

# VALIDEZ Y CONFIABILIDAD DEL INSTRUMENTO EMPLEADO PARA EVALUAR EL DESEMPEÑO DOCENTE EN EL COLEGIO DE POSGRADOS

---

MARÍA ESTHER MÉNDEZ CADENA/ LEÓN R. GARDUÑO ESTRADA  
Colegio de Posgraduados Campus Puebla, Universidad de las Américas, Puebla

**RESUMEN:** Este documento presenta una serie de análisis estadísticos realizados al conjunto de resultados obtenidos sobre un instrumento de evaluación a profesores por parte de los estudiantes, realizados durante los años del 2007 al 2011 en los diferentes campus del Colegio de Posgraduados que existen en el país. El propósito de estos análisis es determinar la validez y la confiabilidad del instrumento empleado para tales evaluaciones a fin de llegar a una conclusión sobre 1. la continuación en su empleo, 2. la modificación del mismo o, 3. la necesidad de elaborar uno diferente. No obstante la diversidad de modelos y propuestas acerca de la calidad de la educación y su evaluación, todas ellas coinciden en incluir el aspecto pedagógico desempeñado por el profesor. Considerando la literatura nacional e internacional actual sobre el tema, una cantidad de autores están de acuerdo en que los estudiantes pueden evaluar la enseñanza que reciben de manera confiable y válida. Para llevar a cabo este trabajo, se empleó una base de

datos integrada por las respuestas dadas a un cuestionario de 35 reactivos por un total de 8297 estudiantes distribuidos en los cuatro campus de los Colegios de Posgraduados en México que cuentan con programas de maestría y/o doctorado en ciencias. Los procedimientos estadísticos empleados fueron el análisis factorial, correlaciones, y el coeficiente alfa de Cronbach. Los análisis realizados mostraron una ausencia completa de validez y confiabilidad. La propuesta planteada fue la de elaborar un nuevo instrumento.

**Palabras clave:** Evaluación de profesores, instrumentos de evaluación, validez, confiabilidad.

## Introducción

No obstante la diversidad de modelos y propuestas acerca de la calidad de la educación algunas de las cuales destacan dimensiones generales como las de relevancia, eficacia, pertinencia, eficiencia, o equidad (García, 2010; Tristán, 2008), todas ellas coinciden en incluir el aspecto pedagógico. En otras palabras, uno de los elementos centrales en la discusión sobre el tema es el de la calidad de la enseñanza por parte del profesor.

De acuerdo a García (2000) la evaluación de la docencia ha estado presente a fines de la década de los veinte. Considerando la literatura nacional e internacional actual sobre el tema, una cantidad de autores (García, 2000; Garduño, 2002; Garduño, Gutiérrez, Godínez, & Valadez, 2004; Garduño, & Flores -en revisión para publicación-; Montecinos, Rittershausen, Solís, Contreras, & Contreras, 2010; Rueda, 2008; etc.) están de acuerdo en que los estudiantes pueden evaluar la enseñanza que reciben de manera confiable y válida. Sin embargo, a pesar de que la evaluación de profesores en instituciones educativas de diversos niveles ha estado presente desde hace muchos años, éste todavía continúa siendo un tema muy vigente. Se siguen elaborando instrumentos, propuestas de evaluación, aplicando nuevos procedimientos de análisis estadístico, etc. que hacen de evaluación docente asuntos de discusión continua (Barrera, & Myers, 2011; Laudadio, 2012; Ross, 2001; Xu, Yueyu, 2012).

Atender a la valoración de las actividades de los profesores se justifica por una de las finalidades de la evaluación: la orientada al mejoramiento de la enseñanza a través de la retroalimentación proporcionada a los docentes así como sobre sus repercusiones en los programas de actualización. Adicionalmente, la evaluación de los docentes permite a las instituciones identificar las expectativas y satisfacción de los estudiantes con el propósito de generar políticas y estrategias de atracción y retención. La evaluación de la efectividad de la docencia es un aspecto que se considera fundamental en casi todas las IES; determinar la calidad con la que se llevan a cabo las diversas funciones docentes es esencial para realizar una variedad de recomendaciones y decisiones académicas y administrativas.

En la evaluación del desempeño docente predomina la opinión de los estudiantes sobre las cualidades pedagógicas de los profesores (Belcheir, 2001; Davidson-Shivers et al., 2004; Rueda, 2008) misma que se recupera a través de cuestionarios cerrados y/o de

respuesta abierta. Esto permite que quienes están directamente en contacto con los profesores participen y retroalimenten la práctica pedagógica, las prácticas de laboratorio y proporcionen información sobre aspectos de importancia para la administración como la asistencia a clases, entre otros.

Dentro de la muy extensa literatura sobre investigación y evaluación de temas educativos, uno que siempre está presente es el de la validez y confiabilidad de los instrumentos, procedimientos o técnicas empleadas. La determinación de la confiabilidad y validez sobre un instrumento de medición es sumamente importante para cualquier propósito en donde se pretenda contar con información que permita conocer el estado de una variable o de cualquier fenómeno complejo; y más aún, cuando la información obtenida vaya a utilizarse para tomar cualquier tipo de decisión. De manera general, la validez se define como el grado en que un instrumento mide lo que se pretende medir. Por su parte, la confiabilidad se refiere al grado en que un instrumento es consistente en medir aquello que se propone (Ross, 2001).

Este documento presenta una serie de análisis estadísticos realizados al conjunto de resultados obtenidos sobre un instrumento de evaluación a profesores por parte de los estudiantes, realizados durante los años del 2007 al 2011 en los diferentes campus del Colegios de Posgraduados que existen en el país. El propósito de estos análisis es determinar la validez y la confiabilidad del instrumento empleado para tales evaluaciones a fin de llegar a una conclusión sobre 1. la continuación en su empleo, 2. la modificación del mismo o, 3. la necesidad de elaborar uno diferente.

## Método

Para llevar a cabo este trabajo, se empleó una base de datos integrada por las respuestas dadas a 24 reactivos cerrados del cuestionario empleado por un total de 8297 estudiantes distribuidos en los cuatro campus de los Colegios de Posgraduados que cuentan con programas de maestría y/o doctorado en ciencias. Tal base de datos incluye la información de los años 2007 al 2011.

El instrumento se responde con el empleo de dos escalas tipo Likert con cinco opciones de respuesta que van desde Nunca a Siempre o de Malo a Excelente. Los reactivos están agrupados en alguna de tres diferentes categorías: 1. Relativo al profesor(a), 2. Relativo al curso, y 3. Relativo a las prácticas de campo o laboratorio. La primera categoría contiene 11 preguntas, la segunda seis, y la tercera cuatro preguntas. Al final de cada categoría se incluye una pregunta resumen en la que se solicita al estudiante expresar su percepción general sobre el contenido de cada una de las tres categorías mencionadas.

### Análisis estadísticos

Con el propósito de determinar la validez del instrumento, dada la información contenida en la base de datos, se realizaron una serie de análisis factoriales y correlaciones. Para un propósito de determinación de la validez de un instrumento, el análisis factorial permite determinar la congruencia entre los reactivos sobre una variable latente. Para el caso que nos ocupa, el análisis nos permitió determinar si los reactivos en el instrumento están distribuidos de tal manera que indiquen un patrón consistente entre ellos. Dado que el instrumento, de principio, está integrado por un conjunto de reactivos clasificados en tres categorías, el análisis factorial nos posibilita conocer si ciertamente los reactivos se agrupan como están dispuestos o no. No obstante, aunque no existiera esta disposición en categorías, de cualquier manera, el análisis factorial permite conocer la congruencia entre los reactivos dentro de los factores que se identifican con el análisis. Este es un procedimiento empleado ampliamente para este propósito (Aypan, 2009; Barth, 2008; [Guo-Hai](#), 2010; [Murray](#), 2012). Los análisis factoriales realizados se llevaron a cabo sobre las 24 preguntas relacionadas directamente a evaluar al profesor. Todos los análisis factoriales se realizaron con el método de componentes principales y rotación varimax. El valor eigen seleccionado para cada incluir a cada factor debió ser igual a uno o mayor. Tales análisis se realizaron por año, por periodo escolar, así como por nivel educativo: maestría y doctorado.

Adicionalmente, con el propósito de contribuir a la determinación de la validez del instrumento con el empleo de una variable criterio, se realizaron una serie de correlaciones entre la calificación dada a cada uno de los reactivos dentro de cada una de las tres categorías identificadas en el instrumento y la calificación sobre el reactivo “resumen” de la

categoría. Dado que se pretende que el reactivo “resumen” indique la percepción global del estudiante sobre los contenidos de los reactivos dentro de cada categoría, la determinación de la correlación entre cada uno y el mencionado reactivo “resumen” posibilitaría identificar cuáles reactivos son consistentes con tal percepción.

Para determinar la confiabilidad del instrumento se empleó el método de alfa de Cronbach.

## Resultados

En cuanto a los resultados obtenidos por los análisis factoriales se encontró que el número de factores identificados es diferente en los años analizados y, en consecuencia, las varianzas explicadas por cada uno difieren. Aunque la mayoría de los reactivos por factor son los mismos entre los años, no existe un patrón consistente dentro de cada uno. Por ejemplo, sobre los resultados para el año 2011, se observa que los factores contienen prácticamente los mismos reactivos que los identificados para los análisis de los años anteriores, pero, sin embargo, los reactivos no están agrupados de manera coherente dentro de ellos. De forma general podemos decir que en los factores se mezclan reactivos referidos a muy diferentes aspectos: infraestructura, uno o más reactivos resumen, relacionados con prácticas de laboratorio, asistencia a clase, etc. El mismo problema se observa a nivel de periodo escolar.

Sobre los análisis factoriales realizados con los datos de los diferentes periodos escolares, de la misma manera que en los resultados encontrados considerando los diferentes años en que se aplicó el instrumento, se observa variabilidad en el número de factores identificados, falta de congruencia entre los reactivos sobre una variable latente común a cada factor, y diferente número de reactivos dentro de cada factor identificado. De tal forma, se encuentra el mismo problema ya señalado para los análisis realizados por año.

Sobre los análisis factoriales realizados sobre el nivel de maestría y de doctorado se identificaron dos factores para el de maestría y tres para el de doctorado. En cuanto a las varianzas totales, éstas son semejantes. Por otro lado, de igual manera que para los análisis anteriores, no se pudo identificar ninguna consistencia conceptual entre los

reactivos en ninguno de los factores. En otras palabras, cada factor incluye reactivos relacionados a muy diferentes aspectos del curso. Así, en los factores del nivel maestría o doctorado se integran reactivos sobre las prácticas de laboratorio, tres reactivos “resumen” –que por lo tanto pertenecen a tres diferentes categorías-, la actualidad del contenido, los apoyos didácticos, reactivos relacionados con los laboratorios, apoyos didácticos, la utilidad práctica y conceptual de los temas.

Los resultados del coeficiente alfa de Cronbach a nivel de año, se encuentra que éstos son bastante altos y van de .955 a .971. En cuanto a los resultados de los diferentes coeficientes en la misma prueba por periodo, sus valores oscilan de .943 a .971. Estos resultados indican una alta consistencia interna entre los reactivos. Mientras que por grado académico para maestría es de .960 y para doctorado .962.

### Correlaciones entre cada reactivo y el reactivo resumen de cada categoría

Los resultados de las correlaciones entre cada reactivo y el reactivo resumen de cada categoría, de principio, parecen indicar una estrecha relación entre ellos, lo que abundaría sobre la validez del instrumento. Sin embargo, la información obtenida por los diferentes análisis factoriales ya descritos, parecen ser más contundentes en cuanto a la falta de validez del instrumento y, por lo tanto, de los reactivos que lo integran. Como ya se mencionó, los diferentes factores integran reactivos de diferentes categorías y de una variedad de contenidos. Las cargas factoriales obtenidas son bastante altas y los resultados del alfa de Cronbach también son muy altos. De tal manera, sería de esperarse que las correlaciones entre los reactivos y el reactivo “resumen” fueran también altas. No obstante que prácticamente todos los coeficientes de correlación son estadísticamente significativos en los diferentes análisis realizados por año, periodo y nivel, este resultado se puede explicar por el alto número de casos analizados: entre mayor sea el número de casos es más probable que los coeficientes de correlación resulten significativos.

No obstante, se encontraron algunos coeficientes de correlación con valores no muy altos; por ejemplo, algunos de los relacionados con el tema de las prácticas de laboratorio. En los análisis realizados a nivel de periodo escolar, se pueden identificar algunos periodos en que dos reactivos (20 y 21) presentan correlaciones bajas y, en algunos casos, no significativas como en primavera 2008 y 2009 así como en verano 2007, 2008 y, 2009. Sin embargo, cabe hacer notar, que el número de casos fue muy reducido como en el verano del 2009. Por otro lado, los resultados no fueron consistentes a través de los diferentes periodos analizados.

En cuanto a los análisis por nivel educativo, encontramos que todos los coeficientes son altos y significativos.

## Conclusiones

En virtud de los resultados de los análisis factoriales realizados, podemos concluir que el instrumento empleado por el Colegio de Posgraduados en sus diferentes campus que ofrecen programas en ciencias no reúnen los requerimientos de validez necesarios para tomar cualquier tipo de decisión sobre los y las docentes. Como ya se describió, los resultados de tales análisis muestran una completa falta de coherencia entre los reactivos tanto en las categorías en que se agrupan, pero también dentro de algún nuevo patrón(es) que pudiera permitir inferir la presencia de una variable latente que apoyara la validez del instrumento aunque con otro ordenamiento de reactivos.

En cuanto a los resultados del coeficiente alfa de Cronbach, como ya se mencionó, se obtuvieron valores muy altos. Sin embargo, los resultados no contribuyen a poyar la validez del instrumento. Como ya se indicó en un apartado anterior, la validez de un instrumento se refiere al grado en que el instrumento mide aquello para lo que está destinado a medir. En cuanto a la confiabilidad, ésta indica la consistencia del instrumento sobre aquello que se intenta medir. De tal manera, un instrumento requiere de la confiabilidad para ser válido; necesitamos que el instrumento mida de manera consistente lo que pretendemos medir. Sin embargo, un instrumento puede ser confiable sin ser válido.

En otras palabras, podemos medir de forma consistente una variable sin necesariamente ser la variable que queremos medir.

Sobre la serie de correlaciones realizadas entre cada reactivo y el reactivo “resumen” en cada categoría, la gran mayoría de los coeficientes son altos y estadísticamente significativos. Sin embargo, dados los altos valores de las cargas factoriales encontradas en los diferentes análisis factoriales y la completa ausencia de patrones coherentes de reactivos sobre las categorías planteadas o sobre alguna otra variable, tales resultados no contribuyeron a apuntalar la validez del instrumento.

De tal manera, la recomendación que se desprende de este trabajo es la de desarrollar un nuevo instrumento para evaluar el desempeño docente en el aula.

Como ya se apuntó al inicio de este documento, el tema de la calidad de la educación es cada vez más importante en este mundo cada vez más competitivo. La revisión del tema plantea grandes retos que hay que definir y resolver en este planeta con grandes problemas en muchos sentidos. A nivel de los programas de estudio y, de manera significativa, los relacionados con las ciencias, requieren de una redefinición de sus fines para, de ahí, partir a la elaboración de los programas de estudio y de la función del profesor en el aula. Tal afirmación plantea una estrecha relación entre educación, calidad y evaluación. De manera general (Garduño, 1999), podemos decir que la educación implica un proceso de perfeccionamiento del ser humano en todos sentidos. La calidad, por su parte, se refiere a los rasgos o características, o conjunto de ellas, que singularizan y hacen distintivo a un objeto o a un servicio. No obstante esta aparente definición estática, la calidad implica una búsqueda constante de mejoramiento sobre las características y propósito del objeto.

De tal manera, la tarea de desarrollar procedimientos o instrumentos de evaluación docente inicia desde una discusión sobre los fines de la educación, seguida de una definición de calidad sobre los diferentes elementos que participan en el proceso educativo. Sin dejar de lado el tema de la definición de los fines de la educación y del tema de la calidad, el desarrollo de un instrumento de evaluación docente, en tanto que constituye un



elemento clave en la concepción de la calidad, plantea, a su vez, también reto con diferentes facetas: una dimensión política, una dimensión teórica, una dimensión metodológica-procedimental, una dimensión de uso y, una dimensión de meta evaluación.

## Referencias

- Aypan, A. (2009). Teachers' Evaluation of Their Pre-Service Teacher Training. Educational Sciences: Theory & Practice, 9 (3), 1113. Recuperado el 30 de abril de 2013 de la base de datos EBSCO.
- Barrera, I. & Myers, R. (2011). Estándares y evaluación docente en México: el estado del debate. Programa de Promoción de la Reforma Educativa en América Latina y el Caribe (PREAL).
- Barth, M. (2008). Deciphering Student Evaluations of Teaching: A Factor Analysis Approach. Journal of Education for Business, 84(1), 40-46.
- Belcheir, M. J. (2001). Activities and Perceptions of 1998-99 Graduates. Research Report 2001-01. Boise State University, Boise, ID.
- Davidson-Shivers, G. V., Inpornjivit, K., & Sellers, K (2004). Using alumni and student databases for program evaluation and planning. College Student Journal, 38.
- García, J. (2000). ¿Qué factores extraclase o sesgos que afectan la evaluación docente en la educación superior? Revista Mexicana de Investigación Educativa, 5(10). Recuperado el 3 de mayo de 2013, de <http://www.redalyc.org/pdf/140/14001006.pdf>
- García, B. (2010). Modelos teóricos e indicadores de evaluación educativa. Revista Electrónica Sinéctica, 35. Recuperado el 28 de abril de 2013, de [http://www.sinectica.iteso.mx/?seccion=articulo&lang=es&id=498\\_modelos\\_teoricos\\_e\\_indicadores\\_de\\_evaluacion\\_educativa](http://www.sinectica.iteso.mx/?seccion=articulo&lang=es&id=498_modelos_teoricos_e_indicadores_de_evaluacion_educativa)
- Garduño, L. (1999). Hacia un modelo de evaluación de la calidad de instituciones de educación superior. Revista Iberoamericana de Educación, 21. Recuperado el 15 de abril de 2013, de <http://www.rieoei.org/rie21a06.htm>
- Garduño, L. (2002). Importancia relativa de tres enfoques psicológicos: El problema de la validez de la evaluación de profesores. Revista Perfiles Educativos. Vol. 24, 96.
- Garduño, L., Gutiérrez, C., Godínez, G., & Valadéz, A. (2004). Evaluación de los Cursos Nacionales de Actualización en matemáticas I: Factores de éxito y fracaso. Ponencia presentada en el Sexto Foro Nacional de Evaluación Educativa. CENEVAL. Aguascalientes, Ags. Trabajo publicado en las memorias del evento.
- Garduño, L., Flores, A. (En revisión para publicación). College student satisfaction: Specification and

- Estimation of a Structural Model. *Journal of Higher Education*.
- Guo-Hai Chen, (2010). Stability and correlates of student **evaluations** of teaching at a Chinese university. *Assessment & Evaluation in Higher Education*, 35, 6.
- Laudadío, J., (2012). Evaluación de estilos de enseñanza en la universidad: estudio preliminar de las propiedades psicométricas del cuestionario sobre la Orientación Docente del Profesor Universitario (CODPU). *Interdisciplinaria: Revista de Psicología y Ciencias Afines*, 29, 1.
- Montecinos, C., Rittershausen, S., Solís, M., Contreras, I., & Contreras, C. (2010). Standards-based performance assessment for the evaluation of student teachers: a consequential validity study, *Asia-Pacific Journal of Teacher Education*, 38, 4.
- Murray, J. (2012). Development and Psychometric **Evaluation** of the Independent School **Teacher**. *Journal of Experimental Education*, 80, 3.
- Ross, A. (2001). El concepto de validez y otros ensayos / Tr. Genaro R. Carrio, Osvaldo Paschero [et al.]. México: Fontamara.
- Rueda, M. (2008). La evaluación del desempeño docente en las universidades públicas de México. *Revista Iberoamericana de Evaluación Educativa*, 1, 3. Recuperado el 5 de mayo de 2013 de, [http://www.rinace.net/riee/numeros/vol1-num3\\_e/art1.html](http://www.rinace.net/riee/numeros/vol1-num3_e/art1.html)
- Tristán, A. (2008). Análisis multinivel de la calidad educativa en México ante los datos de PISA 2006. México: INEE.
- Xu, Yueyu. (2012). Developing a comprehensive teaching **evaluation** system for foundation courses with enhanced **validity** and reliability. *Educational Technology Research & Development*, 60, 5.

## Notas