Revista de la Educación Superior

RESU 200

Vol. 50 · 2021

ANUIES
resu.anuies.mx

ARTÍCULO

Evaluación docente en una universidad pública mexicana

Teacher's performance evaluation at a Mexican public university

Rodrigo Polanco-Bueno*, Angélica Buendía-Espinosa**
y Eduardo Peñalosa-Castro***

*Departamento de Derecho, División de Ciencias Sociales y Humanidades Universidad Autónoma Metropolitana-Azcapotzalco.
**Departamento de Producción Económica, División de Ciencias Sociales y Humanidades, Universidad Autónoma Metropolitana-Xochimilco
***Departamento de Ciencias de la Comunicación, División de Ciencias de la Comunicación y Diseño, Universidad Autónoma Metropolitana-Cuajimalpa

Recibido el 3 de mayo de 2021; Aprobado el 16 de noviembre del 2021

Resumen

Este artículo reporta los resultados del proceso de diseño y validación psicométrica de un instrumento de evaluación de la docencia en una universidad pública mexicana. El cuestionario fue diseñado por un equipo constituido por directivos, académicos y estudiantes y sometido a un análisis de validez de contenido. El instrumento final, compuesto por 26 ítems distribuidos en cuatro dimensiones se aplicó a una muestra de estudiantes que arrojó 53,192 registros que sirvieron de base para realizar los análisis psicométricos. Los análisis de consistencia interna mediante los métodos de Alfa de Chronbach y mitades equivalentes, arrojaron coeficientes aceptables, tanto a nivel de la prueba total como de cada una de sus dimensiones. El análisis de

0185-2760/© 2016 Asociación Nacional de Universidades e Instituciones de Educación Superior A.C. Este es un artículo Open Access bajo la licencia CC BY-NC-ND (http://creativecommons.org/licenses/by-nc-nd/4.0/).

ítems realizado con los métodos de distribución de frecuencias, correlaciones ítem prueba y contrastación de grupos altos y bajos mostraron un comportamiento adecuado de cada uno de los elementos del cuestionario. Finalmente, la validación de constructo se pudo verificar para tres de las dimensiones con análisis factorial exploratorio y confirmatorio mediante sistema de ecuaciones estructurales (EQS). No fue posible realizarlo para la cuarta dimensión, debido a su limitado número de ítems.

PALABRAS CLAVE: Evaluación docente; Profesores universitarios; Análisis psicométrico; Calidad de la docencia

ABSTRACT This paper reports the results of a process of design and psychometric validation of a teaching assessment instrument in a Mexican public university. A team, composed by directors, faculty and students of the educational institution designed the instrument and submitted it to a process of content validity. The final questionnaire, composed by 26 items distributed into four dimensions, was applied to a student sample of 53,192 records, which were the input for the psychometric analyses. Internal consistency analyses using the Crhonbach'a Alfa and split half reliability methods, yielded satisfactory coefficients, both at the level of the total test and of each of its dimensions. Item analysis was carried out by the frequency distribution methods, item-test and item-dimension correlations, and higher and lower group comparisons. Again, all items showed a proper performance on these tests. Finally, construct validation could be verified for three dimensions with an exploratory factor analysis and was confirmed by structural equation system (EQS). It was not possible to do it for the fourth dimension, due to its limited number of items.

KEYWORDS: Teacher assessment; College teachers; Psychometric analysis; Teaching quality

INTRODUCCIÓN

Con la pandemia causada por la COVID-19, diversos procesos se modificaron en las universidades e Instituciones de Educación Superior. La docencia se constituyó en la función sustantiva que requirió especial atención, dado que se vivió una transición obligada a la educación no presencial, misma que una vez implementada ha requerido de seguimiento

y de mejora continuas. El objetivo de este trabajo es presentar la validez y la confiabilidad de un instrumento para obtener la opinión del alumnado con respecto al desempeño docente, en la Universidad Autónoma Metropolitana (UAM) en México. Aunque en la universidad ya se aplicaba un cuestionario con la misma intencionalidad, éste fue modificado para atender las nuevas condiciones en que operan las actividades docentes. Así, se busca proponer políticas y estrategias institucionales para, con base en la opinión del alumnado como un elemento relevante, mejorar la evaluación y la calidad de la docencia.

La evaluación de la docencia es un proceso complejo. Las acciones para medir y valorar las competencias en el ámbito educativo superior requieren de repensar las medidas y los procedimientos vigentes. Hay algunos elementos clave que favorecerían el perfeccionamiento de prácticas de valoración del desempeño docente. Definir qué se evalúa en relación exclusivamente con la docencia otorga bases para planificar la evaluación de manera ordenada y establecer categorías de análisis, dado que en esta función universitaria interactúan diversos elementos como la infraestructura en la que se imparte la docencia, las especificaciones contractuales y laborales de los docentes, la organización y el código universitario, la interacción profesor-alumno, por mencionar algunos (Rueda y Sánchez, 2018).

La evaluación de la docencia en la UAM es una actividad regulada específicamente en el artículo 215 del Reglamento de Ingreso, Promoción y Permanencia del Personal Académico (RIPPPA), que establece las actividades para la función de docencia, y en el artículo 4 del Reglamento de Alumnos establece los derechos de los alumnos. El primero específica las tareas asociadas a la función docencia que consideran la planeación, organización preparación y conducción del proceso de enseñanza-aprendizaje, de acuerdo con los planes y programas de estudio aprobados por los órganos colegiados correspondientes. Se suman los artículos 218, 218 Bis, 219 al 220 Ter, éste último señala que los consejos divisionales definirán los criterios y procedimientos para la evaluación de las actividades de docencia, investigación y preservación y difusión de la cultura, en función de los objetivos de la propia división. En el caso de la docencia, la evaluación deberá considerar periodos trimestrales, los informes anuales de los profesores y la opinión de los alumnos mediante encuestas idóneas que contemplen los elementos del artículo 215 del RIPPPA (UAM, 2020). El artículo 4 del Reglamento de Alumnos establece los derechos de los alumnos en cuanto a las condiciones institucionales, académicas y materiales en que el alumnado recibe su formación. Entre ellas destaca el derecho a opinar en relación con el desarrollo y con los resultados de los programas de las UEA.

La evaluación del cumplimiento de las funciones sustantivas por los profesores en la UAM ha estado asociada, desde hace por lo menos 30 años, al sistema de estímulos y becas para el reconocimiento y deshomologación salarial que priva en la universidad: beca a la permanencia, beca a los grados académicos, beca a la docencia, estímulo a la

investigación y estímulo a la trayectoria académica. Particularmente, el 10 de junio de 1992, fueron aprobadas las becas al reconocimiento de la carrera docente del personal académico (Arbesú, 2004).

Buendía *et al*. señalan que diversas reflexiones en la comunidad universitaria dan cuenta de las dificultades y retos que la evaluación de la docencia representa para la Universidad. Algunas son: a) asociada más a un proceso de control que de mejora de la función docente; b) es un proceso incompleto en cuanto a sus objetivos, fines y usos por los actores involucrados; y c) responde más al incentivo económico que representa la mejora de la práctica docente que a la colaboración entre los actores principales, el personal académico y los alumnos, para la retroalimentación. Particularmente, el instrumento de opinión estudiantil presenta diversas problemáticas:

a) Carece de una aproximación teórica sobre la enseñanza, pero se acerca más a la teoría conductual del aprendizaje y la transmisión del conocimiento inerte, lo que es contrario a la figura de profesor investigador que sustenta el modelo universitario de la UAM en general y de la diversidad de modelos académicos que promueven cada una de sus unidades académicas (Whitehead, 1929; Renkl, Mandl y Gruber, 1996);
b) Limita a un modelo único de ser docente y responde al "ideal" de tareas asignadas que establece la propia Universidad en el artículo 215 del RIPPPA;
c) Está realizado para efectos de un control académico-administrativo del docente que no se aplica en la gestión de la docencia;
d) Está organizado en tres niveles: autevaluación, organizativo y desempeño. De las 21 preguntas que integran las dos dimensiones a evaluar (organizativa y de desempeño), 16 se relacionan con el desempeño (asistencia y puntualidad; presentación del programa, objetivos y bibliografía; evaluación del mismo; duración de las sesiones; dominio del tema; horas impartidas; asesorías extra clase; cumplimiento del programa y bibliografía); cinco preguntas de "carácter informativo" valoran el entusiasmo del profesor al impartir la clase; el clima de respeto y cordialidad en el aula; los recursos didácticos y pedagógicos empleados por el maestro y la necesidad de formación didáctica, pedagogía y manejo de grupos (Arbesú, 2004);
e) No reconoce diferencias disciplinares y entre niveles educativos (licenciatura y posgrado);
f) Valora el desempeño del profesor desde una perspectiva proceso-producto (Arbesú, 2004); y no permite obtener información sobre la práctica pedagógica en el contexto del modelo universitario de la UAM y de los modelos académicos de sus cinco unidades;
g) No corresponde con las diferencias que existen en la práctica educativa para los distintos niveles educativos, licenciatura y posgrado; así como para las grandes áreas de conocimiento que se cultivan en la Universidad;

h) El tiempo y la forma de comunicar los resultados de la opinión estudiantil, así como del proceso general de evaluación; no contribuyen a establecer estrategias de mejora de la docencia.

En el contexto de la pandemia causada por el Sars-Cov-2, se implementó en la Universidad el Proyecto Emergente de Enseñanza Remota (PEER), el cual permitió continuar con las actividades docentes, pero, al mismo tiempo, visibilizó viejos problemas tales como: desigualdad de los estudiantes y profesores en el acceso a equipo y conexión, condiciones de aprendizaje de estudiantes y profesores, deficiente formación docente en tecnologías digitales, estrategias de enseñanza, aprendizaje y evaluación en educación no presencial, rezago educativo y logros de aprendizaje, establecimiento de comunicación efectiva y apoyo a la salud emocional de las comunidades educativas.

Con base en lo anterior una comisión especial propuso: a) fortalecer la formación docente en estrategias de enseñanza-aprendizaje y de evaluación de los aprendizajes para clases vía remota y presencial; b) fortalecer los conocimientos y habilidades tecnologías digitales; y c) mantener y mejorar los apoyos en todos los niveles de la gestión a la comunidad universitaria (coordinaciones, jefaturas, servicios, en el área técnica, a nivel institucional, apoyos psicológicos, becas, etc.).

Método

Objetivos

La elaboración del instrumento recayó en una comisión integrada por personal académico de la Universidad y ha sido aplicado en dos periodos lectivos en el 2020, trimestres de otoño y de primavera. El método para medir la validez y confiabilidad del instrumento considera calcular la confiabilidad del instrumento: los reactivos en lo individual y las dimensiones como agregados de reactivos. Enseguida se realiza el análisis de datos para obtener el análisis factorial exploratorio del instrumento y confirmatorio del instrumento.

Población y muestra

La población de alumnos activos de la UAM durante el trimestre de otoño, que corresponde a la población objeto de esta investigación fue de 47,371, incluyendo estudiantes de licenciatura y posgrado. La muestra de estudiantes que participó en el estudio fue de 18,992, la cual es adecuada al tamaño de la población, ya que supera el tamaño de muestra

esperado de 12,316 con un nivel de confianza del 99% y un nivel de error del 1%. El hecho de que un mismo alumno o alumna, participó en la evaluación de más de una unidad de enseñanza aprendizaje (UEA)[1] hizo que el total de registros de encuestas que fueron contabilizadas para efectos de los análisis reportados en este artículo, fuera de 53,192.

Una de las estrategias para avanzar en las propuestas fue la renovación del instrumento para obtener la opinión del alumnado con respecto a la docencia remota. El diseño y validación del mismo estuvo a cargo de un equipo de académicos de la UAM, con la asesoría de un experto en evaluación docente de la Universidad Nacional Autónoma de México[2]. La versión final del cuestionario quedó constituida por un reactivo dicotómico (ítem1), un reactivo de escala numérica 1-10 (ítem 25), un reactivo de respuesta abierta (ítem 26), y 23 reactivos de formato Likert (ítems 2 a 24, los cuales fueron clasificados en cuatro dimensiones (ver tabla 2):

1) Organización de la UEA. Cinco ítems: Cuatro escala Likert y uno dicotómico.
2) Práctica docente: 13 ítems de formato Likert.
3) Autoevaluación: Cuatro ítems de formato Likert.
4) Evaluación global: Tres ítems: dos reactivos de formato Likert, uno de escala numérica (1-10) y un reactivo de opción múltiple.

Tabla 2. Instrumento aplicado para la Evaluación de la Docencia

REACTIVO/NIVEL					
Organización de la UEA					
La o el profesor presentó y entregó el programa de la UEA en la primera semana de clase o antes	**No**			**Sí**	
	Totalmente en desacuerdo	**En desacuerdo**	**De acuerdo**	**Totalmente de acuerdo**	**No aplica**
1. El programa incluyó: objetivos, contenidos temáticos, estrategias de enseñanza aprendizaje, bibliografía, otros apoyos didácticos, formas de evaluación y cronograma de actividades.					
2. El programa incluyó actividades realizadas en modalidad sincrónica (tiempo real) y asincrónica (sin interacción simultánea).					
3. Se acordaron normas, criterios de convivencia armónica y respetuosa, así como mecanismos de comunicación en el grupo.					
4. En general, las actividades se han realizado conforme a lo programado.					

1 En la UAM se denomina a las asignaturas Unidades de Enseñanza Aprendizaje (UEA)

2 Los autores agradecen al equipo que diseño y validó el instrumento: Angélica Buendía Espinosa, Alia Balam González, Lidia E. Blásquez Martínez, Tomás Ejea Mendoza, Mercedes Jatziri Gaitán González, José Mariano García Garibay, Carla Garzón Flores, Luis Montaño Hirose, Daniel Montealegre García, Esther Morales Franco, Melina Olivares Juárez y Mario Rueda Beltrán (investigador del Instituto de Investigaciones sobre la Universidad y la Educación de la UNAM, asesor del proyecto).

Práctica docente
5. La o el profesor muestra conocimiento amplio sobre los temas del programa.
6. Se favorece la participación individual y colectiva para el desarrollo de los conocimientos.
7. Se favorece la participación individual y colectiva para el desarrollo de las habilidades (comunicación, uso de lenguaje, pensamiento crítico, resolución de problemas, trabajo en equipo).
8. Se promueve tu aprendizaje autónomo con base en la implementación de actividades, recursos y apoyos didácticos.
9. Las actividades prácticas se han realizado conforme a las necesidades de la UEA y contribuyen al logro de los aprendizajes y experiencias.
10. Se resuelven las dudas con base en explicaciones compresibles y fortalecen los aprendizajes sobre los contenidos abordados.
11. Se incentiva el desarrollo de prácticas de investigación para el fortalecimiento de los aprendizajes.
12. Se impulsa la aplicación práctica y/o analítica de los conocimientos adquiridos.
13. Se promueve un ambiente de respeto, confianza y colaboración
14. En general, todas las actividades se han realizado con pleno respeto a los derechos universitarios, como son, entre otros, la igualdad, diversidad y pluralidad de la comunidad universitaria en general y del alumnado del grupo en particular.
15. La o el profesor imparte asesorías cuando le son solicitadas.
16. Los criterios y formas de evaluación establecidas en el programa se han respetado.
17. Recibes retroalimentación de las modalidades de evaluación implementadas durante el curso.
18. Autoevaluación
19. Me he presentado puntualmente a clases y he permanecido la duración total de las sesiones.
20. He participado en clase expresando dudas, aportando ejemplos, respondiendo preguntas y trabajando en equipo.
21. He cumplido con los requisitos y actividades académicas establecidas en el programa.
22. Hasta el momento he logrado los aprendizajes esperados de acuerdo con los objetivos del programa.
Evaluación global
23.Tomaría otro de los cursos que imparte la o el profesor.
24. En las condiciones extraordinarias del PEER la o el profesor mostró especial interés en el desarrollo de la UEA.
25. Con base en lo anterior califica del 1 al 10 el desempeño del profesor el trimestre 20-P: (1 es nada satisfactorio y 10 es muy satisfactorio).
1 2 3 4 5 6 7 8 9 10
○ ○ ○ ○ ○ ○ ○ ○ ○ ○
26. Comentarios y sugerencias a tu profesor

En esencia, el nuevo instrumento de evaluación se basó en el modelo universitario de la UAM, y su validación permitiría asegurar que cumple con las características suficientes y necesarias para valorar las actividades que el profesor debe realizar en función de lo que este modelo plantea.

Procedimiento de recogida y análisis de datos

El instrumento se aplicó en línea a las y los estudiantes determinados en la muestra. Se realizó el análisis del comportamiento de las respuestas, con el fin de verificar la consistencia interna del mismo (confiabilidad) y estimar el grado en el que mide lo que pretende medir, es decir los rasgos latentes definidos en las cuatro dimensiones a las que se hizo referencia (validez de constructo). El proceso de validación se realizó mediante los siguientes análisis:

1) Validación cualitativa del instrumento por parte de jueces expertos
2) Consistencia interna del cuestionario
   a) Procedimiento de mitades equivalentes (Spearman-Brown)
   b) Alfa de Chronbach

3) Análisis de ítems
   a) Método de distribución de frecuencias
   b) Correlaciones ítem-prueba total
   c) Correlaciones ítem-dimensión
   d) Diferencia de medias entre grupos extremos

4) Validación de constructo mediante análisis factorial exploratorio y factorial confirmatorio a partir de un análisis de ecuaciones estructurales.

Dado que la mayor parte de los análisis requieren datos en escalas ordinales o de intervalo, el ítem de respuesta abierta fue eliminado del análisis, y el ítem dicotómico fue analizado de manera separada. Asimismo, para los análisis factoriales se eliminaron los dos últimos reactivos, dado que optamos por analizar al menos tres reactivos de cada dimensión o factor.

Resultados

1. Validación por jueces expertos

En el diseño del instrumento participó un investigador experto en evaluación de la docencia y externo a la Universidad. Su apoyo consistió en revisar las modificaciones realizadas y en validar con un grupo de expertos su pertinencia.

2. Consistencia interna (prueba total)

Para realizar los análisis de consistencia interna, se trabajó con los 23 ítems de formato Likert y con el ítem de escala. De esta manera, a partir de las respuestas de los y las alumnas, se pudo estimar la consistencia interna a partir de los métodos de mitades equivalentes (Spearman Brown) y alfa de Chronbach.

Los resultados de este análisis para la prueba total se exhiben en la tabla 2, los cuales muestran una consistencia muy buena (superior a 0.9) en los dos tipos de análisis, lo cual refleja que lo que mide el instrumento en su totalidad lo hace de manera consistente.

Tabla 2. Resultados de los procedimientos de mitades equivalentes y a de Chronbach.

N	N ítems	Mitades equivalentes Pearson	Corrección Spearman-Brown	a de Chronbach
53192	24	0.901	0.944	0.962

3. Consistencia interna (dimensiones)

Adicionalmente, se realizaron análisis mediante el método de a de Chronbach, para estimar la consistencia interna de cada una de las dimensiones del cuestionario. El análisis de todas las dimensiones se muestra en la tabla 3.

Tabla 3. Resultados del análisis de consistencia interna por el método a de Chronbach para cada una de las dimensiones

Organización de la UEA		
N	N elementos	a de Chronbach
53192	5	0.898
Práctica docente		
N	N elementos	a de Chronbach
53192	13	0.954
Autoevaluación		
N	N elementos	a de Chronbach
53192	4	0.821
Evaluación global		
N	N elementos	a de Chronbach
53192	3	0.807

Como puede apreciarse en la tabla, todos los índices de consistencia fueron superiores a 0.8, indicando una buena consistencia interna.

4. Análisis de ítems

El análisis del *poder discriminativo* de los ítems, es decir de su capacidad para distinguir opiniones favorables de las desfavorables, se realizó mediante los métodos de distribución de frecuencias, correlaciones ítem-prueba total, correlaciones ítem-dimensión y diferencia de medias entre grupos extremos. Se reportarán los resultados referentes a los ítems 2 a 24, que son los que utilizan formato Likert. Los ítems 1 y 25, que utilizan otros tipos de formato, serán analizados por separado.

a) Distribución de frecuencias de las opciones por ítem

El primer método, de distribución de frecuencias, verifica el sesgo que pudiera tener un ítem, basado en la forma como se distribuyen sus respuestas en sus diferentes niveles. De esta manera, cuando uno de los niveles es elegido por el 80% o más de los respondientes, se considera que el ítem puede tener un sesgo importante y, en consecuencia, debe ser modificado o sustituido por uno nuevo.

Los resultados de este análisis se presentan en la tabla 4, los cuales muestran que en ningún caso los ítems mostraron una concentración mayor al 80% en alguno de sus niveles. El ítem 6 y el ítem 15 fueron los únicos que superaron el 70%, lo cual está aún dentro de los límites aceptables.

Tabla 4. Distribución de frecuencias de los distintos niveles de los ítems

Ítem	Tot. Desacuerdo		Desacuerdo		De acuerdo		Tot. Acuerdo	
	f	%	f	%	f	%	f	%
2	2054	3.9	1814	3.4	12980	24.4	35376	66.5
3	2104	4.0	2536	4.8	14436	27.1	32881	61.8
4	2061	3.9	2047	3.8	14012	26.3	34072	64.1
5	1998	3.8	2339	4.4	14211	26.7	33931	63.8
6	1590	3.0	1667	3.1	11425	21.5	37897	71.2
7	1969	3.7	2801	5.3	14548	27.3	32973	62.0
8	2009	3.8	3019	5.7	15006	28.2	32141	60.4
9	1711	3.2	2424	4.6	15216	28.6	33347	62.7
10	2092	3.9	3023	5.7	15724	29.6	31006	58.3
11	2365	4.4	3342	6.3	13834	26.0	33033	62.1
12	1929	3.6	3246	6.1	16198	30.5	30201	56.8
13	1859	3.5	2960	5.6	16307	30.7	31104	58.5
14	1763	3.3	1686	3.2	12837	24.1	36146	68.0
15	1544	2.9	1218	2.3	12495	23.5	37324	70.2
16	2334	4.4	3127	5.9	14658	27.6	28367	53.3
17	1690	3.2	1671	3.1	14615	27.5	34136	64.2
18	2553	4.8	3932	7.4	15238	28.6	29995	56.4

19	1035	1.9	1471	2.8	13797	25.9	35463	66.7
20	1179	2.2	4266	8.0	22255	41.8	23599	44.4
21	955	1.8	1353	2.5	15752	29.6	34652	65.1
22	2049	3.9	4385	8.2	21033	39.5	25151	47.3
23	4904	9.2	4327	8.1	13602	25.6	29594	55.6
24	2666	5.0	3013	5.7	14007	26.3	32595	61.3

b) Correlaciones ítem-prueba total (o dimensiones)

El segundo método de análisis de reactivos consistió en correlacionar, mediante la prueba Rho de Spearman para variables ordinales el puntaje asignado a cada ítem con la calificación total obtenida en la combinación de todos los ítems de formato Likert, así como con la puntuación obtenida en cada dimensión.

Tabla 5. Correlaciones Rho de Spearman y niveles de significación de cada ítem con el puntaje del cuestionario total.

Ítem	Correlación	Significación
2	0.683	< 0.001
3	0.711	< 0.001
4	0.726	< 0.001
5	0.729	< 0.001
6	0.692	< 0.001
7	0.770	< 0.001
8	0.785	< 0.001
9	0.756	< 0.001
10	0.795	< 0.001
11	0.791	< 0.001
12	0.777	< 0.001
13	0.793	< 0.001
14	0.735	< 0.001
15	0.717	< 0.001
16	0.735	< 0.001
17	0.760	< 0.001
18	0.791	< 0.001
19	0.472	< 0.001
20	0.576	< 0.001
21	0.507	< 0.001
22	0.711	< 0.001
23	0.764	< 0.001
24	0.755	< 0.001

Como puede advertirse en la tabla 5, todas las correlaciones fueron significativas y a excepción de los ítems 19, 20 y 21, superiores a 0.7.

Tabla 6. Correlaciones Rho de Spearman y niveles de significación de cada ítem con el puntaje del la dimensión Organización de la UEA.

Ítem	Correlación	Significación
2	0.823	< 0.001
3	0.862	< 0.001
4	0.835	< 0.001
5	0.835	< 0.001

En la tabla 6 se aprecia un comportamiento similar en las correlaciones de la dimensión Organización de la UEA, con correlaciones significativas y superiores a 0.8.

Tabla 7. Correlaciones Rho de Spearman y niveles de significación de cada ítem con el puntaje de la dimensión Práctica Docente.

Ítem	Correlación	Significación
6	0.712	< 0.001
7	0.798	< 0.001
8	0.814	< 0.001
9	0.783	< 0.001
10	0.823	< 0.001
11	0.817	< 0.001
12	0.818	< 0.001
13	0.825	< 0.001
14	0.756	< 0.001
15	0.736	< 0.001
16	0.787	< 0.001
17	0.776	< 0.001
18	0.823	< 0.001

Los resultados correspondientes a la dimensión Práctica Docente, presentados en la tabla 7, nuevamente manifiestan valores de correlación y niveles de significación adecuados para todos los ítems.

Tabla 8. Correlaciones Rho de Spearman y niveles de significación de cada ítem con el puntaje de la dimensión Autoevaluación.

Ítem	Correlación	Significación
19	0.720	< 0.001
20	0.825	< 0.001
21	0.730	< 0.001
22	0.808	< 0.001

Algo similar ocurre en la dimensión Autoevaluación. Todos los ítems revelan valores aceptables como se puede observar en la tabla 8. Este hecho es interesante, si se toma en consideración que los ítems 19, 20 y 21 habían mostrado correlaciones con el puntaje total inferiores a 0.7.

Tabla 9. Correlaciones Rho de Spearman y niveles de significación de cada ítem con el puntaje de la dimensión Evaluación Global.

Ítem	Correlación de Pearson	Significación
23	0.887	< 0.001
24	0.834	< 0.001

Finalmente, las correlaciones de los ítems pertenecientes a la dimensión Evaluación Global con el puntaje de esta dimensión, evidencian valores y niveles de significación adecuados, como puede apreciarse en la tabla 9.

Tabla 10. Correlación biserial puntual entre el ítem 1 y los puntajes de la prueba total y de dimensión Evaluación Organización de la UEA.

Ítem	Dimensión	Correlación Biserial puntual	Significación
1	Total	0.322	< 0.001
1	Organización	0.415	< 0.001

Dado que los reactivos 1 y 25 utilizaron una escala diferente a la escala Likert, su análisis se realizó por separado. En el caso del ítem 1, se empleó el coeficiente de correlación biserial puntual para correlacionar el puntaje nominal (si/no) del reactivo con el puntaje intervalar de la dimensión Evaluación Organización y del puntaje total del cuestionario. Como puede apreciarse en la tabla 10, aunque en los dos casos las correlaciones fueron inferiores a 0.5, presentaron buenos niveles de significación.

Tabla 11 Correlación Producto Momento de Pearson entre el ítem 25 y los puntajes de la prueba total y de dimensión Evaluación Global.

Ítem	Dimensión	Correlación Pearson	Significación
25	Total	0.940	< 0.001
25	Global	0.683	< 0.001

En el caso del ítem 25, dado que tanto la escala del ítem como de la dimensión global y del puntaje total son intervalares, se utilizó el Producto Momento de Pearson. Como se indica en la tabla 11, el ítem 25 exhibe correlaciones significativas tanto con el puntaje total del instrumento, como con la dimensión global.

c) Contrastación de grupos alto y bajo (diferencias significativas)

Con el fin de analizar el poder discriminativo de los ítems se compararon los puntajes arrojados por cada ítem, en sujetos con una opinión general muy favorable y sujetos con una opinión muy desfavorable. Para este fin, Se conformaron dos grupos a partir de los puntajes totales en el cuestionario: Un grupo *Alto*, constituido por los sujetos cuyo puntaje se ubicó dentro del 30% de los puntajes más altos (opinión más favorable) y un grupo *Bajo*, constituido por los sujetos cuyo puntaje se ubicó dentro del 30% de los puntajes más bajos (opinión más desfavorable).

Una vez conformados estos dos grupos, se procedió a analizar las diferencias, entre los grupos alto y bajo, de los puntajes de *cada reactivo*. Dado que la escala Likert no es, en rigor, una escala de intervalo sino una escala ordinal, el análisis de comparación de medias se realizó con dos pruebas. La prueba no paramétrica U de Mann Whitney, para escalas ordinales, y la prueba t de Student para escalas de intervalo, que es la que con mayor frecuencia se usa, a pesar de las limitaciones mencionadas.

Tabla 12. Diferencias de medias entre los grupos alto y bajo para la prueba total, mediante las pruebas t de Student y U de Mann-Whitney.

Ítem	N	Media altos	Media bajos	t	gl	Sig.	U	Sig.
2	32,000	3.99	2.74	149.701	16438.600	< 0.001	26950744.500	< 0.001
3	32,000	3.98	2.60	163.390	16688.713	< 0.001	20654961.000	< 0.001
4	32,000	3.99	2.63	163.772	16283.484	< 0.001	19852928.500	< 0.001
5	32,000	3.99	2.66	168.803	16396.400	< 0.001	19097933.000	< 0.001
6	32,000	3.99	2.82	145.399	16191.615	< 0.001	30328592.000	< 0.001
7	32,000	3.99	2.54	183.543	16208.605	< 0.001	13195123.000	< 0.001
8	32,000	3.99	2.48	190.682	16217.517	< 0.001	10488684.000	< 0.001
9	32,000	3.99	2.68	180.578	16243.714	< 0.001	15876984.500	< 0.001
10	32,000	3.99	2.46	197.831	16230.535	< 0.001	7731970.000	< 0.001
11	32,000	4.00	2.48	199.100	16145.749	< 0.001	10235796.000	< 0.001
12	32,000	3.99	2.45	190.397	16469.983	< 0.001	9928796.000	< 0.001
13	32,000	3.99	2.52	196.827	16230.792	< 0.001	8474044.000	< 0.001
14	32,000	4.00	2.71	160.935	16084.711	< 0.001	20967076.000	< 0.001
15	32,000	4.00	2.82	152.968	16051.463	< 0.001	24990812.500	< 0.001
16	32,000	3.96	2.13	184.969	16795.687	< 0.001	11569794.000	< 0.001
17	32,000	4.00	2.63	170.950	16098.635	< 0.001	14645895.000	< 0.001
18	32,000	3.98	2.33	199.914	16531.921	< 0.001	8406047.500	< 0.001
19	32,000	3.94	3.00	98.597	17449.608	< 0.001	57946352.500	< 0.001
20	32,000	3.82	2.61	126.830	19845.667	< 0.001	38065053.500	< 0.001
21	32,000	3.96	3.11	111.982	17420.073	< 0.001	50579002.500	< 0.001
22	32,000	3.92	2.53	178.485	18520.585	< 0.001	19190028.500	< 0.001

23	32,000	3.95	2.16	210.345	17343.625	< 0.001	12803099.500	< 0.001
24	32,000	3.98	2.48	186.468	16628.962	< 0.001	14438104.000	< 0.001
25	32,000	9.79	6.72	153.488	17443.069	< 0.001	18563636.500	< 0.001

Los resultados de la comparación de medias para la prueba total, se despliegan en la tabla 12,. Como puede apreciarse en la tabla, las diferencias de medias en todos los ítems fueron significativas en ambas pruebas y para todos los ítems del cuestionario.

Tabla 13. Diferencias de medias entre los grupos alto y bajo para la dimensión Organización de la UEA, mediante las pruebas t de Student y U de Mann-Whitney.

Ítem	N	Media altos	Media bajos	t	gl	Sig.	U	Sig.
2	32,000	4.00	2.57	174.404	15999.000	< 0.001	14000000.000	< 0.001
3	32,000	4.00	2.43	190.520	15999.000	< 0.001	7984000.000	< 0.001
4	32,000	4.00	2.53	179.227	15999.000	< 0.001	12192000.000	< 0.001
5	32,000	4.00	2.58	185.018	15999.000	< 0.001	11960000.000	< 0.001

Los resultados de la comparación de medias de los reactivos pertenecientes a la dimensión de Organización de la UEA se muestran en la Tabla 13. Nuevamente, puede apreciarse que la diferencia entre los grupos alto y bajo fue significativa tanto en la prueba t de Student como en la prueba U de Mann Whitney, evidenciando un adecuado poder discriminativo de estos reactivos en relación con la dimensión a la que pertenecen.

Tabla 14. Diferencias de medias entre los grupos alto y bajo para la dimensión Práctica Docente, mediante las pruebas t de Student y U de Mann-Whitney.

Ítem	N	Media altos	Media bajos	t	gl	Sig	U	Sig
6	32,000	4.00	2.82	148.757	15999.000	< 0.001	28496000.0	< 0.001
7	32,000	4.00	2.53	189.067	15999.000	< 0.001	10880000.0	< 0.001
8	32,000	4.00	2.47	197.167	15999.000	< 0.001	7968000.00	< 0.001
9	32,000	4.00	2.66	186.823	15999.000	< 0.001	13328000.0	< 0.001
10	32,000	4.00	2.43	201.894	15999.000	< 0.001	5880000.00	< 0.001
11	32,000	4.00	2.47	205.214	15999.000	< 0.001	8080000.00	< 0.001
12	32,000	4.00	2.41	198.529	15999.000	< 0.001	6416000.00	< 0.001
13	32,000	4.00	2.49	202.656	15999.000	< 0.001	6040000.00	< 0.001
14	32,000	4.00	2.69	165.036	15999.000	< 0.001	18888000.0	< 0.001
15	32,000	4.00	2.80	156.721	15999.000	< 0.001	23048000.0	< 0.001
16	32,000	4.00	2.09	195.976	15999.000	< 0.001	7072000.00	< 0.001
17	32,000	4.00	2.63	172.667	15999.000	< 0.001	13904000.0	< 0.001
18	32,000	4.00	2.31	206.230	15999.000	< 0.001	6032000.00	< 0.001

Algo similar a lo anterior, ocurrió con respecto a los ítems de la dimensión Práctica docente, cuyos resultados se presentan en la tabla 14. Nuevamente los ítems mostraron un buen poder discriminativo en relación con su dimensión, como lo atestigua el hecho de que todas las pruebas de diferencia de medias realizadas para todos los ítems arrojaron valores significativos.

Tabla 15. Diferencias de medias entre los grupos alto y bajo para la dimensión Autoevaluación, mediante las pruebas t de Student y U de Mann-Whitney.

Ítem	N	Media altos	Media bajos	t	gl	Sig	U	Sig.
19	32,000	4.00	3.68	151.746	15999.000	< 0.001	20944000.000	< 0.001
20	32,000	4.00	2.43	198.307	15999.000	< 0.001	3864000.000	< 0.001
21	32,000	4.00	2.88	165.070	15999.000	< 0.001	21352000.000	< 0.001
22	32,000	4.00	2.49	212.389	15999.000	< 0.001	6528000.000	< 0.001

La tabla 15 permite apreciar los resultados de las comparaciones de medias de los reactivos de la dimensión Autoevaluación. Nuevamente las pruebas t de Student como U de Mann Whitney arrojaron diferencias significativas para cada uno de los ítems.

Tabla 16. Diferencias de medias entre los grupos alto y bajo para la dimensión Evaluación Global, mediante las pruebas t de Student y U de Mann-Whitney.

Ítem	N	Media altos	Media bajos	t	gl	Sig.	U	Sig.
23	32,000	4.00	2.01	256.296	15999.000	< 0.001	3280000.000	< 0.001
24	32,000	4.00	2.37	203.658	15999.000	< 0.001	7576000.000	< 0.001
25	32,000	10.00	6.46	193.296	15999.000	< 0.001		

Finalmente, el análisis correspondiente a los ítems de la dimensión Evaluación global se ilustra en la tabla 16. En este análisis, el ítem 25, que está expresado en una escala de intervalo, solamente se analizó con la prueba t de Student. Nuevamente, este análisis nos permite constatar el poder discriminativo de los tres ítems correspondientes a la dimensión de Evaluación global.

d) Análisis Factorial Exploratorio

Se realizó un Análisis Factorial Exploratorio con el fin de reducir las dimensiones e identificar si las dimensiones se agrupan y conforman factores.

En primera instancia, se calculó la prueba de KMO y Bartlett, y se obtuvo el siguiente resultado:

Tabla 17. Prueba de KMO y Bartlett

Medida Kaiser-Meyer-Olkin de adecuación de muestreo		.971
Prueba de esfericidad de Bartlett	Aprox. Chi-cuadrado	912769.532
	gl	210
	Sig.	.000

Tenemos que el valor es de .971, lo cual habla de una alta correlación entre reactivos, y esto permitirá identificar los factores que se forman.

En la tabla 18, que aparece a continuación, se presentan los resultados del análisis factorial exploratorio; se optó por los buscar la agrupación de los cuatro factores que se indican en la página 7 de este artículo.

De estos cuatro componentes se tiene el 69.085% de la varianza total explicada, como aparece en la tabla 18, a continuación.

Tabla 18. Varianza total explicada

	Autovalores iniciales			Sumas de cargas al cuadrado de la extracción			Sumas de cargas al cuadrado de la rotación[a]
Componente	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado	Total
1	12.193	58.063	58.063	12.193	58.063	58.063	11.453
2	1.352	6.437	64.500	1.352	6.437	64.500	6.175
3	.963	4.585	69.085	.963	4.585	69.085	8.297

Este total de 69.085% acumulado supera lo que algunos autores han indicado como óptimo, que debería ser mayor a 50% en las ciencias sociales (Hair *et al*, 1998).

Por otro lado, tenemos los factores que pueden hacerse en la tabla 19. En ella se observa que cada uno de los tres grupos que se generan como factores superan correlaciones de .460, lo cual es alto de acuerdo con la literatura especializada. El valor más alto que se encuentra es de 0.870, que refleja un acuerdo importante de los participantes por ubicar en este factor a este reactivo.

Tabla 19. Factores que se encuentran en los reactivos 2 al 22

Pregunta 2	0.868	
Pregunta 3	0.759	
Pregunta 4	0.570	
Pregunta 5	0.734	
Pregunta 6		0.649
Pregunta 7		0.808
Pregunta 8		0.828
Pregunta 9		0.733

Pregunta 10	0.746	
Pregunta 11	0.827	
Pregunta 12	0.855	
Pregunta 13	0.838	
Pregunta 14	0.699	
Pregunta 15	0.637	
Pregunta 16	0.749	
Pregunta 17	0.470	
Pregunta 18	0.756	
Pregunta 19		0.870
Pregunta 20		0.799
Pregunta 21		0.779
Pregunta 22		0.480

En esta tabla no se incluyen los reactivos que forman parte del factor de la autoevaluación global, ya que solamente cuenta con dos reactivos Likert, y en esa medida no cumple con criterios cuantitativos de número de reactivos.

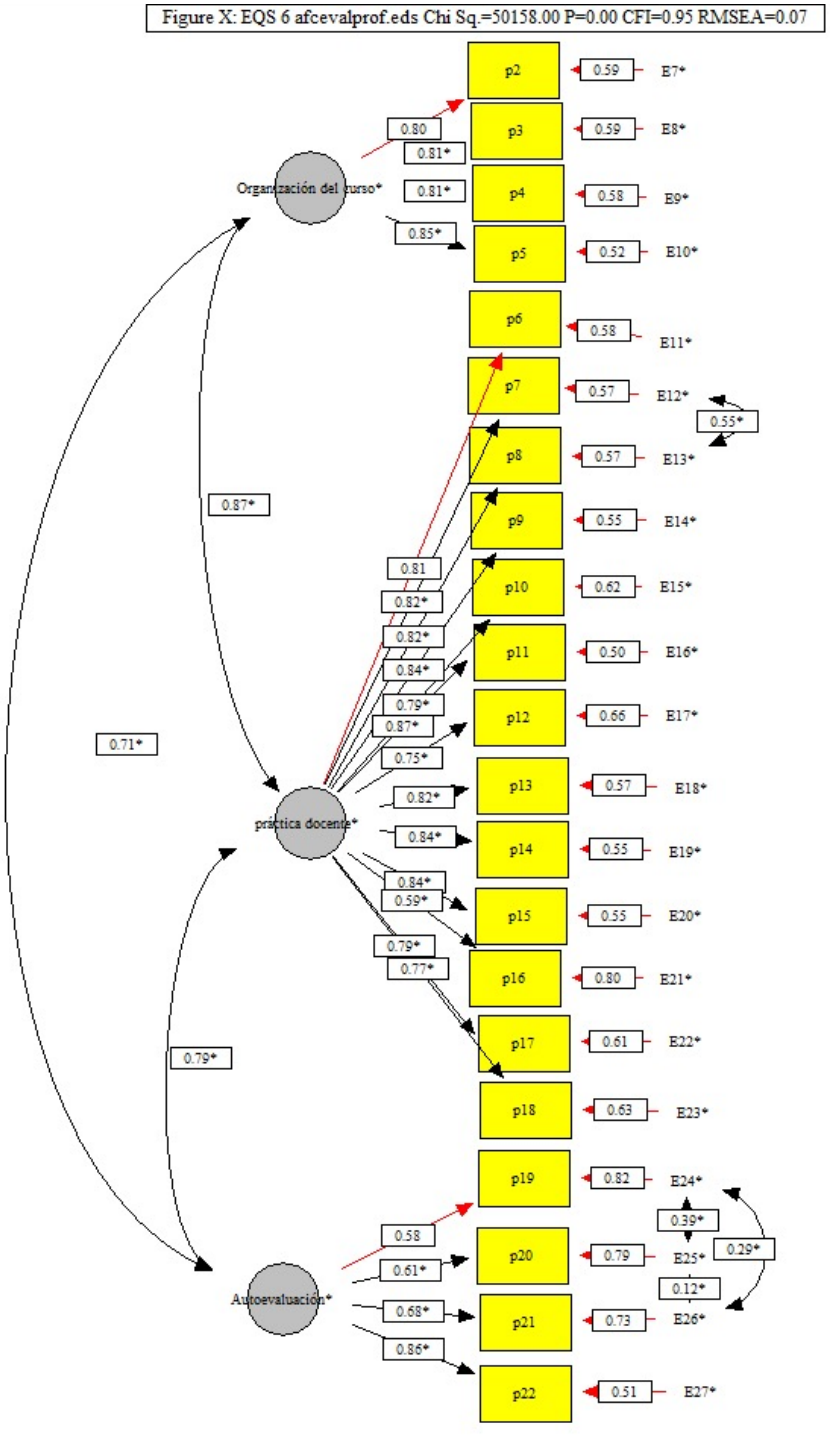
El instrumento final queda conformado por 21 reactivos, prácticamente todos los que contienen escala Likert, y se muestran en la tabla 2. La prueba de esfericidad de Bartlett fue significativa (912769, gl=210, sig > .001), y el indicador de adecuación del tamaño de la muestra Kaiser-Meyer-Olkin fue adecuado (0.971). La varianza explicada total ascendió a 69.085%.

e) Análisis Factorial Confirmatorio

La figura 1 muestra la información y el gráfico resultante del Análisis Factorial Confirmatorio, que se calculó con ayuda del *software* para ecuaciones estructurales EQS. Este análisis permitió confirmar los factores que el Análisis Factorial Exploratorio propuso. Los mismos reactivos que en este análisis agrupan factores también lo hacen en el confirmatorio, con lo cual se valida el instrumento que se aplicó con el fin de evaluar a los profesores del trimestre 2020-O (Trimestre de Otoño del 2020) de la UAM.

En la figura 1 se muestra, en la parte superior, que aun cuando la X^2 es significativa, y no debería serlo, pues el modelo puesto a prueba no debería ser significativamente diferente del modelo de la realidad, es un hecho que cuando la muestra es amplia (que es el caso) los valores identificados pueden ser significativamente diferentes en este parámetro. Descartado esto, los valores que nos parecen importantes son: el CFI, que debe ser mayor o igual a 0.95, y el RMSEA, que debe ser menor o igual que 0.08 (Byrne, 2008). En este caso contamos con un valor de CFI de 0.95 y de RMSEA de 0.07, por lo que el modelo de tres factores se valida.

Figura 1. Análisis Factorial Confirmatorio



Discusión y conclusiones

El campo de la evaluación docente es un tema en permanente cambio y alrededor del cual se han desarrollado un conjunto de conceptos, métodos y alcances muy diversos. Si bien es cierto que desde sus inicios la evaluación docente ha sido considerada como un indicador (o conjunto de indicadores) al servicio de la medición de la calidad de la docencia y, en este contexto, muy vinculada a procesos laborales y de estímulos salariales, cada vez es mayor la insistencia en utilizar sus resultados con fines de retroalimentación y mejora, así como un ingrediente fundamental en la formación y desarrollo profesional docente. En cualquier caso, cada día es mayor el uso y desarrollo de diferentes herramientas evaluativas en las instituciones educativas y, en particular, de las instituciones de educación superior.

La evaluación docente, como todo proceso evaluativo, está estrechamente vinculada a la toma de decisiones. En consecuencia, la calidad de las decisiones está influida, al menos en parte, por la calidad de los instrumentos utilizados en su evaluación. Desafortunadamente, la tarea de valorar el desempeño docente no siempre recae en un equipo de especialistas en educación y/o en evaluación, como lo atestigua Rueda (2008). Por ello, aparentemente son pocas las instituciones educativas que se dan a la tarea de realizar procesos de análisis de validez y confiabilidad, que den razón de la calidad de sus instrumentos evaluativos. La Universidad Autónoma Metropolitana, consciente de esta carencia ha aprovechado la coyuntura de la actualización de su instrumento de evaluación docente para realizar esta tarea, la cual constituyó el propósito de esta investigación, la cual se abocó al análisis tanto de la validez como de la confiabilidad del cuestionario.

En el proceso de diseño del instrumento participó un equipo constituido por directivos, académicos y estudiantes de la Universidad con el objeto de contar con el punto de vista de los diferentes actores que participan en el proceso educativo. Esto garantizó que desde el proceso mismo de construcción del cuestionario se asegurara que el contenido evaluado correspondiera con las intenciones del mismo, en otras palabras, su validez del contenido.

Los resultados del análisis psicométrico realizado a partir de la aplicación del cuestionario muestran una consistencia interna adecuada, con un coeficiente Alfa de Chronbach de 0.96 para la prueba total y coeficientes Alfa parciales superiores a 0,8 en todas las dimensiones, lo cual pone de manifiesto que los constructos evaluados son consistentes con los ítems utilizados para evaluarlos. Por otra parte, el análisis de ítems, realizado mediante los métodos de distribución de frecuencias, correlaciones ítem-prueba total e ítem-dimensión y diferencias de medias entre grupos de puntajes en los dos extremos de la distribución muestra relaciones y diferencias estadísticamente significativas. De esta manera, no parece ser necesario remover o modificar ninguno de los ítems del cuestionario.

Por último, se emplearon dos pruebas multivariadas para realizar la validación de constructo del cuestionario, la cual se refiere al grado de confianza que podemos tener en la estructura del cuestionario. Este instrumento se diseñó para evaluar cuatro dimensiones (o rasgos latentes): Organización de la UEA, Práctica docente, autoevaluación y evaluación global. De esta manera, se esperaba que el comportamiento de los puntajes arrojados por los ítems fuera consistente con dicha estructura. La primera prueba consistió en un análisis factorial exploratorio, el cual arrojó tres factores con un 69.8% de varianza explicada, en el que 21 de los 25 ítems del cuestionario mostraron una distribución de las cargas factoriales acorde con la distribución de los ítems en tres de las cuatro dimensiones del cuestionario original. La cuarta dimensión no pudo ser analizada debido al número reducido de ítems. Las pruebas de esfericidad de Bartlett y el indicador de adecuación de muestreo de Kaiser-Meyer-Olkin mostraron también valores adecuados. La segunda prueba consistió en un análisis factorial confirmatorio, mediante análisis de ecuaciones estructurales realizado con el software EQS. Los resultados confirmaron la estructura de tres dimensiones exhibida en el análisis factorial exploratorio, con índices de bondad de ajuste CFI de 0.95 y RMSEA de 0.07, los cuales confirman que el modelo arrojado por los datos ajusta con el modelo teórico que sirvió de base para el diseño del instrumento. Dado el tamaño amplio de la muestra, los resultados de la prueba X2 cuadrada deben ser tomados con precaución pues pueden generar diferencias significativas erróneas, razón por la cual se recomienda eliminarlos del análisis.

Referencias

Arbesú, María Isabel (2004). Evaluación de la docencia universitaria: Una propuesta alternativa que considera la participación de los profesores. *Revista Mexicana de Investigación Educativa*, 9 (23). 863-890.

Byrne, Barbara. (2008). *Structural Equation Modeling with EQS: Basic Concepts, Applications and Programming*, Nueva York, Routledge.

Buendía, Angélica; García, Susana; Grediaga, Rocío; Landesman, Monique; Rodríguez-Gómez, Roberto; Rondero, Norma; Rueda, Mario; y Vera, Héctor. (2017). Queríamos evaluar y terminamos contando: alternativas para la evaluación del trabajo académico. *Sociológica*, 32(92). 309-326.

Hair, Joseph F.; Anderson, Rolph E.; Tatham, Ronald L.; Black, William C. (1999). *Análisis Multivariante*. Madrid, España: Pearson/Prentice Hall.

Renkl, Alexander; Mandl, Heinz y Gruber, Hans (1996). Inert knowledge: analyses and remedies. *Educational Psychologist*, 31(2), 115-121.

Rueda, Mario (2008). La evaluación del desempeño docente en la universidad. *Revista Electrónica de Investigación Educativa, Especial*. Consultado el 28 de noviembre de 2020, en: http://redie.uabc.mx/NumEsp1/contenido-rueda.html

Rueda, Mario (2018). Los retos de la evaluación docente en la universidad. Publicaciones. *Facultad de Educación y Humanidades del Campus de Melilla*, 48(1), 143–159. doi:10.30827/publicaciones.v48i1.7334

Rueda Beltrán, Mario; Sánchez Mendoza, Marlen Yasmin

Rueda, Mario. y Sánchez, Marlen Y. (2018). Trayectoria de la Red Iberoamericana de Investigadores sobre la Evaluación de la Docencia (RIIED). Revista Educación, Política y Sociedad, 3(2), 76-89.

Whitehead Alfred North (1929). *The Aims of Education and Other Essays*. New York, NY: Free Press.

UAM (2020). *Hacia la renovación de la evaluación de la docencia en la UAM*. En: https://cbi.izt.uam.mx/images/Avisos/Evaluacion-de-la-Docencia-vF-22.pdF