
Análisis multivariado de la evaluación docente estudiantil. Un caso de estudio

Multivariate analysis of student evaluation of teaching. A case study

Alex J. Zambrano^a
alexjzc@gmail.com

Esther Gutiérrez Mora^b
esrural@gmail.com

Resumen

La evaluación docente estudiantil es el método más utilizado en las universidades latinoamericanas para evaluar la calidad de la enseñanza impartida por los docentes. Teniendo en cuenta este referente, el presente artículo da cuenta de la forma en que se aplicó un análisis multivariado al cuestionario utilizado por la Universidad Santo Tomás en la sede de Bogotá durante los períodos 2012-2014. Este trabajo contempló un análisis de fiabilidad, un análisis factorial exploratorio y confirmatorio para evaluar la validez, confiabilidad y dimensionalidad del instrumento con el cual son evaluados los docentes. Además, se evaluó la relación entre el desempeño del docente y el rendimiento del estudiante. Se realizó un análisis de clasificación mixta para caracterizar los resultados obtenidos en diferentes niveles (divisiones, facultades y profesores). Finalmente, se construyó un índice de percepción de desempeño para las facultades según los alumnos. Entre los resultados más destacados encontramos que el instrumento propuesto es válido y confiable, sin embargo, hay que ajustar dos ítems. También se observó que la calificación del estudiante hacia el docente no está relacionada con la calificación final del estudiante. En la clasificación de los docentes se encontraron tres grupos de docentes donde se diferencian los profesores con bajo, medio y alto desempeño según los alumnos. Entre los ítems que mejor discriminan a los tres grupos están que “el docente usa ayudas educativas que potencian el aprendizaje” y que “el docente aplica estrategias metodológicas que favorecen el aprendizaje”. Se evidenció que la percepción del desempeño de los docentes por los alumnos es superior al 78%. Estos resultados permitieron implementar estrategias pedagógicas para mejorar el trabajo docente en la universidad.

Palabras clave: Análisis factorial exploratorio, análisis de fiabilidad, análisis de clasificación, evaluación docente estudiantil, desempeño docente.

^aFacultad de Estadística, Universidad Santo Tomás

^bUnidad de Desarrollo Curricular y Formación Docente. Universidad Santo Tomás

Abstract

Student teacher evaluation is the most commonly used method in latin american universities to evaluate the quality of teaching provided by teachers. Bearing this reference in mind, the present article gives an account of how a multivariate analysis was applied to the questionnaire used by the Universidad Santo Tomás in Bogotá during the 2012-2014 periods. This work included a reliability analysis, an exploratory and confirmatory factorial analysis to evaluate the validity, reliability and dimensionality of the instrument with which teachers are evaluated. In addition, the relationship between teacher performance and student performance was assessed. A mixed classification analysis was carried out to characterize the results obtained at different levels (divisions, faculties and professors). Finally, a performance perception index was constructed for the faculties according to the students. Among the most outstanding results we find that the proposed instrument is valid and reliable, however, two items need to be adjusted. It was also noted that the student's grade toward the teacher is not related to the student's final grade. In the classification of teachers, three groups of teachers were found, differentiating between low, medium and high performing teachers according to the students. Among the items that best discriminate against the three groups are that "the teacher uses educational aids that enhance learning" and that "the teacher applies methodological strategies that favour learning". It was shown that the perception of teachers' performance by students is higher than 78%. These results allowed the implementation of pedagogical strategies to improve teaching work in the university.

Keywords: Exploratory factor analysis, reliability analysis, classification analysis, student teacher evaluation, teacher performance.

1. Introducción

El proceso de evaluación docente actualmente es uno de los principales retos que debe afrontar no solamente las universidades latinoamericanas, también las universidades europeas (Martínez-Gómez et al. 2010). El proceso de mejora de la calidad académica implica necesariamente la evaluación del personal docente, ya que es un elemento importante para desarrollar una adecuada cultura de evaluación interna en las universidades (Tejedor 2003).

Hay muchas maneras de evaluar la actividad educativa y por lo tanto, el personal docente. Según Martínez-Gómez et al. (2010), Berk (2005), realiza una revisión en la cual describe hasta 12 variedades de evaluación. Sin embargo, la forma más utilizada en Colombia para la evaluación de profesores es a través de evaluaciones estudiantiles. Pero la evaluación de la calidad de la enseñanza es un proceso sumamente complejo, dado que este concepto es subjetivo y multidimensional (Vaillant 2007).

La Universidad Santo Tomás de Bogotá (USTA) reconoce la necesidad de diseñar una metodología estadística apropiada para extraer, analizar e interpretar los datos

contenidos en el cuestionario de Evaluación Docente Estudiantil (EDE), con el fin de proporcionar al personal docente una excelente información para realizar actuaciones de mejora. Estudios similares publicados en la literatura revelaron que las dimensiones subyacentes de un cuestionario particular de las evaluaciones de los estudiantes no siempre son las mismas y dependen de la estructura del cuestionario.

El presente trabajo forma parte de una investigación realizada por la USTA en el marco de la mejora continua de la calidad docente. Algunos de los objetivos particulares de este trabajo fueron: Determinar el comportamiento de la EDE de la USTA en los periodos 2012-2014; realizar un proceso de clasificación de las divisiones, facultades, programas y docentes de la USTA de acuerdo a los resultados de la evaluación docente estudiantil en el periodo 2012-2014; determinar el grado de validez del instrumento de evaluación docente en su componente estudiantil a partir de los resultados obtenidos en el periodo 2012-2014; establecer el grado de correlación entre los resultados del desempeño docente y el promedio de desempeño de los estudiantes que evaluaron para el periodo 2012-2014.

De acuerdo a lo anterior, este artículo se organiza de la siguiente forma: En la sección 2 se especifica la metodología, las bases de datos utilizadas y cada uno de los análisis realizados con el fin de conseguir los objetivos mencionados anteriormente; en la sección 3, se presentan cada uno de los resultados de cada una de las metodologías implementadas. Finalmente en la sección 4, se describen las conclusiones y futuros trabajos alrededor del tema.

2. Metodología

2.1. Instrumento

El cuestionario usado en la Universidad Santo Tomás fue creado en el 2005 . Se realizaron entre los años 2005 a 2007 ajustes tanto de validación como de sistematización. En los periodos de 2008 a 2009 se conformó un grupo de trabajo el cual realizó un proceso inicial de meta - evaluación del proceso de evaluación docente con el fin de generar nuevas y mejores formas de abordar el proceso de evaluación docente. A partir del 2010 de las mejoras tenidas en cuenta se tienen la evaluación la cual contiene la estructura de la Tabla 1 (USTA 2010).

La Tabla 2, presenta la escala de grado de cumplimiento con la cual se evalúa cada afirmación según los estudiantes.

2.2. Muestra y procedimiento

Se utilizaron tres bases de datos en el presente trabajo. La primera base consistía en los puntajes individuales de cada alumno y está compuesta por 21 variables, donde cada una de las observaciones correspondía a una puntuación individual

Tabla 1: Instrumento de evaluación docente estudiantil Universidad Santo Tomás

Ítem	Afirmación
Dimensión 1: Programa	
p_1	El docente dio a conocer el programa de la asignatura o espacio académico previamente al desarrollo de la misma.
p_2	El docente hace explícita la importancia de la asignatura o espacio académico en la formación de los estudiantes.
p_3	El desarrollo de la asignatura o espacio académico a la fecha está acorde con la planeación realizada.
Dimensión 2: Disciplina	
p_5	El docente hace uso de la producción académica de su autoría (libros, artículos, prototipos, módulos, material audiovisual, guías, entre otros) para el desarrollo de la asignatura o espacio académico.
p_6	El docente hace uso de la producción académica institucional para el desarrollo de la asignatura o espacio académico.
p_7	El docente demuestra dominio de las temáticas de la asignatura o espacio académico.
p_8	El docente desarrolla en la asignatura o espacio académico, actividades que relacionan diversas áreas del conocimiento.
p_9	El docente estimula el uso de materiales en una segunda lengua para el desarrollo de la asignatura o espacio académico.
Dimensión 3: Pedagógico	
p_{10}	Las explicaciones del docente son comprensibles.
p_{11}	El docente usa ayudas educativas que potencian el aprendizaje.
p_{12}	El docente hace acompañamiento al proceso de aprendizaje.
p_{13}	El docente aplica estrategias metodológicas que favorecen el aprendizaje.
p_{14}	El docente promueve la consulta de los desarrollos disciplinares recientes en la asignatura o espacio académico.
p_{15}	El docente hace retroalimentación oportuna de las evaluaciones que realiza.
p_{16}	El proceso de evaluación se realiza con estrategias diversas (pruebas escritas, exposiciones, tutorías, etc.)
p_{17}	Los criterios usados para evaluar son justos y transparentes.
Dimensión 4: Compromiso Institucional	
p_{19}	El docente establece relaciones humanas basadas en el respeto.
p_{20}	El docente implementa estrategias que favorecen la convivencia dentro y fuera del aula.
p_{21}	El docente desarrolla actividades que estimulan la capacidad del trabajo en equipo.
p_{22}	El docente asiste puntualmente a sus clases y a las demás actividades programadas.
p_{24}	El docente promueve las actitudes éticas, creativas y críticas derivadas de la misión institucional.

Tabla 2: Escala de medición

Escala	Calificación
Se cumple plenamente	5
Se cumple en alto grado	4
Se cumple medianamente	3
Se cumple en bajo grado	2
Se cumple insuficientemente	1
No se cumple	0

en escala likert para los ítems del cuestionario, realizado por un alumno en una asignatura impartida por un profesor en un curso académico específico. Esta base se le denominó IS (por sus siglas en inglés Individual Scores). Para la segunda base se consideraron los promedios de cada uno de los 21 ítems para cada uno de los profesores según el número de estudiantes que lo evaluó en las diferentes asignaturas impartidas. A esta base se le llamó puntajes promedio y se le denominó MS (por sus siglas en inglés Mean Scores). Por lo cual se etiquetó cada una de las variables como “promedio p_i ”, donde i era el número de ítem. La tercera base consistió en agregar los resultados anteriores por facultad, teniendo en cuenta que cada docente pertenece a una sola facultad. A esta base se le llamó puntajes promedio de facultad y se denominó MSF (por sus siglas en inglés Mean Scores Faculty).

Cada una de estas bases permitieron alcanzar los objetivos propuestos en este trabajo. La Tabla 3 muestra el número de preguntas observaciones obtenidas en cada base para cada semestre académico.

Tabla 3: Estructura de las bases de datos

Periodo académico	Estudiantes - Profesor (base IS)	Docentes evaluados (base MS)	Facultades (base MSF)
20121	33702	1110	21
20122	31877	1137	23
20131	35009	861	23
20132	37284	884	26
20141	38080	900	26
20142	34743	911	26

Si bien el proyecto corresponde a la meta-evaluación docente desde el 2012 al 2014, donde se evaluaron 6 periodos semestrales, en este trabajo solo se presentarán resultados del segundo periodo del 2014, pues se evidenció similitud en todos los resultados de los diferentes periodos.

2.3. Fiabilidad, validez y multidimensionalidad del instrumento

Con el fin de determinar la fiabilidad y validez de un instrumento se utiliza el análisis factorial exploratorio (AFE) y el índice alfa de Cronbach. Sin embargo, según Batista-Foguet et al. (2004) propone realizar un análisis factorial confirmatorio (AFC). Esta prueba permite determinar si el modelo propuesto tiene la misma estructura que el modelo teórico con el cual se construyó el instrumento. Según Tejero-González & Castro-Morera (2011) se define tres criterios para determinar que una escala sea robusta, consistente y unipolar para evaluar la validez, fiabilidad y multidimensionalidad de un instrumento. Estos criterios son los siguientes:

1. Un instrumento no tiene mayor calidad técnica por tener más ítems, sino por garantizar la máxima explicación de varianza sin perder validez de contenido. Por lo cual un instrumento de cuatro factores como es propuesto el instrumento es pertinente a la hora de determinar la validez.
2. El modelo factorial debe garantizar al menos el 60% de la varianza de las puntuaciones, que los ítems saturasen en su factor de pertenencia por encima de 0.5, sin cargar de forma estadísticamente significativa en otros factores y que todos los ítems alcanzaran una comunalidad mínima de 0.5.
3. El instrumento sea cual fuera su estructura factorial, debía alcanzar una fiabilidad de 0.8, donde todos y cada uno de los ítems covarisen entre ellos con correlaciones superiores a 0.3, sin implicar un crecimiento de la fiabilidad global de la escala en el caso de que alguno de los ítems fura eliminado.

Además de lo anterior, se analiza la estructura del modelo mediante 4 factores, considerando factores correlacionados utilizando Análisis Factorial Confirmatorio (AFC).

Se analizaron los 21 ítems del cuestionario considerando que se correlacionaban con las variables latentes correspondientes según cada modelo para evitar errores de especificación. Se estimaron los parámetros utilizando mínimos cuadrados ponderados probando la normalidad de multivariada, todo esto utilizando el software R (R Core Team 2016) y utilizando la librería lavvan (Rosseel 2012) y la librería psych (Revelle 2017).

En este artículo se presenta los resultados de la base que se utilizó la cual fue IS, para el periodo 20142 ya que es la información más reciente que se utilizó.

2.4. Correlación entre la nota definitiva y la nota de la calificación docente

Se realizó una correlación entre la nota definitiva del estudiante y la calificación de la percepción de la evaluación docente dada por estudiante. Para ello se identificó la definitiva de cada estudiante en una materia y se asoció con la calificación promedio del docente en el cuestionario dado por el estudiante. Esta información se filtró teniendo en cuenta que el estudiante haya calificado al docente y a su vez que el docente le haya dado una nota al estudiante en su definitiva. La finalidad de realizar este análisis era poder determinar si existe correlación entre estas dos notas. Para esto se utilizó la base MS y se solicitó a registro y control de la universidad la nota de los estudiantes. Lo anterior se presenta para el periodo 20142.

2.5. Clasificación de los docentes

Usamos la clasificación mixta realizando un análisis de componentes principales (ACP) y una clasificación jerárquica con el fin de identificar la percepción de la evaluación docente por parte de los estudiantes. Para ello se utilizó la base MS. Si la clasificación mixta se hubiera aplicado a la base de datos de IS, habríamos obtenido grupos de estudiantes con categorías de respuestas similares, lo cual no es de interés para nuestro estudio. Mediante esta clasificación se logró clasificar a los docentes según sus resultados en docentes percibidos con desempeño bajo, medio y alto.

La finalidad de esta clasificación era identificar cuáles son las competencias más fuertes que tienen los docentes percibidos con desempeño alto y que competencias deben mejorar los docentes percibidos con desempeño bajo según los estudiantes. Finalmente, se realiza una clasificación supervisada construyendo un árbol de clasificación con el fin de identificar a que grupo debería pertenecer cada docente. Este modelo de árbol, también permite identificar las preguntas más importantes que se deberían tener en cuenta al momento de clasificar, permitiendo así identificar cuáles son las preguntas más importantes que deberían mejorar los docentes para quedar mejor evaluados. Para realizar esta clasificación utilizamos el paquete FactoMineR (Lê et al. 2008) y para la clasificación supervisada el paquete rpart (Therneau et al. 2017). Este proceso de clasificación se realizó para todos los periodos, sin embargo, notamos que esta clasificación no cambió en ninguno de los periodos por lo cual en este artículo solo presentamos los resultados del periodo 20142.

2.6. Construcción del indicador de desempeño según facultades

Se propone construir un indicador sintético denominado “Indicador percepción de desempeño de las facultades según los Estudiantes” usando la base MSF. Este indicador se construyó realizando un ACP y reescalando los resultados entre 0 y 100. Para esto se tuvo en cuenta las metodologías propuestas en Becerra (2010). Este indicador clasificó en tres grupos a las facultades, en los cuales se identifican las mejores facultades, las facultades con percepción intermedia y las facultades que presentan baja percepción de desempeño por sus estudiantes según los resultados de los docentes. Este indicador se calcula para identificar cuáles deben ser las variables más importantes para tener presente a la hora de mejorar la percepción. El fin de la clasificación es identificar cuáles son las competencias mejor percibidas y las que deben mejorar las facultades para que sean mejor percibidas por parte de sus estudiantes. Para ello se evalúa cuáles son las preguntas que más influyen en la construcción del indicador utilizando los coeficientes de una regresión lineal. La clasificación se realiza utilizando el paquete FactoMineR (Lê et al. 2008). Este análisis se realizó en todos los periodos del 2012 al 2014 y se hizo un comparativo de los resultados.

3. Resultados

3.1. Resultados de la validación del instrumento

3.1.1. Análisis descriptivo

Observando los estadísticos descriptivos, observamos que todos los ítems son asimétricos (Skew) a izquierda y presente poca variabilidad (kurtosis), por lo cual nos da indicio que los ítems no son normales de forma univariada. Por otro lado, los promedios (mean) en cada uno de los ítems están entre 3.8 y 4.5. El ítem 9, según el coeficiente de variación (cv) presenta una variabilidad de 0.34 la cual es la más alta. (Tabla 4).

Tabla 4: Estadísticos descriptivos

Item	vars	n	mean	sd	median	trimmed	mad	skew	kurtosis	cv
p_1	1	34743	4.33	1	5	4.53	0	-2.02	4.79	0.23
p_2	2	34743	4.33	0.92	5	4.51	0	-1.85	4.4	0.21
p_3	3	34743	4.3	0.98	5	4.48	0	-1.86	4.22	0.23
p_5	4	34743	4.13	1.12	4	4.33	1.48	-1.67	2.98	0.27
p_6	5	34743	4.11	1.08	4	4.3	1.48	-1.6	2.95	0.26
p_7	6	34743	4.36	0.93	5	4.54	0	-1.89	4.38	0.21
p_8	7	34743	4.22	1	4	4.4	1.48	-1.67	3.37	0.24
p_9	8	34743	3.92	1.35	4	4.19	1.48	-1.45	1.51	0.34
p_{10}	9	34743	4.2	1.05	5	4.4	0	-1.64	2.94	0.25
p_{11}	10	34743	4.18	1.03	4	4.36	1.48	-1.59	2.89	0.25
p_{12}	11	34743	4.19	1.05	4	4.38	1.48	-1.64	2.99	0.25
p_{13}	12	34743	4.17	1.05	4	4.36	1.48	-1.6	2.89	0.25
p_{14}	13	34743	4.2	1.01	4	4.38	1.48	-1.65	3.26	0.24
p_{15}	14	34743	4.19	1.08	4	4.39	1.48	-1.72	3.19	0.26
p_{16}	15	34743	4.23	1.02	5	4.42	0	-1.76	3.6	0.24
p_{17}	16	34743	4.26	1.04	5	4.46	0	-1.85	3.81	0.25
p_{19}	17	34743	4.38	0.94	5	4.58	0	-2.08	5.29	0.21
p_{20}	18	34743	4.32	0.96	5	4.51	0	-1.91	4.46	0.22
p_{21}	19	34743	4.35	0.95	5	4.54	0	-1.99	4.87	0.22
p_{22}	20	34743	4.34	1	5	4.55	0	-2.04	4.78	0.23
p_{24}	21	34743	4.33	0.98	5	4.52	0	-2	4.86	0.23

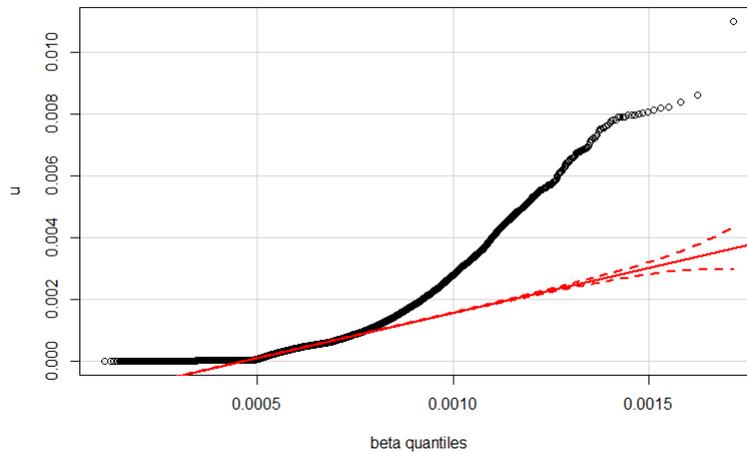
Ahora probando para contrastar si los datos son normales de forma multivariada, se procede a realizar una prueba gráfica la cual consiste en determinar si la distancia de Mahalanobis para cada observación x_i al centroide de los datos \bar{x} ,

$$D_i^2 = (x_i - \bar{x})^t S^{-1} (x_i - \bar{x}).$$

Si los x_i proceden de una distribución normal multivariada, se tiene que

$$u_i = \frac{nD_i^2}{(n-1)^2} \sim \text{Beta}(\alpha, \beta),$$

Figura 1: Prueba qqplot para determinar normalidad multivariada



para $\alpha = (p - 2)/2p$ y $\beta = (n - p - 2)/(2(n - p - 1))$, donde n es el número de observaciones y p el número de variables (Díaz & Morales 2012).

Realizando un gráfico de qqplot observamos que los datos no siguen una distribución normal multivariada (Figura 1). Sin embargo, Batista-Foguet & Coenders (2000) aseguran que a pesar de que los datos no son normales multivariados, existen métodos robustos los cuales evitan la falta de normalidad multivariada de los datos.

3.1.2. Fiabilidad

Para determinar la fiabilidad del instrumento utilizamos el coeficiente alpha de cronbach. Nótese que la fiabilidad alcanzada por el instrumento de forma global es de 0.97, lo cual nos indica que el instrumento es fiable. Por otro lado, calculando la fiabilidad en cada una de las dimensiones del instrumento encontramos fiabilidad entre 0.88 y 0.96, lo cual también no corrobora que la fiabilidad del instrumento en cada una de las dimensiones resulta ser alta (Tabla 5).

Tabla 5: Fiabilidad del instrumento

Dimensión	Confiabilidad
Global	0.97
Programa	0.9
Disciplina	0.88
Pedagogía	0.96
Compromiso institucional	0.93

3.1.3. Análisis factorial exploratorio

Para evaluar el instrumento se realizó el análisis Factorial Exploratorio (AFE), con el fin de identificar si la estructura determinada por el instrumento era la indicada. Para evitar el problema de la no normalidad multivariada se realizaron las estimaciones utilizando mínimos cuadrados ponderados ya que es un método libre de distribución el cual es adecuado para este análisis (Browne 1984). Por otro lado, como las variables tenidas en cuenta son de escala Likert, entonces se sugiere calcular la matriz de correlaciones policóricas el cual es conveniente para variables ordinales como se define en este trabajo (Batista-Foguet & Coenders 2000). El método rotación se utilizó el oblimin el cual es un método oblicuo adecuado para este trabajo ya que los factores deben ser correlacionados (Morales 2013).

Tabla 6: Análisis Factorial Exploratorio del Instrumento

Item	Factor 1	Factor 2	Factor 3	Factor 4	Comunalidad	Correlación ítem-total	Alfa sin ítem
p_1			0.94		0.83	0.73	0.97
p_2			0.72		0.83	0.8	0.97
p_3			0.85		0.86	0.8	0.97
p_5				0.81	0.76	0.7	0.97
p_6				0.84	0.83	0.76	0.97
p_7	0.34				0.74	0.78	0.97
p_8	0.42				0.8	0.84	0.97
p_9				0.58	0.58	0.61	0.97
p_{10}	0.87				0.83	0.82	0.97
p_{11}	0.9				0.86	0.85	0.97
p_{12}	0.92				0.86	0.85	0.97
p_{13}	1				0.89	0.86	0.97
p_{14}	0.76				0.83	0.85	0.97
p_{15}	0.73				0.77	0.8	0.97
p_{16}	0.62				0.78	0.81	0.97
p_{17}	0.59				0.78	0.81	0.97
p_{19}		0.94			0.87	0.78	0.97
p_{20}		0.93			0.88	0.81	0.97
p_{21}		0.79			0.81	0.79	0.97
p_{22}		0.66			0.71	0.72	0.97
p_{24}		0.87			0.86	0.81	0.97
Porcentaje de varianza explicada	32 %	22 %	14 %	13 %			
Porcentaje de varianza global de la escala	81 %						

Observando los resultados en la Tabla 6 encontramos lo siguiente:

- Se obtuvo una estructura dimensional el cual recoge el 81 % de la variabilidad.
- El primer factor recoge todos los ítems de la dimensión pedagógica. Sin embargo, los ítems 7 y 8 que son de la dimensión disciplinar, son percibidos por parte de los estudiantes en la dimensión pedagógica con comunalidades por debajo de 0.5. Esto indica que se debe estudiar estos ítems para reescribirlos, o cambiarlos de dimensión o en el peor de los casos eliminarlos.
- El segundo factor recoge los ítems de la dimensión Compromiso institucional.
- El tercer factor recoge los ítems de la dimensión Programa.
- El cuarto factor recoge los ítems de la dimensión Disciplina.

Al observar las correlaciones entre los factores, notamos que las correlaciones son bastante altas (Tabla 7), lo que nos indica que entre los factores observados se observa muy buena realización.

Tabla 7: Matriz de correlaciones entre los factores

	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	1	0.86	0.84	0.84
Factor 2		1	0.82	0.77
Factor 3			1	0.81
Factor 4				1

3.1.4. Análisis factorial confirmatorio

En el Análisis Factorial Confirmatorio (AFC) se procede a ajustar tres modelos con cuatro factores. El modelo 1 es el modelo propuesto en la construcción del instrumento, el modelo 2 es el modelo propuesto por el AFE y el modelo 3 es eliminando los ítems 7 y 8 los cuales presentaron problemas en el AFE, debido a que no se incluyeron en el factor propuesto en el instrumento inicial. Para el modelo factorial confirmatorio, se utilizó los mínimos cuadrados ponderados diagonalizados (DWLS), el cual es un método robusto cuya matriz de correlaciones es la policóricas la cual se ajusta para variables de tipo ordinal.

A continuación, en la Tabla 8, se observa un resumen de las diferentes estadísticas que se calculan para determinar el ajuste del modelo. Si el lector desea conocer acerca de cada uno de estos estadísticos se sugiere la lectura de Hair et al. (1999).

La Tabla 9, presenta los diferentes índices de ajuste de los modelos propuestos. Según las estadísticas de ajuste absoluto (χ^2 , RMSEA, ECVI), el mejor modelo es el 3, sin embargo, todos los modelos indican buen ajuste. Las medidas de ajuste

Tabla 8: Estadísticas de ajuste

Acrónimo	Definición	Criterio de ajuste.
χ^2	Estadístico-ratio de verosimilitud chi-cuadrado	p-valores por debajo de 0.05 se recomiendan como mínimo aceptado.
RMSEA	Error de aproximación cuadrática media.	Se sugiere valores aceptables por debajo de 0.08.
ECVI	Índice de validación cruzada esperada.	Se sugiere valores cercanos a 0.
NFI	Índice de ajuste normado.	Se sugiere que esté por encima de 0.9.
CFI	Índice de ajuste comparativo.	Valores por encima de 0.9 indican un buen ajuste del modelo.
TLI	Índice de Tucker-Lewis	Se sugiere que esté por encima de 0.9.
PGFI	Índice de bondad de ajuste de la parsimonia.	Se sugiere valores cercanos a 1.
PNFI	Índice de ajuste normado de parsimonia.	Valores elevados indica un menor ajuste.

incremental (NFI, CFI, TLI), calculadas para todos los modelos permiten encontrar que el ajuste es bastante bueno. Las medidas de ajuste de parsimonia (PGFI, PNFI) indica que el modelo 1 es el mejor, sin embargo, todos los modelos indican buen ajuste en cuanto a parsimonia. Los resultados anteriormente mencionados permiten observar que los modelos propuestos son adecuados. Por simplicidad se prefiere el modelo 1, ya que se ajusta al modelo propuesto, teniendo en cuenta que se debe ajustar los ítems 7 y 8.

Tabla 9: Índices de ajuste de los diferentes modelos

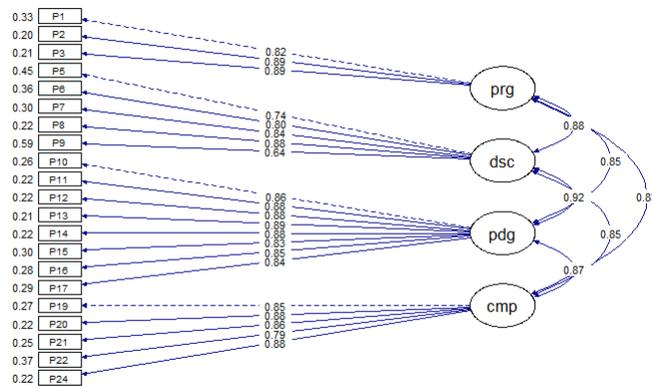
Modelos evaluados	χ^2	df	p-valor	RMSEA	ECVI	NFI	CFI	TLI	PGFI	PNFI
modelo 1	893.563	183	0	0.011	0.028	0.999	0.999	0.999	0.791	0.87
modelo 2	766.243	183	0	0.01	0.025	0.999	0.999	0.999	0.792	0.871
modelo 3	443.388	146	0	0.008	0.015	0.999	1	0.999	0.768	0.853

Teniendo en cuenta los resultados anteriores, la solución a partir de los coeficientes estandarizados se observan en la Figura 2. Nótese que los coeficientes estandarizados están por encima de 0.5, lo cual nos indica que todos ítems se deben tener en cuenta. Por otro lado, los factores propuestos indican buena correlación entre ellos. Lo cual nos indica que, si un docente presenta buenos resultados en cualquiera de los factores (programa, disciplina, pedagogía, compromiso institucional) también presenta buenos resultados en cualquiera de los otros.

3.2. Resultados del análisis de correlación

Para realizar el análisis de correlación entre la nota definitiva del estudiante y cada una de las preguntas de la evaluación docente dada por el estudiante, se tuvo en cuenta que las variables del cuestionario se respondieron en forma nominal,

Figura 2: Análisis Factorial Confirmatorio



prg : Dimensión del programa; dsc: Dimensión disciplinar; pdg: Dimensión pedagogía; cmp: Dimensión compromiso institucional.

mientras que la nota se califica de forma continua en el intervalo 0 a 5 por lo cual se realizó el cálculo de las correlaciones de sperman, las cuales son adecuadas para este tipo de situaciones y no depende de supuestos. Al realizar esto encontramos que las correlaciones están por debajo de 0.17 lo cual no indica que son muy pequeñas (Tabla 10).

Tabla 10: Correlaciones entre la nota definitiva del estudiante y cada una de las preguntas de la evaluación docente

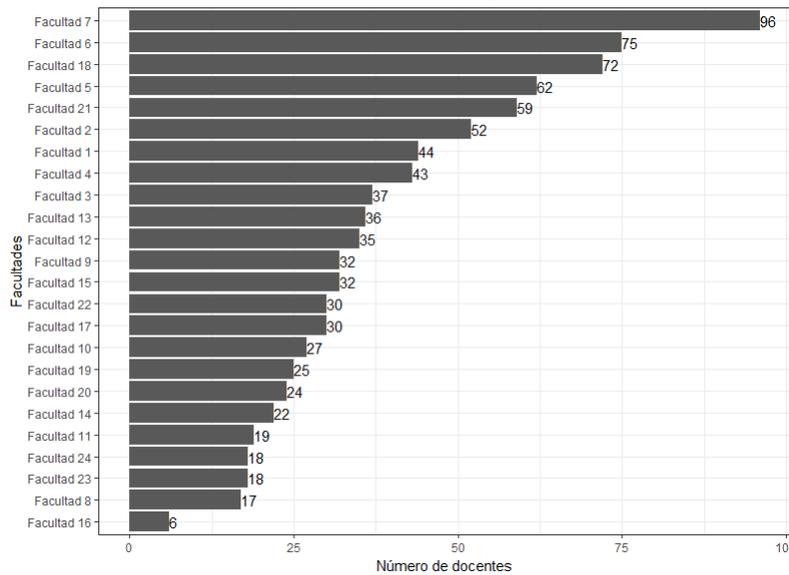
Preguntas	P1	P2	P3	P5	P6	P7	P8	P9	P10	P11	P12
Correlaciones	0.1	0.13	0.1	0.09	0.09	0.12	0.13	0.06	0.17	0.16	0.16
Preguntas	P13	P14	P15	P16	P17	P19	P20	P21	P22	P24	
Correlaciones	0.15	0.13	0.13	0.15	0.16	0.17	0.16	0.16	0.11	0.15	

Incluso al identificar la correlación entre la nota definitiva del estudiante y cada una de las dimensiones de la evaluación docente éstas correlaciones no superan el 0.17, lo cual representa un valor muy bajo (Tabla 11).

Tabla 11: Correlación entre la nota definitiva del estudiante y cada una de las dimensiones de la evaluación docente

Dimensión	Programa	Disciplina	Pedagógico	Compromiso Institucional
Correlaciones	0.12	0.1	0.17	0.17

Figura 3: Distribución de docentes evaluados según facultad en el periodo 20141



Al realizar una prueba de hipótesis entre la nota definitiva del estudiante y el promedio de la evaluación docente dada por el estudiante encontramos una correlación de 0.16 y un p-valor muy pequeño ($2.2e-16$), lo cual no rechaza la hipótesis nula de que existe una correlación, sin embargo, esta correlación es muy pequeña lo que no hace pensar que el estudiante evalúa al docente independiente de la nota definitiva que el docente le coloca en su materia.

3.3. Resultados de clasificación de los docentes

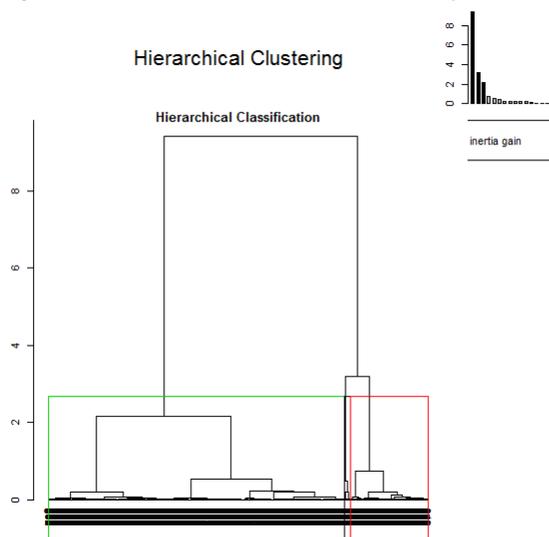
Para la clasificación de los docentes en el periodo 20141 se tuvo en cuenta los resultados promedios de los 911 profesores los cuales se encuentran distribuidos en la Figura 3. Nótese que las primeras 6 facultades son las facultades más grandes que representa el 46% de todos los docentes evaluados¹.

Para realizar la clasificación, se utiliza la clasificación mixta (Lebart et al. 2006). Esta clasificación consiste en realizar primero un análisis factorial (para este trabajo se hizo un análisis de componentes principales); posteriormente teniendo la información de las primeras coordenadas factoriales, se realiza una clasificación jerárquica; finalmente para reafirmar los grupos obtenidos se hace una clasificación k-means.

La utilización de las primeras coordenadas factoriales es un proceso de pretratamiento de los datos con el fin de recoger la mayor variabilidad posible en el

¹Debido a que la siguiente información es prioritaria para la universidad se anonimiza la información.

Figura 4: Resultados de la clasificación jerárquica



conjunto de datos, esto permite eliminar la cantidad de ruido debido a la aleatoriedad de los mismos. Para nuestro estudio se tomaron los primeros tres ejes los cuales recogieron el 90 % de toda la variabilidad de los datos. Según la clasificación jerárquica se escogen toman tres grupos (Gráfico 4).

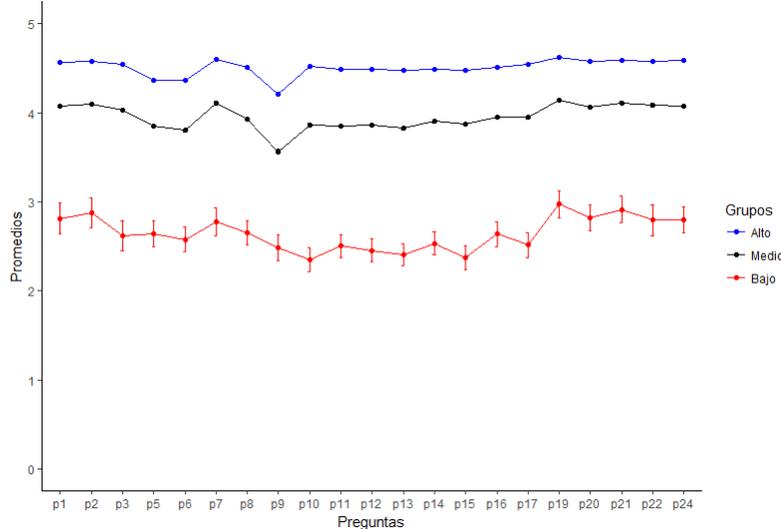
Tabla 12: Estadísticos descriptivos de la clasificación de los docentes según la evaluación docente por parte de los estudiantes

Grupos	n	mean	sd	median	trimmed	mad	min	max	range	skew	kurtosis	cv
Bajo	29	2.64	0.71	2.92	2.74	0.43	0	3.25	3.25	-1.85	4.01	0.27
Medio	341	3.95	0.22	4.02	3.97	0.2	3.32	4.22	0.91	-0.83	-0.12	0.06
Alto	541	4.51	0.17	4.5	4.5	0.18	4.22	5	0.78	0.5	-0.15	0.04

A continuación, se describen los resultados generales de la clasificación (Tabla 12):

- El grupo de resultados bajos, lo conforma el 3 % de los docentes cuyos resultados en la evaluación docente por parte de los estudiantes está por debajo de 3.3.
- El grupo de resultados medios lo conforman el 37 % de los docentes cuyos resultados en la evaluación docentes por parte de los estudiantes están entre 3.3 y 4.2.
- El grupo de resultados altos lo conforman el 59 % de los docentes cuyos resultados en la evaluación docentes por parte de los estudiantes están por encima de 4.2.

Figura 5: Resultados de los promedios de las preguntas de la evaluación docente estudiantil según la clasificación de los grupos



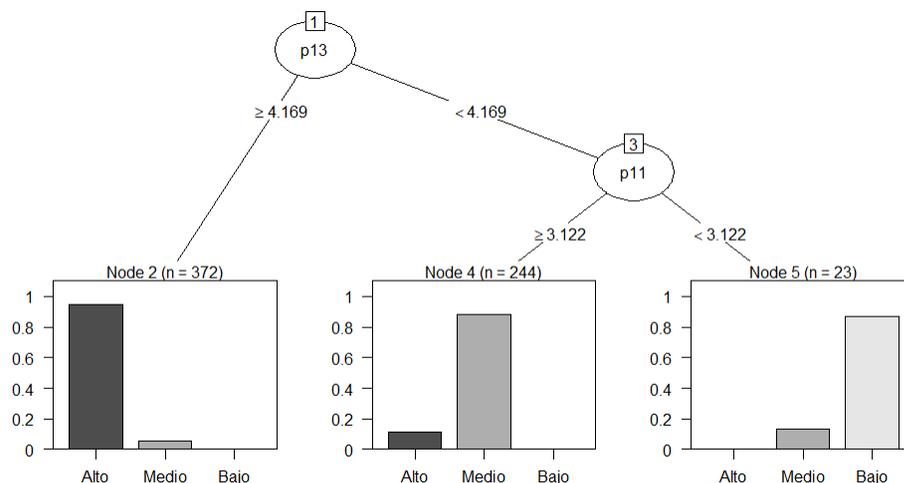
Analizando los perfiles de los promedios de cada una de las preguntas en la evaluación docente estudiantil (Figura 5), notamos lo siguiente:

- Los docentes que se clasifican en el grupo bajo, tuvieron resultados por debajo de 3 en todas las preguntas.
- Los docentes en el grupo medio, tuvieron resultados entre 3.5 y 4.1 en todas las preguntas.
- Los docentes en el grupo alto, tuvieron resultados por encima de 4.2.
- Las preguntas más bajas para los docentes del grupo bajo se encuentran en la pregunta 9 (uso de material en una segunda lengua) y las preguntas de la dimensión pedagógica (preguntas de la 10 a la 17).
- Una constante en todos los grupos en el resultado bajo en la pregunta 9.

Ahora para determinar qué podría hacer un docente para mejorar su calificación docente según la clasificación obtenida, se realiza un árbol de clasificación (Figura 6). Para ello se utilizó el 70% de la muestra como muestra de entrenamiento y el 30% restante como muestra de prueba para validar el modelo y se obtiene una tasa de mala clasificación del 7%.

Observando el árbol (Figura 6) encontramos que las preguntas que más influyen en la clasificación son las preguntas p_{11} (El docente usa ayudas educativas que potencian el aprendizaje) y p_{13} (El docente aplica estrategias metodológicas que favorecen el aprendizaje). Según el gráfico anterior encontramos lo siguiente:

Figura 6: Resultados del árbol de clasificación para los promedios de las preguntas de la evaluación docente estudiantil según la clasificación de los grupos

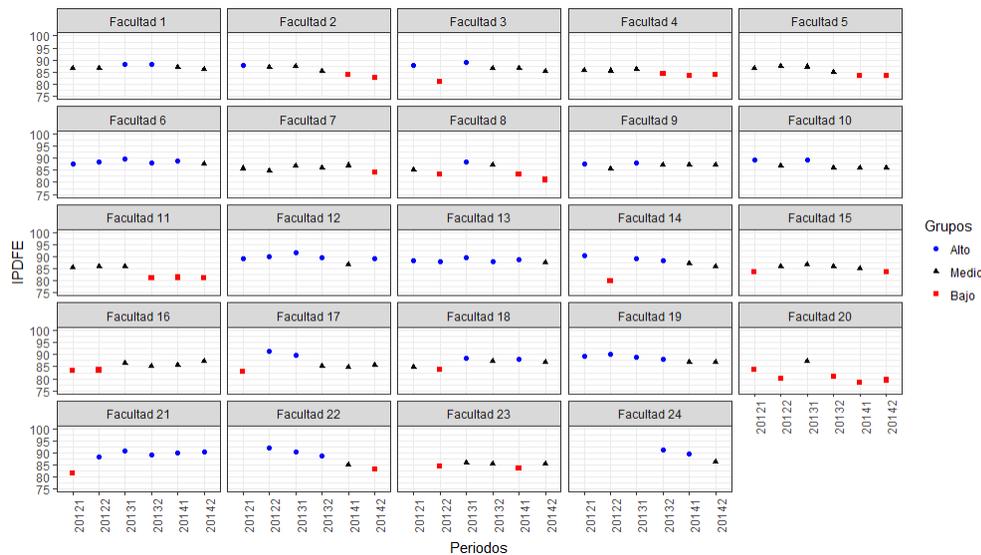


- Si el docente obtiene una calificación por encima de 4.2 en la pregunta p_{13} tiene una alta probabilidad de pertenecer al grupo de los docentes con calificación alta.
- Si el docente obtiene una calificación por debajo de 4.2 en la pregunta p_{13} , pero una calificación por encima de 3.1 en la pregunta p_{11} , tiene una alta probabilidad de pertenecer al grupo de los docentes con calificación media.
- Si el docente obtiene una calificación por debajo de 4.2 en la pregunta p_{13} y una calificación por debajo de 3.1 en la pregunta p_{11} , tiene una alta probabilidad de pertenecer al grupo de los docentes con calificación baja.

3.4. Resultados del índice percepción de desempeño para las facultades según los estudiantes

Se propone construir un indicador sintético denominado “*Indicador de percepción de desempeño de las facultades según los estudiantes*” (IPDFE) usando la base MSF. Este indicador se construyó realizando un ACP y reescalando los resultados obteniendo entre 0 y 100. Para esto se tuvo en cuenta las metodologías propuestas en Becerra (2010). Este indicador se clasificó en tres grupos en los cuales se identifican las mejores, facultades percibidas por parte de los estudiantes y las facultades que presentan bajo desempeño percibido por sus estudiantes según los docentes. El fin de la anterior clasificación es identificar cuáles son las competencias mejor percibidas y las que deben mejorar las facultades para que sean mejor percibidas por parte de sus estudiantes. Para ello utilizamos el paquete FactoMineR (Lê

Figura 7: Resultados del IPDFE para cada facultad en cada periodo y según la clasificación de los grupos



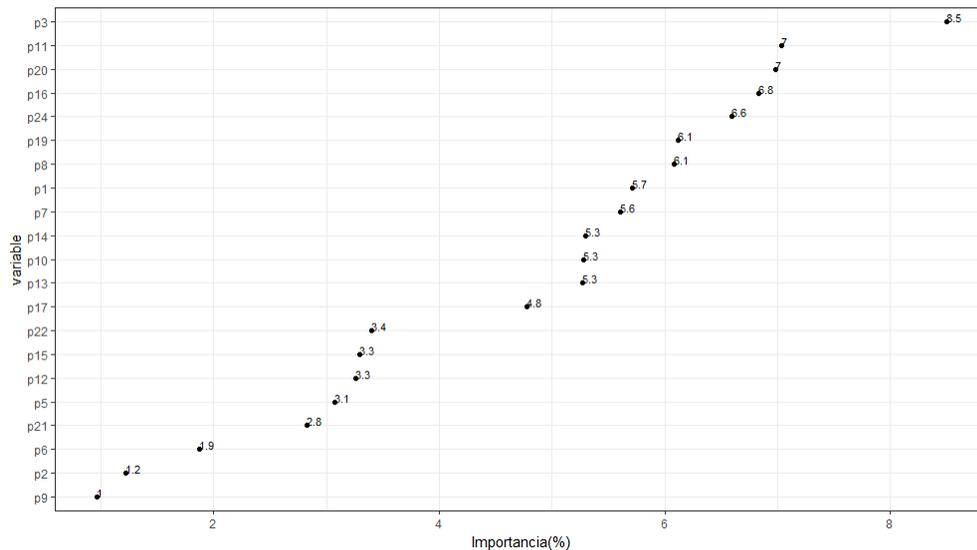
et al. 2008). Este análisis se realizó en todos los periodos del 2012 al 2014 y se hizo un comparativo de los resultados.

Analizando los resultados del IPDFE en cada periodo (Gráfico 7), notamos lo siguiente²

- En general todas las facultades tienen un resultado por encima de 78 %, lo que nos indica que las facultades tienen una buena percepción de desempeño de los docentes por parte de los estudiantes.
- Las facultades con mejor IPDFE en casi todos los periodos son la Facultad 21, Facultad 13, Facultad 12 y la Facultad 6.
- Las facultades con peor IPDFE en casi todos los periodos es la Facultad 20.
- Si una facultad tiene un IPDFE por debajo de 84.6 % se clasifica con resultados bajos.
- Si una facultad tiene un IPDFE entre 84.6 % y 87.7 % se clasifica con resultados medios.
- Si una facultad tiene un IPDFE mayor de 87.7 % se clasifica con resultados altos.

²Las facultades 22 y 24 son las más jóvenes por ello no tienen mediciones completas en todos los periodos de medición.

Figura 8: Importancia de las variables para el IPDFE



Estimando los coeficientes de regresión de un modelo lineal general, para determinar como quedaría el IPDFE, teniendo las preguntas de la evaluación docente el índice queda de la siguiente manera:

$$\begin{aligned}
 IPDFE = & 1.14p_1 + 0.24p_2 + 1.70p_3 + 0.61p_5 + 0.37p_6 + 1.12p_7 + 1.22p_8 + 0.19p_9 \\
 & + 1.05p_{10} + 1.41p_{11} + 0.65p_{12} + 1.05p_{13} + 1.06p_{14} + 0.66p_{15} + 1.37p_{16} \\
 & + 0.96p_{17} + 1.22p_{19} + 1.40p_{20} + 0.56p_{21} + 0.68p_{22} + 1.32p_{24}
 \end{aligned}$$

Para identificar las preguntas que son más importantes en la construcción del IPDFE, se determina teniendo en cuenta el efecto total de cada una de las preguntas con las cuales se construye el indicador. Para ello se consideran los pesos de cada variable utilizando los coeficientes del IPDFE de la siguiente manera:

$$w_i = \frac{\beta_i}{\sum_{i=1}^{21} \beta_i}$$

$$\sum_{i=1}^{21} w_i = 1$$

En el Gráfico 8 se observan las preguntas p_3 (El desarrollo de la asignatura o espacio académico a la fecha está acorde con la planeación realizada), p_{11} (El docente usa ayudas educativas que potencian el aprendizaje), p_{20} (El docente implementa estrategias que favorecen la convivencia dentro y fuera del aula), p_{16} (El proceso de evaluación se realiza con estrategias diversas (pruebas escritas, exposiciones, tutorías, etc.)), p_{24} (El docente promueve las actitudes éticas, creativas y críticas

derivadas de la misión institucional), p_{19} (El docente establece relaciones humanas basadas en el respecto) y p_8 (El docente desarrolla en la asignatura o espacio académico, actividades que relacionan diversas áreas del conocimiento) son las variables más importancia en la construcción del IPDFE.

4. Conclusiones

- Al realizar un análisis descriptivo para validar el instrumento de la evaluación docente encontramos que los promedios de cada una de las preguntas de la evaluación docente están entre 3.8 y 4.5. La pregunta que presenta mayor variabilidad es el ítem 9 la cual hace referencia al uso de materiales en segunda lengua. Este primer resultado nos indica que hay una percepción muy variable sobre las actividades de la evaluación.
- Al realizar un análisis de confiabilidad encontramos que el instrumento es fiable.
- El análisis factorial exploratorio y confirmatorio permite reconocer que los ítems 7 y 8 se deben ajustar en el instrumento ya que los alumnos perciben éstas preguntas en un factor diferente al que pertenecen. Sin embargo, el modelo presenta indicadores bastante buenos.
- En el análisis de correlación entre la nota definitiva del estudiante y la calificación del docente, encontramos que existe correlación pero ésta es muy baja, por lo cual podemos indicar que la calificación docente dada por los estudiantes es independiente de la nota definitiva que el docente le asigna.
- Al realizar la clasificación de los docentes se encontraron tres grupos de docentes entre los cuales se diferencian aquellos con bajo, medio y alto desempeño según los alumnos. Entre los ítems que mejor discriminan los tres grupos se encuentra que el docente debe utilizar ayudas educativas y también que el docente debe aplicar estrategias metodológicas que favorezcan el aprendizaje del docente.
- Por último, se construyó un índice de percepción de desempeño para las facultades según los estudiantes en el cual se observa que en general la percepción del desempeño de los docentes por parte de los estudiantes se encuentra por encima de 78 %.
- Estos resultados permitieron implementar estrategias institucionales para contribuir al mejoramiento del desempeño docente en la universidad, ejemplo de ello el “Plan de mejoramiento Docente 2006-2011”; el “Plan de formación docente 2014-2020” y el “Plan de acompañamiento para fortalecer el desempeño docente en el aula”, iniciativas lideradas desde la Vicerrectoría Académica General a través de la Unidad de Desarrollo Curricular y Formación Docente (Gutiérrez & Zambrano 2017).

Agradecimientos

Este trabajo hace parte de los resultados del proyecto titulado “Metaevaluación del proceso de Evaluación docente Institucional USTA 2012-2014” adscrito a la Unidad de Investigación de la USTA y desarrollado en forma conjunta con la Facultad de Estadística y el equipo de profesionales de la Unidad de Desarrollo Curricular y Formación Docente quienes acompañaron y enriquecieron el proceso en sus diferentes etapas.

Recibido: 28 de agosto de 2017

Aceptado: 16 de mayo de 2018

Referencias

- Batista-Foguet, J. M. & Coenders, G. (2000), *Modelos de Ecuaciones Estructurales*, La Muralla, Madrid.
- Batista-Foguet, J. M., Coenders, G. & Alonso, J. (2004), ‘Análisis factorial confirmatorio. su utilidad en la validación de cuestionarios relacionados con la salud’, *Medicina Clínica* **122**(1), 21–27.
- Becerra, M. O. (2010), Comparación del análisis factorial múltiple (afm) y del análisis de componentes principales para datos cualitativos (prinqual), en la construcción de índices, Tesis de maestría, Universidad Nacional de Colombia, Bogotá.
- Berk, R. A. (2005), ‘Survey of 12 strategies to measure teaching effectiveness’, *Int. J. Teach. Learn. High. Educ.* **17**, 48–62.
- Browne, M. W. (1984), ‘Asymptotically distribution-free methods for the analysis of covariance structures’, *British Journal of Mathematical and Statistical Psychology* **37**(1), 62–83.
- Díaz, L. G. & Morales, M. A. (2012), *Análisis estadístico de datos multivariados*, Universidad Nacional de Colombia, Bogotá.
- Gutiérrez, E. & Zambrano, A. J. (2017), La dimensión contextual en la comprensión de la evaluación docente -la Universidad Santo Tomás y su experiencia en el proceso de metaevaluación del desempeño docente-. Documento en proceso de publicación.
- Hair, J. F., Anderson, R. E., Tatham, R. L. & Black, W. C. (1999), *Análisis Multivariante*, Prentice Hall, Madrid.
- Lê, S., Josse, J. & Husson, F. (2008), ‘FactoMineR: A package for multivariate analysis’, *Journal of Statistical Software* **25**(1), 1–18.

- Lebart, L., Piron, M. & Morineau, A. (2006), *Statistique exploratoire multidimensionnelle: visualisation et inférences en fouilles de données*, Dunod.
- Martínez-Gómez, M., Carot Sierra, J. M., Jabaloyes, J. & Zarzo, M. (2010), ‘A multivariate method for analyzing and improving the use of student evaluation of teaching questionnaires: a case study’, *Quality & Quantity* **45**(6), 1415–1427.
- Morales, P. (2013), El análisis factorial en la construcción e interpretación de test, escalas y cuestionarios, Documento técnico, Universidad Pontificia Comillas, Madrid.
- R Core Team (2016), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
*<https://www.R-project.org/>
- Revelle, W. (2017), *psych: Procedures for Psychological, Psychometric, and Personality Research*, Northwestern University, Evanston, Illinois. R package version 1.7.5.
*<https://CRAN.R-project.org/package=psych>
- Rosseel, Y. (2012), ‘lavaan: An R package for structural equation modeling’, *Journal of Statistical Software* **48**(2), 1–36.
*<http://www.jstatsoft.org/v48/i02/>
- Tejedor, T. (2003), ‘Un modelo de evaluación del profesorado universitario’, *Revista de Investigación Educativa* **21**(1), 157–182.
- Tejero-González, C. M. & Castro-Morera, M. (2011), ‘Validación de la escala de actitudes hacia la estadística en estudiantes españoles de ciencias de la actividad física y del deporte’, *Revista Colombiana de Estadística* **34**(2), 1–14.
- Therneau, T., Atkinson, B. & Ripley, B. (2017), *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-11.
*<https://CRAN.R-project.org/package=rpart>
- USTA (2010), Dimensión de la política docente, Documento técnico, Universidad Santo Tomás, Bogotá.
- Vaillant, D. (2007), ‘Mejorando la formación y el desarrollo profesional docente en latinoamérica’, *Revista Pensamiento Educativo* **41**(2), 207–222.