RAES Revista Argentina de Educación Superior

Validez sustantiva de las evaluaciones docentes basadas en opiniones de estudiantes

Edith J. Cisneros-Cohernour
Universidad Autónoma de Yucatán
cchacon@uady.mx

Resumen

Este trabajo presenta los resultados de un estudio exploratorio acerca del significado de la buena docencia desde la perspectiva de los estudiantes de una universidad Mexicana y una Española. Se utilizó el marco teórico de Messick (1990) para examinar la validez sustantiva de las evaluaciones de la docencia basadas en opiniones de estudiantes. La recolección de datos incluyó análisis documental de los instrumentos de evaluación y grupos focales con estudiantes de licenciatura. El estudio es parte de una investigación para validar el Modelo MECD de García, Loredo, Luna y Rueda, 2008).

Palabras clave: evaluación, docencia, educación superior

Abstract

This article presents the findings of an exploratory study about the meaning of good teaching from the perspective of college students in a Mexican and a Spanish university. Messick's theoretical framework was used in order to examine the substantive validity of student ratings of teaching. Data collection involved document analysis of evaluation instruments and focus groups with undergraduate students. The study was part of validation study of the MECD Model of García, Loredo, Luna and Rueda (2008).

Keywords: evaluation, teaching, higher education.

Introducción

Aunque la evaluación formal de la docencia se ha llevado a cabo en México desde los años 80s, el interés por la misma se ha incrementado en los últimos años debido a presiones externas de rendición de cuentas y principalmente como mecanismo de control (Rueda, 2011; Aguilar, 2009).

Al incrementarse el interés por la evaluación docente, también se ha

incrementado la preocupación por su validez y por las consecuencias del uso que se da a los resultados para la toma de decisiones administrativas. Esto se debe a que la mayor parte de las evaluaciones del desempeño docente se llevan a cabo por medio de cuestionarios de opinión que se administran a los estudiantes al final de los cursos. Esta tendencia tan común en países desarrollados y en vías en desarrollo ha sido constante en varios países de Norteamérica y Europa y se ha incrementado también en América Latina.

Aunque algunos autores en la literatura de la evaluación docente apoyan la confiabilidad de los resultados de la evaluación basada en opiniones de estudiantes, tales como Marsh, 1987; Centra, 1993; Costin, Greenough, y Menges, 1971; McKiechie y Lin, 1979; Cashin, 1988, 1995, otros investigadores, como Ryan y Johnson, (1998), afirman que aunque la investigación puede ser defendible desde el punto de vista lógico y psicométrico, existe preocupación por su validez cuando la evaluación se utiliza con propósitos administrativos.

Esta preocupación también se debe también a que la revisión de la literatura realizada por Brodie (1988) encontró que estudios previos sobre la evaluación basada en opiniones de estudiantes reportan que los profesores que otorgan notas o calificaciones más altas y solicitan menos tarea de sus estudiantes, tienden a obtener mejores evaluaciones por su docencia. Por su parte Braskamp y Ory (1994), señalan que tanto la naturaleza del curso (optativo y obligatorio) como la etapa en la que se realiza la evaluación son factores que afectan los resultados, por lo que deben ser tomados en cuenta en el análisis e interpretación de los resultados.

Si bien se ha encontrado que la opinión de los estudiantes es sumamente importante, el hecho de incluirlos como la única fuente de información y excluir de la evaluación toda la complejidad de la docencia y su contexto, pone en entredicho la validez de los resultados para representar la calidad del desempeño docente. Como afirma Stake y Burke (2000), la calidad de la docencia solo puede ser juzgada propiamente en el contexto en que ésta tiene lugar y tomando en consideración todos los factores que en ella influyen. Si no se hace un esfuerzo por examinarla en su complejidad, la evaluación será inválida.

Por otra parte, Ory and Ryan (2000), afirman que aun cuando se han realizado estudios sobre las evaluaciones de la docencia realizadas en Norteamérica, existe necesidad de mayor investigación debido a éstas no han considerado las implicaciones

del nuevo marco teórico de validez, principalmente en cuanto a la validez del constructo en los aspectos conceptual, substancial y de consecuencias. Un estudio de esta naturaleza es particularmente importante en México y otros países como España, debido a la tendencia de adoptar instrumentos de evaluación y procesos desarrollados en otro contexto y cultura para evaluar la calidad de la docencia. El examinar aspectos de validez sustantiva en dos universidades en diferentes contextos, contribuye también a nuestra comprensión acerca de la validez de las evaluaciones de la docencia en contextos internacionales.

Planteamiento del problema

Debido a la importancia de realizar investigación en México sobre la evaluación de la docencia basada en opiniones de estudiantes, y en particular con base en el nuevo marco de validez, esta investigación tuvo como propósito:

Examinar la validez sustantiva de la evaluación de la docencia basada en opiniones de estudiantes, para lo cual se analizó el significado de la buena docencia desde la perspectiva de los alumnos y los procesos que éstos siguen cuando evalúan a sus profesores. El estudio se realizó en una universidad Mexicana y una Español. El trabajo es parte de un estudio de la validación del Modelo de Evaluación de Competencias Docentes (MECD) de García, Loredo, Luna y Rueda (2008).

Marco teórico

A principios de los años noventa surge un cambio en la literatura sobre evaluación que dio lugar a una nueva conceptuación de la validez. El responsable de este cambio fue Samuel Messick con su famoso capítulo sobre validez (1989), seguido del trabajo de Shepard (1993), Lane, Park y Stone, (1998); Reckase, (1998); Yen, (1998); Cronbach, (1989), y Moss (1992, 1996; 1998).

Este marco teórico se mueve de un concepto fragmentado de validez a un concepto unificado que integra: “consideraciones de conceptual, criterio y consecuencias dentro del marco de constructo para la prueba empírica de hipótesis racionales sobre el significado de los puntajes y de las relaciones teóricamente relevantes, incluyendo aquellas de naturaleza aplicada y científica (Messick, 1995, p. 751).

La validez deja de ser vista como una propiedad de una prueba, y pasa a ser “un juicio global de la medida en que la evidencia empírica y teórica apoya cuán adecuadas

y apropiadas son las interpretaciones basadas en la evaluación " (Messick, 1995, p. 741). Más aún, la validez se refiere no solo a los significados y la interpretación de los resultados de las pruebas, sino también a las inferencias y consecuencias sociales que resultan de la evaluación. Por lo tanto, significados y consecuencias son esenciales para la validez (Messick, 1989, 1995).

Messick (1989, 1995) presenta seis aspectos importantes de la validez de constructo que deben ser utilizados por todas las evaluaciones educativas y psicológicas para identificar fuentes de invalidez: constructo, substantiva, estructural, externa, generalizabilidad y de consecuencias. Las cuestiones críticas y fuentes de evidencia que se enfatizan en cada uno de estos aspectos son[1]:

- *Conceptual*: incluye evidencia de la relevancia del contenido, su representatividad y calidad técnica (Lennon, 1956; Messick, 1989).
- *Sustantivo*: se refiere a la fundamentación teórica de las consistencias observadas en las respuestas a las pruebas, incluyendo los modelos de procesos de desempeño de tareas (Ebreton, 1983), así como la evidencia empírica de los procesos teóricos en los que realmente se involucran los estudiantes cuando llevan a cabo la tarea de evaluación.
- *Estructural*. Aprecia la fidelidad de la estructura de calificación a la estructura del dominio del constructo en cuestión (Loevinger, 1957)
- *Externo*: incluye evidencia convergente y discriminante de comparaciones de múltiples características (Campbell & Fiske, 1959), así como evidencia de relevancia de criterio y utilidad aplicada (Crombach & Glesser, 1965).
- *Generalizabilidad*: examina la medida en la cual las propiedades del puntaje y sus interpretaciones pueden generalizarse para y a lo largo de grupos poblacionales, y tareas (Cook y Campbell, 1979; Shulman, 1970), incluyendo la generalización de la validez de las relaciones del criterio-prueba (Hunter, Schmidt, y Jackson, 1982).
- *Consecu*encias: Aprecia el valor de las implicaciones de la interpretación de los puntajes sobre la base de las acciones, así como de las consecuencias potenciales del uso de las pruebas, especialmente en relación con fuentes de invalidez relacionadas al sesgo, justicia, y justicia distributiva (Messick, 1980, 1989).

En su análisis de cómo las evaluaciones de la docencia en educación superior son consistentes con el nuevo marco de validez, Ory y Ryan (2001) encontraron que aunque

[1] Messick, 1994, p. 11-12.

existen estudios sobre algunos aspectos de validez, hay necesidad de mayor investigación en esta área. Aunque algunas investigaciones se han llevado a cabo sobre la validez de las evaluaciones basadas en opiniones de estudiantes sobre la generalización y validez externa de estas evaluaciones, los estudios sobre los aspectos de validez conceptual, sustantiva y de consecuencias son incipientes. Asimismo, existe necesidad de realizar estudios de este tipo para examinar con mayor profundidad si la evaluación representa justamente la calidad de la docencia en su contexto y sobre cómo las personas encargadas de tomar decisiones y maestros usan los resultados de la evaluación para el desarrollo profesional y para la toma de decisiones administrativas.

Metodología

El estudio de validación se centró en el aspecto de validez sustantiva. Ésta se enfoca en las preguntas ¿a qué se deben las diferencias en los puntajes? ¿Qué sabemos acerca de los procesos de respuesta en diferentes situaciones? Si los estudiantes evalúan de forma positiva, ¿están respondiendo sinceramente? en relación con la naturaleza de la evaluación, ¿equivale al constructo evaluado?

Es importante comprender por qué los cambios tienen lugar. También es importante entender cómo los estudiantes usan las escalas para responder, y si el significado de la escala que le han dado los estudiantes es el apropiado, dado el significado que se intenta medir.

Solo puede existir validez sustantiva si todos los estudiantes siguen procesos similares cuando responden los instrumentos de evaluación. Asimismo, es importante saber si algunos subgrupos de estudiantes responden de forma diferente que otros; si la evaluación es apropiada para diferentes grupos de estudiantes de antecedentes étnicos y culturales diversos, y cómo los estudiantes usan e interpretan las categorías de respuesta.

También se requiere saber si cuándo otorgan una calificación intermedia, esto indica su inhabilidad de responder, indica que otorgan un puntaje en medio de la escala, o se trata de falta de interés. Además, es necesario saber si algunos estudiantes son más renuentes que otros a utilizar los puntajes extremos de la escala y cuál es el significado de los puntajes extremos. Para hacer inferencias válidas de los puntajes otorgados por los estudiantes, se requiere determinar si encajan correctamente el significado de la escala que le dan los estudiantes y el significado que ésta intenta medir.

Evidencia de validez sustantiva se encuentra cuando lo que se evalúa y lo que se

mide encajan correctamente. Como Ory y Ryan (2001), afirman: cuando alguien utiliza pensamiento crítico para responder a ítems que evalúan su pensamiento crítico, hay evidencia de la validez sustantiva de los puntajes (p. 14). A menos que se tenga certeza sobre el significado de los puntajes, no puede decirse que los resultados de la evaluación son representaciones válidas de la calidad instruccional.

Para obtener evidencia de la validez sustantiva de las evaluaciones realizadas por los estudiantes, se llevó a cabo trabajo de campo en dos universidades públicas, una en México y la otra en España. La universidad mexicana está integrada por 15 facultades y dos escuelas preparatorias. De acuerdo con su sitio de internet, la institución imparte 49 carreras a nivel licenciatura, 17 a nivel diplomado, 28 a nivel especialización, 27 a nivel maestría y 4 a nivel doctorado en las áreas de: Ciencias Biológicas y Agropecuarias; Ciencias Exactas e Ingenierías; Ciencias de la Salud; Ciencias Sociales, Económico-Administrativas y Humanidades; y Arquitectura, Arte y Diseño. Asimismo, esta institución sirve las necesidades educativas de más de 11,000 estudiantes y cuenta con una planta docente de más de 800 profesores de tiempo completo.

La universidad española cuenta con una larga historia y, por consecuencia, una larga tradición. Sin embargo, el centro docente en el cual se recogieron los datos, es relativamente reciente, ya que como tal fue creado en 1978. Actualmente, acceden a este centro estudiantes que tienen un promedio de calificaciones de secundaria respecto a carreras de reconocido prestigio social, como medicina, arquitectura, ingeniería, etc. Además, gran parte de los estudiantes provienen de una provincia del sur de España, y en menor número de otras provincias o regiones.

Para realizar la recogida de datos, se llevaron a cabo grupos focales con estudiantes de licenciatura en dos universidades. Como afirman Massol, Dorio y Sabariego (2004), los grupos focales son una "técnica cualitativa que recurre a la entrevista realizada a todo un grupo de personas para recopilar información relevante sobre el tema de investigación" (, p. 343).

Durante los grupos focales se utilizó una guía de entrevista, basada en el estudio: The Evaluation of Teaching in the Context of a Research University: Meanings, Tradeoffs, and Equity Concerns realizado por la autora en 2001. La guía de entrevista se centró en dos aspectos principales: el significado de la buena docencia, la importancia de algún elemento de la docencia por sobre todo lo demás y los procesos que siguen los estudiantes cuando evalúan a sus profesores.

Los participantes del grupo focal fueron estudiantes de licenciatura de una Facultad de Ciencias sociales en cada una de las dos universidades. En el caso de la universidad mexicana participaron 45 estudiantes, en el caso de la española fueron 15 estudiantes.

Para realizar el análisis de los datos se utilizó un diagrama de afinidad de Jiro Kawakita (Hirata, 2005).

Resultados

Los resultados del estudio indican que la evaluación de la docencia se lleva a cabo por medio de cuestionarios de opinión que los estudiantes responden al final de los cursos. En el caso de la universidad española existe un solo instrumento a nivel institucional que fue elaborado con base en otros instrumentos encontrados en la literatura sobre evaluación docente. En el caso de la universidad mexicana, se está desarrollando un sistema de evaluación de la docencia a nivel institucional porque hasta el momento cada facultad desarrolla su propio instrumento, basado principalmente en opiniones de los profesores de la institución sobre cuestionarios desarrollados en otros contextos.

Es importante notar que en las dos instituciones se ha incrementado la importancia al puntaje numérico que obtienen los profesores en la evaluación, principalmente en el caso de la universidad mexicana, en la que la evaluación de la práctica docente se ha convertido en uno de los indicadores para determinar la permanencia del personal y como requisito para participar en el programa de estímulos al desempeño del personal académico en la institución.

A continuación se describen con mayor detalle los resultados de la validez sustantiva de la evaluación con base en las opiniones de los estudiantes:

Significado de la buena docencia

Independientemente del contexto, los estudiantes indicaron que deciden quien es un buen docente cuando lo comparan con un ideal, solo algunos estudiantes en el contexto español afirmaron que lo hacen cuando comparan a sus profesores con sus pares. Asimismo, un estudiante mexicano agregó que en su percepción de la buena docencia influyen los comentarios de sus pares: “si tenemos buenas referencias de un maestro, nos inscribimos a su curso y eso también influye en la evaluación”.

En general, los estudiantes españoles definieron la docencia de calidad como

aquella que da mayor importancia a la actuación profesional y ética del docente y al respeto a las diferencias entre los estudiantes.

Sin embargo, no todos estuvieron de acuerdo en relación con los elementos que componen estos aspectos. En relación con una actuación profesional y ética, algunos estudiantes españoles indicaron que esperan que su profesor que posea un auténtico interés en su trabajo, otros que sepa motivar a los estudiantes, explique correctamente la materia que imparte, en un lenguaje comprensible para los alumnos, enseñe con claridad y coherencia con el programa de estudios, que sea consistente con los criterios de evaluación previamente acordados, y un estudiante indicó que espera que sea un ejemplo dentro y fuera del aula y se mantenga actualizado. Todos esperan que el docente tenga buen sentido de humor y estabilidad emocional.

En relación con el respeto a las diferencias entre los estudiantes, éstos expresaron que esperan que un buen docente sea respetuoso de sus estudiantes, conozca a sus alumnos para comprender sus necesidades específicas y sepa escucharlos.

Además de lo anterior, algunos estudiantes indicaron que un buen docente debe mantener una buena relación y comunicación con sus estudiantes, debe saber responder a sus preguntas y vincular la teoría con la práctica".

En cuanto a los estudiantes mexicanos, éstos afirmaron que la buena docencia es "aquella que te permite aprender, que llama a estudiar", aunque difirieron en sus opiniones acerca de los elementos que la componen.

Los estudiantes de octavo semestre afirmaron que la buena docencia enfatiza una actuación profesional y ética, lo que involucra, para algunos el usar estrategias variadas para que el alumno aprenda; para otros implica ser proactivo, con vocación, entusiasta y justo. Asimismo, esperan que un buen docente sea respetuoso de las diferencias entre los estudiantes por lo que debe ser empático, comprender a su alumnado y ayudarlo.

Además de lo anterior, otro grupo de estudiantes de octavo semestre indicaron que la buena docencia debe ser activa y vincular la teoría con la práctica, así como promover altas expectativas, esto es debe retar a los estudiantes a dar más de sí mismos. Estos estudiantes agregaron que prefieren a un docente en cuya formación predomine la formación pedagógica que el dominio de contenido y que les gustaría tener un profesor que "sea creativo y los deslumbre".

Por su parte los estudiantes de primer semestre, fueron los que más difirieron entre ellos acerca de lo que constituye la buena docencia. Un grupo enfatizó la actuación

profesional y ética del docente, mientras que otros le dieron más importancia al respeto a la diversidad entre estudiantes, las altas expectativas, la enseñanza activa y la relación profesor alumno.

En cuanto a la actuación profesional y ética, los estudiantes tuvieron variadas opiniones acerca de lo que esta implica: que el profesor sabe mucho, le apasiona lo que estudia y lo comparte, domine el contenido y use diferentes metodologías. En cuanto al respeto a la diversidad entre los estudiantes, los alumnos afirmaron que esperan que su profesor sea empático, que evite hacer que los estudiantes se sientan de menos o mal, los comprenda, y que les tenga paciencia.

En cuanto a las altas expectativas, unos estudiantes indicaron que esperan que un buen docente sea firme al ejercer la disciplina, tenga altas expectativas, enseñe la responsabilidad y el valor del trabajo. Otros definieron la docencia de calidad como aquella en la que el docente promueve un aprendizaje activo en el que vincula la teoría con la práctica. Unos más indicaron que la calidad docente involucra que el profesor dé más de su tiempo y tenga una disposición para compartir con sus estudiantes aspectos de su vida personal.

Elemento más importante de la calidad docente

Se les preguntó a los estudiantes si existe algún elemento de buena docencia que es más importante, que en caso de faltar podría influir negativamente en la evaluación que realizan de sus profesores.

Se encontró que para los estudiantes españoles, el elemento fundamental de la buena docencia es que el profesor mantenga una buena relación con los estudiantes, muestre continuidad en su humor y demuestre estabilidad emocional. Por su parte, los estudiantes mexicanos indicaron que para ellos el elemento más importante es que el docente sea justo al calificar.

Cuando se les preguntó a los estudiantes lo que significa "ser justo al calificar", difirieron en sus opiniones: unos esperan que el docente tome en cuenta el contexto y las diferentes formas de aprender al evaluar y que sea flexible dependiendo de las circunstancias personales de los alumnos. Otros que establezca criterios de evaluación y no los cambie, no tenga favoritos, distinga entre el trabajo y la persona y no haga excepciones. También hubo un estudiante que indicó que ser justo significa que ella obtenga una buena calificación: "Si me califica mal, yo lo califico mal".

Además de lo anterior, otro estudiante indicó que para él es fundamental que el

profesor use diferentes metodologías: "si el método no es adecuado se pierde todo"

Procesos que siguen los estudiantes para evaluar a los profesores

En relación a cómo deciden los estudiantes otorgar el puntaje más alto, intermedio o más bajo a un profesor en la evaluación de su docencia, también se encontraron diferencias entre los estudiantes.

En el caso de los *puntajes más altos*, se encontró que en el caso de los estudiantes españoles, algunos dan más peso a factores relacionados con el proceso mientras que otros a los resultados de aprendizaje. Entre los que dan más peso al proceso, indicaron que toman en cuenta si el maestro lo hace todo bien, emplea métodos, recursos y materiales apropiados, hace sentir cómodos a sus estudiantes en el aula e imparte sus clases de forma amena y utiliza ejemplos de su propia experiencia. Entre los estudiantes que dan más peso a los resultados, estos indicaron que para ellos es importante aprender, si están satisfechos otorgan el puntaje más alto, sino el más bajo.

Por su parte, los estudiantes mexicanos, también difirieron entre ellos sobre lo que toman en cuenta para otorgar el puntaje más alto al evaluar a sus profesores. Un grupo de estudiantes indicó que toma en cuenta el proceso de enseñanza, en especial si el profesor los motivó para aprender el curso. Otro grupo indicó que se basa en los resultados de aprendizaje, si aprenden, otorgan el puntaje más alto. Otros indicaron que se basan en si su profesor es una persona que posee habilidades docentes y dominio de contenido, mientras que otros afirmaron que la parte personal es la que influye más en su decisión de otorgar el puntaje más alto, relacionada con la relación profesor alumno.

En el caso de los *puntajes intermedios,* los estudiantes españoles indicaron que ellos otorgan este puntaje cuando hay cosas que han funcionado bien y otras no y cuando perciben que el profesor ni los favorece ni nos perjudica, mantiene una actitud correcta o da la impresión de estar dispuesto a escuchar a los estudiantes, pero en realidad no lo está.

Por su parte los estudiantes mexicanos difirieron entre ellos sobre cuando otorgan un puntaje intermedio. Unos estudiantes indicaron que otorgan este puntaje "a quien da bien sus clases, pero no tiene control de grupo", en tanto que otros afirmaron que lo otorgan "a quien se esfuerza pero no lo logra" La mayoría de los estudiantes indicó que la puntuación la otorgan con base en el efecto que el maestro causa en el alumno, "cada quien responde de diferente manera"

En el caso del *puntaje más bajo*, un grupo de los estudiantes españoles indicaron que

otorgan este puntaje si el maestro demuestra escaso respeto hacia los alumnos, mientras que otros indicaron que ellos otorgan este puntaje si en general se encuentran insatisfechos con los resultados obtenidos en el curso y no han podido "asimilar los conocimientos propuestos en la asignatura"

Los estudiantes mexicanos, en su mayoría indicaron que otorgan el puntaje más bajo a profesores que no tienen control de grupo o se desvían del contenido del curso para abordar otras temáticas. Por su parte, un estudiante agregó que otorga este puntaje: "a quién no sabemos qué hace aquí"

Finalmente, se preguntó a los estudiantes si toman en serio la evaluación docente. Tanto los estudiantes españoles como mexicanos indicaron que no toman en serio la evaluación de sus profesores. En el caso de los estudiantes españoles, dijeron que esto se debe a que el instrumento de evaluación "no es claro, no permite respuestas matizadas, sino que más bien lleva a posicionarse en respuestas alternativas, del tipo sí/no" y porque la administración no les da el tiempo necesario para responderlo, porque "habitualmente se rellena en pocos minutos".

En el caso de los estudiantes mexicanos, estos indicaron que ellos no toman en serio el proceso de evaluación porque los obligan a realizarla bajo presión: "la evaluación se lleva a cabo después del curso y es requisito para inscribirse al siguiente semestre." Los estudiantes también opinaron que el proceso no es confidencial "porque es en línea", ya que antes de evaluar deben identificarse con su número de matrícula. Otros estudiantes indicaron que no toman en serio la evaluación porque no saben qué pasa con los resultados.

Conclusión

La validez es un elemento esencial de una evaluación, ya que nos permite saber si realmente evaluamos lo que deseábamos evaluar. El interés por la validez de las evaluaciones de la docencia se ha incrementado a medida que los resultados han comenzado a utilizarse para la toma de decisiones administrativas, tales como renovaciones de contratos e incluso para decidir la permanencia del personal académico.

Esto es altamente preocupante, debido a las limitantes de la investigación en esta área y en particular a la tendencia de basar la evaluación únicamente en cuestionarios de opinión de estudiantes. Como afirma Adams (1997), la literatura claramente revela que las conclusiones basadas en las evaluaciones que los estudiantes hacen de sus profesores

deben ser tomadas con cautela, y que si se utilizan deberían ser solo una entre muchas fuentes de información sobre el desempeño del profesor (Cashin, 1988; Seldin, 1993, Blundt, 1991, Haskell, 1997). Como afirman McKeachie & Kaplan (1996: 7), fallamos en ética cuando permitimos que decisiones importantes sobre el personal se tomen con base en datos potencialmente engañosos". Eso sin contar con las cuestiones legales que pueden surgir cuando se les utilizan los resultados de la evaluación para decisiones administrativas.

Con la intención de obtener una mayor comprensión de la validez de las evaluaciones basadas en opiniones de estudiantes, esta investigación exploratoria se centró en la validez sustantiva de las evaluaciones de la docencia basadas en opiniones de los estudiantes. La investigación se realizó en el contexto de dos facultades de ciencias sociales en dos universidades públicas de México y España, lo que contribuye a mejorar nuestra comprensión acerca de la validez de las evaluaciones de la docencia en contextos internacionales.

Los resultados indican que las dos universidades donde se realizó el estudio evalúan la docencia por medio de cuestionarios de opinión, que los estudiantes responden al final de sus cursos. En el caso de la universidad mexicana, después que el profesor ha entregado las calificaciones del curso, lo cual no solo afecta la validez de los resultados sino que es contrario a los Personnel Evaluation Principles (Stufflebeam, 2008).

Se encontró consistencia con el trabajo de Aguilar (2009), en cuanto a que las organizaciones educativas mantienen en el discurso que la evaluación de la docencia es "un proceso que permite obtener información clara, precisa y confiable sobre las acciones que se emprenden en una institución educativa, lo cual lleva a un círculo de retroalimentación y toma de decisiones para alcanzar niveles de eficiencia y eficacia que permitan elevar la calidad de la enseñanza y fortalecer la formación docente" (p.1).

Sin embargo, el análisis de las políticas e instrumentos de evaluación claramente muestra que existe una separación entre la teoría y la práctica de la evaluación docente en las instituciones de educación superior participantes porque la forma en que los profesores son evaluados sigue llevándose a cabo de forma simplista y bajo una visión reduccionista de la docencia que reduce la complejidad de ésta y de su contexto a un puntaje numérico. El énfasis que la institución pone en este puntaje numérico puede llevar a consecuencias no intencionadas, tales como que las personas se centren más en incrementar sus puntajes que en mejorar su práctica docente.

En relación con el significado de la calidad de la docencia, se encontró que los estudiantes definen la buena docencia en relación con un perfil ideal, integrado por diferentes características personales y conductas del profesor, así como de aspectos instruccionales. Se encontró que los estudiantes difieren entre sí acerca de las características que debe poseer el profesor, pero indican que ciertas conductas del mismo pueden influir negativamente en la evaluación, tal como mantener una buena relación con los estudiantes, poseer buen humor y mantener estabilidad emocional (España), y ser justo al calificar (México). En general, se encontraron diferencias entre las dos universidades y dentro de los grupos de estudiantes.

El análisis de las características identificadas por los estudiantes hace evidente que su concepción de calidad docente centrada en el maestro y no en el estudiante, por lo que es necesario examinar con mayor detalle si el sistema de evaluación puede o no estar otorgando evaluaciones negativas a profesores que enfatizan una docencia centrada en el alumno.

En relación con los procesos que siguen los estudiantes al evaluar a sus profesores, se encontró que los alumnos varían en la forma en que asignan los puntajes a sus maestros. Algunos otorgan bajas evaluaciones a quienes carecen de habilidades pedagógicas, no están motivados o son insensibles a las necesidades de los estudiantes, pero hay estudiantes que dan más peso a las metodologías de enseñanza y sobretodo con base en la evaluación que hace el profesor de su aprendizaje. Los resultados encontrados con los estudiantes mexicanos fueron similares a estudios previos con estudiantes norteamericanos realizados por la autora.

Es importante notar que en las dos universidades, los estudiantes reconocieron que están conscientes de estos sesgos y de que estos pueden influir en la evaluación. Asimismo, se encontró que los sesgos pueden ser tanto por parte de quienes evalúan como de quienes han diseñado los instrumentos, interpretan y usan los resultados de la docencia.

Con base en los resultados podemos concluir que existe evidencia de que los resultados basados en las evaluaciones de los estudiantes no son indicadores válidos de la calidad de la docencia porque el puntaje numérico que se obtiene de la evaluación, por sí mismo puede tener diferentes significados, ya que los estudiantes dan peso a diferentes aspectos al momento de evaluar, no necesariamente al aspecto evaluado, y en general no toman con seriedad el proceso de evaluación.

Futuros estudios deben explorar de manera más profunda los significados de la evaluación para quienes hacen el juicio acerca de su calidad, de la validez conceptual y sobre las consecuencias de las interpretaciones de los resultados de la evaluación para diferentes actores, así como las implicaciones y consecuencias para los evaluados y para el mejoramiento de la calidad de la formación de los estudiantes.

Asimismo y debido a las limitaciones que surgen al emplear una sola fuente de información en la evaluación de la docencia, se recomienda utilizar un enfoque que involucre múltiple fuentes y metodologías para obtener una mejor comprensión de la calidad de la docencia, así como del contexto en que la enseñanza tiene lugar. Algunos de estos datos pueden incluir, pero no exclusivamente:

- Portafolios de enseñanza y artefactos para ilustrar cómo el maestro desarrolla y enseña (objetivos, rúbricas que relacionan objetivos con actividades en el aula y tareas, y herramientas para evaluar el aprendizaje de los estudiantes).
- Observaciones de aula
- Encuestas y entrevistas con el personal académico
- Encuestas y entrevistas con estudiantes, entre otros.
- Asimismo, debe tomarse en cuenta el contexto institucional, la misión de los programas, la carga docente y los recursos institucionales entre otros (AERA, 2013).

Este enfoque incluirá a los estudiantes en el proceso de evaluación docente, pero tomando en consideración sus limitaciones y conjuntamente con otras fuentes de información.Los resultados de la evaluación deberán servir como base para un estudio más profundo de la práctica docente de los profesores evaluados, especialmente los profesores que no obtienen un buen desempeño en la evaluación.

Finalmente, es importante que tanto decisores políticos, administradores y evaluadores reconozcan las limitantes y problemas de validez de esta fuente de información para la toma de decisiones administrativas y den preferencia al uso de los resultados de la evaluación para la mejora de la práctica docente.

Referencias

ADAMS, Jhon, W. (1997). Student evaluations: The ratings game. Inquiry, 1, 2, 10-16.

AGUILAR, Virginia. (2009). Sistema de evaluación docente: Hacia un modelo formativo e integral en educación superior. Ponencia presentada en el X

Congreso Nacional de Investigación Educativa del Consejo Mexicano de Investigación Educativa (COMIE).

BLUNT, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review*, 18 (Summer), 48-53.

BRODIE, Dalbert, A. (1998). Do students report that easy professors are excellent teachers? *The Canadian Journal of Higher Education*, 23, 1, 1-20.

CAMPBELL, Donald. T. and FISKE, Donald. W. Convergent and discriminant validation by the multitrait matrix. *Psychological Bulletin*, 1959, 56, 81-105.

CASHIN, William. E. (1988). Student ratings of teaching: A summary of the research. IDEA paper No. 20. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

CASHIN, William. E. (1995). Student ratings of teaching: The research revisited. IDEA Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

CENTRA, John. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass Inc.

COOK, Thomas. D., and CAMPBELL, Donald. T. *Quasi-experimentation: Design and Analysis for Field Settings.* Chicago: Rand McNally, 1979.

COSTIN, Frank, GREENOUGH, William y MENGES, Robert, J. (1971). Student ratings of college teaching: reliability, validity, and usefulness. Review of Educational Research, 41, 5, 511-535.

CRONBACH, Lee J. "Construct Validation after Thirty Years." In R.L. Linn (Ed.), *Intelligence: Measurement, Theory, and Public Policy* (pp.147-171). Chicago: University of Illinois Press, 1989.

CROMBACH, Lee. J., and GLESSER, Goldine. C. (1965). *Psychological and Personnel Decisions* (2nd. Ed.). Urbana, IL., University of Illinois Press.

EMBRESTON, Susan E. Construct validity: Construct representation versus nomothetic span. *Psychological bulletin*, 1983, 93, 179-197.

GARCÍA, Benilde, LOREDO, Javier LUNA, Edna y RUEDA, Mario. (2008). Modelo de Evaluación de Competencias Docentes para la Educación Media y Superior. *Revista Iberoamericana de Evaluación Educativa, 1*(3e), pp. 124-136.

Haskell, R.E. (1997). Academic freedom, tenure, and student evaluations of faculty: Galloping polls in the 21st century.Education Policy Analysis Archives, 5(16), 1-36. Available: http://olam.ed.asu.edu/epaa/v5n6.html

HIRATA, Ricardo. (2003). *Curso: 7 nuevas herramientas para el control de la calidad.* Material de trabajo. Kiesen Consultores S.A. de C.V. Mérida, Yucatán.

HUNTER, John. E., SCHMIDT, Frank. L., (2004). *Methods of meta-analysis: Correcting error and bias in research findings*. San Francisco: Sage.

LANE, Suzanne, PARKE, Carol. S., and STONE, C.A. (1998). "A Framework for Evaluating the Consequences of Assessment Programs." *Educational Measurement: Issues and Practice*, 17(2), 24-28.

LENNON, Roger. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16, 294-304.

LOEVINGER, Jane. "Objective Tests as Instruments of Psychological Theory (Monograph), *Psychological Reports*, 1957, 3, 635-694.

MARSH, H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research*, 11, 253-388.

MCKEACHIE, W. J., & LIN, Y. G. (1979). A Note on Validity of Student Ratings of Teaching. *Educational Research Quarterly, 4*, 3, 45-47.

MCKEACHIE, W.J., KAPLAN, M. (1996). Persistent problems in evaluating college teaching. *AAHE Bulletin*, (February), 5-9.

MASSOL, Inés, Dorio, Inma, Sabariego, Marta. (2004). Metodología de la investigación Educativa. De la Colección: *Manuales de Metodología de Investigación Educativa*, bajo la dirección de Juan Etxeberría y Javier Tejedor. Con la coordinación de Rafael Bisquerra. Madrid: Editorial La Muralla

MESSICK, Samuel. (1980). Test Validity and the Ethics of Assessment. *American Psychologist*, 35, 1012-1027.

MESSICK, Samuel. Validity. In R. L. Linn (Ed.), *Educational Measure*ment (3rd Ed.). New York: American Council on Education and Macmillan, 1989.

MESSICK, Samuel. *Alternative Modes of Assessment: Uniform Standards of Validity*. Paper presented at a conference on Evaluating Alternatives to Traditional Testing for Selection, 1994, Bowling Green: OH., Octubre.

MESSICK, Samuel. Validity of Psychological Assessment: Validation of Inferences from Persons Responses and Performances as Scientific Inquiry into Score Meaning. *American Psych*ologist, 1995, 50(9), 741-749.

MOSS, Pamela A. "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment," *Review of Educational Research*, 1992, 62, 229-258.

ORY, John y RYAN, Katherine. How do Student Ratings Measure Up to a New Validity Framework? *New Directions for Institutional Research,* 2001.

RECKASE, Mark.D. (1998). Consequential Validity from the Test Developer's Perspective." *Educational Measurement: Issues and Practice,* 17(2), 13-16.

RYAN, Katherine y JOHNSON, Trav. (2000). Democratizing evaluation: Meanings and methods from practice. New Directions for Evaluation, Chicago, IL, November.

SELDIN, Peter. (1993, July 21). The use and abuse of student ratings of instruction. *The Chronicle of Higher Education,* A-40.

SHEPARD, Lorrie, A. (1993). Evaluating test validity. In L. Darling-Hammond (Ed.). Review of Research in Education, 19, 405-450

SHULMAN, Lee. S. Reconstruction of educational research. *Review of Educational Research,* 1970, 40, 371-390.

STAKE, Robert E. y BURKE, Mayra (2000). *Evaluating teaching*. Research report. Center for Instructional Research and Curriculum Evaluation (CIRCE). USA.

STAKE, R. E. (2014). Evaluating teaching. Ponencia presentada en el II Coloquio de Evaluación Educativa. Secretaría de Educación, Mérida: Yucatán.

STUFFLEBEAM, D. (2008). The Personnel Evaluation Standards: How to assess systems for evaluating educators. Corwin Press, USA.

YEN, Wendy.M. "Investigating the Consequential Aspects of Validity: Who is Responsible and What Should They Do?" *Educational Measurement: Issues and Practice,* 1998, 17(2), 5-6.