.1027/1015-5759/a000243

Marsh, H. W. (2007a). *Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness*. In R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319-383). New York: Springer. 10.1007/1-4020-5742-3_9

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology 99*, 775-790. 10.1037/0022-0663.99.4.775

Montoya, O. (2015). Aplicación del análisis factorial a la investigación de mercados. *Scientia Et Technica Scientia et Technica*, Año XIII 35, 281-286.

Mortelmans, D. and Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies* - EDUC STUD. 35. 547-552.

Mortelmans, D. and Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies* - EDUC STUD, 35, 547-552. 10.1080/03055690902880299

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research 83*(4), 1-45. 10.3102/0034654313496870

Toland, M. D., and De Ayala, R. J. (2005). A Multilevel Factor Analysis of Students' Evaluations of Teaching. *Educational and Psychological Measurement 65*(1), 272–296. 10.1177/0013164404268667

Uttl, B., White, C. A., and 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. 10.1016/j.stueduc.2016.08.007

frontiers in Psychology | Educational Psychology

Sánchez, T.F., León, J., Gilar-Corbi, R., Castejón, J.L. (2021). Validation of a Short Scale for Student Evaluation of Teaching Ratings in a Polytechnic Higher Education Institution. *Front. Psychol*., Aceptado para su publicación el 24-5-2021 https://doi.org/10.3389/fpsyg.2021.635543

Validation of a Short Scale for Student Evaluation of Teaching Ratings in a Polytechnic Higher Education Institution

Tarquino F. Sanchez[1], Jaime Leon[2]*, Raquel Gilar-Corbi[3], Juan L. Castejón[3]

[1]College of Science and Engineering, Escuela Politécnica Nacional, Ecuador, [2]University of Las Palmas de Gran Canaria, Spain, [3]University of Alicante, Spain

Submitted to Journal:

Frontiers in Psychology

Specialty Section:

Educational Psychology

Article type:

Original Research Article

Manuscript ID:

635543

Received on:

30 Nov 2020

Revised on:

10 May 2021

Accepted on:

24 May 2021

Journal website link:

www.frontiersin.org

Article

Validation of a Short Scale for Student Evaluation of Teaching Ratings in a Polytechnic Higher Education Institution

Tarquino Sánchez

Escuela Politécnica Nacional, Ecuador

Departamento de Electrónica, Telecomunicaciones y Redes de la Información

Ladrón de Guevera E11-253, Quito 170517, Ecuador

[PO·Box 17-01-2759]

Telephone: +593 02-2976300 Ext 2251Quito · Ecuador

Mail: tarquino.sanchez@epn.edu.ec

Jaime León (Corresponding author)

Universidad de Las Palmas de Gran Canaria

Facultad de Formación del Profesorado

Juana de Arco, 1

35004 Las Palmas de Gran Canaria (Islas Canarias, España)

Telephone: (+34) 928451761

Mail: jaime.leon@ulpgc.es

Raquel Gilar-Corbi

Department of Developmental Psychology and Didactics. University of Alicante. Campus San Vicente del Raspeig. Ap. 99 E-03080 Alicante (Spain)

e-mail: raquel.gilar@ua.es

Telephone: +34965902911; Fax: +34965903860

Juan-Luis Castejón

Department of Developmental Psychology and Didactics. University of Alicante. Carretera de San Vicente del Raspeig. Ap. 99 E-03080 Alicante (Spain)

Tel: +34965903861

E-mail: jl.castejon@ua.es

Received: date; Accepted: date; Published: date

Abstract: The general purpose of this work is twofold, to validate scales and to present the methodological procedure to reduce these scales to validate a rating scale for the student evaluation of teaching in the context of a Polytechnic Higher Education Institution. We explored the relationship between the long and short versions of the scale; examine their invariance in relation to relevant variables such as gender. Data were obtained from a sample of 6110 students enrolled in a polytechnic higher education institution, most of whom were male. Data analysis included descriptive analysis, intraclass correlation, exploratory structural equation modelling (ESEM), confirmatory factorial analysis, correlations between the short and long form corrected for the shared error variance, gender measurement invariance, reliability using congeneric correlated factors, and correlations with academic achievement for the class as unit with an analysis following a multisection design. Results showed four highly correlated factors that do not exclude a general factor, with an excellent fit to data; configural, metric, and scalar gender measurement invariance; high reliability for both the long and short scale and subscales; high short and long-form scale correlations; and moderate but significant correlations between the long and short versions of the scales with academic performance, with individual and aggregate data collected from classes or sections. To conclude, this work shows the possibility of developing student evaluation of teaching scales with a short form scale, which maintains the same high reliability and validity indexes as the longer scale.

Keywords: criterion validity, reliability, scale validation, short scale development, construct validity, student evaluation of teaching.

1. Introduction

The academic failure and dropout rates in higher education in Ecuador, especially in Engineering studies, are very high. Sandoval, Naranjo, Vidal and Gilar (2020) find a dropout rate in the first year of university studies at the National Polytechnic School of around 70%. Braxton et al. (2000) and Kuh (2001) point out the quality of teaching as one of the determining aspects of academic failure and dropout. Likewise, instructional factors are one of the key factors in explaining academic success and dropout. Schneider and Preckel (2017) highlights the effect on academic readiness of the teacher-student interaction, the type of communication, the preparation, organization, and presentation of content by the teacher, the teacher's planning, and the feedback provided to the student, are some of the aspects.

Student evaluation of teaching (SET) ratings is a generalized procedure in the institutions of higher education (Huybers, 2014; Richardson, 2005; Zabaleta, 2007). SET is a useful tool for formative aims, such as, feedback for the improvement of instruction, and for administrative decision-making about recruitment, career progress or economic incentives (Linse, 2017). A systematic review on the subject shows that there are very few publications on the validation of student evaluation of university teaching scales -SET- in South America, collected in the most important databases such as Scopus and WoS -Web of Science- (Andrade -Abarca, Ramón-Jaramillo, and Loaiza-Aguirre, 2018; Pimienta, 2014), and some more when the scope of the search is expanded (Fernández and Coppola, 2008; Montoya, Arbesú, Contreras, and Conzuelo, 2014).

In the Ecuadorian context, there are the works of Aguilar and Bautista (2015) and Andrade-Abarca et al. (2018), who validate questionnaires in the field of an Ecuadorian polytechnic university. While in the review by Loor, Gallegos, Intriago, Guillén (2018) on the evaluation of university teaching staff, the need to improve the quality of the evaluation process is concluded.

Student Evaluation of Teaching Ratings Scales

The instruments normally used to measure students' evaluation of their teachers, programs, and students' satisfaction with their instruction are known as standard rating scales. However, research on student evaluation of teaching ratings has not yet provided clear answers to some questions about their validity (Hornstein, 2017; Marsh, 2007a, b; Spooren, Brockx, & Mortelmans, 2013; Uttl, White, & Gonzalez, 2017).

Many evaluation instruments have been constructed and validated within the home institution itself, and the results of such validation have not always been published, and in some instances they have not even been tested for psychometric quality (Richardson, 2005). In addition, there is a lack of consensus on the number and type of dimensions (Spooren et al., 2013), due to conceptual problems related to the lack a theoretical framework about what effective teaching is, and methodological problems concerning the measurement of these dimensions as a data-driven process (in which different post hoc analytic techniques are used). It seems necessary to use the most common dimensions, which are associated with greater teaching effectiveness.

A question concerning construct validity that arises in relation to student evaluation of teaching rating scales is whether it has a one-dimensional (Abrami, d'Apollonia, and Rosenfield, 1997), Cheung, 2000) or multidimensional structure. Marsh et al. (2009) defended the application of exploratory structural equation modelling (ESEM) methods integrating confirmatory (CFA) and exploratory factor analyses (EFA) to analyse issues related to multidimensional student evaluations of university teaching (SETs), on the basis of the measures that can be obtained both of the specific dimensions and a general factor of the quality of teaching.

An open and controversial question related to the criterion validity is the relationship of SET scores to student academic achievement. To answer this question, a series of revision and meta-analytical studies have been carried out (Cohen, 1981; Feldman, 1989; Clayson, 2009; Uttl, White, & Gonzalez, 2017). Taken together, the results regarding the relation between SET and academic performance, when multiple sections are included and the previous academic achievement is controlled, show that SET is moderately related to academic achievement; however, the effect of SET on academic performance is smaller than that found in some previous meta-analytic studies (Cohen, 1981; Feldman, 1989), at around only 10%.

Another methodological question concerns evaluation systemactic-bias. This problem is present when a confirmed characteristic of students habitually influences their evaluations of teachers (e.g. gender) (Badri, Abdulla, Kamali, & Dodeen, 2006; Basow, Phelan, & Capotosto, 2006; Boring, 2015; Centra & Gaubatz, 2000; Darby, 2006). A possible source of bias is the discipline. If the evaluation of teaching is situational and is affected by academic disciplines, being higher in studies in the field of education and the liberal arts and less in other areas such as business and engineering (Clayson, 2009), it seems necessary to carry out new studies in areas different from the previous ones, such as the technical areas where there are fewer studies on the subject.

The present study was carried out in a different context to most previous studies (Clayson, 2009), the student evaluations of teaching in a higher education institution, the National Polytechnic School of a South American country, Ecuador, where students study technical subjects, such as engineering, architecture, and biotechnology. Unfortunately, in South Americathere is a shortage of reliable and valid SET scales in polutechnic higher education institution, although it is a widespread procedure in these institutions since the early 1980s (Pareja, 1986).

The Council of Ecuadorian Higher Education establishes the obligatory nature of the evaluation of the teaching staff of higher education institutions, both for their entry and for their promotion, in the Career and Ladder Regulations of the Professor and Researcher of the Higher Education System, and they may even be dismissed from teaching in case of performance evaluations of less than 60%

twice consecutively, or four comprehensive evaluations of performance less than 60% during their career (CES, 2017).

The evaluation of the quality of teaching in the National Polytechnic School of Ecuador uses different procedures, including self-assessment, evaluation by peers and managers, and evaluation by students through evaluation questionnaires. The elaboration of this questionnaire is based on the criteria proposed by the institution itself and the guidelines suggested by the Higher Education Council (Consejo de Educación Superior, CES).

The instrument of student evaluation of teaching used in the National Polytechnic School is the 'Cuestionario de Evaluación de la Enseñanza del Profesor de la Escuela Politécnica Nacional del Ecuador' (Teacher Evaluation Questionnaire of the National Polytechnic School). The elaboration of the questionnaire was based on previous SET literature (Marsh, 2007a; Mortelmans & Spooren, 2009; Toland and De Ayala, 2005) and consists in the proposal of several effective teaching criteria. Next, a teaching committee, part of the management team of the National Polytechnic School, developed a set of items. This committee consisted of 5 main tenured professors with extensive experience in teaching quality, and a representative from the administrative sector and a student. The aspects to be evaluated and the specific items that make up the questionnaire are approved each academic year by the management team of the National Polytechnic School. The items are grouped theoretically into the following four factors. 1. Planning, mastery, and clarity in the explanation of the subject matter (i.e. The teacher conveniently expresses the class objectives and contents, indicating their relationship with the student's training). 2. Methodology and resources (i.e. The teacher prepared teaching material apart from the textbook and made it known). 3. Teacher-student relationship (i.e. The teacher created a climate of trust and productivity in class). 4. Evaluation (i.e. The evaluation events are related to the teaching given). Although the number and dimensions of effective teaching remains an open question (Spooren et al, 2013), these four dimensions are present in the most of SET literature (Feldman, 1989; Huybers, 2014; Richardson, 2005).

Thus, face and content validity are taken into account during the process of developing an instrument. Face validity indicates whether an instrument seems appropriate, that is, face validity does not analyze what the instrument measures but what it appears to measure; i.e., the extent to which the items of a SET instrument appear relevant to a respondent (Rispin, Davis, Sheafer & Wee, 2019; Spooren, Brockx, & Mortelmans, 2013). Content validity refers to whether the content of an instrument has been included in an exhaustive and representative way, that is, if the content has been included in an appropriate way. Content validity is obtained from the consensus based on informed opinion of experts; it is recommended to include at least 5 experts for the evaluation of content validity (Yaghmale 2009). However, the empirical validation is minimal and is limited to a descriptive analysis of the items individually considered. It lacks a complete process of construct and criterion validity, as well as an estimation of the reliability of the scale and/or the subscales that make up these questionnaires.

Although many studies have been developed on the subject of the validation of student evaluation of teaching scales in higher education, few have done so in the specific scope of polytechnic institutions and SEM studies; there are also very few examples of rigorous development of short teacher assessment scales. For this reason, our work tries to contribute to filling this gap.

Scale Reduction

Currently, a line of work has been developed to reduce the length of scales already used or elaborate scales with a reduced number of items. The lack of time for the application of scales, fatigue, and possible stereotyped responses in scales that are too long or that are part of a set of scales that are applied within the same study, etc., has led to proposals of short scales (Goetz et al., 2014; Lafontaine et al., 2016). These scales have to be small enough to allow for a rapid assessment of purposed constructs, but large enough to ensure appropriate reliability, validity, and accurate parameter estimation.

Short scales are considered to present psychometric inconveniences in comparison to long scales with regard to both reliability and validity, as they can be more affected by random measurement errors (Credé, Hamrms, Niehorster, & Gaye-Valentine, 2012; Lord & Novick, 1968).

In the short-form scales, the number of items per factor proposed varies from one to four items. Thus, several authors propose scales and subscales in which each factor should include four items (Marsh, Hau, Balla, & Grayson, 1998; Marsh, Martin, & Jackson, 2010; Marsh et al., 2009; Poitras, Guay, & Ratelle, 2012). Moreover, other authors, such as Credé et al. (2012), point out the loss of psychometric qualities when the scales have between one and three items. On the other hand, Kline (2016) points out that construct validation procedures, such as confirmatory factor analysis and other modelling methods, require at least three indicators per factor for a model to be identified. From a point of view that combines theoretical demands with practical interest, the PISA study of 2000 and the German PISA study of 2003 use short scales with three items (Brunner et al., 2010).

Another group of studies propose the use of short scales based on the finding that reliability and validity of short measures is similar to those of the corresponding longer scales measures, and have high correlation with long scales (Christophersen & Konradt, 2011; Gogol et al., 2014; Nagy, 2002). Gogol et al. (2014) compared the reliability and validity of three-item and single-item measures to those of the corresponding longer scales, finding satisfactory reliability and validity indices in all short forms and a high correlation with long scales; however, single-item measures showed the lowest reliability indices and correlations with the longer scales. Based on these results, the authors defended the use of short scales.

In sum, there are empirically founded reasons to propose short scales of three or four items. Although three items seem sufficient to guarantee the reliability and validity of the measure, in some cases, such as when additional assumptions are made about the psychometric properties of the items and factors (variables error variances, factor variances, etc.) or the hierarchical nature of the data is taken into account in multilevel analysis, four items per factor are recommended for accurate parameter estimation (Marsh, Hau, Balla, & Grayson, 1998).

Research Objectives

Hence, in this work, the following objectives are established:

1. Validate a Student Evaluation of Teaching Rating Scale and a short version of the corresponding long scale, including four items for each measured dimension, in a large sample of higher education students enrolled in a polytechnic higher education institution.
2. Test alternative structures of the dimensions of the Student Evaluation of Teaching Rating Scale.
3. Find the relationship between the long and short forms of the scale and academic achievement.
4. Examine whether the scores are invariant with respect to relevant variables such as the gender of the students in the context of scientific-technological studies.
5. Considering the hierarchical nature of the data, determine the ratings of the teaching of individual students located in different groups, classes, or sections, as well as where each group evaluates a different teacher.

2. Materials and Methods

Participants

The sample comprised 6110 students of the National Polytechnic School of Ecuador who rated the teaching of their teachers. These students were enrolled in eight different faculties in 28 different degree programs and attended 358 different classes. 68.3% of the students were male and 31.7% female. The higher percentage of male students is representative of the population of students of polytechnic studies. The average age was 22.6 years old (SD = 3.2). These students rated the teaching of their teachers during the 2016–17 academic year.

The sample of teachers was composed of 310 teachers, most of which were males (62.8%), aged between 26 and 57 years (mean = 43.7), belonging to all professional categories, from assistant professor to principal, with a majority (42%) of full professors, and extensive teaching experience (mean = 18,6 years).

This sample of participants corresponds to the students enrolled in the aforementioned studies, who took part in the evaluation process of the teaching staff of their institution, the EPN, at the end of a semester.

Measures

Students' evaluations of teaching ratings were obtained from the 'Cuestionario de Evaluación de la Enseñanza del Profesor de la Escuela Politécnica Nacional del Ecuador' [Teacher Evaluation Questionnaire of the National Polytechnic School], approved by the teaching staff for the 2016–17 academic year. This scale comprises 32 items grouped theoretically into the following four factors. 1. Planning, mastery, and clarity in the explanation of the subject matter (items 1–9). 2. Methodology and resources (items 10–15). 3. Evaluation (items 16–23) 4. Teacher–student relationship (items 24–32). Response scale ranges from 1 to 5; 1: do not agree at all; 2: little agreement, 3: moderately agree; 4: strongly agree; and 5: totally agree. The full and reduced scales with the items grouped into the four theoretical dimensions are included in the Appendix A.

The measures of student academic performance were obtained for a subsample of 1538 students. This subsample consisted of those students for whom data on their academic performance were available in the university's administrative computerized records. There is no known evidence that this subsample is biased with respect to the total sample used in this study. This measure of academic performance at the end of the semester was operationalized by the grade awarded by the teacher, based on a final exam: a written examination, both theoretical and practical. These final exams were the same across sections in some cases and were different for different sections in others. Different sections follow the same program and have the same assessment criteria that are specified in the study program of each course. Therefore, the exams, although different, can be considered quite equivalent. There are also common general rules for all exams in the National Polytechnic School of Ecuador. The scores of final grades ranged from 0 to 40 for all courses.

Students' age and gender, as well as teachers' age, gender, and experience, were collected from administrative records.

Procedure

The data were collected from the existing computer records in the administration of the Polytechnic School, and permission for access to them was granted to the academic staff of the Institution. The data provided by the institution were anonymous, with only one identification code for each student.

The application of the evaluation of teaching scale by the students was carried out towards the end of the semester, before they knew their final grades. All the teachers were evaluated by the students in a similar period of time. All the students had to evaluate the teachers to be able to access their final grades. The student evaluation of teaching was conducted through an electronic platform on which the data were recorded.

The impact that faculty procedures of student evaluations of teaching have on response rates has been analysed by several authors in special electronic evaluations. Thus, Young, Joines, Standish, and Gallagher (2019) found that evaluations made by students were considerably higher when faculty gave in-class time to students to complete student evaluation of teaching, compared to an electronic form issued by the administration. However, other studies of this issue did not find differences between the evaluations made with electronic questionnaires and paper and pencil questionnaires, or when a more representative sample responded instead of a smaller, more biased sample (Nowell, Gale, & Kerkvliet, 2014).

As response rates to electronic administration are lower than to paper-and-pencil questionnaires, the procedure followed in this work consisted in requiring all the students to answer the evaluation survey in order to access their final grades. This procedure has proved useful and valid in some higher education institutions (Leung & Kember, 2005; Nair & Adams, 2009).

Data Analysis

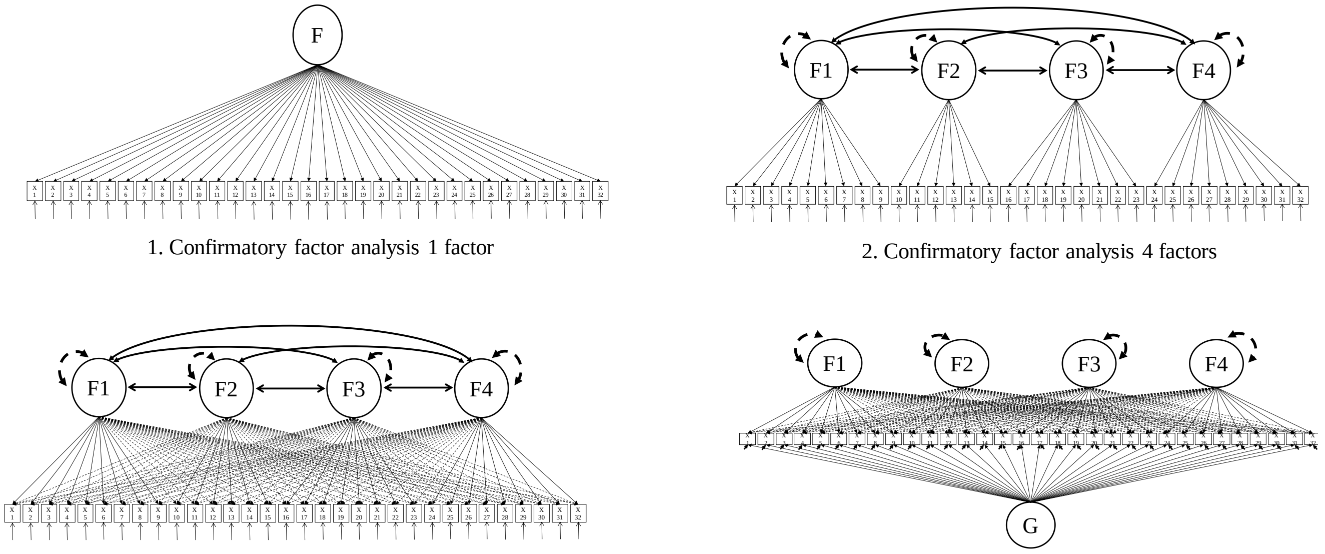
Preliminary analyses.

We explored means, standard deviations, skewness, and intraclass correlations (ICCs) for all items. Skewness indicates the asymmetry of the distribution, while ICC gives information about the non-independence of data, that is, the similarity of students' responses in the same class.

Construct validity.

To gather evidence of the scale's construct validity, we followed the recommendations of Schmitt, Sass, Chappelle, and Thompson (2018). There are different methods to retain the 'best' factor structure; for instance, exploratory factor analysis (EFA), confirmatory factor analysis (CFA), or exploratory structural equation model (ESEM). EFA has the disadvantage of the difficulty to replicate results with different samples, while CFA leads to biased loadings and correlations between factors because it requires that cross-loadings be 0 in the non-target factors (Garn, Morin, & Lonsdale, 2018). ESEM combines EFA and CFA, provides goodness of fit indices, and allows testing for multiple-group measurement invariance (Xiao, Liu, & Hau, 2019). Schmitt et al. (2018) recommend using EFA when there is no a priori theory, using CFA when there is a strong theory and evidence of the scale structure, and using ESEM when the a priori theory is sparse. Howard, Gagné, Morin, and Forest (2018) add that ESEM should be retained over CFA when correlations are different between factors are different in these two methods.

Another interesting issue in factor analysis, specifically in multidimensional structures, is bi-factor models (Morin, Katrin Arens, & Marsh, 2016). Bi-factor models are used to divide covariance between a global factor (i.e., teachers' style) and specific factors (i.e., Methodology and resources or Teacher-student relationship).



1. Confirmatory factor analysis 1 factor

2. Confirmatory factor analysis 4 factors

3. Exploratory structural equation model 4 factors

4. Bi-factor exploratory structural equation model 4 factors

Therefore, in view of the above information and our data, we can test the following models: one-factor via CFA, four-factor via CFA, four-factor via ESEM, and four- and bi-factor via ESEM (See Figure 1). To select the factor structure, we relied on the adjusted χ^2-difference tests and changes in CFI and RMSEA. The estimation method was Robust Maximum Likelihood because the data were non-normal; moreover, as responses were not independent, we corrected χ^2 and standard errors using a sandwich estimator (Muthen & Satorra, 1995; Muthén & Muthén, 2018). All analyses were conducted with Mplus 8.4 (Muthén & Muthén, 2020).

Short version

To choose items for a short version, we account for factor loadings, corrected for item-test correlations, reliability, and the item theoretical significance (Marsh, Martin, & Jackson, 2010). To test the agreement of both versions, we relied on the Levy correction of the short vs long form correlation. This correction accounts for the shared error variance between both forms due to the subset of items (Barrett, 2015; Levy, 1967). Moreover, because correlation only considers the monotonicity between both forms, we also relied on the Gower index (Barrett, 2012; Gower, 1971), whose values range between 0 and 1, where values close to 1 indicate agreement.

Gender measurement invariance.

To test whether male and female students interpret the scale similarly, we performed a measurement invariance test (Vandenberg & Lance, 2000). Specifically, we compared three models: configural, metric, and scalar (Muthén & Muthén, 2018). The configural model has factor loadings, intercepts, and residual variances free across groups and factor means fixed at zero in all groups. In the metric model, factor loadings are held equal across groups, while intercepts and residual variances are free across groups, and factor means are fixed at zero in all groups. Finally, in the scalar model, factor loadings and intercepts are equal across groups, while residual variances are free across groups, and factor means are constrained to zero in one group and free in the other group. For model comparisons, we used the adjusted χ^2-difference tests and changes in CFI and RMSEA.

Reliability

Finally, to test the reliability of the short and long form, we did not use Cronbach's alpha because there is increasing evidence of its lack of accuracy and the difficulty of meeting its assumptions: the parallelism and tau-equivalence of the items (McNeish, 2018; Zhang & Yuan, 2016). Cho (2016) proposes different formulas to estimate reliability whenever items lack parallelism, tau-equivalence, or both, not only for unidimensional structures but also for multidimensional structures.

Criterion validity: Relation with academic achievement.

To analyse the relationships between student ratings of teaching and academic performance, the data were taken individually and grouped into sections. Initially, the validity of students' ratings might be evidenced by the correlation between SET and academic achievement. Nevertheless, students' grades cannot be supposed to constitute a simple measure of teaching effectiveness because each group could have different evaluations (Richardson, 2005). The key evidence cited in support of student evaluations of teaching as a measure of a teacher's instructional effectiveness is multisection studies, in which different professors teach the same subject following the same outline, and at the end of the semester, all the sections have the same exam or equivalent ones (Cohen, 1981; Uttl, White, & Gonzaléz, 2017). To find the correlation between scale scores and academic performance, the data were taken individually and treated as a typical multisection study in which the average class was used as the unit of analysis.

3. Results

Preliminary Analyses

Means varied between 3.85 for Item 15 and 4.07 for Item 9, while standard deviations ranged from 1.02 for Item 2 to 1.16 for Item 11. Skewness varied from −.840 for Item 15 to −1.120 for Item 1. More information can be found in Appendix B.

Construct Validity

We compared the four proposed models. We observed that the probability that a four-factor CFA had the same fit as a one-factor CFA was $p < .001$ ($\Delta\chi^2 = 10217.93$, df = 8). Similarly, the probability that a four-factor ESEM had the same fit as a four-factor CFA was $p < .001$ ($\Delta\chi^2 = 1272.977$, df = 84). Finally, the probability that a four-factor ESEM had the same fit as a bi-factor four-factor ESEM was $p < .001$ ($\Delta\chi^2 = 1143.317$, df = 28).

Table 1

χ^2 test and fit indices for different structures

Model	χ^2		RMSEA	CFI
	Value	DF		
CFA 1F	16679.456	464	.076	.873
CFA 4F	6461.526	458	.046	.953
ESEM 4F	5188.549	374	.046	.962
Bi-ESEM 4F	4045.232	346	.042	.970

The structure with the best fit was the bi-factor four-factor ESEM. However, to retain this structure, moderate-high factor loadings were required in the global factor (Howard et al., 2018), and in this case, the factor loading absolute values were between .024 and .228, with an average value of .093. Therefore, we discarded the bi-factor four-factor ESEM and proceeded to explore the four-factor ESEM structure. This structure provided moderate to high loadings and low cross-loadings (See Appendix 1). Specifically, for Planning, mastery, and clarity in the explanation of the subject matter (Factor 1), the loadings ranged between .508 and .857, for Methodology and resources (Factor 2) between .601 and .856, for Evaluation (Factor 3) between .385 and .885, and for Teacher-student relationship (Factor 4) between .629 and .958. Thus, we decided to retain this structure.

Short Version

Construct validity

Following Marsh and colleagues' recommendations (2010), we selected four items of each subscale. Next, we proceeded to test the selected structure via ESEM. The chi square test result and fit indices were: $\chi^2(6110, 62) = 509.115$ ($p < .001$), CFI = .992, RMSEA = .034 (90% C. I. = .032, .037). For Planning, mastery, and clarity in the explanation of the subject matter, the loadings ranged between .676 and .898, for Methodology and resources between .572 and .916, for Evaluation between .672 and .864, and for Teacher-student relationship between .675 and .946 (See Appendix C).

Agreement between both versions

As shown in Table 2, Levy's corrected correlation and the Gower index revealed a high concurrence between both forms, ranging from r = .893 to r = .974.

Table 2

Agreement between the long and short forms

Factor	Levy's correlation	Gower index
Planning, mastery, and clarity	.893	.963
Methodology and resources	.901	.974
Evaluation	.919	.972
Teacher-student relationship	.918	.969

Gender measurement invariance

32-item scale

Multiple-group analyses to examine potential gender differences in the model results showed that the probability of the same fit between the configural and the metric model was $p < .902$ ($\Delta\chi^2$ = 93.127, df = 112). Similarly, the comparison between the metric and the scalar model yielded $p < .902$ ($\Delta\chi^2$ = 126.335, df = 140). Thus, we found no gender differences in loadings, thresholds, or factor means in the long form scale (See Table 3).

Table 3

χ^2 test and fit indices for invariance testing

Model	χ^2		RMSEA	CFI
	Value	DF		
Configural	2301	748	.051	.959
Metric	2321	860	.046	.961
Scalar	2374	888	.046	.960

16-item scale

The comparison between the configural and the metric models revealed that the probability that the model fits would be the same was $p < .847$ ($\Delta\chi^2$ = 38.043, df = 48). Similarly, the comparison between the metric and the scalar model yielded $p < .629$ ($\Delta\chi^2$ = 55.838, df = 60). Thus, we did not find gender differences in loadings, thresholds, or factor means in the short form either (See Table 4).

Table 4

χ^2 test and fit indices for invariance testing (short form)

Model	χ^2		RMSEA	CFI
	Value	DF		
Configural	274.8	124	.039	.980
Metric	305.6	172	.031	.991
Scalar	327.8	184	.032	.990

Reliability

32-item scale

The reliability of the scale was assessed using the Congeneric Correlated Factors formula. Reliability for the whole scale was .980, for Planning, mastery, and clarity in the explanation of the subject matter .949, for Methodology and resources .901, for Evaluation .948, and for Teacher-student relationship .947.

16-item scale

The reliability for the whole scale was .972, for Planning, mastery, and clarity in the explanation of the subject matter .904, for Methodology and resources .901, for Evaluation .920, and for Teacher-student relationship .919.

Correlation with Academic Achievement

Table 5 shows the correlations between the long and short versions of the scale of evaluation of teaching with academic performance, taking individual and aggregate data in sections. As we can see, all the correlations were statistically significant with moderate-low values. Both the subscales and the total scale showed significant correlations with academic performance. The values of the correlations of the reduced scale were very similar to those of the long scale. In addition, the correlations in the aggregated data in classes or sections were slightly higher than in the individual data.

Table 5

Correlations between the long and short versions of the scale of evaluation of teaching with academic performance, taking individual and aggregate data in sections

	Individual data		Aggregate data in sections	
Subscales	Long	Short	Long	Short
1. Planning, explanation, and presentation of subject	.21**	.21**	.21**	.23**
2. Method and materials	.23**	.22**	.26**	.26**
3. Evaluation	.23**	.22**	.24**	.23**
4. Teacher-student relationship	.21**	.20**	.26**	.23**
Total scale	.23**	.23**	.25**	.26**

Note. All correlations showed a $p < .01$.

4. Discussion

The results clearly show the structural validity of the student evaluation of teaching ratings elaborated in the National Polytechnic School of Ecuador. Given that the main objective of this study is to propose a short scale that shows reliability and validity, AFC and Exploratory Structural Equation Modelling were used.

Results showed a multidimensional model with four highly correlated factors that do not exclude a general factor, with an excellent fit to data, both in the long scale and in the short version of the scale. The structure with the best fit was the bi-factor four-factor ESEM; however, the factor loadings on the global factor were low (Howard et al., 2018) and, thus, the four-factor ESEM structure was retained.

Based on a sample of 26,746 students who took the Program for International Student Assessment (PISA) of 2012, Scherer, Nilsen, and Jansen (2016), found that bi-factor exploratory structural equation modelling outperformed alternative approaches with respect to model fit.

The researchers are divided on the basis of the existence of a second-order general factor (Abrami, et al, 1997; Cheung, 2000) or different first-order correlated factors (Marsh, 1991b, 2007a). As for the practical implications of this issue, perhaps the most accurate conclusion is the one provided as early as 1991 by Marsh (1991a) himself: 'I have chosen a middle ground recommending the use of both specific dimensions and global ratings' (p. 419).

The use of academic performance measures as an external criterion validity of the student evaluation of teaching (SET) rating scales is very common in validation works, which has been called a strong test for criterion validity. However, the meta-analyses (Cohen, 1981; Feldman, 1989; Clayson, 2009; Marsh, 2007a Uttl, White, & Gonzalez, 2017) shows the existence of moderate (.50- 20) to small (.20-.00) positive correlations between SET scores and student achievement. Although these results provide relative evidence of the convergent validity of SET scales; due to the variety of views concerning good teaching, and due to the variety in the measurement and predictors of student achievement, (Schneider, & Preckel, 2017; Spooren, Brockx, & Mortelmans, 2013), academic achievement should not be the only indicator of SET scales criterion validity.

Student Evaluation of Teaching rating scales are multidimensional, many researchers defend the use of single, global scores (Apodaca & Grad, 2005). For this reason, even when recognizing the multidimensional and hierarchical structure of the dimensions evaluated in the scales on student evaluation of teaching, many works studying this issue use global scores; meanwhile, the feedback provided to teachers for the improvement of teaching practice includes a profile of the scores in the different dimensions, which show the strengths and weaknesses of each teacher's methods.

Given the existence of student gender bias in student evaluation of teaching, configural, metric, and scalar gender measurement invariance were tested. Previous research has shown that female subjects are likely to score higher in SET ratings (e.g. Badri et al., 2006; Darby 2006). Bonitz (2011) found that gender variations in SET scores could be due to gender variations in traits such as agreeableness that correlate with the SET scores. However, the results of this study showed configural, metric, and scalar gender measurement invariance in the context of scientific-technological studies.

Although the literature on gender bias in SET shows that male students express a bias in favor of male professors, (American Sociological Association, 2019; Boring 2017; Centra & Gaubatz 2000; Mitchell, & Martin, 2018), the extensive review by Kreitzer, and Sweet-Cushman (2021), shows that the effect of gender is conditional upon other factors. Other works show that the gender bias against perceived female instructors disappears (Uttl and Violo, 2021). The results of Rivera and Tilcsik

(2019) even show that these gender differences can disappear in scales with 6 points or less, like those of our scale.

The results of this work also show the concurrent validity of the reduced scale of 16 items, which showed a high correlation with the full scale of 32 items. Levy's corrected correlation and the Gower index revealed high concurrence between both forms, with values above .90. These results are slightly higher than those obtained in other studies that also showed a high degree of agreement between long and short forms of such scales (Gogol et al., 2014, Lafontaine et al., 2016).

The high values of the reliability coefficients, estimated according to the assumptions of the SEM model used, are also striking for both the long and short whole scales and subscales. These values were higher than .90 and reached values of .98 and .97 for the whole scales. The Congeneric Correlated Factors procedure (Cho, 2016) was applied in consideration of there being different factor loadings to obtain the values of multidimensional reliability coefficients apart from Cronbach's alpha, which supposes that all factor loadings are equal (i.e. tau-equivalents), and thus underestimates the reliability.

On the other hand, the results also showed moderate, significant correlations between both the long and short versions of the scale with academic performance, taking individual and aggregate data in classes or sections.

The evidence in support of student evaluations of teaching as a measure of teachers' instruction effectiveness comes from studies showing correlations between measures of student evaluation and student achievement, a strong test for criterion validity.

The results obtained with aggregate data, taking the section as the unit of analysis, showed a moderate and statistically significant correlation (.26) between student ratings and final performance. This result is expected from studies of instructors' teaching effectiveness, in which it is considered that multisection studies are more appropriate for apprehending the true relationship between student evaluations of teaching and academic performance (Cohen, 1981; Uttl et al., 2017).

However, the relationship of the students' evaluation of teaching with their academic performance is lower than that found in some previous meta-analytic studies (Cohen, 1981), but higher than that found in the meta-analysis of Uttl et al. (2017) of the studies published to that date, when small study size effects and prior academic achievement were considered. Taken together, the results demonstrated the good psychometric qualities of the Teacher Evaluation Questionnaire of the National Polytechnic School and its construct and criterion validity, as well as its high reliability. In addition, the psychometric indices of the short version of this scale suggest the possibility of developing short scales of three or four items that are equally reliable and valid.

In addition, the relationships obtained between the long and short versions of the new instrument with academic performance have practical implications for teacher teaching. This instrument may help teachers to adapt their teaching to student needs and preferences in the context of specific characteristics of polytechnic studies.

However, we must not lose sight of the open controversy between students' perceptions of the quality of the teaching, or perceptions of leaning, and their actual learning. In the context of STEM -Science, Technology, Engineering and Mathematics - instruction Deslauriers, McCarty, Miller, Callaghan, and Kestin (2019) find that students in active classrooms learned more, but their perception of learning was lower than that of their peers in passive instruction.

Regarding the limitations of this study and possible future studies, given that the long and short forms were administered as part of the full scale, and despite the correction of Levy and Gower for the calculation of the correlation between the two version, it would be necessary to administer the long and short scales to the same sample independently. In addition, it would be convenient to examine the factorial structure of the short scale in an independent representative sample of students. In this study, we analyzed the relationship with academic achievement, it might be of interested to explore the relationship with higher education engagement (Vizoso, Rodríguez, & Arias-Gundín, 2018) or general pedagogical knowledge (Klemenz, König, & Schaper. 2019). Finally, obtaining longitudinal data in the same and different samples of the National Polytechnic School could serve to deepen the validity of the scale developed in this work.

It should also be taken into account that these results have been obtained in a single institution, which limits the generality of the results; however, it is the largest institution of polytechnic studies (science, biotechnology, engineering, architecture, etc.), the largest in Ecuador that collects students from all over the country.

In sum, this work provides evidence of the validity of a teaching evaluation scale in the setting of a polytechnic institution of higher education, as well as a rigorous methodological procedure for the validation of short versions of teaching evaluation scales.

Author Contributions: All authors contributed to the study conception and design, material preparation, data collection, analysis, and writing of the manuscript. All authors read and approved the final manuscript.

References

Abrami, C.P., d'Apollonia, S., & Rosenfield, S. (1997). The dimensionality of student ratings of instruction: what we know and what we do not. In R. E. Perry and J.C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 321–367). New York: Agathon Press.

Aguilar R.M., and Bautista, M.J. (2015). Teacher profiles and excellence: A study at the Universidad Técnica Particular of Loja, Ecuador. *Revista Iberoamericana de Educación a Distancia, 18*(2), 225-250.

American Sociological Association (2019). Statement on Student Evaluations of Teaching. https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_feb132020.pdf

Andrade-Abarca, P.S., Ramón-Jaramillo, L.N., and Loaiza-Aguirre, M.I. (2018). Application of the SEEQ as an instrument to evaluate university teaching activity. *Revista de Investigación Educativa, 36*(1), 259-275. http://dx.doi.org/10.6018/rie.36.1.260741

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

Badri, M. A., Abdulla, M., Kamali, M. A., & Dodeen, H. (2006). Identifying potential biasing variables in student evaluation of teaching in a newly accredited business program in the UAE. *International Journal of Educational Management, 20*(1), 43-59.

Barrett, P. (2012). *Gower program help file*. Auckland, New Zealand: Advanced Projectes R&D Ltd.

Barrett, P. (2015). *Levy's short vs long form corrected correlation*. Auckland, New Zealand: Advanced Projectes R&D Ltd.

Basow, S. A., Phelan, J. E., & Capotosto, L. (2006). Gender patterns in college students' choices of their best and worst professors. *Psychology of Women Quarterly*, *30*(1), 25-35.

Bonitz, V.S. (2011). Student evaluation of teaching: Individual differences and bias effects. Iowa State University. Digital Repository. https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=3183&context=etd

Boring, A. (2015). Gender Biases in student evaluations of teachers. Documents de Travail de l'OFCE 2015-13. Paris: Observatoire Francais des Conjonctures Economiques (OFCE). Available from: http://www.anneboring.com/uploads/5/6/8/5/5685858/aboring_gender_biases_in_set_april_2014.pdf

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

Brunner, M., Keller, U., Dierendonck, C., Reichert, M., Ugen, S., Fischbach, A., et al (2010). The structure of academic self-concepts revisited: The nested Marsh/Shavelson model. *Journal of Educational Psychology, 102*(4), 964–981. https://doi.org/10.1037/a0019644

Centra, J.A., & Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? The Journal of Higher Education,71(1), 17-33. https://doi.org/10.1016/j.jpubeco.2016.11.006

Centra, J.A., and Gaubatz, N.B. (2000). Is there gender bias in student evaluations of teaching? *The Journal of Higher Education*, *71*(1), 17-33. https://doi.org/10.2307/2649280

Cheung, D. (2000). Evidence of a single second-order factor in student ratings of teaching effectiveness. *Structural Equation Modeling*, *7*(3), 442-460, https://doi.org/10.1207/S15328007SEM0703_5

Cho, E. (2016). Making reliability reliable: A systematic approach to reliability coefficients. *Organizational Research Methods*, *19*(4), 651-682. https://doi.org/10.1177/1094428116656239

Christophersen, T., & Konradt, U. (2011). Reliability, validity, and sensitivity of a single-item measure of online store usability. *International Journal of Human–Computer Studies*, *69*(4), 269–280.

Clayson, D. E. (2009). Student evaluations of teaching: Are they related to what students learn? A meta-analysis and review of the literature. *Journal of Marketing Education*, *31*, 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/10.3102/00346543051003281

Consejo de Educación Superior (CES) (2017). Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior. [Career and Ladder Regulations of the Professor and Researcher of the Higher Education System]. Retrieved on the 20 April 2019, from: https://bit.ly/2Y6Jc0w

Credé, M., Harms, P., Niehorster, S., & Gaye-Valentine, A. (2012). An evaluation of the consequences of using short measures of the Big Five personality traits. *Journal of Personality and Social Psychology*, *102*(4), 874–888. https://doi.org/10.1037/a0027403

Darby, J. A. (2006). Evaluating courses: An examination of the impact of student gender. *Educational Studies, 32*(2), 187-199.

Deslauriers, L., McCarty, L.S., Miller, K., Callaghan, K., and Kestin, G. (2019). Measuring actual learning versus feeling of leaning in response to being actively engaged in the classroom. *PNAS, 116* (39) 19251-19257. https://doi.org/10.1073/pnas.1821936116

Feistauer, D., & Richter, T. (2017). How reliable are students' evaluations of teaching quality? A variance components approach. *Assessment & Evaluation in Higher Education*, *42*(8), 1263–1279.

Feldman, K.A. (1989). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, *30*(6), 583-645.

Fernández, N., y Coppola,N. (2008). An approach to evaluation of academic teaching in some Iberan-American Countries. A Comparative perspective between resemblances, differences, and convergence. *Perspectivas em Políticas Públicas*, *1* (2), 131-163.

Garn, A. C., Morin, A. J. S., & Lonsdale, C. (2018). Basic psychological need satisfaction toward learning: A longitudinal test of mediation using bifactor exploratory structural equation modeling. *Journal of Educational Psychology*, *111*(2), 354-372. https://doi.org/10.1037/edu0000283

Gogol, K, Bunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischhach, A., & Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational-affective constructs with three-item and single-item measures. *Contemporary Educational Psychology*, *39*(3), 188-205.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857-871.

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

Howard, J. L., Gagné, M., Morin, A. J. S., & Forest, J. (2018). Using bifactor exploratory structural equation modeling to test for a continuum structure of motivation. *Journal of Management*, *44*(7), 2638-2664. https://doi.org/10.1177/0149206316645653

Huybers, T. (2014) Student evaluation of teaching: The use of best–worst scaling, *Assessment & Evaluation in Higher Education, 39*:4, 496-513. https://doi.org/10.1080/02602938.2013.851782

Klemenz, S., König, J., & Schaper, N. (2019). Learning opportunities in teacher education and proficiency levels in general pedagogical knowledge: New insights into the accountability of teacher education programs. *Educational Assessment, Evaluation and Accountability*, *31*(2), 221–249. https://doi.org/10.1007/s11092-019-09296-6

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4th ed.). New York: Guilford Press

Kreitzer, R.J., Sweet-Cushman, J. (2021). Evaluating Student Evaluations of Teaching: a Review of Measurement and Equity Bias in SETs and Recommendations for Ethical Reform. *J Acad Ethics* (2021). https://doi.org/10.1007/s10805-021-09400-w

Lafontaine, M.-F., Brassard, A., Lussier, Y., Valois, P., Shaver, P.R., & Johnson, S.M. (2016). Selecting the best items for a short-form of the experiencies in close relationships questionnaire. *European Journal of Psychological Assessment*, *32*(2), 140-154.

Leung, D.Y.P., & Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the Internet. *Research in Higher Education*, *46*(5), 571-591.

Levy, P. (1967). The correction for spurious correlation in the evaluation of short form tests. *Journal of Clinical Psychology*, *23*(1), 84–86.

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin*, *69*(6), 410.

Linse, A.R. (2017). Interpreting and using student rating data: Guidance for faculty serving as administrators and on evaluation committees. *Studies in Educational Evaluation*, *54*, 94-106.

Loor K.J., Gallegos, M.R., Intriago, M.M., Guillén, X. (2018). University faculty evaluation: Ibero-America trends. *Educación Médica Superior, 32*(1), 239-252.

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

Marsh, H. W. (1991a). A multidimensional perspective on student's evaluations of teaching effectiveness: Reply to Abrami and d'Apollonia (1991. *Journal of Educational Psychology*, *83*,(3), 416–421. https://doi.org/10.1037/0022-0663.83.2.285

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology*, *83*, 285–296. https://doi.org/10.1037/0022-0663.83.2.285

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on students' evaluations of teaching. *American Educational Research Journal*, *38*, 183–212. https://doi.org/10.3102/00028312038001183

Marsh, H. W. (2007a). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319–383). New York: Springer.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology*, *99*, 775–790. https://doi.org/10.1037/0022-0663.99.4.775

Marsh, H. W., Martin, A. J., & Jackson, S. E. (2010). Introducing a short version of the physical self description questionnaire: New strategies, short-form evaluative criteria, and applications of factor analyses. *Journal of Sport & Exercise Psychology*, *32*(4), 438-482.

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

Marsh, H.H., Hau, K.T., Balla, J.R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, *33*, 181-230.

McNeish, D. M. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*(3), 412-433. https://doi.org/10.1037/met0000144

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

Montoya, J., Arbesú, I. , Contreras, G., y Conzuelo S. (2014). Evaluation of university teaching in Mexico, Chile and Colombia: analysis of the experiences. *Revista Iberoamericana de Evaluación Educativa*, *7*(2e), 15-42.

Morin, A. J. S., Katrin Arens, A., & Marsh, H. W. (2016). A bifactor exploratory structural equation modeling framework for the identification of distinct sources of construct-relevant psychometric multidimensionality. *Structural Equation Modeling*, *23*(1), 116-139. https://doi.org/10.1080/10705511.2014.961800

Mortelmans, D, & Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies*, *35*, 547-552. https://doi.org/10.1080/03055690902880299

Muthen, B. O., & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, *25*, 267-316. https://doi.org/10.2307/271070

Muthén, L. K., & Muthén, B. O. (2020). *Mplus user's guide* (8.ª ed.). Los Angeles, CA: Muthén & Muthén.

Nagy, M. S. (2002). Using a single-item approach to measure facet job satisfaction. *Journal of Occupational and Organizational Psychology*, *75*(1), 77–86. https://doi.org/10.1348/096317902167658

Nair, C.S., & Adams, P. (2009) Survey platform: A factor influencing online survey delivery and response rate. *Quality in Higher Education*, *15*(3), 291-296. https://doi.org/10.1080/13538320903399091

Nowell, C., Gale, L.R., Kerkvliet, J. (2014). Non-response bias in student evaluations of teaching. *International Review of Economic Education*, *17*, 30-38.

Pareja, F. (1986). *La educación superior en el Ecuador* [The higher education in Ecuador]. Caracas: Regional Center for Higher Education in Latin America and the Caribbean (CRESALC)-UNESCO.

Pimienta, J.H. (2014). Development and validation of an instrument for measuring teacher performance based on competencies**.** *Revista de Docencia Universitaria, 12*(2), 231-250.

Poitras, S.C., Guay, F., & Ratelle, C.F. (2012). Using the self-directed search in research: Selecting a representative pool of items to measure vocational interest. *Journal of Career Development*, *39*, 186-207. https://doi.org/10.1177/0894845310384593

Richardson, J.T.E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assessment & Evaluation in Higher Education*, *30*:4, 387-415. https://doi.org/10.1080/02602930500099193

Richardson, J.T.E. (2012). The role of response bias in the relationship between students' perceptions of their courses and their approaches to studying in higher education. *British Educational Research Journal*, *38*, 399-418. https://doi.org/10.1080/01411926.2010.548857

Rispin, K., Davis, A.B., Sheafer, V.L. & Wee, J. (2019). Development of the Wheelchair Interface Questionnaire and initial face and content validity. *African Journal of Disability 8*(0), a520. https://doi.org/ 10.4102/ajod.v8i0.520

Rivera, L. A., and Tilcsik, A.. (2019). Scaling Down Inequality: Rating Scales, Gender Bias, and the Architecture of Evaluation. *American Sociological Review*, *84*(2):248–274. DOI:10.1177/0003122419833601

Robalino, M., Körner, A., Murillo, F.J. (2007). *Evaluación del desempeño y carrera profesional docente: un estudio comparado entre 50 países de América y Europa*. UNESCO Office Santiago and Regional Bureau for Education in Latin America and the Caribbean. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000152934

Sánchez, T., Gilar-Corbi, R., Castejón, J-L., Vidal, J., & León, J. (2020). Students' evaluation of teaching and their academic achievement in a higher education institution of Ecuador, *Frontiers in Psychology, 11*. https://doi.org/10.3389/fpsyg.2020.00233

Sandoval-Palis, I.; Naranjo, D.; Vidal, J.; Gilar-Corbi, R. (2020). Early dropout prediction model: A case study of university levelling course students. *Sustainability, 12*, 9314.

Scherer R, Nilsen T., & Jansen M. (2016). Evaluating individual students' perceptions of instructional quality: An investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Frontiers in Psychology*, *7*:110. https://doi.org/10.3389/fpsyg.2016.00110

Schmitt, T. A., Sass, D. A., Chappelle, W., & Thompson, W. (2018). Selecting the “best” factor structure and moving measurement validation forward: An illustration. *Journal of Personality Assessment*, *100*(4), 345-362. https://doi.org/10.1080/00223891.2018.1449116

Schneider, M.; Preckel, F. (2017). Variables associated with achievement in higher education: A systematic review of meta-analyses. *Psychological Bulletin*, *143*, 565-600, https://doi.org/10.1037/bul0000098.

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), *1–45***.** https://doi.org/10.3102/0034654313496870

Toland, M., & De Ayala, R.J. (2005). A multilevel factor analysis of students’ evaluations of teaching. *Educational and Psychological Measurement*, *65*, 272-296.

Uttl, B., and Violo, V.C. (2021). Small samples, unreasonable generalizations, and outliers: Gender bias in student evaluation of teaching or three unhappy students? ScienceOpen Research, 22 January 2021. https://www.scienceopen.com/document/read?vid=b1353421-2e05-4a79-9c6f-4ac5a44dcc03

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

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*(1), 4-70. https://doi.org/10.1177/109442810031002

Vizoso, C., Rodríguez, C., & Arias-Gundín, O. (2018). Coping, academic engagement and performance in university students. *Higher Education Research & Development*, *37*(7), 1515–1529. https://doi.org/10.1080/07294360.2018.1504006

Xiao, Y., Liu, H., & Hau, K.-T. (2019). A comparison of CFA, ESEM, and BSEM in test structure analysis. *Structural Equation Modeling, 26*(5), 665-677. https://doi.org/10.1080/10705511.2018.1562928

Yaghmale, F. (2009). Content validity and its estimation. *Journal of Medical Education 3*(1), 25-27

Young, K., Joines, J., Standish, T., & Gallagher, V. (2019) Student evaluations of teaching: The impact of faculty procedures on response rates, *Assessment & Evaluation in Higher Education*, *44* (1), 37-49. https://doi.org/10.1080/02602938.2018.1467878

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

Zhang, Z., & Yuan, K. H. (2016). Robust coefficients alpha and omega and confidence intervals with outlying observations and missing data: Methods and software. *Educational and Psychological Measurement*, *76*(3), 387-411. https://doi.org/10.1177/0013164415594658

Zhoc, K. C., Webster, B. J., King, R. B., Li, J. C., & Chung, T. S. (2019). Higher Education Student Engagement Scale (HESES): Development and psychometric evidence. *Research in Higher Education*, *60*(2), 219-244. https://doi.org/10.1007/s11162-018-9510-6

frontiers
in Psychology | Educational Psychology

Sánchez T, Gilar-Corbi R, Castejón J-L, Vidal J and León J (2020) Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador. *Front. Psychol*. 11:233. https://doi.org/10.3389/fpsyg.2020.00233

frontiers
in Psychology

ORIGINAL RESEARCH
published: 06 March 2020
doi: 10.3389/fpsyg.2020.00233

Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador

Tarquino Sánchez[1], Raquel Gilar-Corbi[2], Juan-Luis Castejón[2], Jack Vidal[1] and Jaime León[3]*

[1] National Polytechnic School, Quito, Ecuador, [2] Developmental and Educational Psychology Department, University of Alicante, Alicante, Spain, [3] Department of Education, University of Las Palmas de Gran Canaria, Las Palmas, Spain

This paper addresses the relationship between student evaluation of teaching (SET) and academic achievement in higher education. Meta-analytic studies on teaching effectiveness show a wide range of results, ranging from small to medium correlations between SET and student achievement, based on diverse methodological approaches, sample size studies, and contexts. This work aimed to relate SET, prior academic achievement, and academic achievement in a large sample of higher education students and teachers, using different methodological procedures, which consider as distinct units of analysis the group class and the individuals, the variability between students within classes, and the variability between group-class means, simultaneously. The data analysis included the calculation of group-class means and its relationship with the group-class mean academic achievement, through correlation and hierarchical regression techniques; additionally, a multilevel path analysis was applied to the relationship between prior academic achievement, SET, and their academic achievement, considering the variability among group classes. A multisection analysis was also carried out in those course disciplines in which there was more than one class group (section). The results of individual and group-class analysis revealed that SET was moderately low but related to academic achievement in a significant way once the effect of previous academic achievement was controlled. In addition, multilevel path analysis revealed the effect of SET on achievement, both within and between group-class levels. The results of the analysis carried out in the course disciplines with different sections, according to a multisection design, yielded similar results to the individual and aggregated data analyses. Taken together, the results revealed that SET was low related to academic achievement, once the effect of previous academic achievement was controlled. From these results, it follows that the use of SET as a measure of teachers' effectiveness for making administrative decisions remains controversial.

Keywords: student evaluation of teaching ratings, academic achievement, teaching effectiveness, multisection study, multilevel analysis

OPEN ACCESS

Edited by:
José Jesús Gázquez,
University of Almería, Spain

Reviewed by:
Olivia López Martínez,
University of Murcia, Spain
Efstathios Stefos,
The National University of Education (Ecuador), Ecuador

****Correspondence:***
Raquel Gilar-Corbi
raquel.gilar@ua.es

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Psychology

Received: *02 November 2019*
Accepted: *31 January 2020*
Published: *06 March 2020*

Citation:
Sánchez T, Gilar-Corbi R, Castejón J-L, Vidal J and León J (2020) Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador. Front. Psychol. 11:233. doi: 10.3389/fpsyg.2020.00233

INTRODUCTION

Student evaluation of teaching (SET) is a generalized practice in almost every institution of higher education around the world (Richardson, 2005; Zabaleta, 2007; Huybers, 2014) – from European countries (Husbands and Fosh, 1993) to Australian and North American universities (Richardson, 2005) and South American higher education institutions (Pareja, 1986).

However, this issue is contemporary and is a topic still open to question in higher education. Researchers working on SET have not yet provided a clear answer about some critical questions on the validity and utility of evaluations (Marsh, 2007; Spooren et al., 2013). Although the use of student evaluation as feedback for teachers is not so controversial, the utilization of student evaluation for measuring teaching effectiveness, based on the assumption that students learn better with highly rated teachers, is very controversial.

One of the central controversial points is the relation of SET ratings to their learning outcomes, such as academic achievement (Uttl et al., 2017). The evidence in support of SET as a measure of teachers' instruction effectiveness comes from the studies showing a correlation between measures of student evaluation and student achievement.

Methodological Concerns/Questions

Initially, the validity of students' judgments might be proven by the correlation between SET and academic achievement. However, the evaluation criteria for distinct course units may differ, and students' grades cannot be considered a simple measure of teaching effectiveness (Richardson, 2005).

The key evidence provided in favor of SET as a measure of the effectiveness of teachers' instruction is multisection studies (Uttl et al., 2017). Leventhal (1975) and Cohen (1981) defend that the stronger SET validation design implicates the designation of students to different sections of a multisection course. If the designation is random, between-section differences in student performance can be caused by differences in teachers. When students self-select into sections, it can be difficult to infer rating/achievement relationship. If this is the case, Marsh and Overall (1980) consider that, in these studies, they should provide adequate controls/measures for initial ability or prior achievement.

Some researchers (Cohen, 1981; Clayson, 2009; Uttl et al., 2017) point out that student achievement is highly dependent on factors such as intelligence or prior achievement and that to fully control these factors, it is necessary to randomly assign students to classes and teachers or, alternatively, use other control procedures of initial student ability or achievement, such as analysis of covariance using measures of prior academic achievement or capacity as covariates; using the change in grades based on pretest and posttest moments; or regressing individual students' performance scores on measures of students' prior achievement and using residual gains in performance, averaged across students within sections, as measures of learning. It is advisable to use a statistical procedure in which both ratings and performance are adjusted for initial student ability or performance.

An ideal multisection study design entails a course discipline or subject matter with many comparable group class – sections – taking the same program and assessment guidelines, in which students are randomly assigned to sections, with a different teacher in each section; all teachers are assessed through ratings before a final exam; and student academic achievement is evaluated by employing the same or an equivalent final exam. If a student shows better academic achievement due to highly rated teachers, a correlation between sections' average SET and sections' average final exam should be observed (Uttl et al., 2017).

This leads us to consider the appropriate unit of analysis in these types of studies (Cohen, 1981). Some researchers utilize the student as the unit of analysis, relating the student's academic achievement with his/her teacher rating. Other researchers utilize the group class as the unit of analysis, correlating mean group-class achievement with mean class SET. Researchers using individual student data follow a design that allows them to establish whether students who perform better, regardless of the class they attend, score the teachers better. To analyze the association between SET and student academic achievement for respective teachers, the group class (or teacher) must be used as the unit of analysis in the validity design (Cohen, 1981; Abrami et al., 1990; Marsh and Roche, 2000; Clayson, 2005; Richardson, 2005).

Although this solution is widely accepted, criticism has recently emerged. It is argued that the variability between students, despite being averaged, could confuse the variation between group means. Consequently, it may be found that there are no relationships between SET and achievement for individual students, even as the between-class mean data show a significant relationship (Clayson, 2007; Weinberg et al., 2009). It is necessary to use statistical methods that consider both the individual variability within the group class and the variability between group-class means.

Another methodological issue that can affect the results on the relationship between SET and student academic achievement is the number of sections (Cohen, 1981; Uttl et al., 2017). Kulik and McKeachie (1975) indicated that big correlations often appear with small sample sizes, suggesting that to find a stable validity coefficient, at least 30 sections are needed in a multisection study. More recently, Uttl et al. (2017) presented specific results on this topic in their meta-analysis of faculty's teaching effectiveness.

Revision Studies

To answer the question on the relationship between SET and academic achievement, a series of revision and meta-analytical studies have been carried out.

As early as the seventies, many researchers analyzed the association between SET and student achievement. However, as Kulik and McKeachie (1975) pointed out, "the most impressive thing about studies relating class achievement to class ratings of instructors is the inconsistency of the results" (p. 235).

Cohen (1981) performed the first meta-analysis based on 68 multisection studies, in which various equivalent sections/classes follow the same outline and the same or equivalent assessments; each section is instructed by a different professor, and these professors are evaluated using students' evaluation of teaching ratings. Cohen's (1981) results indicated that SET scores correlated moderately with academic achievement (r = 0.43), concluding that these results support the validity of SET as a measure of teaching effectiveness. However, recent studies have questioned some aspects of Cohen's (1981) meta-analysis, referring to the repeatable search strategy followed by Cohen or the sample size of sections on which Cohen's meta-analysis studies are based, with as few as five sections (Uttl et al., 2017).

The primary objective of Feldman's (1989) meta-analysis was to extend Cohen's analysis of the correlation between several specific dimensions of the evaluation of the teacher's instruction. The four dimensions most correlated with academic achievement were, in this order, preparation and organization, clarity and understandableness, perceived outcome, and teacher's stimulation of interest in the course and its subject matter. Feldman's (1989) results showed that the correlation between preparation and organization, the dimension most strongly correlated with academic achievement, ranged from 0.36 to 0.57. However, this meta-analysis did not account for the size of individual studies, so the moderate to high correlations may be an artifact of small-study effects.

The objectives of Clayson's (2009) meta-analysis were to address situational questions and methodological questions. Criteria for including studies were related to college instruction, data based on multiple sections of the same course discipline, a measure of learning common across sections, a learning measure based on actual testing results and not on student perception, and SET conducted before the students took their final exam. Overall, 17 articles were included, containing 42 studies and 1,115 sections. Considering the situational dependence of previous meta-analysis on educational and/or psychological disciplines, studies were coded according to the subject matter of study.

The raw averaged correlation coefficient between SET and academic achievement was 0.33, whereas the weighted average correlation was 0.13, using between-group-class data. When within-class individual student data were used, this correlation was found to be very close to zero (-0.03). Furthermore, their results also showed a negative relation between Z-transformed r and the size of the sample, indicating that as the number of sections increases, the value of the correlation decreases. A moderator variable was identified; the association was greater in education and liberal arts disciplines, but lower in business classes. The more control was used – for example, considering the effect of previous academic achievement – the less association was found. Clayson (2009) concluded that "a small average relationship exists between learning and the evaluations but that association is situational and not applicable to all teachers, academic disciplines, or levels of instruction" (p. 16).

One of the criticisms of Clayson's (2009) work is that the number of articles included in the previous meta-analysis by Cohen (1981) exceeded 40 articles, while Clayson used 17 articles with 42 multisection studies. In addition, Clayson's meta-analysis was based on different individual multisection studies, mixed in as if it were a multisection study (Uttl et al., 2017).

The most extensive revision work on the relationship between the results of SET and their academic achievement is the one recently carried out by Uttl et al. (2017). On the one hand, they reanalyzed the previous meta-analyses of Cohen (1981); Feldman (1989), and Clayson (2009); on the other hand, they updated the previous meta-analyses of SET/achievement correlations included in multisection studies to date.

Both in the reanalysis of the previous meta-analyses and in Uttl et al.'s (2017) meta-analysis, special attention is paid to the effects of small study size or small number of sections. Furthermore, in this study, correlations weighted by sample size were used, instead of averaged correlations. The third objective was to analyze the effects of prior achievement on the relation between SET and final achievement.

The results of the reanalysis carried out by Uttl et al. (2017) indicate that, in these studies, the moderate SET/achievement correlations are close to zero when the small-study-size effects are considered. As noted by Kulik and McKeachie (1975), large correlations usually appear with small sample sizes; more low correlations are found when larger samples are used.

In the reanalysis of Cohen's (1981) data, Uttl et al. (2017) found that the SET/achievement correlation estimated by using only studies with 30 or more sections was 0.27. The reanalysis of Cohen's (1981) data did not support Cohen's conclusion that SET explains 18–25% of academic achievement variability (mean r = 0.47); instead, Uttl et al. (2017) conclude that SET explains at best 10% of variance in academic performance.

According to Uttl et al. (2017), the reanalysis of Feldman's (1989) meta-analysis also showed that Feldman's results were dependent on small-study effects and that the specific student rating dimensions do no correlate with achievement. Similarly, the reanalysis of Clayson's (2009) work also points out that the correlations estimated were lower than reported, once the small-study effects were considered.

In the updated meta-analysis carried out by Uttl et al. (2017), the overall SET/achievement means correlation was 0.23. The values for correlations adjusted for prior achievement/ability were 0.16 and 0.25, eliminating two studies considered as outliers. In addition, when small sample bias is into account and after outliers are removed, the SET/achievement correlation was 0.08 for all correlations and −0.03 for correlations adjusted for prior ability. Thus, individual differences in knowledge, ability, and motivation influence the academic performance more than teaching ratings did.

In sum, the different analyses carried out by Uttl et al. (2017) – with the assumption of fixed and random effects, with and without prior achievement, with outliers eliminated, and considering or not considering the effect of size – found correlations that varied approximately between 0.08 and 0.30, which were significantly lower than the values found in previous studies.

The Present Study

The present study aimed to check the relationships between SET and academic achievement, starting from the knowledge offered by previous studies. This study is carried out in a different context to most previous works. It is based in the South American country Ecuador and analyzes SET in the National Polytechnic School—a higher education institution for the study of technical subjects, such as engineering, architecture, and biotechnology. If the association between SET and academic performance is situational and not applicable to all academic disciplines, appearing stronger in studies in the field of education and the liberal arts and less in other areas such as business classes (Clayson, 2009), it seems necessary to carry out new studies, focusing on technical areas different to previous studies where there are fewer studies on the subject.

Although there are no records on the beginning of the evaluation of teachers in higher education in Ecuador, this has been a widespread practice in Ecuadorian higher education institutions since the early 1980s (Pareja, 1986).

The Council of Ecuadorian Higher Education obligates the evaluation of the teaching staff of higher education institutions, both for their entry and for their promotion, in the Career and Ladder Regulations of the Professor and Researcher of the Higher Education System. Teachers' professorships may even be removed if they obtain a negative SET twice consecutively or if they obtain four negative evaluations throughout their careers (Consejo de Educación Superior [CES], 2017).

The variable prior knowledge/ability is found to be a powerful moderator of the relation between SET and academic achievement (Cohen, 1981; Clayson, 2009; Uttl et al., 2017). When prior academic achievement/ability is considered, the correlations between SET and achievement correlation decrease, even coming close to zero. The present study includes a measure of previous academic achievement and statistical procedures that adjust both measures of SET and achievement for prior student achievement. Although prior achievement is one of the variables that most influence the final achievement, this study examines whether SET makes a significant contribution to the final achievement, after the effect of previous achievement is controlled for.

An open methodological question, which seeks to address this study, is the unit of analysis. Most of the researchers in this field use the group-class average as the unit of analysis, arguing that the individual differences within the group class are eliminated and the differences between the means of the group classes, sections, or teachers (Cohen, 1981; Abrami et al., 1990; Marsh and Roche, 2000; Clayson, 2005; Richardson, 2005; Uttl et al., 2017) are clearly reflected; other researchers defend the need to account for the individual variability within the group classes (Clayson, 2007; Weinberg et al., 2009). Some studies in this field have considered both aspects separately (Clayson, 2009), but to our knowledge, none have considered the variability within and between group classes or teachers jointly. In this study, we will use methods that consider both sources of variability, the students and the group class, for multilevel analysis.

In addition, since multisection designs are the ones that offer the most valuable estimate of the relationship between SET and academic achievement, an aggregated data analysis is carried out following the procedure of a multisection design, using the data from course disciplines with two or more sections.

From this theoretical context, the following objectives were established:

(1) Correlate the individual students' teacher ratings and their academic achievement.
(2) Correlate the average of SET in the class-group means with the academic achievement means of each group class.
(3) Examine the relationship of SET with the final academic achievement, once the effect of the prior academic achievement has been controlled for, establishing the specific contribution of SET to the final academic achievement, using the group averages as the unit of analysis.
(4) Evaluate the joint contribution of the individual student and the group class evaluations of teaching to the final academic achievement, considering the previous achievement.
(5) Analyze the relationships between SET, academic achievement, and prior academic achievement, following the procedure of a multisection design, considering those course disciplines or subjects matters in which there are different sections.

MATERIALS AND METHODS

Participants

The sample included 1,538 students of the National Polytechnic School from Ecuador, enrolled in eight different faculties and schools and studying 28 different degrees. Of these students, 68.6% were male and 31.4% were female. The higher percentage of male students is representative of the population of students of polytechnic studies. The average age was 22.3 years (*SD* = 3.2). This sample was chosen from a larger sample of 6,100 students who rated the teachers during the 2016/2017 academic year. These 1,538 students attended 343 different course disciplines and were distributed into 453 class groups. Most of these course disciplines had only one class group, while 48 course disciplines had more than one class group or section (with 776 students in total). The number of sections ranged from 2 to 10, with a total of 158 sections across different course disciplines. The total number of students in the different sections was 776. The teachers' sample consisted of 310 teachers, who represented a varied sample in terms of age, category, and teaching experience. More than half of these teachers were male (62.8%).

Measures

Student evaluation of teaching was obtained from the "Cuestionario de Evaluación de la Enseñanza del Profesor de la Escuela Politécnica Nacional del Ecuador" (Teacher Evaluation Questionnaire of the National Polytechnic School), approved by the teacher staff for the 2016/2017 academic year. The scale consisted of 33 items grouped theoretically into four factors: planning, mastery, and clarity in the explanation of the subject; methodology and resources; teacher – student relationship; and evaluation.

The results of the validation of this questionnaire in a large sample of 6,100 students (Sánchez et al., 2019) showed the permanence of these four theoretical factors in an exploratory factor analysis, with a high reliability of internal consistency – Cronbach's alpha ranged between 0.94 and 0.86 and was 0.96 for the total scale. The results also show a high correlation between the four factors (0.78–0.88).

Two measures of student academic achievement were taken: previous academic achievement and academic achievement at the end of the semester. Previous accumulated achievements were a measure of the mean academic achievement reached by students on all previous subject matters, among those who were enrolled until the beginning of the current semester. This measure was obtained from computerized administrative records. Although strictly it cannot be considered a measure of prior performance in the particular subject matter, it can be seen as being indicative of the general knowledge or ability with which the student begins the study of the subject.

The measure of academic achievement at the end of the semester was operationalized by grades awarded by the teacher, based on a final exam, consisting of theoretical and practical written examinations. These final exams in some cases were the same across sections and in others were different for different sections. The different sections follow the same program and have the same assessment criteria. These criteria are specified in the study program of each course. There are also common general rules for all exams in the Polytechnic School. The measures of previous accumulated academic achievement and the final grades ranged from 0 to 40 for all courses.

Students' age and gender as well as teachers' age, gender, and experience were collected from administrative records.

Procedure

The data were collected from the existing computer records in the administration of the Polytechnic School and permission was granted for access to the records by the academic staff of the institution. The data provided by the institution were anonymous, with an identification code for each student.

The application of the SET scale was carried out at the end of the semester, before the students knew their final grades. All teachers were evaluated by the students in the same term. All students had to evaluate the teachers to be able to access their final grades. The SET was made through an electronic platform, in which the data were recorded.

The impact of faculty procedures of SET on response rates has been studied by several authors, especially focusing on electronic evaluations. A high response rate is important, which in the field of evaluation in higher education is estimated at 70% (Richardson, 2005). Young et al. (2019) found that the number of responses was significantly higher when students had time in class to complete the evaluation of teaching compared to the electronic form of administration. When the response rate in electronic administration was lower than that with paper-and-pencil questionnaires, this work followed the procedure of forcing all students to answer the evaluation survey in order to access their final grades. This procedure has proved useful and valid in some higher education institutions (Leung and Kember, 2005; Nair and Adams, 2009).

Data Analysis

The data analysis was performed according to the design and goals of this research.

On the one hand, average class group was employed as a unit of analysis; on the other, the individual data of the students were analyzed.

When the class-group average was employed as the unit of analysis, a correlation analysis and a hierarchical regression analysis were performed. Correlation analysis was calculated with Pearson's product–moment correlation technique. The linear hierarchical multiple regression analysis included, in the first step, prior academic achievements and, in the second step, SET. This methodological approach establishes the specific contribution of a variable, which enters last in the analysis, to the prediction of the dependent variable – in this case, the academic achievement at the end of the semester. In addition, the extra amount of variance accounted for in the final academic achievement by SET can be estimated (Cohen and Cohen, 1983).

A multilevel path analysis was performed on the individual data, grouped into sections. This analysis accounts jointly for the variability among individual students within the class groups (level 1) and the variability between groups, taught by different teachers (level 2). A path analysis is established in which the influence of previous academic achievement on the final academic achievement and on SET is examined and in which the relation of SET with the final academic achievement is also included. All variables were observed; no latent variable was defined.

The program used was the structural equation modeling (EQS) by Bentler (2005). Parameter estimation was conducted on the basis of maximum likelihood (ML); ML estimation is based on the characteristics of multivariate normality that are used to produce optimal estimates of the population parameters, and thus, it requires relatively large sample sizes. Implementation of a diversity of fit indices is recommended when evaluating the model fit, including chi-square, chi-square relative to the degree of freedom, standardized root mean square residual (SRMR), root mean square error of approximation (RMSEA), and the comparative fit index (CFI) (Hu and Bentler, 1999).

The analysis of grouped data, although it may be considered more appropriate than the analysis of individual data (Cohen, 1981), raises some important methodological questions. An analysis of class groups mixing different course disciplines or subject matter and sections of the same courses raises questions about the validity of correlation coefficients estimated from a pooling of heterogeneous microarray data (Hassler and Thadewald, 2003; Almeida-de-Macedo et al., 2013). The effect of heterogeneous variance–covariances across a pool of data causes less efficient estimates of Pearson correlation coefficients across groups than does the approach of combining correlation coefficients of individual groups.

To overcome this question, an aggregated data analysis is carried out following the procedure of a multisection design, using the data from course disciplines with two or more sections. To consider the small-sample bias effect, correlations weighted by simple size were used.

RESULTS

The results presented are divided into two sections – those related to the aggregated data and those related to individual data – that consider the hierarchical nature of the data for the multilevel path analysis.

Average Group as Unit of Analysis

The data of the 1,538 students were averaged across the 453 class groups, from the same or different course disciplines.

Table 1 shows correlations between the mean group prior academic achievement, the mean group SET, and the mean group final academic achievement.

As **Table 1** shows, statistically significant correlations between mean prior academic achievement and mean final academic achievement were identified, as well as between mean SET ratings and final academic achievement. Prior academic achievement was not statistically correlated with SET.

To determine the specific contribution of SET on final academic achievement, a hierarchical multiple regression analysis was performed, in which independence of residuals was estimated (Durbin–Watson = 2.02).

A hierarchical linear regression analysis (see **Table 2**) was conducted in which prior academic achievement was entered in step 1 and SET in step 2.

Model 1 was significant (R^2 = 0.27, F = 145.95), and prior academic achievement significantly predicted the final academic achievement (β = 0.52, $p < 0.001$). In the second step (model 2), SET significantly predicted final academic achievement (β = 0.26, $p < 0.001$), beyond the effect of prior academic achievement. This model explained 34% of the variance of final performance.

The change between model 1 and model 2 was statistically significant (ΔR^2 = 0.07, F = 100.42, $p < 0.001$), indicating that the specific proportion of variance in final academic achievement accounted for by SET was 7%, and it is statistically significant.

TABLE 1 | Correlations between variables with data grouped into class groups.

Variable	1	2	3	*M*	*SD*
1. Prior achievement	1			25.48	4.31
2. SET ratings	03	1		4.01	0.70
3. Final achievement	0.52**	0.28**	1	27.79	6.71

*N = 453. **p < 0.01.*

TABLE 2 | Hierarchical regression of prior academic achievement and student evaluation of teaching (SET) on academic achievement.

Variable	*B*	*SE B*	β
Step 1			
Constant	7.26	1.72	
Prior achievement	0.81	0.07	0.52**
R^2			0.27
F			145.95**
Step 2			
Constant	−2.42	1.72	
Prior achievement	0.79	0.06	0.51**
SET	2.51	0.40	0.26**
R^2			0.34
ΔR^2			0.07
ΔF			100.42**

*N = 453. **p < 0.001.*

Individual Student as Unit of Analysis

Correlations between student prior academic achievement, SET, and student final academic achievement are shown in **Table 3**.

The results of individual students were similar, although slightly lower, to those averaged by groups. Statistically significant correlations were found between individual students' prior achievements and individual students' final academic achievement, as well as between SET and final academic achievement. Prior academic achievement was not statistically correlated with SET.

As individual students were grouped into class groups, a multilevel structural equation analysis with observed variables was performed, with individual students within the section as level 1 and the difference between groups as level 2. The total student sample was 1,538, distributed into 453 class groups.

The model tested the influence of previous academic achievement on final academic achievement and SET, as well as the influence of SET on final academic achievement. **Figure 1** shows the model and results of the multilevel structural analysis.

The ML method was employed for parameter estimation. This method assumes multivariate normal distributions, although the method of ML is robust for departures from normality, especially if the sample is large and the skewness is <2 and kurtosis <7, in absolute terms (West et al., 1995) – values that are below those obtained in this work.

Once the model displayed in **Figure 1** includes relationships between all the variables, it is a saturated model in which the number of parameters to estimate is equal to the data; since it makes theoretical sense to consider the similarity of the individual (within) and section (between) parameters, the three path coefficients were constrained to be equals.

This model provided a very good fit to the data (Bentler CFI = 0.996, χ^2 = 4.89, df = 3, p = 0.18; McDonald's MFI = 0.999; SRMR = 0.020: RMSEA = 0.030) (see **Table 4**).

Furthermore, for the test of equivalence of path coefficients across levels, the EQS reported a cumulative multivariate Lagrange multiplier (LM) test (χ^2) and an incremental univariate χ^2 value, along with their probability values, for each constraint. To find non-invariant parameters across groups, the probability associated with the incremental univariate χ^2 values of <0.05

TABLE 3 | Correlations between student individual variables.

Variable	1	2	3	*M*	*SD*	Skewness	Kurtosis
1. Prior achievement	1			5.88	5.17	−0.98	1.95
2. SET ratings	0.03	1		3.99	0.70	−1.23	2.16
3. Final achievement	0.50**	0.23**	1	28.08	6.71	−1.60	0.96

*N = 1,538. **p < 0.01.*

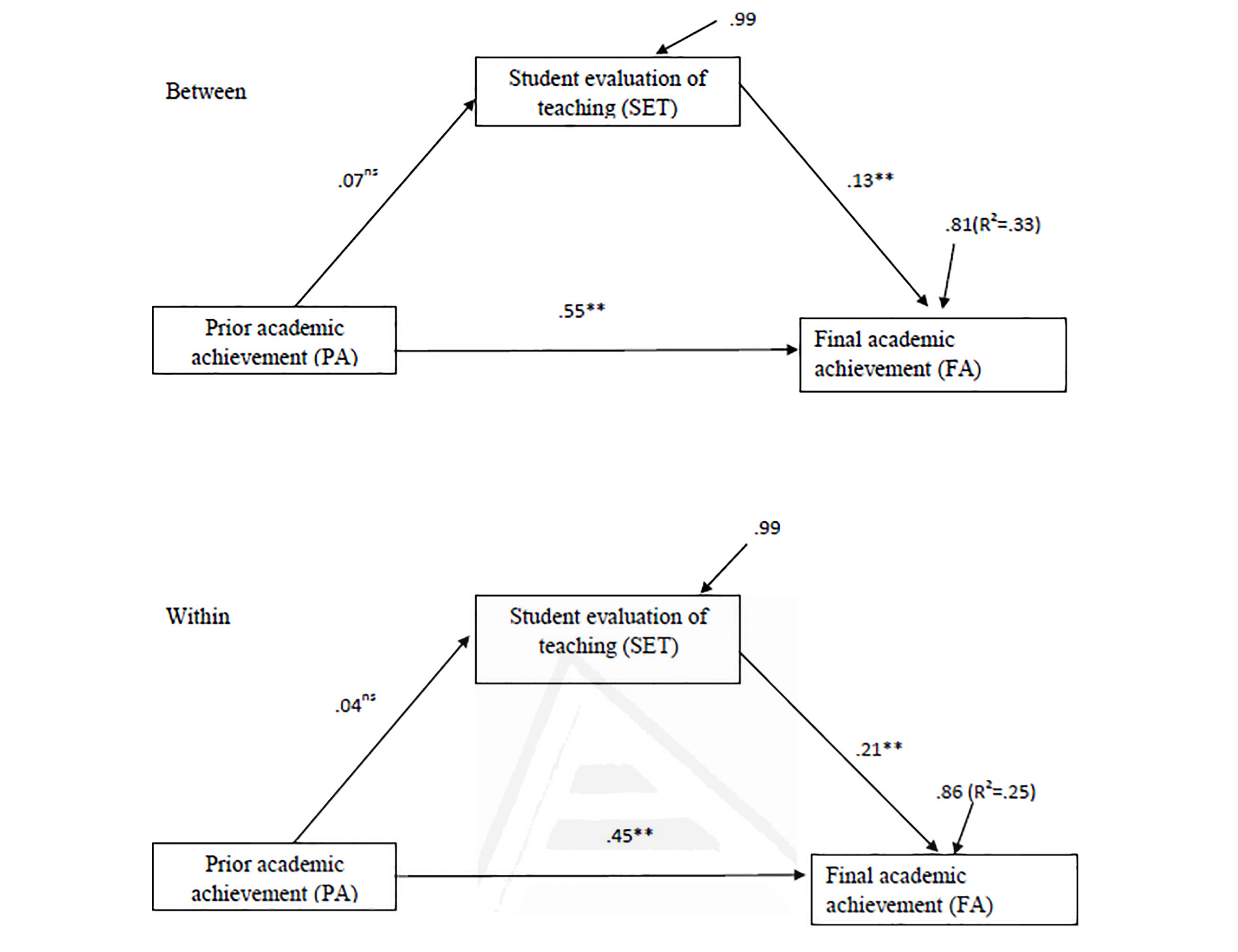

Between
Student evaluation of teaching (SET)
.99
.07ns
.13**
.81(R²=.33)
Prior academic achievement (PA)
.55**
Final academic achievement (FA)
Within
Student evaluation of teaching (SET)
.99
.04ns
.21**
.86 (R²=.25)
Prior academic achievement (PA)
.45**
Final academic achievement (FA)

FIGURE 1 | Standardized coefficients of the multilevel model with between prior academic achievement (V1), SET (V3), and academic achievement (V2). Path coefficients constrained to be equal. ***p* < 0.01; ns = not significant.

(Byrne, 2008) was checked; none of the equality constraints were significant (V3, V2, p = 0.30; V3, V4, p = 0.36; and V4, V2, p = 0.46), indicating the equivalence of the three coefficients across levels.

The relationships between the observed variables proposed in the model were significant ($p < 0.05$), except for the effects generated by prior academic achievement on SET. Both at the individual (within) and at the section levels (between), the highest regression coefficient was prior academic achievement on final academic achievement (β = 0.45, $p < 0.01$ for level 1; β = 0.55, $p < 0.01$ for level 2). SET also has an effect on final academic achievement, at both the student level (β = 0.21, $p < 0.01$) and group level (β = 0.13, $p < 0.01$). Conversely, prior academic achievement was not statistically related to SET, either at the individual level (β = 0.04, $p > 0.05$) or at the group level (β = 0.07, $p > 0.05$).

The total percentage of variance explained from the final academic achievement at the level of the students was 25%, while at the level of the sections, it was 33%.

TABLE 4 | Mean correlations between variables estimated with data grouped into sections.

Variable	1	2	3
1. Prior achievement	1		
2. SET	0.09	1	
3. Final achievement	0.16*	0.26**	1

*N = 150. *p < 0.05; **p < 0.01.*

Multisection Design Analysis

Sections' average SET, sections' average prior achievement, and sections' average academic achievement were correlated for each course discipline; then the mean of the correlations weighted by the sample size was estimated. Specifically, we transform *rs* to Fisher's *Z* scores, calculating average Fisher's *Z* scores across all course disciplines and weighing *Z*s by each sample size, and transformed average Fisher *Z* scores back to *r*. These results are shown in **Table 4**.

Again, there will be a moderate but significant correlation between the previous academic achievement and the final academic achievement, although in this case with a lower value; there was also an average positive correlation between the means of the SET and academic achievement, based on the means of each section weighted by the sample size. SET averages were not related to the prior performance.

To control for the effect of prior achievement on the relationship between SET and academic achievement, the partial correlation between the means of the sections within each course discipline of the SET and the final achievement means was estimated, considering the means of the previous achievement. Then the mean of the partial correlations weighted by the sample size was estimated. The average value of the partial correlation coefficient between SET and final achievement, considering the effect of the previous achievement, estimated in the average of the different course disciplines, was $r = 0.22$.

To examine the effect of small samples in multisection studies, the correlation between the number of sections and the absolute value of the correlation between SET and final achievement was calculated, obtaining $r = -0.18$, indicating that there is a tendency to obtain higher correlations when these correlations are based on a smaller number of sections.

DISCUSSION

This work aimed to clarify several of the issues raised about SET as a measure of teacher effectiveness. For this, a large number of individual students and group class were included; a multisection design was used when course disciplines had more than one class group; previous academic performance was considered, since the random allocation of students to the sections was not assured; and statistical methods were used which consider both the individual student variability within sections and the variability between sections. Furthermore, the study was carried out in a geographical and disciplinary context different from that of most previous studies.

The results obtained with aggregated data, taking the group class as the unit of analysis, showed a moderate but statistically significant correlation (0.28) between SET and final academic achievement. This value corresponds to the value obtained in the meta-analysis of Uttl et al. (2017) when the data of Cohen (1987) were reanalyzed considering small-sized studies and effects (i.e., only the studies with a number of 30 or more sections).

These results also showed a moderately high correlation between prior academic achievement and final academic achievement. This finding is in accordance with previous meta-analytic studies on the variables associated with achievement in higher education, in which prior knowledge/abilities appear as one of the main determinants of academic achievement (Schneider and Preckel, 2017).

However, the correlation between prior achievement and SET was not statistically significant, suggesting that SET is not affected by previous academic achievements.

Control for prior academic achievement with the hierarchical regression analysis procedure continued to show a significant effect of SET on academic achievement; this effect was around 7%, which corresponds to a correlation of 0.27, similar to that found in the reanalysis of Cohen's (1981) data, and is slightly higher than the value obtained in the meta-analysis of Uttl et al. (2017) based on nearly 100 multisection studies published to that date, which stood at 0.23.

The results obtained with the individual student data showed a statistically significant correlation (0.23) between SET and final academic achievement, which was a bit lower than that obtained with the data aggregated in sections. This result is consistent with previous studies about instructor's teaching effectiveness, in which it is considered that multisection studies that use the grouped data of the sections are more appropriate to apprehend the true relationship between SET and academic achievement (Cohen, 1981; Uttl et al., 2017).

The results of individual data showed again a moderately high correlation between prior academic achievement and final academic achievement, as well as a non-significant relation of SET with prior academic achievement.

Following the suggestion of several authors regarding these types of studies, both the individual variability within the sections and the variability between sections (Clayson, 2007; Weinberg et al., 2009) of the data of the present work included a multilevel structural equation analysis.

The results of the multilevel analysis showed that there was a significant effect of SET on the final academic achievement, at both the individual and the section levels, even after controlling the effect of prior academic achievement. In addition, the magnitude of the effect was similar in both levels. The total percentage of variance explained from the final academic achievement at the level of the sections was 33%, while at the level of the individual students, it was 25%, with 8% of the explained variance of final academic achievement attributable to the sections: that is, to the effect of the teacher.

The results obtained with aggregated data, taking the section as the unit of analysis, following the guidelines of a multisection design, show that a significant, although low, relationship remains between SET and academic achievement when the sample size effect is considered ($r = 0.26$), even when the effect of the prior academic achievement is controlled ($r = 0.22$). Therefore, the results of the individual and the group analyses do not differ substantially from the results obtained in the analysis of the sections, supporting partially the results of the individual analysis and aggregated group analysis, in which biased correlations could appear due to pooling of heterogeneous samples, when the analysis of the data is carried out following the guidelines of a multisection design.

These results were similar to those found in studies carried out in different geographical and disciplinary contexts. The study was conducted in a Higher Polytechnic School of Ecuador, which teaches scientific and technological disciplines, which are different from the humanistic and social disciplines rated in most of the studies on teaching effectiveness (Clayson, 2009).

On the basis of the large-scale datasets from Australia, Canada, and the United States (N = 26,746 students) in the Programme for International Student Assessment (PISA), 2012, Scherer et al. (2016) find support for significant relations to the educational outcomes. Students' achievement could be best predicted by perceived classroom management (β = 0.20 to 0.31).

Together, the results show the relation between SET and academic achievement, in a study where multiple sections are included, controlling previous academic achievement and considering both the student variability within sections and the variability between sections with different teachers, in subject matters of a scientific – technological nature.

However, the amount of influence of SET on academic achievement is lower than that found in some previous meta-analytic studies (Cohen, 1981; Feldman, 1989), but higher than that found in the meta-analysis of Uttl et al. (2017) carried out on the multisection studies published to that date; when small-study-size effects and prior academic achievement were considered, it was close to zero.

Although university student academic achievement depends mainly on various intellectual and non-intellectual factors (Richardson et al., 2012; Schneider and Preckel, 2017), the results of this work support the conclusion that SET has a modest, around 5%, but significant influence on academic achievement and is therefore related to teacher effectiveness.

However, taking into consideration our results and the results of previous meta-analyses, especially the comprehensive meta-analysis of Uttl et al. (2017), the influence of SET on academic achievement seems to be sufficiently limited to make relevant administrative decisions. Although use of SET as a feedback for teachers' use and as a measure of student satisfaction is not problematic (Spooren et al., 2013; Uttl et al., 2017), the use of SET as a measure of teachers' effectiveness for making administrative decisions about teachers' hiring, firing, promotions, and merit pay is controversial (Uttl et al., 2017, 2019; American Sociological Association, 2019).

Limitations

The analysis that takes into account individual student and average group as units of analyses, mixing different subject courses and sections of the same courses, raises questions about the validity of correlation coefficients estimated from pooling heterogeneous microarray data, given that it causes less efficient estimates of Pearson correlation coefficients than does the approach of combining correlation coefficients of individual groups, as is done in the analysis that follows a multisection design, although, on the other hand, and the results obtained from the multisection analysis are consistent with the individual and group analyses.

Final exams in some cases were the same across sections; however, in others, they were not identical for different sections; although different sections follow the same program and have the same assessment criteria, the exams should be identical or equivalent, as required for a multisection study.

This study uses a low number of sections, ranging from 2 to 10, which can lead to the small section size effect, given the tendency to obtain higher correlations when these correlations are based on a smaller number of sections.

Prior academic achievement in the subject matter was not measured; the measure was of the accumulated academic performance in all subject matters in which the student had been enrolled before the beginning of the semester. However, in scientific–technological disciplines, the academic achievement accumulated previously is a measure that is usually related to the final achievement, and it also seems to be an adequate measure to study the possible influence on SET.

Another question that arises in relation to this study is the procedure of obtaining the SET. Although research shows that, in general, electronic evaluation procedures are as valid as traditional procedures (Spooren et al., 2013), more research is necessary on this procedure of forcing all the students to answer the evaluations of teaching, in terms of social desirability, acquiescence, and stereotyped answers, etc. From a methodological perspective, in the path analysis, all the variables are observed variables and not latent; therefore, the measurement error could not be estimated.

DATA AVAILABILITY STATEMENT

The datasets generated for this study are available on request to the corresponding author.

ETHICS STATEMENT

Data collection was made from existing computer records of the Polytechnic School administration, and the academic staff of the institution granted access to them. The data provided by the institution were anonymous, with only one identification code for each student.

AUTHOR CONTRIBUTIONS

TS: theoretical review of the study. RG-C: quantitative methods and theoretical review of the study. J-LC and JL: quantitative methods. JV: data collection and review of the references.

FUNDING

This research was supported by National Secretariat of Higher Education, Science and Technology, SENESCYT (PIC-18-INE-EPN-002).

REFERENCES

Abrami, P. C., d'Appolonia, S., and Cohen, P. A. (1990). Validity of student ratings of instruction: what we know and what we do not. *J. Educat. Psychol.* 82, 219–231. doi: 10.1007/s10459-017-9783-0

Almeida-de-Macedo, M., Ransom, N., Feng, Y., Hurst, J., and Wurtele, E. S. (2013). Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. *BMC Bioinformatics* 14:214. doi: 10.1186/1471-2105-14-214

American Sociological Association, (2019). *Statement on Student Evaluations of teaching*. Avaliable at: https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_sept52019.pdf (Septembre 9, 2019).

Bentler, P. M. (2005). *EQS 6 Structural Equations Program Manual*. Encino, CA: Multivariate Software.

Byrne, B. M. (2008). *Structural Equation Modeling with EQS: Basic Concepts, Applications and Programming*. London: Routledge.

Clayson, D. E. (2005). Within-class variability in student-teacher evaluations: example and problems. Decision. *Sci. J. Inno. Educ.* 3, 109–124. doi: 10.1111/j.1540-4609.2005.00055.x

Clayson, D. E. (2007). Conceptual and statistical problems of using between class data in educational research. *J. Mark. Educ.* 27, 122–129. doi: 10.1002/mono.12060

Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn? A meta-analysis and review of the literature. *J. Mark. Educ.* 31, 16–30. doi: 10.1177/0273475308324086

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies. *Rev. Educ. Res.* 51, 281–309. doi: 10.3102/00346543051003281

Cohen, P. A. (1987). "A Critical analysis and reanalysis of the multisection validity meta-analysis," in *Paper Presented at the Annual Meeting of the American Educational research Association*, Washington, DC). doi: 10.3102/00346543051003281

Consejo de Educación Superior [CES], (2017). *Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior. [Career and Ladder Regulations of the Professor and Researcher of the Higher Education System]*. Avaliable at: https://procuraduria.utpl.edu.ec/sitios/documentos/NormativasPublicas/Reglamento%20de%20Carrera%20y%20Escalaf%C3%B3n%20del%20Profesor%20e%20Investigador%20del%20Sistema%20de%20Educaci%C3%B3n%20Superior%202018.pdf (20 April 2019).

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. *Res. Higher Educ.* 30, 583–645. doi: 10.1007/bf00992392

Hassler, U., and Thadewald, T. (2003). Nonsensical and biased correlation due to pooling heterogeneous samples. *Statistician* 52, 367–379. doi: 10.1111/1467-9884.00365

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model.* 6, 1–55. doi: 10.1080/10705519909540118

Husbands, C. T., and Fosh, P. (1993). Students' evaluation of teaching in higher education: experiences from four european countries and some implications of the practice. *Assess. Eval. Higher Educ.* 18, 95–114. doi: 10.1080/0260293930180202

Huybers, T. (2014). Student evaluation of teaching: the use of best–worst scaling. *Assess. Eval. Higher Educ.* 39, 496–513. doi: 10.1080/02602938.2013.851782

Kulik, J. A., and McKeachie, W. J. (1975). "The evaluation of teachers in higher education," in *Review of Research in Education*, Vol. 3, ed. F. N. Kerlinger, (Itasca: Peacock), 201–240.

Leung, D. Y. P., and Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the Internet. *Res. Higher Educ.* 46, 571–591. doi: 10.1007/s11162-005-3365-3

Leventhal, L. (1975). Teacher rating forms: critique and reformulation of previous validation designs. *Can. Psychol. Rev.* 16, 269–276. doi: 10.1037/h0081814

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *J. Educ. Psychol.* 99, 775–790. doi: 10.1037/0022-0663.99.4.775

Marsh, H. W., and Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria. *J. Educ. Psychol.* 72, 468–475. doi: 10.1037/0022-0663.72.4.468

Marsh, H. W., and Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *J. Educ. Psychol.* 92, 202–228. doi: 10.1037/0022-0663.92.1.202

Nair, C. S., and Adams, P. (2009). Survey PLATFORM: a factor influencing online survey delivery and response rate. *Q. Higher Educ.* 15, 291–296. doi: 10.1080/13538320903399091

Pareja, F. (1986). *La Educación Superior en el Ecuador [The higher education in Ecuador]*. Caracas: Regional Center For Higher Education in Latin America And the Caribbean (CRESALC)-UNESCO.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assess. Eval. Higher Educ.* 30, 387–415. doi: 10.1080/02602930500099193

Richardson, M., Abraham, C., and Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review and meta-analysis. *Psychol. Bull.* 138, 353–377.

Sánchez, T., Sandoval, I., Salazar, D., Gilar, R., and Castejón, J. L. (2019). "Validation of the teacher evaluation questionnaire of the National Polytechnic School, applying the method of factor analysis with extraction of principal components," in *17th LACCEI International Conference for Engineering, Education, and Technology, 24-26 July 2019*, Jamaica.

Scherer, R., Nilsen, T., and Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Front. Psychol.* 7:110. doi: 10.3389/fpsyg.2016.00110

Schneider, M., and Preckel, F. (2017). Variables associated with achievement in higher education: a systematic review of meta-analyses. *Psychol. Bul.* 43, 565–600. doi: 10.1037/bul0000098

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. *Rev. Educ. Res.* 83, 598–642. doi: 10.3102/0034654313496870

Uttl, B., Cnude, K., and White, C. A. (2019). Conflict of interest explain the size of student evaluation of teaching and learning correlations in multidrction studies: a meta-analysis. *PeerJ* 7, e7225. doi: 10.7717/peerj.7225

Uttl, B., White, C. A., and Gonzalez, D. W. (2017). Meta-analysis of faculty's teaching effectiveness: student evaluation of teaching ratings and student learning are not related. *Stud. Educ. Eval.* 54, 22–42. doi: 10.1016/j.stueduc.2016.08.007

Weinberg, B. A., Hashimoto, M., and Fleisher, B. M. (2009). Evaluating teaching in higher education. *J. Econ. Educ.* 40227– 61,

West, S. G., Finch, J. F., and Curran, P. J. (1995). "Structural equation models with non-normal variables," in *Structural Equation Modeling: Concepts, Issues, and Applications*, ed. R. H. Hoyle, (Thousands, CA: Sage), 56–75.

Young, K., Joines, J., Standish, T., and Gallagher, V. (2019). Student evaluations of teaching: the impact of faculty procedures on response rates. *Assess. Eval. Higher Educ.* 44, 37–49. doi: 10.1186/s12909-015-0387-1

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. *Teach. Higher Educ.* 12, 55–76. doi: 10.1080/13562510601102131

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Sanchez, T., Naranjo, D., Vidal, J., Salazar, D., Pérez, C., & Jaramillo, M. (2021). Analysis of academic performance based on sociograms: A case study with students from at-risk groups. *Journal of Technology and Science Education*, *11*(1), 167-179. https://doi.org/10.3926/jotse.1110

OmniaScience

Journal of Technology and Science Education
JOTSE, 2021 – 11(1): 167-179 – Online ISSN: 2013-6374 – Print ISSN: 2014-5349
https://doi.org/10.3926/jotse.1110

ANALYSIS OF ACADEMIC PERFORMANCE BASED ON SOCIOGRAMS: A CASE STUDY WITH STUDENTS FROM AT-RISK GROUPS

Tarquino Sanchez, David Naranjo, Jack Vidal, Diego Salazar, Cristina Pérez, Marianela Jaramillo

Escuela Politécnica Nacional (Ecuador)

tarquino.sanchez@epn.edu.ec, david.naranjo@epn.edu.ec, vidal.edu.ec@gmail.com, diegovdj@gmail.com
cristina.perez@epn.edu.ec, marijaramillo20@yahoo.es

Received September 2020
Accepted November 2020

Abstract

The present work analyzes the academic performance of students from at-risk groups from the perspective of Social Network Analysis (SNA), studying the academic and interaction information of 45 students belonging to at-risk groups who attended a pilot socio-academic course during one academic term. This information was used to create a sociogram, which served as the basis for determining the centrality metrics of the SNA. The relationships between these metrics and the academic variables were then studied by means of correlation analysis and linear regression with LASSO standardization. As a preview of the results, it was determined that the academic performance of the students in the pilot course was influenced, on the one hand, by their academic knowledge prior to being admitted to the university, represented by the score on the Mathematics and Geometry section of the diagnostic test, and on the other hand, by the dynamics of the social network in which they interacted in the classroom, represented by the eigenvector centrality. These results have significant potential for explaining the academic performance according to SNA metrics, and they provide evidence to support the implementation of practices that promote a healthy social environment in an academic context.

Keywords – Social network analysis, Sociogram, Academic performance.

To cite this article:

Sanchez, T., Naranjo, D., Vidal, J., Salazar, D., Pérez, C., & Jaramillo, M. (2021). Analysis of academic performance based on sociograms: A case study with students from at-risk groups. *Journal of Technology and Science Education,* 11(1), 167-179. https://doi.org/10.3926/jotse.1110

----

1. Introduction

Education is a crucial and determining factor in the development of a nation, as it has a direct influence on the progress of both people and societies. The educational process provides the competences that allow progress to be made in different areas of society, such as culture, productivity and economic competitiveness, scientific and technological innovation and processes intended to ensure higher levels of social well-being. As a result, all educational processes seek to optimize, constantly and permanently, the performance of students in the different stages of their academic training (Bhardwaj, 2016; Smith, Fraser, Chykina, Ikoma, Levitan et al., 2017).

In this sense, one of the primary indicators in any analysis concerning education is academic performance, in which two basic components are considered: the learning process and evaluation (García, Lamos-Duarte, Vargas-Rivera, Camargo-Villalba & Capacho, 2019). According to Pizarro (1985), academic performance is a measurable indicator of the responding or indicative skills that allows us to estimate what an individual has learned during a certain educational process. Along the same lines, Caballero D., Abello and Palacio (2007) emphasize that academic performance, besides being the result of institutionalized training, also includes aspects of non-institutionalized training, and thus the presence of other underlying components of this indicator is relevant. In this context, Navarro (2003) states that when academic performance is studied, other factors that might influence it are analyzed on par with it. These factors encompass the environments of the educational system itself, as well as the teaching methodologies that are applied, the previous knowledge of the students and the amplitude of the curricula, among other factors. However, other components are also considered, such as socioeconomic factors, motivational aspects, emotional factors and social skills.

In light of the fact that academic performance is inherently multifactorial in nature, several different focuses can in turn be proposed to analyze it and to study the factors that converge in this complex component of the educational process.

1.1. Previous Works

Traditionally, academic performance at the university level has been associated with factors such as previous academic preparation, access to technology and scores on entrance exams (Callejas, Griol & Lázaro-Álvarez, 2020; Goodchild & Bjørkestøl, 2020; Ismail, Mahmood & Abdelmaboud, 2018). Likewise, socioeconomic variables have also been studied, such as monthly family income, residence in rural areas, housing type and gender, as determining factors in academic performance.

Parallel to this, the influence of student-student and student-teacher interaction has been studied with a qualitative focus (Amo & Santelices, 2017; Sandoval, Sánchez, Velasteguí & Naranjo, 2018; Yang & Tang, 2003; Sánchez, Gilar-Corbi, Castejón, Vidal & León, 2020). However, in spite of the fact that very interesting results have been obtained in qualitative studies in terms of the influence of social skills on academic performance, the importance of interactions in the classroom has become especially relevant from the perspective of social network analysis (SNA), the methodology of which is based on building a sociogram to establish the relationships between the actors and to analyze the structures that result from the recurrence of these relationships (Abbasi, Altmann & Hossain, 2011; Jain & Langer, 2014).

On the sociogram, each of the actors in the social network is represented by a node, and each of the interactions that exist between these actors is shown by segments that link said nodes. Based on the sociogram, different metrics are determined to quantify the properties at the whole network and node levels, which are known as centrality metrics (Gomes Jr., 2019). Among these metrics is degree centrality, which represents the number of nodes that are directly related to the node in question; closeness centrality, which represents the capacity of a node to reach the remaining nodes on the network by following the shortest distances that separate them (geodesic path); betweenness centrality, which represents the number of times that a node is found on a geodesic path; and eigenvector centrality, which represents the relevance of a node on the network, based on the node degree values and, recursively, on the relevance of its neighboring nodes (Abbasi et al., 2011; Rizzuto, Ledoux & Hatala, 2009).

In addition, among the metrics that quantify the properties at the network level is density, which represents the number of real relationships between the nodes of a network with regard to the total number of possible relationships; and the average degree, which represents the average number of relationships for each of the nodes on a network (Abbasi et al., 2011; Rizzuto et al., 2009).

Different studies have found that the structural properties of the social networks are associated not only with aspects exclusively concerning socialization, such as the formation of friendship bonds or the creation of a sense of belonging to an institution, but also with metrics such as density. Centrality indicators can be analyzed in conjunction with the variables traditionally considered in the study of academic performance and, consequently, be directly and substantially linked to phenomena inherent to the educational system, such as dropping out and academic performance (De-Marcos, Garciá-López, Garciá-Cabot, Medina-Merodio, Domínguez, Martínez-Herraíz et al., 2016; Gomes Jr., 2019; Helal, Li, Liu, Ebrahimie, Dawson, Murray et al., 2018; Jain & Langer, 2014; Mihaly, 2011).

Likewise, it has been found that the structure of a social network in a group of students can be considered an explanatory component, which can be harnessed to varying degrees of performance in a formal instruction process in an educational setting (Helal et al., 2018; Jain & Langer, 2014).

Parallel to this, the inclusion of SNA metrics in the analysis of academic performance has made it possible to indirectly associate the influence of the so-called soft skills, which are the qualities that people develop in order to improve their relationships with peers; these are learned through daily experience, and they have a great impact on a personal, professional, labor and social level (Patacsil & Tablatin, 2017).

To summarize, SNA can be used to generate numeric indicators that describe some of the characteristics of the behavior of the actors of a social network, as well as the structure of the network itself. These indicators can be linked to other variables that describe a phenomenon of interest, such as academic performance, in order to determine correlations and thus obtain measurable results based on an abstract concept, which in this case, comes from social interaction.

1.2. The Present Study

The present study proposes to analyze, from the perspective of SNA, the influence of social interactions on the academic performance of a group of students belonging to at-risk sectors. This research is developed within a context that is relatively different from that of other previous works, as it analyzes the information on students of similar socioeconomic and academic characteristics, known in Ecuador as beneficiaries of the Quota Policy. The students who make up this group are selected by the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación in Ecuador (SENESCYT, according to its acronym in Spanish) during each admissions process.

Since 2014, the SENESCYT has conducted a socioeconomic characterization process on applicants to institutions of higher education (IHE) with the aim of identifying those who are most at-risk socioeconomically and promoting their access to higher education through an affirmative action policy known as the Quota Policy. Students benefiting from this program are given beneficial access to 15% of the total positions offered by institutions of higher education in Ecuador (Sandoval et al., 2018; Sandoval, Sánchez, Naranjo & Jiménez, 2019).

Starting in 2017, the SENESCYT has included beneficiaries of the Quota Policy among the students given preferential access to the Leveling Course of the Escuela Politécnica Nacional (EPN). However, when comparing them to their peers from the general population, it was evident that the students from the at-risk groups had certain deficiencies in their previous academic preparation, which led to high failure and dropout rates. At the same time, it was observed that these students showed high levels of demotivation, disorganization and had difficulties interacting in a classroom setting with both classmates and their instructors (Ramos, Sánchez, Reina & Franco-Crespo, 2020; Sandoval, Sánchez, Naranjo et al., 2019). In response to these precedents, the EPN implemented a pilot socio-academic intervention course during the second term of 2019 with a sample of beneficiaries of the Quota Policy. This course was focused on giving students from at-risk groups the academic skills they need and developing their abilities for classroom interaction and participation.

Previous works on this group of students have investigated the relationship between academic achievement and traditionally studied variables, such as socioeconomic factors and academic background (Sandoval, Sánchez, Velasteguí, et al., 2018; Sandoval-Palis, Naranjo, Vidal & Gilar-Corbi, 2020). However, this work also proposed the study of the influence of social interactions on this phenomenon and applied a regression technique based on automatic learning in order to determine the set of variables with the greatest explanatory potential. It should be stressed that in the context of this study, the component of interest that represents academic performance concerns evaluation, which is expressed in the form of scores that measure student performance (García et al., 2019).

The precedents presented suggest the possibility of studying the relationship between social interactions and academic performance, one component of the multifactorial educational process, and thus the following objectives have been established:

(1) To determine in a quantifiable manner the interaction between students in the pilot socio-academic intervention course through the calculation of SNA metrics.

(2) To analyze the correlation between the variables that represent social interaction and those representing academic performance.

(3) To establish the set of variables, both academic and SNA-related, which have the greatest influence on academic performance.

2. Materials and Methods

This research is framed within a descriptive and correlative case study approach.

2.1. Participants

A total of 257 students benefiting from the Quota Policy were admitted to the EPN during the second term of 2019, of which 70% were male and 30% were female. 69% of the students applied to the leveling course for the Engineering, Science and Administrative Sciences programs, while the remaining 31% applied to the leveling course for the Advanced Technology program .63.6% came from the province of Pichincha, 22.9% came from other provinces in the Andes Region, 10.3% came from the provinces of the Amazon Region and the remaining 3.2% came from the provinces of the Coastal Region. The students were given a diagnostic test that evaluated their knowledge and skills in two sections: Mathematics and Geometry and Language and Communication. The average score out of 10 possible points obtained on each section was 4.28 ± 1.9090 and 4.70 ± 1.4975, respectively.

For the selection of participants in the pilot socio-academic invention course, an orientation workshop was held in which the students were informed about the option to participate in said program for one term prior to the EPN leveling course. Given that the pilot socio-academic intervention course was not mandatory and that it was necessary for it to have a similar number of students as a parallel ordinary leveling course, a sample of 45 students were selected from among those who agreed to participate voluntarily in the program. An attempt was made to maintain the populational proportions with regard to gender, type of leveling course (Engineering, Sciences and Administrative Sciences or Advanced Technology) and province of origin. Likewise, through an inferential analysis, it was determined that the average of scores of the 45 students on each of the sections of the diagnostic test showed no statistically significant differences from the corresponding populational averages. Table 1 shows the characteristics of the study participants.

Variable	Distribution
Gender	65% Male 35% Female
Type of Leveling Course	69% Engineering, Sciences and Administrative Sciences 31% Advanced Technology
Province of Origin	65% Pichincha 35% Others
Score on the Mathematics-Geometry Diagnostic Exam	4.52 ± 1.5842
Score on the Language and Communication Diagnostic Exam	5.15 ± 1.1162

Table 1. Characteristics of the study participants

2.2. Measurements

The previous knowledge and skills in Mathematics, Geometry and Language and Communication were evaluated according to a maximum of 10 points on a diagnostic test consisting of 80 multiple choice questions. The test was designed and validated by professionals from the Basic Science Department of the EPN (Sandoval, Sánchez, Velasteguí, et al., 2018; Sandoval, Sánchez, Naranjo et al., 2019).

The final score out of 10 possible points is the simple average of the scores obtained by the students following the culmination of the pilot socio-academic intervention course in the Mathematics, Geometry and Reading/Writing subjects. In turn, the scores on each of the subjects were made up by grades on homework, quizzes, a mid-term exam and another final exam at the end of the term.

The SNA metrics were determined based on an interactive survey administered to the students at the end of the term.

2.3. Procedure

Information on gender, the score on each of the sections of the diagnostic test and the final score were obtained from the academic records from the pilot intervention course. In this study, only the sum of anonymous information is presented. Students provided informed consent to voluntarily participate in this study and they were told that their information would be used for research purposes.

The 45 students who made up the sample of this study attended a pilot socio-academic intervention course for one term. During this period, they received classes in Mathematics (10 hours/week), Geometry (10 hours/week), and Reading and Writing (10 hours/week). They also received instruction on the use of computer tools (2 hours/week), study techniques and strategies (2 hours/week), and motivational workshops and coaching (2 hours/week). In the latter, the students explored different aspects concerning the development of their soft skills and emotional intelligence, as well as how they could apply these skills in the teaching-learning process. Parallel to the training process, they were continuously monitored by the social work department created exclusively to work with these students.

The interaction survey was administered by means of an electronic form sent to the students by email. This form consisted of a list with the names of all 45 students, with the instructions to select the two people with whom they had interacted the most in the academic setting (performing tasks, classroom work and study groups) during their experience in the course.

2.4. Data Analysis

With the results of the interactive survey, a sociogram was created to represent the social network of the pilot intervention course. Since the intent was to study the potential of the interactions between students rather than their direction, all relationships were considered to be symmetric (undirected) and binary (Newman, 2003).

The density, average score and the degree centrality (DC), closeness centrality (CC), betweenness centrality (BC) and eigenvector centrality (EC) metrics were determined based on the sociogram. Next, they were integrated into a SNA metrics matrix with information on the academic performance of the students. Those missing interaction or performance information (as the result of having dropped out of the pilot course) were excluded from the set of data.

T tests and correlation analyses were performed on the variables studied, and finally, a linear regression was estimated to determine the variables with the greatest influence on the academic performance of the students after having finished the pilot course.

The linear regression model was constructed according to the following equation (1):

$$FS = \beta_0 + \beta_1\, Gender + \beta_2\, MDS + \beta_3\, LDS + \beta_4\, DC + \beta_5\, BC + \beta_6\, CC + \beta_7\, EC \qquad (1),$$

Where:

FS: final score,

β_i: regression coefficients,

Gender: student's gender,

MDS: score on the Mathematics-Geometry diagnostic exam,

LDS: score on the Language and Communication diagnostic exam,

DC: degree centrality,

BC: betweenness centrality,

CC: closeness centrality,

EC: eigenvector centrality.

The regression coefficients β_i were determined by means of the *Least Absolute Shrinkage and Selection Operator (*LASSO*)* automatic learning standardization. This technique was used to obtain a series of iterations which provided, first of all, the reduction of the variability of the estimates by reducing the regression coefficients, and parallel to this, the selection of variables to build a simplified model, since some coefficients are reduced to zero (Gauraha, 2018).

All analyses were carried out on RStudio version 1.2.1335.

3. Results

The mean of the final score obtained by students was 5.33, with a standard deviation of 1.5052. Two students were excluded from the analysis because they dropped out of the course before it ended.

Figure 1 shows the sociogram of the pilot course. Each student is represented by a node, the size of which is proportional to the final score. It is observed that apparently both men and women are distributed in a relatively homogeneous manner throughout the network; however, it can be seen that some of the nodes representing men form subnetworks with an interconnection that is relatively greater than that shown for women. On the other hand, it is observed that a large part of the smaller nodes (corresponding to lower final scores) are located in peripheral areas, while the larger nodes (higher final scores) are found in areas where the network has greater cohesion. This is a preliminary indicator of the potential relationship that exists between academic performance and the interaction among students.

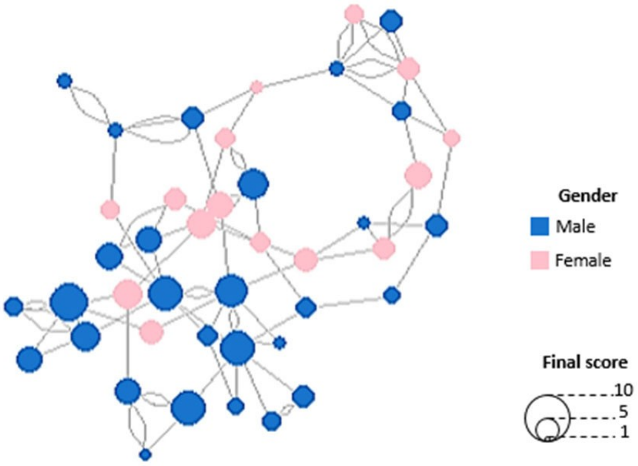


Figure 1. Sociogram of the pilot course

Table 2 shows that the density of the network is 0.094. Furthermore, the value of the mean score was 3.95, indicating that the students in the pilot course had social interactions with approximately 4 classmates in the academic setting.

Parameter	Value
Number of nodes	43
Number of relationships	85
Density	0.094
Mean score	3.95

Table 2. SNA metrics on a network level

Table 3 shows a summary of the SNA metrics on the node level obtained from the sociogram of the pilot course. In order to determine the potential relationships between these metrics and the other variables, a t test for equal means was first carried out between women and men; these results are presented in Table 4.

Metric	Mean	Minimum	Maximum	Standard Deviation
Degree centrality	3.95	2.00	9.00	1.9018
Closeness centrality	6.76×10^{-3}	4.98×10^{-3}	9.17×10^{-3}	9.917×10^{-4}
Betweenness centrality	54.56	0.00	237.84	59.1329
Eigenvector centrality	0.25	0.03	1.00	0.2415

Table 3. SNA metrics on a node level

It is observed that for a level of significance of 0.05, only the eigenvector centrality is statistically different between women and men. This means that the relevance of men in order to establish relationships on the pilot course network was greater than that of women. This agrees with what was previously observed on the sociogram, which showed subnetworks with greater interconnection formed by men.

Variable	Difference of means (Women-Men)	p Value
Final score (FS)	-0.116	0.788
Score on the Mathematics-Geometry Diagnostic Exam (MDS)	-0.579	0.251
Score on the Language and Communication Diagnostic Exam (LDS)	0.0870	0.832
Degree centrality (DC)	-0.441	0.426
Betweenness centrality (BC)	12.848	0.453
Closeness centrality (CC)	1.19×10^{-4}	0.711
Eigenvector centrality (EC)	-0.150	0.013

Table 4. Results of the t test for equal means between women and men in the pilot course

On the other hand, Table 5 shows the correlation matrix considering both the SNA metrics and the academic variables. It is observed that, for a 0.01 level of significance, both the diagnostic score in Mathematics and Geometry (MDS) and the eigenvector centrality (EC) show a moderate correlation with the final score (FS). Likewise, a moderate correlation is observed between EC and the remaining SNA metrics, and a strong correlation is revealed between closeness centrality (CC) and betweenness centrality (BC).

Table 6 shows the results of the regression coefficients by means of LASSO standardization. It is observed that the simplified final score model (FS) consists of the score from the Mathematics and Geometry diagnostic test (MDS) and the eigenvector centrality (EC), as shown in Equation 2; likewise, the coefficients associated with each of these variables are significant at a 0.001 and 0.01 level, respectively.

Variable	FS	MDS	LDS	DC	BC	CC	EC
FS	1.00						
MDS	0.666**	1.00					
LDS	0.214	0.280	1.00				
DC	0.467**	0.357*	0.222	1.00			
BC	0.408**	0.293	0.153	0.622**	1.00		
CC	0.487**	0.185	0.094	0.398**	0.764**	1.00	
EC	0.596**	0.573**	0.134	0.615**	0.551**	0.618**	1.00

**The correlation is significant at the 0.01 level (2-tailed).
The correlation is significant at the 0.05 level (2-tailed).

Table 5. Correlation matrix

Variable	Coefficient	p Value
(Intercept)	3.266	2.68×10^{-7}
Gender	0.000	-
MDS	0.408	2.85×10^{-3}
LDS	0.000	-
DC	0.000	-
BC	0.000	-
CC	0.000	-
EC	1.996	1.82×10^{-2}

Table 6. Regression coefficients by means of LASSO standardization

$$FS = 3{,}266 + 0{,}408\, MDS + 1{,}996\, EC \tag{2}$$

Figure 2 shows the dispersion diagram of the final score (FS) according to the Mathematics and Geometry diagnostic score (MDS); the size of the dots is proportional to the eigenvector centrality.

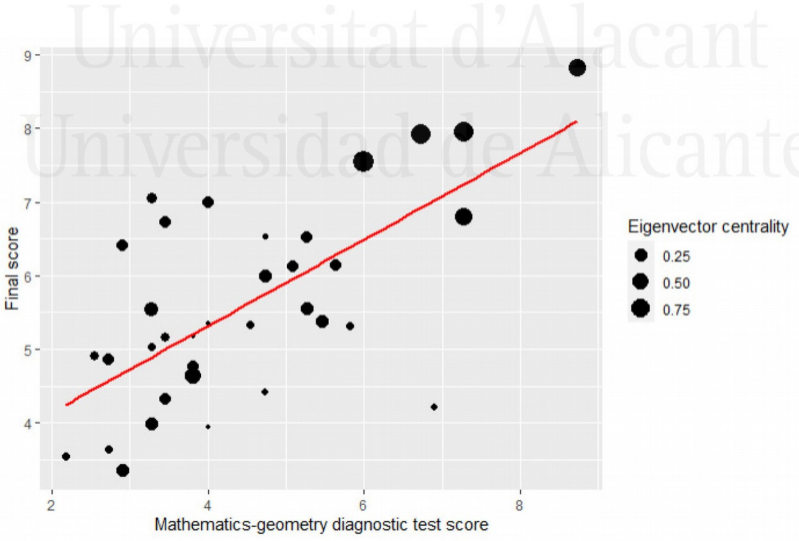


Figure 2. Final score according to the Mathematics and Geometry diagnostic test score and eigenvector centrality

4. Discussion

The present study intended to analyze the academic performance of students belonging to at-risk groups based on the SNA metrics obtained from the sociogram.

The apparently low resulting value for density could be explained by the fact that the students selected only two classmates with whom they interacted the most, which results in a systematic reduction of the total number of real relationships on the network. However, it is likely that having allowed the students to select the number of classmates with whom they most interacted without any restrictions would have provided a similar result, since as Ramírez Ortiz, Caballero Hoyos and Ramírez López (2004) explain, the interactions among students are not expected to be completely homogeneous, since students develop preferences for working with certain classmates over the course of the academic term. This study also analyzed the information on students from at-risk groups and the network densities fell within a range of 0.095-0.169. Even so, the previous values only serve as a guide, since in order for them to be comparable to the results obtained in the present study, the number of nodes on the network must be the same in both cases (Mihaly, 2011; Ramírez Ortiz et a.l 2004; Rizzuto et al., 2009).

Furthermore, the fact that the interactions observed for men demonstrate greater relevance than those of women could be a consequence of the fact that, even today, collaborative work by women is often limited by an environment in which gender bias is present. This behavior has been observed by Jain & Langer (2014), who studied the relevance of social interactions on academic performance.

With regard to the correlation analysis, the results of the present study are similar to those obtained by Gomes Jr. (2019) and Mihaly (2011), who found strong and moderate correlations between the eigenvector centrality value and the mean final score obtained by the students. This result, in turn, suggests that the students with the best connections tend to earn the best grades.

Similarly, some SNA studies, such as those conducted by Abbasi et al. (2011) and Rizzuto et al. (2009), have found that the correlation between the SNA metrics is determined, on the one hand, by the very definition of a particular metric, and on the other hand, by the specific structure of the network. As a result, contrasting results between different students can be obtained. Along these lines, Gomes Jr. (2019), for example, found a weak correlation between closeness centrality (CC) and betweenness centrality (BC), while in the present study, a strong correlation was observed between these two metrics, which indicates that those students who acted as intermediaries among their classmates (greater BC) have a greater potential to interact with classmates with whom they did not have a direct relationship (greater CC). On the other hand, as in this study, De-Marcos et al. (2016) observed a moderate correlation between eigenvector centrality and the other centrality metrics, which is explained by the inherent recursiveness that defines the eigenvector centrality.

The linear regression model showed a clear relationship, on the one hand, between the academic performance of the students in the pilot course and their prior knowledge, and on the other hand, between academic performance and the power of the students' social connections. This result is in agreement with what was mentioned by Sawyer (2013), in the sense that the skills and knowledge students have when they enter the university are reliable predictors of academic performance, especially during the first year; however, an increasing number of studies like this one include variables that describe the social capital, considered from the perspective of Sociology.

In the case of this study, the variable describing the social capital in the model is the eigenvector centrality. This metric has higher values at the nodes that are linked to highly connected nodes (recursively). In this sense, it can be established that the resulting relationship between social capital and academic performance is, in fact, a property of the complex social dynamics represented on a sociogram (Pulgar, Candia & Leonardi, 2020; Ramírez Ortiz et al., 2004).

This complexity is evidenced by studying the distribution of the final score according to the eigenvector centrality, for example. If we consider a low score to be one less than 5 points and an acceptable score to be between 6 and 7 points, Figure 2 shows that there are groups of students with a low final score and an acceptable final score, even though they have similar eigenvector centrality values. This behavior could be due to the fact that, in the case of students with a low final score, their performance was negatively influenced by their academic background, in spite of having established academic bonds with their classmates. Meanwhile, in the case of students with an acceptable final score, in spite of their inadequate academic preparation, having built relationships with their classmates allowed them to perform better than expected in the pilot course. Nevertheless, it is also reasonable to expect that there are additional factors that have an influence on their classroom performance, and since information is not available to determine the underlying reasons, it cannot be ruled out that these data show an antagonistic behavior, even if this procedure would strengthen the correlation.

On the other hand, while the optimization of an academic performance model goes beyond the scope of this study, the adjusted coefficient of determination for the model presented was 0.5016; this value is comparable to the results obtained by De-Marcos et al. (2016), Gomes Jr. (2019) and Mihaly (2011), who have modeled academic performance with SNA metrics; for this reason, the results from this study were considered to have significant explanatory potential, given that the variables considered merely represent a small subset of the factors that influence academic performance.

5. Conclusions

Previous works have established a certain degree of association between student friendship circles and academic performance. In this study, these associations and correlations were explored with a more precise and measurable focus, through the metrics resulting from the sociogram of a pilot socio-academic intervention course for students from at-risk groups. Using a linear regression model, it was determined that, in addition to academic knowledge prior to university admission, the academic performance of a student is influenced by the dynamics of the social network on which they interact within the classroom.

The prior academic knowledge of the students was described solely by the score on the Mathematics and Geometry section of the diagnostic test; in this sense, the score on the Language and Communication section of the diagnostic test did not have a great effect on the score earned by students at the end of the pilot course. Furthermore, the dynamics of the social network were described by the eigenvector centrality value.

The linear model, constructed based on the score on the Mathematics and Geometry section of the diagnostic test, and the eigenvector centrality, has significant explanatory potential, since in general, the academic performance of a student is influenced by many more factors, such as social, emotional and economic aspects, which in turn have an influence on classroom interaction.

This study could serve as the basis for guiding educators at institutions of higher education in evaluating the role of social interactions on academic performance, since the results obtained provide evidence to support the implementation of practices that promote a healthy social environment in an academic context.

Despite the fact that metrics were determined in the SNA which describe social interactions, their dynamic nature could not be expressed over time, which represents one limitation of this study. Similarly, the study sample is relatively small and is subject to the limitations of a case study, particularly with regard to the voluntary decision to participate in the pilot socio-academic intervention course. It will be necessary to conduct further studies with a larger sample that also consider how social interactions change over time, in order to obtain much more general results and to apply models that evaluate the fixed effects inherent to complex social phenomena.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

This research was supported by the Secretaría Nacional de Educación Superior, Ciencia, Tecnología e Innovación (SENESCYT; PIC-18-INE-EPN-002). This paper is part of Tarquino Sánchez's Ph.D thesis.

References

Abbasi, A., Altmann, J., & Hossain, L. (2011). Identifying the effects of co-authorship networks on the performance of scholars: A correlation and regression analysis of performance measures and social network analysis measures. *Journal of Informetrics,* 5(4), 594-607. https://doi.org/10.1016/j.joi.2011.05.007

Amo, C., & Santelices, M.V. (2017). Trayectorias universitarias: más que persistencia o deserción. *Congresos CLABES*. http://revistas.utp.ac.pa/index.php/clabes/article/view/1676/2412

Bhardwaj, A. (2016). Importance of Education in Human Life: a Holistic Approach. *International Journal of Science and Consciousness,* 2(2), 23-28. www.ijsc.net

Caballero, C.C., Abello, R.Ll., & Palacio, 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. https://www.redalyc.org/articulo.oa?id=799/79925207

Callejas, Z., Griol, D., & Lázaro-Álvarez, N. (2020). Predicting Computer Engineering Students' Dropout in Cuban Higher Education with Pre-Enrollment and early performance data. *Journal of Technology and Science Education,* 10(2), 241-258.

De-Marcos, L., Garciá-López, E., Garciá-Cabot, A., Medina-Merodio, J.A., Domínguez, A., Martínez-Herraíz, J.J., et al. (2016). Social network analysis of a gamified e-learning course: Small-world phenomenon and network metrics as predictors of academic performance. *Computers in Human Behavior,* 60, 312-321. https://doi.org/10.1016/j.chb.2016.02.052

García, M.I.B., Lamos-Duarte, A.F., Vargas-Rivera, O.I., Camargo-Villalba, G.E., & Capacho, N.S. (2019). Learning approaches, academic performance and related factors; in students that curve last year of the programs of the faculty of health sciences. *Educacion Medica,* 20, 10-17. https://doi.org/10.1016/j.edumed.2017.11.008

Gauraha, N. (2018). Introduction to the LASSO. *Resonance,* 23(4), 439-464. https://doi.org/10.1007/s12045-018-0635-x

Gomes Jr., L. (2019). In-class social networks and academic performance: how good connections can improve grades. *Anais do XXXIV Simpósio Brasileiro de Banco de Dados* (25-36). https://doi.org/10.5753/sbbd.2019.8805

Goodchild, S., & Bjørkestøl, K. (2020). Assessing First-Year Engineering Students' Pre-University Mathematics Knowledge: Preliminary Validity Results. *Journal of Technology and Science Education,* 10(2), 259-270.

Helal, S., Li, J., Liu, L., Ebrahimie, E., Dawson, S., Murray, D.J., et al. (2018). Predicting academic performance by considering student heterogeneity. *Knowledge-Based Systems,* 161, 134-146. https://doi.org/10.1016/j.knosys.2018.07.042

Ismail, A.O.A., Mahmood, A.K., & Abdelmaboud, A. (2018). Factors influencing academic performance of students in blended and traditional domains. *International Journal of Emerging Technologies in Learning,* 13(2), 170-187. https://doi.org/10.3991/ijet.v13i02.8031

Jain, T., & Langer, N. (2014). Does Who You Know Matter? Unraveling the Influence of Student Networks on Academic Performance. *SSRN Electronic Journal.* https://doi.org/10.2139/ssrn.2425477

Mihaly, K. (2011). Do More Friends Mean Better Grades?: Student Popularity and Academic Achievement. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.1371883

Navarro, R. (2003). El rendimiento académico: concepto, investigación y desarrollo. *REICE. Revista Iberoamericana sobre Calidad, Eficacia y Cambio en Educación,* 1(2). https://revistas.uam.es/index.php/reice/article/view/5354

Newman, M., (2003). The Structure and Function of Complex Networks. *Society for Industrial and Applied Mathematics Review,* 45(2), 167-256. https://doi.org/10.1137/S003614450342480

Patacsil, F.F., & Tablatin, C.L.S. (2017). Exploring the importance of soft and hard skills as perceived by it internship students and industry: A gap analysis. *Journal of Technology and Science Education,* 7(3), 347-368. https://doi.org/10.3926/jotse.271

Pizarro, R. (1985). *Rasgos y Actitudes del Profesor Efectivo. Tesis para optar al Grado de Magister en Ciencias de la Educación.* Pontificia Universidad Católica de Chile.

Pulgar, J., Candia, C., & Leonardi, P.M. (2020). Social networks and academic performance in physics: Undergraduate cooperation enhances ill-structured problem elaboration and inhibits well-structured problem solving. *Physical Review Physics Education Research,* 16(1), 10137. https://doi.org/10.1103/physrevphyseducres.16.010137

Ramírez Ortiz, M.G., Caballero Hoyos, J.R., & Ramírez López, M.G. (2004). The social networks of academic performance in a student context of poverty in Mexico. *Social Networks,* 26(2), 175-188. https://doi.org/10.1016/j.socnet.2004.01.010

Ramos, V., Sánchez, T., Reina, J. & Franco-Crespo A. (2020). Differences Between Vulnerable and Non-Vulnerable Students Regarding the Psychological Abilities and Self-Control Skills within the Development of Learning. *INTED2020 Proceedings* (8388-8394). https://doi.org/10.21125/inted.2020.2282

Rizzuto, T.E., Ledoux, J., & Hatala, J.P. (2009). It's not just what you know, it's who you know: Testing a model of the relative importance of social networks to academic performance. *Social Psychology of Education,* 12(2), 175-189. https://doi.org/10.1007/s11218-008-9080-0

Sánchez, T., Gilar-Corbi, R., Castejón, J.L., Vidal, J., & León, J. (2020). Students' Evaluation of Teaching and Their Academic Achievement in a Higher Education Institution of Ecuador. *Frontiers in Psychology,* 11, 1-10. https://doi.org/10.3389/fpsyg.2020.00233

Sandoval, I., Sánchez, T., Velasteguí, V., & Naranjo, D. (2018). Factores Asociados Al Abandono En Estudiantes De Grupos Vulnerables. Caso Escuela Politécnica Nacional. *Congresos CLABES* (132-141). https://revistas.utp.ac.pa/index.php/clabes/article/view/1907

Sandoval, I., Sánchez, T., Naranjo, D., & Jiménez, A. (2019). Proposal of a mathematics pilot program for engineering students from vulnerable groups of Escuela Politécnica Nacional. Proceedings of the *LACCEI International Multi-Conference for Engineering, Education and Technology.* https://doi.org/10.18687/LACCEI2019.1.1.387

Sandoval-Palis, I., Naranjo, D., Vidal, J. & Gilar-Corbi, R. (2020). Early Dropout Prediction Model: A Case Study of University Leveling Course Students. *Sustainability,* 12(22), 1-17. https://doi.org/10.3390/su12229314

Sawyer, R. (2013). Beyond Correlations: Usefulness of High School GPA and Test Scores in Making College Admissions Decisions. *Applied Measurement in Education,* 26(2), 89-112. https://doi.org/10.1080/08957347.2013.765433

Smith, W.C., Fraser, P., Chykina, V., Ikoma, S., Levitan, J., Liu, J., et al. (2017). Global citizenship and the importance of education in a globally integrated world. *Globalisation, Societies and Education,* 15(5), 648-665. https://doi.org/10.1080/14767724.2016.1222896

Yang, H.L., & Tang, J.H. (2003). Effects of social network on students' performance: A web-based forum study in Taiwan. *Journal of Asynchronous Learning Network,* 7(3), 93-107. https://doi.org/10.24059/olj.v7i3.1848

Published by OmniaScience (www.omniascience.com)

Journal of Technology and Science Education, 2021 (www.jotse.org)

8. DISCUSIÓN

Con el propósito de verificar que se cumplan los objetivos planteados en esta tesis doctoral, se plantea a continuación una sección de Discusión de los hallazgos encontrados en los artículos científicos en el mismo orden presentado en la sección Resultados.

8.1 Validación del cuestionario de evaluación docente en la Escuela Politécnica Nacional, aplicando el método de Análisis Factorial con extracción de componentes principales

Los primeros objetivos del presente trabajo fueron analizar la validez de constructo del cuestionario de evaluación de la enseñanza/aprendizaje. El Análisis Factorial reveló que la escala estaba compuesta por dos factores. Sin embargo, cuando el Análisis Factorial se vio obligado a ampliar a 4 factores, la estructura teórica del cuestionario inicial se reprodujo exactamente. El segundo objetivo del presente trabajo fue proponer una reducción del cuestionario de evaluación docente. Es difícil reducir un cuestionario manteniendo los aspectos fundamentales de la docencia. Sin embargo, si el objetivo es reducir aún más el cuestionario original de 33 ítems para condensarlo en 14 ítems, por ejemplo, se recomienda eliminar el ítem 2, que cubre el aspecto Planificación, así como los ítems 14 y 15 que hacen referencia a la metodología de calificación. Además, sería óptimo eliminar el ítem 16.

Estos cambios se proponen teniendo en cuenta que el cuestionario mantendría el margen de confiabilidad deseado. Los ítems con saturaciones más altas son los que mejor definen el factor, mientras que los ítems con saturaciones bajas definen el factor con menor precisión. En base a esto, para el cuestionario original de 33 ítems, el Factor 1 tiene un alto nivel de saturación -dentro del rango de 0.780 a 0.689 y determina una relación positiva profesor-alumno, así como un buen ambiente de aprendizaje. De igual forma, el Factor 2 tiene un alto nivel de saturación y describe la planificación, el dominio y la claridad en la explicación del tema, dejando el resto con bajos niveles de saturación. Con base en pruebas de confiabilidad con coeficientes de consistencia interna de Cronbach (α) y prueba de esfericidad de Bartlett, se concluye que los dos tipos de análisis sobre la relevancia y validez de los datos de la matriz se verifican satisfactoriamente, lo que significa que los datos de la matriz original son confiables. Además, todas las preguntas tienen información relevante para el análisis de comunalidades. Los resultados obtenidos satisfacen todos los objetivos establecidos en este trabajo de investigación y ofrecen una propuesta de herramienta para la evaluación de los estudiantes del profesorado de la universidad, a partir de la opinión de profesores y estudiantes. El aporte en esta sección que pretende hacer este trabajo es presentar un instrumento disponible para ser utilizado por universidades y escuelas politécnicas, especialmente en la Escuela Politécnica Nacional, para validar y reducir los cuestionarios de evaluación docente. Los resultados positivos de este estudio confirman que es posible entrar en una nueva etapa de evaluación docente mediante una nueva y bien definida encuesta. Una limitación del estudio es que no se siguió el supuesto de aleatoriedad para el Análisis Factorial, porque las preguntas no están dispuestas en un orden aleatorio. Por otro lado, otra limitación sería que se examinó la validez de constructo, pero en menor medida la validez de criterio, por ejemplo, correlacionando las puntuaciones del cuestionario con algún criterio externo. Además del análisis de validación del instrumento de evaluación docente, se recomienda realizar un análisis multidimensional que incluya aspectos de género, expediente académico, puntaje en exámenes de admisión, asignaturas, titulaciones, entre otros, con el fin de relacionar las puntuaciones en las escalas con otras variables y sus correlaciones.

8.2 Validación de una escala corta para la evaluación de estudiantes de Enseñanza de calificaciones en una Institución de Educación Superior Politécnica.

Los resultados muestran claramente la validez estructural de la encuesta de evaluación docente por parte de los estudiantes elaborada en la Escuela Politécnica Nacional del Ecuador. Dado que el segundo objetivo general de este estudio es proponer una escala corta que muestre confiabilidad y validez, se utilizaron AFC y ESEM.

Los resultados mostraron un modelo multidimensional con cuatro factores altamente correlacionados que no excluyen un factor general, con un excelente ajuste a los datos, tanto en la escala larga como en la versión corta. La estructura con el mejor ajuste fue el ESEM de cuatro factores de dos factores; sin embargo, las cargas factoriales sobre el factor global fueron bajas (Howard et al., 2018) y, por lo tanto, se mantuvo la estructura ESEM de cuatro factores.

Con base en una muestra de 26 746 estudiantes que tomaron el Programa Internacional de Evaluación de los Estudiantes (PISA) del 2012, Scherer, Nilsen y Jansen (2016) encontraron que el modelado de ecuaciones estructurales exploratorias de dos factores superaba a los enfoques alternativos con respecto al ajuste del modelo.

Los investigadores dividen con base a la existencia de un factor general de segundo orden (Abrami, et al, 1997; Cheung, 2000) o diferentes factores correlacionados de primer orden (Marsh, 1991b, 2007a). En cuanto a las implicaciones prácticas de este tema, quizás la conclusión más precisa sea la que ya proporcionó el propio Marsh (1991a) en 1991: "He elegido un término medio recomendando el uso de dimensiones específicas y calificaciones globales" (p. 419).

Aunque se acepta que las escalas de calificación de la evaluación de la enseñanza por parte de los estudiantes son multidimensionales, muchos investigadores defienden el uso de puntajes globales únicos (Apodaca y Grad, 2005). Por ello, aún reconociendo la estructura multidimensional y jerárquica de las dimensiones evaluadas en las escalas de evaluación de la docencia por parte de los estudiantes, muchos trabajos que estudian este tema utilizan puntajes globales; en tanto, la retroalimentación brindada a los docentes para la mejora de la práctica docente incluye un perfil de las puntuaciones en las distintas dimensiones, que muestran las fortalezas y debilidades de los métodos de cada docente.

Dada la existencia de sesgo de género de los estudiantes en la evaluación de la enseñanza por parte de los estudiantes, se probó la invarianza de medición de género configuracional, métrica y escalar. Investigaciones anteriores han demostrado que es probable que las mujeres obtengan puntuaciones más altas en las calificaciones de SET (por ejemplo, Badri et al., 2006; Darby 2006). Bonitz (2011) encontró que las variaciones de género en los puntajes SET podrían deberse a variaciones de género en rasgos como la amabilidad que se correlacionan con los puntajes SET. Sin embargo, los resultados de este estudio mostraron invariancia de medición de género configuracional, métrica y escalar en el contexto de estudios científico-tecnológicos.

Los resultados de este trabajo también muestran la validez concurrente de la escala reducida, que mostró una alta correlación con la escala total de 32 ítems. La correlación corregida de Levy y el índice de Gower revelaron una alta concurrencia entre ambas formas, con valores superiores a .90. Estos resultados son ligeramente superiores a los obtenidos en otros estudios que también mostraron un alto grado de concordancia entre las formas largas y cortas de dichas escalas (Gogol et al., 2014, Lafontaine et al., 2016).

Los altos valores de los coeficientes de confiabilidad, estimados según los supuestos del modelo SEM, también llaman la atención para las escalas y subescalas completas largas y cortas. Estos valores fueron superiores a .90 y alcanzaron valores de .98 y .97 para todas las escalas. Se aplicó el procedimiento *Congeneric Correlated Factors* (Cho, 2016) considerando que existen diferentes cargas factoriales para obtener los valores de los coeficientes de confiabilidad multidimensional además del alfa de Cronbach, lo que supone que todas las cargas factoriales son iguales (es decir, equivalentes de tau), y por lo tanto subestima la fiabilidad.

Por otro lado, los resultados también mostraron correlaciones significativas y moderadas entre las versiones larga y corta de la escala con el rendimiento académico, tomando datos individuales y agregados en clases o secciones.

La evidencia en apoyo de las evaluaciones de la enseñanza por parte de los estudiantes como medida de la efectividad de la instrucción de los maestros proviene de estudios que muestran correlaciones entre las evaluaciones de los estudiantes y el rendimiento de estos mismo estudiantes, una prueba sólida para la validez de criterio (Cohen, 1981; Clayson, 2009).

En conjunto, los resultados demostraron las buenas cualidades psicométricas del cuestionario de evaluación docente de la Escuela Politécnica Nacional y su validez de constructo y criterio, así como su alta confiabilidad. Además, los índices psicométricos de la versión corta de esta escala sugieren la posibilidad de desarrollar escalas cortas de tres o cuatro ítems igualmente fiables y válidos.

Además, las relaciones obtenidas entre las versiones larga y corta del con el rendimiento académico tienen implicaciones prácticas para la docencia docente. Este instrumento puede ayudar a los profesores a adaptar su enseñanza a las necesidades y preferencias de los estudiantes en el contexto de las características específicas de los estudios politécnicos. Sin embargo, no debemos perder de vista la abierta controversia entre las percepciones de los estudiantes sobre la calidad de la enseñanza, o las percepciones de la inclinación, y su aprendizaje real. En el contexto de la STEM – Enseñanza de la Ciencia, Tecnología, Ingeniería y Matemáticas - Deslauriers, McCarty, Miller, Callaghana y Kestin (2019) encuentran que los estudiantes en aulas activas aprendieron más, pero su percepción del aprendizaje fue menor que la de sus compañeros en la instrucción pasiva.

Respecto a las limitaciones de este estudio y posibles estudios futuros, dado que los formularios largo y corto se administraron como parte de la escala completa, y a pesar de la corrección de Levy y Gower para el cálculo de la correlación entre las dos versiones, sería necesario administrar las escalas larga y corta a la misma muestra de forma independiente. Además, sería conveniente examinar la estructura factorial de la escala corta en una muestra representativa independiente de estudiantes. En este estudio analizamos la relación con el rendimiento académico, podría ser de interés explorar la relación con el compromiso en la educación superior (Vizoso, Rodríguez, & Arias-Gundín, 2018) o el conocimiento pedagógico general (Klemenz, König, & Schaper. 2019). Finalmente, la obtención de datos longitudinales en la misma y diferentes muestras de la Escuela Politécnica Nacional podría servir para profundizar en la validez de la escala desarrollada en este trabajo.

También debe tenerse en cuenta que estos resultados se han obtenido en una sola institución, lo que limita la generalidad de los resultados; sin embargo, es la mayor institución de estudios politécnicos (ciencia, biotecnología, ingeniería, etc.), la más grande del Ecuador que reúne a estudiantes de todo el país.

8.3 Evaluación de la docencia y rendimiento académico de los estudiantes en una Institución de Educación Superior del Ecuador.

Este trabajo tuvo como objetivo aclarar varias de las cuestiones planteadas sobre el SET como medida de la eficacia docente. Para ello se incluyó un gran número de alumnos individuales y grupos de clase; se utilizó un diseño multisección cuando las disciplinas del curso tenían más de un grupo de clase; se consideró el desempeño académico previo, ya que no se aseguró la asignación aleatoria de los estudiantes a las secciones; y se utilizaron métodos estadísticos que consideran tanto la variabilidad de cada alumno dentro de las secciones como la variabilidad entre las secciones.

Los resultados obtenidos con datos agregados, tomando la clase grupal como unidad de análisis, mostraron una correlación moderada pero estadísticamente significativa (0.28) entre SET y logro académico final. Este valor corresponde al valor obtenido en el metaanálisis de Uttl et al. (2017) cuando se volvieron a analizar los datos de Cohen (1987) considerando estudios y efectos de pequeño tamaño (es decir, solo los estudios con un número de al menos 30 secciones).

Estos resultados también mostraron una correlación moderadamente alta entre el rendimiento académico previo y el rendimiento académico final. Este hallazgo concuerda con estudios meta-analíticos previos sobre las variables asociadas al rendimiento en la educación superior, en los que los conocimientos/habilidades previas aparecen como uno de los principales determinantes del rendimiento académico (Schneider y Preckel, 2017).

Sin embargo, la correlación entre el rendimiento previo y el SET no fue estadísticamente significativa, lo que sugiere que el SET no se ve afectado por los logros académicos anteriores.

El control del rendimiento académico previo con el procedimiento de análisis de regresión jerárquica continuó mostrando un efecto significativo de SET sobre el rendimiento académico; este efecto fue de alrededor del 7%, lo que corresponde a una correlación de 0.27, similar a la encontrada en el reanálisis de los datos de Cohen (1981), y es ligeramente superior al valor obtenido en el metaanálisis de Uttl et al. (2017) en base a cerca de 100 estudios multisectoriales publicados hasta esa fecha, que se situaban en 0,23.

Los resultados obtenidos con los datos individuales de los estudiantes mostraron una correlación estadísticamente significativa (0.23) entre el SET y el rendimiento académico final, que fue un poco menor que la obtenida con los datos agregados por secciones. Este resultado es consistente con estudios previos sobre la efectividad de la enseñanza del profesor, en los cuales se considera que los estudios multisección que utilizan los datos agrupados de las secciones son más apropiados para explicar la verdadera relación entre SET y rendimiento académico (Cohen, 1981; Uttl et al., 2017).

Los resultados de los datos individuales mostraron nuevamente una correlación moderadamente alta entre el rendimiento académico previo y el rendimiento académico final, así como una relación no significativa de SET con el rendimiento académico previo.

Siguiendo la sugerencia de varios autores sobre este tipo de estudios, tanto la variabilidad individual dentro de las secciones como la variabilidad entre secciones (Clayson, 2007; Weinberg et al., 2009) de los datos del presente trabajo incluyeron un análisis de ecuaciones estructurales multinivel.

Los resultados del análisis multinivel mostraron que hubo un efecto significativo de SET en el rendimiento académico final, tanto a nivel individual como de sección, incluso después de controlar el efecto del rendimiento académico previo. Además, la magnitud del efecto fue similar en ambos niveles. El porcentaje total de varianza explicada del rendimiento académico final a nivel de las secciones fue del 33%, mientras que a nivel de los estudiantes individuales fue del 25%, con un 8% de la varianza explicada del rendimiento académico final atribuible a las secciones: es decir, al efecto del maestro.

Los resultados obtenidos con datos agregados, tomando la sección como unidad de análisis, siguiendo los lineamientos de un diseño multisección, muestran que se mantiene una relación significativa, aunque baja, entre el EET y el rendimiento académico cuando se considera el efecto tamaño muestral (r = 0.26), incluso cuando se controla el efecto del rendimiento académico previo (r = 0.22). Por tanto, los resultados de los análisis individuales y grupales no difieren sustancialmente de los resultados obtenidos en el análisis de las secciones, apoyando parcialmente los resultados del análisis individual y grupal agregado, en los que podrían aparecer correlaciones sesgadas por el agrupamiento de muestras heterogéneos, cuando el análisis de los datos se realiza siguiendo las pautas de un diseño multisección.

Estos resultados fueron similares a los encontrados en estudios realizados en diferentes contextos geográficos y disciplinarios. El estudio se realizó en la Escuela Politécnica Nacional del Ecuador, que imparte disciplinas científicas y tecnológicas, las cuales son distintas a las disciplinas humanísticas y sociales calificadas en la mayoría de los estudios sobre efectividad docente (Clayson, 2009).

Sobre la base de los conjuntos de datos a gran escala de Australia, Canadá y los Estados Unidos (N = 26 746 estudiantes) en el Programa Internacional para la Evaluación de los Estudiantes (PISA), 2012, Scherer et al. (2016) se encontró apoyo para las relaciones significativas con los resultados educativos. El rendimiento de los estudiantes se puede predecir mejor mediante la gestión percibida en el aula (b = 0.20 a 0.31).

En conjunto, los resultados muestran la relación entre SET y rendimiento académico, en un estudio donde se incluyen múltiples secciones, controlando el rendimiento académico previo y considerando tanto la variabilidad de los estudiantes dentro de las secciones como la variabilidad entre las secciones con diferentes profesores, en disciplinas de carácter científico- tecnológica.

Sin embargo, la influencia de SET sobre el rendimiento académico es menor que la encontrada en algunos estudios meta-analíticos previos (Cohen, 1981; Feldman, 1989), pero mayor que la encontrada en el metaanálisis de Uttl et al. (2017) realizado sobre los estudios multisección publicados hasta esa fecha; cuando se consideraron los efectos de estudios pequeños y el rendimiento académico previo, fue cercano a cero.

Si bien el rendimiento académico de los estudiantes universitarios depende principalmente de varios factores intelectuales y no intelectuales (Richardson et al., 2012; Schneider y Preckel, 2017), los resultados de este trabajo apoyan la conclusión de que la SET tiene un modesto, alrededor del 5%, pero significativo factor que influye en el rendimiento académico y, por lo tanto, se relaciona con la eficacia docente.

Sin embargo, teniendo en cuenta nuestros resultados y los resultados de metaanálisis anteriores, especialmente el de Uttl et al. (2017), la influencia de SET con el rendimiento académico parece ser suficientemente limitada para tomar decisiones administrativas relevantes. Aunque el uso de SET como una retroalimentación para uso de los profesores y como una medida de la satisfacción de los estudiantes no representa una dificultad (Spooren et al., 2013; Uttl et al., 2017). Además, el uso de SET como una medida de la efectividad de los profesores para tomar decisiones administrativas sobre la contratación de profesores, promociones y pago por mérito es controvertido Uttl et al., 2017, 2019; American Sociological Association, 2019).

9. CONCLUSIONES

El primer objetivo del presente trabajo fue analizar la validez de constructo del cuestionario de enseñanza-aprendizaje. El Análisis Factorial reveló que la escala estaba compuesta por dos factores. Sin embargo, cuando el Análisis Factorial se vio obligado a reducir a 4 factores, la estructura teórica del cuestionario inicial se reprodujo exactamente.

Los ítems con saturaciones más altas son los que mejor definen el factor, mientras que los ítems con saturaciones bajas definen el factor con menor precisión. En base a esto, para el cuestionario original de 33 ítems, el Factor 1 tiene un alto nivel de saturación dentro del rango de 0.780 a 0.689 y determina una Relación profesor/estudiante, así como un buen ambiente de aprendizaje. De igual forma, el Factor 2 tiene un alto nivel de saturación y describe la *Planificación, el dominio y la claridad en la explicación del tema*, dejando el resto con bajos niveles de saturación.

Con base en las pruebas de confiabilidad con coeficientes de consistencia interna de alfa (α) de Cronbach y la prueba de esfericidad de Bartlett, se concluye que los dos tipos de análisis tanto la relevancia como la validez de los datos de la matriz original se verifican satisfactoriamente, lo que significa que los datos son confiables. Además, todas las preguntas tienen información relevante para el análisis de comunalidades.

Los resultados iniciales obtenidos satisfacen los objetivos 1 y 2 establecidos en este trabajo de investigación y ofrecen una propuesta válida para la evaluación del profesorado por parte de los estudiantes de una universidad, a partir de la opinión de profesores y estudiantes. El aporte que se pretende hacer es presentar un instrumento disponible para ser utilizado por universidades y escuelas politécnicas, especialmente en la Escuela Politécnica Nacional, para validar y reducir los cuestionarios de evaluación docente. Los resultados positivos de este estudio confirman que es posible entrar en una nueva etapa de evaluación docente utilizando una encuesta bien definida.

Una limitación del estudio es que no se siguió el supuesto de aleatoriedad para el Análisis Factorial, porque las preguntas no están dispuestas en un orden aleatorio. Otra limitación, es que se examinó la validez de constructo, pero no el criterio de validez, por ejemplo, correlacionando las puntuaciones del cuestionario con algún criterio externo.

Parte del segundo objetivo del presente trabajo fue proponer una reducción del cuestionario de evaluación docente. Es difícil reducir un cuestionario manteniendo los aspectos fundamentales de la docencia. Sin embargo, si el objetivo es reducir aún más el cuestionario para condensarlo en 14 ítems, por ejemplo, se recomienda eliminar de la escala propuesta en el primer artículo científico, de esta tesis doctoral, el ítem 2, que cubre el aspecto Planificación, así como los ítems 14 y 15 que hacen referencia a la Metodología de Calificación y Evaluación Global del profesor. Además, sería óptimo eliminar el ítem 16. Estos cambios se proponen teniendo en cuenta que el cuestionario mantendría el margen de confiabilidad deseado (ver Anexo 1).

Sobre las correlaciones entre las escalas larga y corta, los resultados muestran la validez concurrente de la escala reducida de 14 ítems, que mostró una alta correlación con la escala total de 32 ítems. La correlación corregida de Levy y el índice de Gower revelaron una alta concurrencia entre ambas formas, con valores superiores a .90.

Los altos valores de los coeficientes de confiabilidad, estimados según los supuestos del modelo utilizado SEM, también llaman la atención para las escalas y subescalas completas largas y cortas. Estos valores fueron superiores a .90 y alcanzaron valores de .98 y .97 para todas las escalas. Se aplicó además del alfa (α) de Cronbach el procedimiento de Factores Congenéricos Correlacionados (Cho, 2016) considerando que existen diferentes cargas factoriales para obtener los valores de los coeficientes de confiabilidad multidimensional, lo que supone que todas las cargas factoriales son iguales (es decir, equivalentes tau), y por lo tanto subestima la fiabilidad.

Los resultados obtenidos con datos agregados, tomando la sección como unidad de análisis, mostraron una correlación moderada y estadísticamente significativa (.26) entre las calificaciones de los estudiantes y el desempeño final. Este resultado se espera de los estudios de efectividad docente de los instructores, en los que se considera que los estudios multisectoriales son más apropiados para aprehender la verdadera relación entre las evaluaciones de los estudiantes sobre la docencia y el desempeño académico (Cohen, 1981; Uttl et al., 2017).

Dada la existencia de sesgo de género de los estudiantes en la evaluación de la enseñanza por parte de los estudiantes, se probó la invarianza de medición de género configuracional, métrica y escalar. Investigaciones anteriores han demostrado que es probable que las mujeres obtengan puntuaciones más altas en las calificaciones de SET (por ejemplo, Badri et al., 2006; Darby 2006). Bonitz (2011) encontró que las variaciones de género en los puntajes SET podrían deberse a variaciones de género en rasgos como la amabilidad que se correlacionan con los puntajes SET. Sin embargo, los resultados de este estudio mostraron invariancia de medición de género configuracional, métrica y escalar en el contexto de estudios científico-tecnológicos.

Finalmente, para concluir sobre el tercer objetivo general de este trabajo, los resultados obtenidos con datos agregados, tomando la clase grupal como unidad de análisis, mostraron una correlación moderada pero estadísticamente significativa (0.28) entre SET y logro académico final. Este valor corresponde al valor obtenido en el metaanálisis de Uttl et al. (2017) cuando se volvieron a analizar los datos de Cohen (1987) considerando estudios y efectos de pequeño tamaño (es decir, solo los estudios con un número de 30 o más secciones).

Estos resultados también mostraron una correlación moderadamente alta entre el rendimiento académico previo y el rendimiento académico final. Este hallazgo concuerda con estudios meta analíticos previos sobre las variables asociadas al rendimiento en la educación superior, en los que los conocimientos/habilidades previas aparecen como uno de los principales determinantes del rendimiento académico (Schneider y Preckel, 2017).

Sin embargo, la correlación entre el rendimiento previo y el SET no fue estadísticamente significativa, lo que sugiere que el SET no se ve afectado por los logros académicos anteriores.

El control del rendimiento académico previo con el procedimiento de análisis de regresión jerárquica continuó mostró un efecto significativo de SET sobre el rendimiento académico; este efecto fue de alrededor del 7%, lo que corresponde a una correlación de 0,27, similar a la encontrada en el reanálisis de los datos de Cohen (1981), y es ligeramente superior al valor obtenido en el metaanálisis de Uttl et al. (2017) en base a cerca de 100 estudios multisectoriales publicados hasta esa fecha, que se situaban en .23.

Los resultados obtenidos con los datos individuales de los estudiantes mostraron una correlación estadísticamente significativa (.23) entre el SET y el rendimiento académico final, que fue un poco menor que la obtenida con los datos agregados por secciones. Este resultado es consistente con estudios previos sobre la efectividad de la enseñanza del instructor, en los cuales se considera que los estudios multisección que utilizan los datos agrupados de las secciones son más apropiados para aprehender la verdadera relación entre SET y rendimiento académico (Cohen, 1981; Uttl et al., 2017).

Los resultados de los datos individuales mostraron nuevamente una correlación moderadamente alta entre el rendimiento académico previo y el rendimiento académico final, así como una relación no significativa de SET con el rendimiento académico previo.

Siguiendo la sugerencia de varios autores sobre este tipo de estudios, tanto la variabilidad individual dentro de las secciones como la variabilidad entre secciones (Clayson, 2007; Weinberg et al., 2009) de los datos del presente trabajo incluyeron un análisis de ecuaciones estructurales multinivel.

Los resultados del análisis multinivel mostraron que hubo un efecto significativo de SET en el rendimiento académico final, tanto a nivel individual como de sección, incluso después de controlar el efecto del rendimiento académico previo. Además, la magnitud del efecto fue similar en ambos niveles. El porcentaje total de varianza explicada del rendimiento académico final a nivel de las secciones fue del 33%, mientras que a nivel de los estudiantes individuales fue del 25%, con un 8% de la varianza explicada del rendimiento académico final atribuible a las secciones: es decir, al efecto del maestro.

10. REFERENCIAS

Abrami, P. C., d'Appolonia, S., and Cohen, P. A. (1990). Validity of student ratings of instruction: what we know and what we do not. *J. Educat. Psychol. 82*, 219-231. https://doi.org:10.1007/s10459-017-9783-0

Almeida-de-Macedo, M., Ransom, N., Feng, Y., Hurst, J., and Wurtele, E. S. (2013). Comprehensive analysis of correlation coefficients estimated from pooling heterogeneous microarray data. BMC *Bioinformatics 14*:214. Doi 10.1186/1471-2105-14-214

Alaminos, A. and Castejón, J. L. (2006). *Elaboración, análisis e interpretación de encuestas, cuestionarios y escalas de opinión*. Alicante: Universidad de Alicante. http://hdl.handle.net/10045/20331

Aparicio, E. (2014). Validación de un cuestionario de evaluación de la docencia universitaria. (*Doctoral thesis*, *Universidad de Alicante*, Alicante, Spain). http://hdl.handle.net/10045/45168

American Sociological Association, (2019). Statement on Student Evaluations o teaching. Avaliable at: https://www.asanet.org/sites/default/files/asa_statement_on_student_evaluations_of_teaching_sept52019.pdf (Septembre 9, 2019).

Bentler, P. M. Basic Structural Equation Models in EQS 6 Structural Equations Program Manual, Encino, Multivariate Software, Inc., 2006, pp. 21-58. http://www.econ.upf.edu/~satorra/CourseSEMVienna2010/EQSManual.pdf

Bentler, P. M., & Wu, E. J. (2005). *EQS 6.1 for Windows*. Encino, CA: Multivariate Software INC.

Browne, M. (1982). Covariance structures. En D.M. Hawkins, D.M (Ed.) *Topics in applied multivariate analysis* (pp. 72-141). Cambridge: Cambridge University Press

Brown, T. (2006). *Confirmatory Factor Analysis for Applied Research*, New York: The Guilford Press.

Byrne, B. M. (2008). *Structural Equation Modeling with EQS: Basic Concepts, Applications and Programming*. London: Routledge.

Carvajal, A., Centeno, C., Watson, R., Martínez, M., and Sanz Rubiales, A´ . (2011). ¿Cómo validar un instrumento de medida de la salud?. *Anales del Sistema Sanitario de Navarra 34*(1), 63-72. 10.4321/S1137- 66272011000100007

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

Clayson, D. E. (2005). Within-class variability in student-teacher evaluations: example and problems. *Decision. Sci. J. Inno. Educ*. 3, 109–124. https://doi.org/10.1111/j.1540-4609.2005.00055.x

Clayson, D. E. (2007). Conceptual and statistical problems of using between class data in educational research. *J. Mark. Educ*. *27*, 122–129. https://doi.org/10.1002/mono.12060

Clayson, D. E. (2009). Student evaluations of teaching: are they related to what students learn? A meta-analysis and review of the literature. *J. Mark. Educ*. *31*, 16–30. https://doi.org/10.1177/0273475308324086

Cohen, J., and Cohen, P. (1983). *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, Inc.

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

Cohen, P. A. (1987). A Critical analysis and reanalysis of the multisection validity meta-analysis, in Paper Presented at the *Annual Meeting of the American Educational Research Association*, Washington, DC). https://doi.org/10.3102/00346543051003281

Coba, M.(2006). *Modelización de Ecuaciones Estructurales*, Quito: Ediciones EPN. http://bibdigital.epn.edu.ec/handle/15000/227.

Consejo de Educación Superior (2017). Reglamento de Carrera y Escalafón del Profesor e Investigador del Sistema de Educación Superior, *Gaceta Oficial del Consejo de Educación Superior*, Quito, 2017. https://gaceta.ces.gob.ec/inicio.html

Consejo de Educación Superior (2018). *Ley Organica De Educacion Superior*, LOES. http://www.ces.gob.ec/documentos/Normativa/LOES.pdf

Dattalo, P. (2013). *Structural Equation Modeling*. New York: Oxford University Press, 2013, pp. 109-148.

https://books.google.com.ec/books?id=9qZoAgAAQBAJ&printsec=frontcover&dq=Dattalo&hl=en&sa=X&ved=0ahUKEwiK-tO-udvnAhWx1VkKHRA7DvkQ6AEIOTAC#v=onepage&q=Dattalo&f=false

Deslauriers, L., McCarty, L., Miller, K., Callaghana, K., y Kestin, G. (2019). Measuring actual learning versus feeling of learning inresponse to being actively engaged in the classroom. *PNAS*, *116* (39), 19251–19257.

Education, and Technology, 24-26 July 2019, Jamaica. Scherer, R., Nilsen, T., and Jansen, M. (2016). Evaluating individual students' perceptions of instructional quality: an investigation of their factor structure, measurement invariance, and relations to educational outcomes. *Front. Psychol. 7*:110. https://doi.org/10.3389/fpsyg.2016.00110

Feldman, K. A. (1989). The association between student ratings of specific instructional dimensions and student achievement: refining and extending the synthesis of data from multisection validity studies. *Res. Higher Educ. 30*, 583–645. https://doi.org/10.1007/bf00992392

Gogol, K, Bunner, M., Goetz, T., Martin, R., Ugen, S., Keller, U., Fischhach, A., and Preckel, F. (2014). "My questionnaire is too long!" The assessments of motivational affective constructs with three-item and single-item measures. *Contemporary Educational Psychology 39*, 188-205.

https://doi.org/10.1016/j.cedpsych.2014.04.002

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, *27*(4), 857-871.

Harman, H. (1968) *Modern Factor Analysis*, (2nd Ed.) Chicago: The University of Chicago Press, Ltd. http://archive.org/details/ModernFactorAnalysis/mode/2up

Harrington, D. (2009). Creating a Confirmatory Factor Analysis Model. New York, Oxford University Press, Inc. .

Hornstein, H. A. (2017). Student evaluations of teaching are an inadequate assessment tool for evaluating faculty performance. *Cogent Education 4*, https://doi.org/1.10.1080/2331186X.2017.1304016

Hassler, U., and Thadewald, T. (2003). Nonsensical and biased correlation due to pooling heterogeneous samples. *Statistician 52*, 367–379.

https://doi.org/10.1111/1467-9884.00365

Hu, L. T., and Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Struct. Equ. Model*. *6*, 1–55. https://doi.org/10.1080/10705519909540118

Husbands, C. T., and Fosh, P. (1993). Students' evaluation of teaching in higher education: experiences from four european countries and some implications of the practice. *Assess. Eval. Higher Educ. 18*, 95–114. https://doi.org/10.1080/0260293930180202

Huybers, T. (2014). Student evaluation of teaching: the use of best–worst scaling. *Assess. Eval. Higher Educ. 39*, 496–513. https://doi.org/10.1080/02602938.2013.851782

Herrero, J. El Análisis Factorial Confirmatorio en el estudio de la Estructura y Estabilidad de los Instrumentos de Evaluación: Un ejemplo con el Cuestionario de Autoestima CA-14, *Psychosocial Intervention*, *19*(3), 289-300, 2010. http://scielo.isciii.es/scielo.php?script=sci_arttext&pid=S1132-55920100003000 09

Jöreskog K. and D. Sörbom (1984). LISREL - *VI User´s guide*, Mooresville, IN: Scientific Software.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika 23*(3), 187-200. https://doi.org/10.1007/BF02289233

Kulik, J. A., and McKeachie, W. J. (1975). The evaluation of teachers in higher education. *Review of Research in Education*, 3 201–240.

Lafontaine, M.-F., Brassard, A., Lussier, Y., Valois, P., Shaver, P. R., and Johnson, S. M. (2016). Selecting the best items for a short-form of the experiencies in close relationships questionnaire. *European Journal of Psychological Assessment 32*(1), 140-154. https://doi.org/10.1027/1015-5759/a000243

Lara-Hormigo, A. (2014). Introducción a las Ecuaciones Estructurales en AMOS y R, Granada: Universidad de Granada,. https://masteres.ugr.es/moea/pages/curso201314/tfm1314/tfm-septiembre1314/memoriamasterantonio_lara_hormigo/!

Leung, D. Y. P., and Kember, D. (2005). Comparability of data gathered from evaluation questionnaires on paper and through the Internet. *Res. Higher Educ. 46*, 571–591. https://doi.org/10.1007/s11162-005-3365-3

Leventhal, L. (1975). Teacher rating forms: critique and reformulation of previous validation designs. *Can. Psychol. Rev. 16*, 269–276. https://doi.org/10.1037/h0081814

Levy, P. (1968). Short-form tests: A methodological review. *Psychological Bulletin*, *69*(6), 410.

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

Marsh, H. W. (2007a). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases and usefulness. En R. P. Perry and J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319-383). New York: Springer.

Marsh, H. W. (2007b). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. *Journal of Educational Psychology 99*, 775-790. https://doi.org/10.1037/0022-0663.99.4.775

Malkanthie, A. (2015). *Structural Equation Modeling with AMOS*, https://www.researchgate.net/publication/278889068_Structural_Equation_Modeling_with_AMOS.

Montoya, O. (2007). Aplicacioón del análisis factorial a la investigación de mercados. Caso de estudido. *Scientia et Technica Scientia et Technica*, *XII* (35), 281-286.

Mortelmans, D. and Spooren, P. (2009). A revalidation of the SET37 questionnaire for student evaluations of teaching. *Educational Studies*, 35, 547-552. https://doi.org/10.1080/03055690902880299

MacCallum, R. (1996). Power Analysis and Determination of Sample Size for Covariance Structure Modeling. *Psychological Methods*, *1*(2), 130-149, 1996. http://ww.w.statpower.net/Content/312/Handout/MacCallumBrowneSugawara96.pdf

Marsh, H. W. (2007). Do university teachers become more effective with experience? A multilevel growth model of students' evaluation of teaching over 13 years. J. *Educ. Psychol. 99*, 775–790. https://doi.org/10.1037/0022-0663.99.4.775

Marsh, H. W., and Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: cognitive and affective criteria. *J. Educ. Psychol. 72*, 468–475. https://doi.org/10.1037/0022-0663.72.4.468

Marsh, H. W., and Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluations of teaching: popular myth, bias, validity, or innocent bystanders? *J. Educ. Psychol. 92*, 202–228. https://doi.org/10.1037/0022-0663.92.1.202

Nair, C. S., and Adams, P. (2009). Survey PLATFORM: a factor influencing online survey delivery and response rate. *Q. Higher Educ. 15*, 291–296. https://doi.org/10.1080/13538320903399091

Pareja, F. (1986). *La Educación Superior en el Ecuador [The higher education in Ecuador]*. Caracas: Regional Center For Higher Education in Latin America And the Caribbean (CRESALC)-UNESCO.

Richardson, J. T. E. (2005). Instruments for obtaining student feedback: a review of the literature. *Assess. Eval. Higher Educ. 30*, 387–415. https://doi.org/10.1080/02602930500099193

Richardson, M., Abraham, C., and Bond, R. (2012). Psychological correlates of university students' academic performance: a systematic review an metaanalysis. *Psychol. Bull. 138*, 353–377.

Sánchez-Almeida, T., Sandoval-Palis, I., Gilar-Corbi, R., Castejón-Costa, L. and Salazar-Orellana, D. (2020) Teaching evaluation questionnaire validation in the Escuela Politécnica Nacional, applying the method of Factor Analysis with extraction of principal components., *Ingenieria e Investigación*, . *40*(1), 70-77. https://doi.org/10.15446/ing.investig.v40n1.79634

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the Validity of Student Evaluation of Teaching: The State of the Art. *Review of Educational Research 83*(4), 1-45. https://doi.org/10.3102/0034654313496870

Schneider, M., and Preckel, F. (2017). Variables associated with achievement in higher education: a systematic review of meta-analyses. *Psychol. Bull*. 43, 565–600.

https://doi.org/10.1037/bul0000098

Spooren, P., Brockx, B., and Mortelmans, D. (2013). On the validity of student evaluation of teaching: the state of the art. Rev. *Educ. Res. 83*, 598–642. https://doi.org/10.3102/0034654313496870

Toland, M. D., and De Ayala, R. J. (2005). A Multilevel Factor Analysis of Students' Evaluations of Teaching. *Educational and Psychological Measurement 65*(1), 272–296. https://doi.org/10.1177/0013164404268667

Uttl, B., White, C. A., and Gonzalez, D. W. (2017). Metaanalysis 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

Uttl, B., Cnudde, K., & White, C. A. (2019). Conflict of interest explains the size of student evaluation of teaching and learning correlations in multisection studies: a meta-analysis. *PeerJ*, *7*, e7225. https://doi.org/10.7717/peerj.7225

Weinberg, B. A., Hashimoto, M., and Fleisher, B. M. (2009). Evaluating teaching in higher education. *J. Econ. Educ*. 40227– 61,

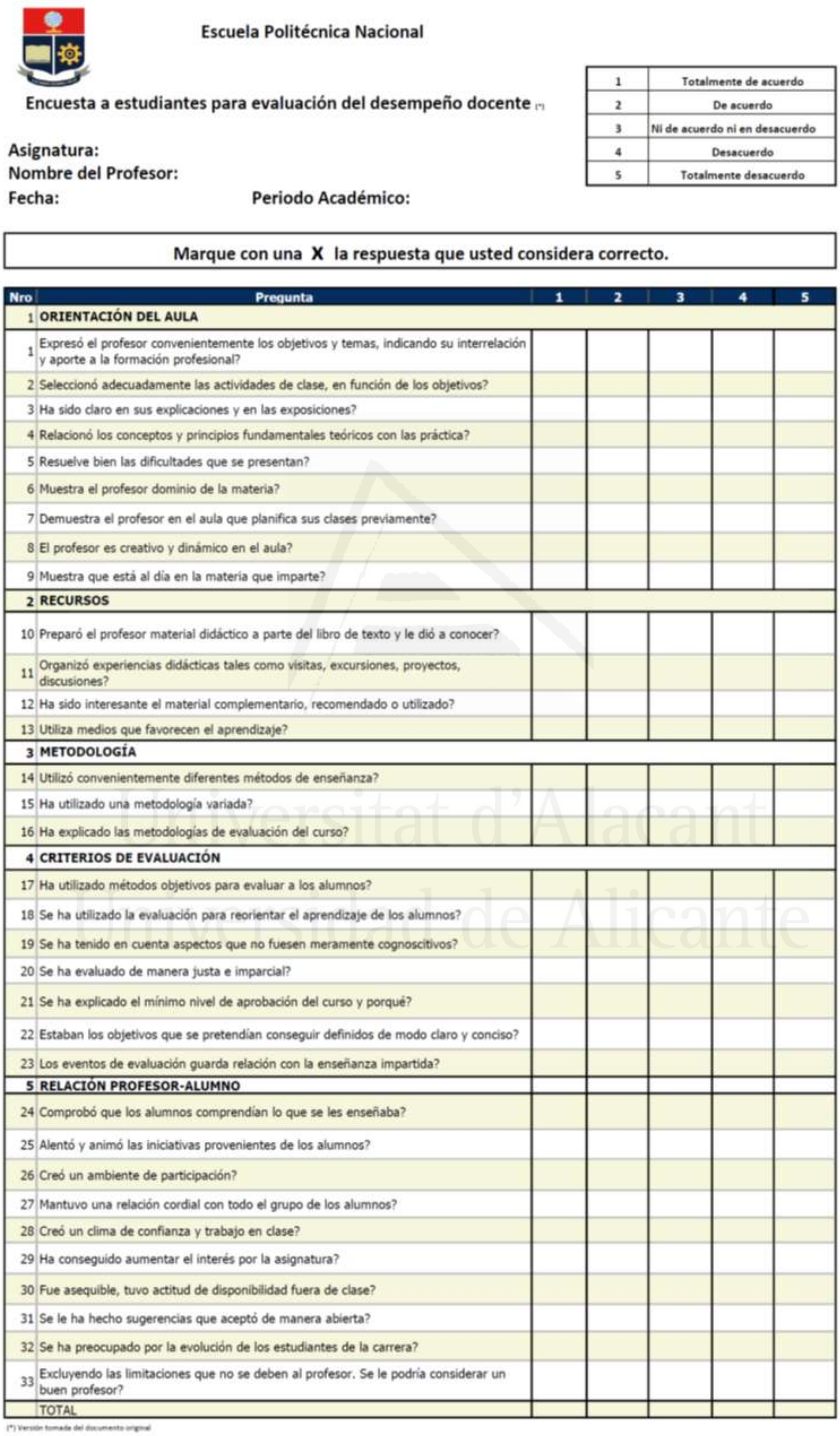
West, S. G., Finch, J. F., and Curran, P. J. (1995). Structural equation models with non-normal variables. En R.H. Hoyle (Ed.,) *Structural Equation Modeling: Concepts, Issues, and Applications* (56-75). Thousands, CA: Sage.

Young, K., Joines, J., Standish, T., and Gallagher, V. (2019). Student evaluations of teaching: the impact of faculty procedures on response rates. *Assess Eval. Higher Educ. 44*, 37–49. https://doi.org/10.1186/s12909-015-0387-1

Zabaleta, F. (2007). The use and misuse of student evaluation of teaching. Teach. *Higher Educ. 12*, 55–76. https://doi.org/10.1080/13562510601102131

11. ANEXOS

11.2 Anexo 1: Escala original de 33 ítems.



Escuela Politécnica Nacional

Encuesta a estudiantes para evaluación del desempeño docente (*)

1	Totalmente de acuerdo
2	De acuerdo
3	Ni de acuerdo ni en desacuerdo
4	Desacuerdo
5	Totalmente desacuerdo

Asignatura:

Nombre del Profesor:

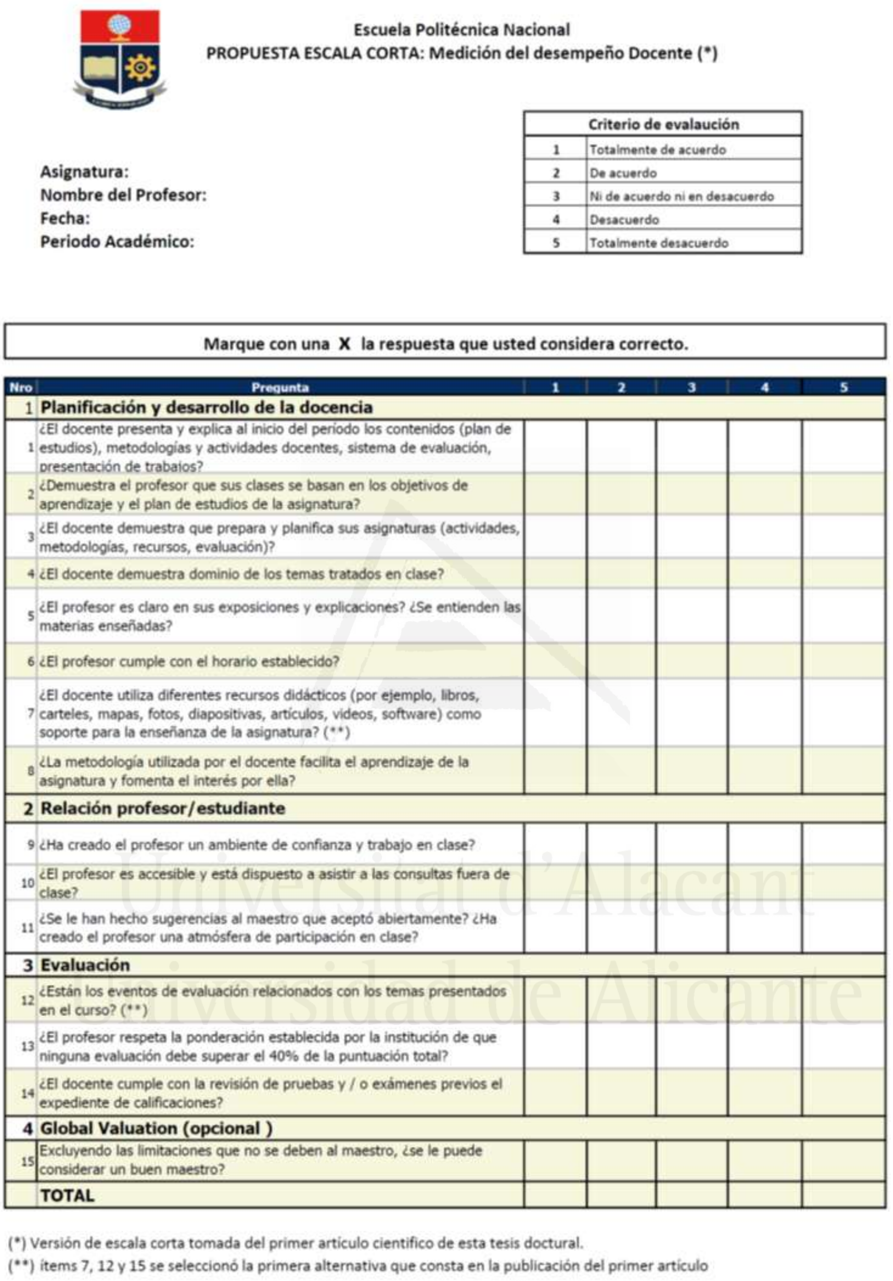
Fecha: Periodo Académico:

Marque con una **X** la respuesta que usted considera correcto.

Nro	Pregunta	1	2	3	4	5
1	**ORIENTACIÓN DEL AULA**					
1	Expresó el profesor convenientemente los objetivos y temas, indicando su interrelación y aporte a la formación profesional?					
2	Seleccionó adecuadamente las actividades de clase, en función de los objetivos?					
3	Ha sido claro en sus explicaciones y en las exposiciones?					
4	Relacionó los conceptos y principios fundamentales teóricos con las práctica?					
5	Resuelve bien las dificultades que se presentan?					
6	Muestra el profesor dominio de la materia?					
7	Demuestra el profesor en el aula que planifica sus clases previamente?					
8	El profesor es creativo y dinámico en el aula?					
9	Muestra que está al día en la materia que imparte?					
2	**RECURSOS**					
10	Preparó el profesor material didáctico a parte del libro de texto y le dió a conocer?					
11	Organizó experiencias didácticas tales como visitas, excursiones, proyectos, discusiones?					
12	Ha sido interesante el material complementario, recomendado o utilizado?					
13	Utiliza medios que favorecen el aprendizaje?					
3	**METODOLOGÍA**					
14	Utilizó convenientemente diferentes métodos de enseñanza?					
15	Ha utilizado una metodología variada?					
16	Ha explicado las metodologías de evaluación del curso?					
4	**CRITERIOS DE EVALUACIÓN**					
17	Ha utilizado métodos objetivos para evaluar a los alumnos?					
18	Se ha utilizado la evaluación para reorientar el aprendizaje de los alumnos?					
19	Se ha tenido en cuenta aspectos que no fuesen meramente cognoscitivos?					
20	Se ha evaluado de manera justa e imparcial?					
21	Se ha explicado el mínimo nivel de aprobación del curso y porqué?					
22	Estaban los objetivos que se pretendían conseguir definidos de modo claro y conciso?					
23	Los eventos de evaluación guarda relación con la enseñanza impartida?					
5	**RELACIÓN PROFESOR-ALUMNO**					
24	Comprobó que los alumnos comprendían lo que se les enseñaba?					
25	Alentó y animó las iniciativas provenientes de los alumnos?					
26	Creó un ambiente de participación?					
27	Mantuvo una relación cordial con todo el grupo de los alumnos?					
28	Creó un clima de confianza y trabajo en clase?					
29	Ha conseguido aumentar el interés por la asignatura?					
30	Fue asequible, tuvo actitud de disponibilidad fuera de clase?					
31	Se le ha hecho sugerencias que aceptó de manera abierta?					
32	Se ha preocupado por la evolución de los estudiantes de la carrera?					
33	Excluyendo las limitaciones que no se deben al profesor. Se le podría considerar un buen profesor?					
	TOTAL					

(*) Versión tomada del documento original

11.3 Anexo 2: Escala corta de 14 ítems.



Escuela Politécnica Nacional
PROPUESTA ESCALA CORTA: Medición del desempeño Docente (*)

Criterio de evalaución	
1	Totalmente de acuerdo
2	De acuerdo
3	Ni de acuerdo ni en desacuerdo
4	Desacuerdo
5	Totalmente desacuerdo

Asignatura:
Nombre del Profesor:
Fecha:
Periodo Académico:

Marque con una **X** la respuesta que usted considera correcto.

Nro	Pregunta	1	2	3	4	5
1	**Planificación y desarrollo de la docencia**					
1	¿El docente presenta y explica al inicio del periodo los contenidos (plan de estudios), metodologías y actividades docentes, sistema de evaluación, presentación de trabajos?					
2	¿Demuestra el profesor que sus clases se basan en los objetivos de aprendizaje y el plan de estudios de la asignatura?					
3	¿El docente demuestra que prepara y planifica sus asignaturas (actividades, metodologías, recursos, evaluación)?					
4	¿El docente demuestra dominio de los temas tratados en clase?					
5	¿El profesor es claro en sus exposiciones y explicaciones? ¿Se entienden las materias enseñadas?					
6	¿El profesor cumple con el horario establecido?					
7	¿El docente utiliza diferentes recursos didácticos (por ejemplo, libros, carteles, mapas, fotos, diapositivas, artículos, videos, software) como soporte para la enseñanza de la asignatura? (**)					
8	¿La metodología utilizada por el docente facilita el aprendizaje de la asignatura y fomenta el interés por ella?					
2	**Relación profesor/estudiante**					
9	¿Ha creado el profesor un ambiente de confianza y trabajo en clase?					
10	¿El profesor es accesible y está dispuesto a asistir a las consultas fuera de clase?					
11	¿Se le han hecho sugerencias al maestro que aceptó abiertamente? ¿Ha creado el profesor una atmósfera de participación en clase?					
3	**Evaluación**					
12	¿Están los eventos de evaluación relacionados con los temas presentados en el curso? (**)					
13	¿El profesor respeta la ponderación establecida por la institución de que ninguna evaluación debe superar el 40% de la puntuación total?					
14	¿El docente cumple con la revisión de pruebas y / o exámenes previos el expediente de calificaciones?					
4	**Global Valuation (opcional)**					
15	Excluyendo las limitaciones que no se deben al maestro, ¿se le puede considerar un buen maestro?					
	TOTAL					

(*) Versión de escala corta tomada del primer artículo cientifico de esta tesis doctural.
(**) ítems 7, 12 y 15 se seleccionó la primera alternativa que consta en la publicación del primer artículo