Revista Electrónica "Actualidades Investigativas en Educación"
E-ISSN: 1409-4703
revista@inie.ucr.ac.cr
Universidad de Costa Rica
Costa Rica

Acevedo Alvarez, Raziel; Olivares Miranda, Maritza
FIABILIDAD Y VALIDEZ EN LA EVALUACIÓN DOCENTE UNIVERSITARIA
Revista Electrónica "Actualidades Investigativas en Educación", vol. 10, núm. 1, enero-abril, 2010, pp. 1-38
Universidad de Costa Rica
San Pedro de Montes de Oca, Costa Rica

**Actualidades Investigativas en Educación**
Revista Electrónica publicada por el
Instituto de Investigación en Educación
Universidad de Costa Rica
ISSN 1409-4703
http://revista.inie.ucr.ac.cr
COSTA RICA

# FIABILIDAD Y VALIDEZ EN LA EVALUACIÓN DOCENTE UNIVERSITARIA

RELIABILITY AND VALIDITY ASSESSMENT IN THE UNIVERSITY TEACHING

Volumen 10, Número 1
pp. 1-38

Este número se publicó el 30 de abril de 2010

Raziel Acevedo Alvarez
Maritza Olivares Miranda

*La revista está indexada en los directorios:*

**LATINDEX, REDALYC, IRESIE, CLASE, DIALNET, DOAJ, E-REVIST@S,**

*La revista está incluida en los sitios:*

**REDIE, RINACE, OEI, MAESTROTECA, PREAL, HUASCARAN, CLASCO**

# FIABILIDAD Y VALIDEZ EN LA EVALUACIÓN DOCENTE UNIVERSITARIA

RELIABILITY AND VALIDITY ASSESSMENT IN THE UNIVERSITY TEACHING

*Raziel Acevedo Alvarez*[1]
*Maritza Olivares Miranda* [2]

***Resumen**: El presente artículo expone una revisión bibliográfica sobre la fiabilidad y la validez de los cuestionarios de opinión estudiantil, utilizados para evaluar la competencia docente universitaria, a fin de reunir información sobre el intenso debate, su complejidad, su análisis y sobre su permanente actualidad; porque la docencia se transforma continuamente, los conocimientos y las habilidades de hoy se desactualizan mañana. Por ello, es necesario evaluar la actividad permanentemente, pero con instrumentos que cumplan cabalmente estos dos componentes psicométricos. Sin embargo, es importante recalcar que muchas de estas evaluaciones no han sido estudiadas en profundidad y en muchos casos carecen de estudios estadísticos; de ahí nacen la gran cantidad de inconvenientes y problemas que enfrentan estas evaluaciones. Además, los cuestionarios de opinión han sido contrastados con variables individuales, olvidando que la docencia es un fenómeno multidimensional integrado por un conjunto de elementos contextuales y, de esa forma, ha de estudiarse, pues algunos de ellos pueden estar asociados a factores ajenos a la docencia universitaria.*

***Palabras claves**: EVALUACIÓN UNIVERSITARIA, EFICACIA DOCENTE, INSTRUMENTOS DE EVALUACIÓN, CUESTIONARIOS, FIABILIDAD, VALIDEZ*

***Abstract**: The present article, exposes an extensive bibliographical review on the reliability and validity of the questionnaires of student opinion, to evaluate the educational university competition, to bring together information of the intense debate, of It complexity and of the permanent current importance; because the teaching transforms constant, the knowledge and today skills change tomorrow, for this reason, it is necessary to evaluate the activity, with instruments that fulfill these two component psychometrics. Nevertheless, it is important to stress that many of these evaluations have not been studied in depth and lack these analyses, of there is born the great quantity of problems that face these evaluations. In addition, the questionnaires of opinion have been resisted by individual variables, forgetting that the teaching is a multidimensional phenomenon integrated to a set of elements, not isolated and of this form they have to be studied, since some of them can be associated with other factors foreign to the university teaching.*

***Key words**: UNIVERSITY EVALUATION, TEACHER COMPETENCE, VALIDITY, RELIABILITY.*

[1] *Doctor en Métodos de Investigación, Diagnóstico y Evaluación Educativa de la Universidad Complutense de Madrid. Profesor Catedrático de la Universidad de Costa Rica. Director de la Sede de Guanacaste, U.C.R. Ha participado en conferencias y congresos en: Italia, España, Austria, Polonia, Colombia, Argentina, Chile, Perú, Nicaragua y Venezuela. Dirección electrónica: heraceve@yahoo.com*

[2] *Magister en Educación con énfasis en Administración Educativa, Licenciada en Educación con énfasis en Curriculum, Bachiller en Educación con énfasis en Preescolar, todos los títulos por la Universidad de Costa Rica; Bachiller en Educación con énfasis en I y II ciclo de la Universidad Estatal a Distancia. Actualmente es profesora en las carreras de Educción Primaria y Preescolar de la Sede de Guanacaste de la Universidad de Costa Rica, también es la Coordinadora de Acción Social, de la Sede de Guanacaste. Dirección electrónica: mariolivares1@gmail.com*

***Artículo recibido**: 25 de junio, 2009*
***Aprobado**: 26 de abril, 2010*

## INTRODUCCION

La universidad, como institución de enseñanza superior, asume un carácter complejo al tratar de buscar el equilibrio entre la investigación, la acción social y la docencia. Esta última es, quizá, la más compleja y la más debatida, porque es un elemento fundamental en el quehacer académico, al constituirse en el área de impacto social responsable de entregar, a las comunidades urbanas y rurales, profesionales en los diversos ámbitos del quehacer humano.

La docencia es vital en la academia, pero en sí misma conlleva un proceso en constante cambio. Día con día, el profesorado ha de buscar nuevas destrezas y habilidades que le permitan una mejor posición de interacción entre y con sus discípulos. Su formación debe ser permanente y renovada, para enfrentar los retos que impone la enseñanza y la evaluación de su actividad, siendo ésta una necesidad imperiosa de los sistemas educativos. En este sentido, baste como justificación, y punto de partida de estas líneas, la necesidad de estudiar, conocer y analizar la evaluación docente, pues es un tema de permanente actualidad.

Ahora bien, para aproximarse a la evaluación docente, se requiere de métodos que permitan valorar e incentivar esta labor, debido a la importancia de contar con herramientas apropiadas para tomar las decisiones oportunas sobre su acción. En este sentido, una de las técnicas más utilizadas para ello ha sido: "el cuestionario de opinión estudiantil", porque el estudiantado es la razón de ser de la docencia, por ello, es vital saber, conocer y entender su pensamiento, su posición frente a la enseñanza que reciben. Ellos son los receptores de todos los conocimientos, las actitudes y las actividades docentes. Con este elemento encontramos otro argumento que destaca la importancia de nuestro estudio, convirtiéndolo en novedoso día con día, mientras existan estudiantes y docentes será indispensable estudiar la validez y la fiabilidad de los instrumentos utilizados para evaluarles.

En otro tema, hablando del uso de los cuestionarios utilizados para evaluar la acción docente, sus resultados se utilizan para tomar múltiples decisiones, acarreando numerosas consecuencias en el personal docente, que pueden incentivar o desmotivar su accionar; por ello, es obligatorio contar con instrumentos fiables y válidos para reunir información precisa y exacta sobre el fenómeno, pero su evaluación debe estar libre de prejuicios e improvisaciones, la cuales nacen cuando no existe un proceso riguroso de trabajo y de análisis. Lamentablemente, aunque se han visto muchos cuestionarios, son pocos los que

han sido analizados con la dedicación necesaria para desestimar arbitrariedades e imprecisiones.

Por tanto, este artículo tiene como propósito fundamental profundizar en los estudios realizados sobre la fiabilidad y la validez aplicados en distintas épocas a numerosos cuestionarios de evaluación, utilizados, principalmente, en Europa, Norteamérica y Australia, a fin de observar sus métodos, sus procesos de análisis y sus conclusiones para mostrar un camino a los interesados en esta actividad; dado que en América Central aún no es común encontrar estos análisis en los cuestionarios de evaluación de las Universidades o entidades de gobierno. Es más, en algunos casos ni siquiera se evalúa la docencia. La realización de una revisión bibliográfica profunda permite mostrar el panorama que ha emergido en cuanto a la confiabilidad y a la validez de este tipo de evaluación. A partir de este propósito surgen las siguientes interrogantes:

- ¿Qué significa validez y confiabilidad?
- ¿Qué métodos se han empleado para medir estas propiedades psicométricas en los instrumentos de evaluación docente?
- ¿Puede realmente el cuerpo docente, confiar en los instrumentos utilizados para su evaluación?

## 1.- LOS CUESTIONARIOS EN LA MEDIDA DE LA COMPETENCIA DOCENTE

El cuestionario utilizado para conocer la opinión de la población estudiantil, respecto de la acción docente, es comúnmente empleado de acuerdo con diferentes intereses de los involucrados, llámese administradores, cuerpo docente o estudiantado. Por ejemplo: 1) la administración lo tiene en cuenta, como parte del proceso de toma de decisiones relacionadas con la selección de su personal, los incentivos de promoción económicos o académicos y con la permanencia de sus docentes dentro del sistema universitario; 2) al profesorado le suministra información actualizada acerca de la visión que tienen sus estudiantes de la forma y de la estructura de sus cursos, de los elementos por mejorar y de la calidad de su enseñanza; 3) a la población estudiantil le proporciona herramientas que le permite seleccionar los cursos y las personas de su interés; y 4) a las personas interesadas en la investigación les es vital, para indagar sobre la eficacia de la enseñanza y el aprendizaje en la universidad. En fin, los resultados aportados por los cuestionarios han sido utilizados para una gran variedad de fines, por lo tanto, es necesario tener la certeza de que estos instrumentos son totalmente fiables y válidos.

Pese a la importancia de los cuestionarios, según Marsh (2003), muy pocos de ellos han sido estudiados profundamente en cuanto a su fiabilidad y a su validez, es más, cuentan con una insuficiente estructura teórica; lo cual incrementa la incertidumbre y el rechazo por parte del cuerpo docente universitario, sobre la calidad, la rigurosidad del instrumento de medida y la relevancia de los datos aportados. En este sentido, la medida de la competencia docente deja una gran cantidad de interrogantes iniciales relacionadas con la fiabilidad y la validez.

## 2.- LA FIABILIDAD, UN PASO INICIAL DE ESTUDIO

La fiabilidad en psicometría se puede conceptualizar de distintas maneras que, a su vez, se traducen en métodos distintos de cálculo, los cuales son totalmente diferentes al proceso de validez[2]. En ese sentido, cuando hablamos en términos propios de la fiabilidad de estas evaluaciones, nos referimos, sobre todo, a dos elementos: *unanimidad* y *estabilidad*:

*La unanimidad* indica directamente al punto en que el estudiantado, es consistente o *unánime* en su juicio cuando hace diferencias entre el profesorado. En otras palabras, pueden observar muy claro cada una de las características de uno u otro docente y de esa forma emitir un juicio global que se agrupa en la escala, no se dispersa la opinión hacia todos los puntos de ella, sino que se reúne alrededor de las características presentadas.

*La estabilidad* expresa el punto en el que las evaluaciones no varían notablemente al pasar del tiempo, sino que el cuerpo estudiantil mantiene los mismos criterios de valoración del profesorado con el transcurrir de los años. De esa forma, la evaluación del cuerpo docente, no cambiará mucho a través del tiempo y los resultados de su evaluación serán similares varios años después.

Estos dos aspectos de la evaluación de la *docencia* universitaria, *unanimidad* y *estabilidad,* son considerados por Overall y Marsh (1980), quienes en un estudio longitudinal muy importante, explican que la fiabilidad en las encuestas del estudiantado se entiende como el acuerdo relativo (unanimidad) entre las valoraciones de diferentes estudiantes dentro de la misma clase, bajo la asunción de que cualquier varianza específica del grupo estudiantil es aleatoria y debería ser considerada como varianza de error. Es estable a lo largo de un período de tiempo de varios años, separando los dos conjuntos de estas

[2] Una exposición detallada sobre los distintos métodos de cálculo puede consultarse en Feldman (1977).

valoraciones, ello podrá ser incluido como la varianza sistemática cuando los coeficientes de estabilidad a largo plazo se basen sobre sus respuestas individuales.

No obstante, sobre este tema Cruse (1987) señala que los coeficientes de *fiabilidad indican* que el estudiantado puntúa de forma consistente, de la misma manera en ocasiones diferentes al profesorado, pero ello no significa que estos evalúan exactamente la docencia universitaria, pues a los estudiantes se les plantea un modelo tradicional de docente, el cual se ha ajustado a ciertos factores característicos del "buen profesor". Este argumento es un tema debatido por mucho tiempo, principalmente, porque los rasgos encontrados por los investigadores pueden o no ser apropiados para un tipo de docencia específica, entonces si eso sucede se dejarían sin valorar otros rasgos especiales de la población docente universitaria.

El problema de todo este trabajo radica en que por la multidimensionalidad y complejidad no solo de la labor docente, sino de la misma universidad en su estructura, se hace difícil encontrar un modelo global que ajuste perfectamente a todas las características involucradas en el campo específico del saber universitario. Sin embargo, vale la pena resaltar que las investigaciones realizadas alrededor del mundo indican la existencia de ciertos factores característicos de la docencia universitaria que pueden ser usados para medirla en su totalidad.

En ese sentido, Marsh (1987) afirma que en los instrumentos bien construidos, la consistencia interna es normalmente alta, aunque "proporciona una estimación inflada de la fiabilidad, porque ignora la proporción sustancial de error debido a la falta de acuerdo entre diferentes estudiantes y por lo tanto no deberían ser usados en general" (pág. 275); quizá puede ser apropiado utilizarlas para determinar hasta qué punto, las correlaciones entre facetas múltiples se han hecho tan grandes que las facetas separadas no pueden ser distinguidas, como con el procedimiento multirasgo multimétodo. Por lo tanto, el autor recomienda que una de las medidas por utilizar es la fiabilidad de la respuesta media de clase basada en el acuerdo entre el estudiantado de la misma clase, o sea, la unanimidad.

## 2.1. La unanimidad

Comúnmente cuando se habla de fiabilidad se hace referencia a *unanimidad* que, en su concepción más simple, es el grado de acuerdo de la población estudiantil al valorar al cuerpo docente, en un ítem cualquiera del cuestionario. Por lo tanto, la fiabilidad puede calcularse en cualquier ítem o en cualquier factor y no globalmente en todo el cuestionario,

como si se tratara de un test convencional. Evidentemente lo anterior considera la existencia de fiabilidad, si las diferencias se deben, fundamentalmente, a que en el cuerpo docente son diferentes y así lo percibe el estudiantado; no a que los segundos son distintos en su manera de evaluar a la población docente. Un ítem o factor será fiable si existe un grado de acuerdo entre el estudiantado al evaluar al profesorado, no si los ítems ordenan a la población docente de manera semejante.

Villa y Morales (1993) manifiestan que operacionalmente se trata de los coeficientes de correlación denominados intra – clase, relacionados con el análisis de varianza y con el coeficiente de fiabilidad de Cronbach utilizados comúnmente en los test. Estos coeficientes se valoran de 0 a 1[3] , y su magnitud depende, fundamentalmente, del número de alumnos que responden en cada clase. Una observación importante es que no va a existir fiabilidad detectable si no hay diferencias entre el profesorado: no se puede clasificar bien a los que son semejantes, por eso, estos cálculos requieren disponer, por lo menos, de más de treinta estudiantes. Normalmente en estas evaluaciones se integran muchos más. En un estudio relacionado con la fiabilidad y el número de estudiantes, Marsh (1982) encuentra el valor medio de la fiabilidad en torno a 0.90, cuando son unos 25 estudiantes los que responden a los cuestionarios de evaluación y cuando son 10 los que responden la fiabilidad baja a 0.74, por esa razón, la muestra en la evaluación debe superar las 30 personas.

El grado de unanimidad presenta un elevado nivel en cuanto a la consistencia interna de las escalas, entre sujetos, en diferentes momentos del curso y en diferentes cursos de la misma tipología. Lo anterior expone que en la clase predomina la unanimidad de opinión frente a la divergencia, cuando se trata de evaluar al cuerpo docente. En otras palabras, la población estudiantil tiene opinión *unánime* o similar, en torno a lo que en grupo se considera la "buena" docencia universitaria y puede, unánimemente, dirigirse con similar opinión a aquellos aspectos considerados deficientes. Esto supone que el mayor grado de unanimidad entre el estudiantado determina un buen nivel de fiabilidad en los ítems que componen los cuestionarios de opinión.

## 2.2. La estabilidad

La *estabilidad* en los cuestionarios de evaluación docente se comprueba a través del tiempo, cuando se evalúa la instrucción con la opinión del estudiantado, o se evalúa al

[3] El valor cercano a uno es considerado como muy bueno y caso contrario en el cercano a cero.

funcionario docente, en dos segmentos separados por meses, semestres u años. Este tipo de estudio se diferencia por el tiempo trascurrido entre las aplicaciones, variando ya sea la mitad y en el final del curso o, en ocasiones, pueden transcurrir meses e incluso años entre estas. Claro, las condiciones ideales para este tipo de investigación radican en comparar las puntuaciones dadas por los mismos sujetos, un tiempo después de haber evaluado a un docente. Los diseños longitudinales de investigación y evaluación han sido una fuerte herramienta para demostrar la estabilidad de las puntuaciones en los cuestionarios de la evaluación del docente universitario.

Así mismo, los diseños transversales o transeccionales también han aportado información al respecto, comparando las puntuaciones de estudiantes actuales con un grupo de exalumnos, asumiendo que son similares. Según Guthrie (1954) la *evidencia* sobre la estabilidad de la clase para valorar al docente data de 1924, aunque para Albanese (1991) y Hativa (1996), los datos son más recientes. Si bien existen diferencias entre los investigadores en cuanto a la fecha de inicio de estos estudios, lo cierto e importante del caso es que los resultados demuestran estabilidad de las puntuaciones del estudiantado.

No obstante, una de las críticas hechas a estos estudios radica en la imposibilidad de decidir quiénes son los evaluadores más importantes, si los antiguos estudiantes o los de nuevo ingreso. La visión de los antiguos estudiantes ofrece una perspectiva que permite diferenciar si recibieron una enseñanza eficaz cuando más adelante en el tiempo han tenido que llevarla a la práctica; esto podría crear diferencias sistemáticas de juicios con respecto a quienes están cursando estudios. El problema de la muestra es quizá el componente más complejo en los estudios longitudinales, pues el anonimato es siempre indispensable. En este sentido, tendríamos que sumar los estudiantes repitentes, los desertores y los factores involucrados en esas acciones, lo cual integraría un grado mayor de complejidad al estudio.

Sin entrar en la discusión de la muestra de estudio, se puede afirmar que las evaluaciones de los estudiantes actuales son importantes, porque son los beneficiarios inmediatos de la enseñanza, son los que en este momento reciben al docente en sus aulas y comparten su enseñanza. Ahora bien, un problema podría surgir respecto a su visión de la actividad la cual puede condicionar la eficacia de la enseñanza, por ende, el cuerpo docente ha de ser lo suficientemente maduro como para mantener una estructura cognitiva y actitudinal capaz de enfrentar estos retos. . En otro sentido, y desde nuestra perspectiva, las evaluaciones de los antiguos estudiantes son también necesarias: tienen más años en la Universidad, a veces hasta han concluido sus estudios universitarios y han recibido en sus

aulas a una gran cantidad de docentes, involucrando múltiples campos y con diferentes competencias; un hecho relevante que les permite contar con una visión más amplia del rol docente, de su competencia, de los aciertos y de los desaciertos. Por lo tanto, son capaces de valorar al docente desde un plano mayor.

Ahora bien, sintetizando algunos resultados obtenidos en los estudios sobre la *estabilidad* de la clase, observamos que el cuestionario de evaluación del profesorado es estable en considerables períodos de tiempo. El estudiantado mantiene su opinión del docente aún después de muchos años de haber concluido sus estudios. Carson (1999), en un estudio longitudinal desarrollado entre 1964-1999, advierte que la población estudiantil, recuerda muy bien a sus pésimos profesores universitarios. Marsh y Overall (1981) encontraron una estabilidad media de r = 0.83 entre 100 cursos estudiados a intervalo de un año, lo cual es un grado de asociación notablemente alto. McKeachie (1987) examinó que las opiniones del estudiantado correlacionan alto (r=0.94), aún después de haber pasado un año o más entre cada aplicación. Como observamos en los estudios internacionales, la estabilidad en los cuestionarios de evaluación docente ha resultado en una correlación positiva muy alta, lo cual demuestra solamente la posición del grupo estudiantil.

Por su parte, Costin, Greenough y Menges (1971) y Gillmore (1973) encuentran correlación, entre 0.79 y 0.87, en las valoraciones del estudiantado de un mismo docente y el grado del curso, demostrando con valores muy altos que las evaluaciones no varían mucho entre los estudiantados nuevos y los de viejo ingreso, evidenciando con ello que la discusión de la muestra es irrelevante, pues existe un alto grado de asociación entre antiguos y nuevos estudiantes. En estudios anteriores sobre el tema como el de Drucker y Remmers (1951), hallaron que el grupo estudiantil, que había permanecido fuera de la institución por cinco o diez años, promediaban al docente de la misma manera que cuando eran estudiantes regulares de la Universidad. A todas luces parece ser que la tendencia internacional de las investigaciones demuestra, claramente, la estabilidad entre las medidas de evaluación docente.

Hativa y Raviv (1993), en un estudio similar, concluyen que el mismo curso ofrecido por el mismo profesor, en diferentes semestres, ofrece índices relativamente estables, a menos que se realice una intervención instruccional específica donde sea implantada una nueva estrategia educativa, pero si se mantiene la estructura de enseñanza que ha utilizado durante todos sus años de labor, la evaluación se conservará estable.

*La estabilidad* de las respuestas en los cuestionarios ha sido investigada, además, para determinar la influencia de las modas en la población estudiantil u otros efectos posibles. Las encuestas dadas por el mismo número de educandos en un período corto de tiempo, según Centra (1972), producen resultados altamente estables. Overall y Marsh (1980) recogieron las opiniones medias cinco semanas después de haber concluido el curso y correlacionaban en un 0.77. Un año después, aplicaron nuevamente el instrumento a los mismos estudiantes y, de igual forma, correlacionaban significativamente. Marsh (1992b) encuentra que hay un perfil de valoración estable a lo largo del tiempo y no muestra cambios sistemáticos, lo que evidencia la estabilidad mostrada por estas evaluaciones construidas adecuadamente.

De todo este cuerpo científico internacional, se puede concluir que el estudiantado universitario dice después (con los años o meses) más o menos lo mismo que decía antes: su opinión sobre un mismo docente se mantiene a través del tiempo con muy poca variación. Aunque quizá se puede pensar que con los años el grupo estudiantil evaluará de manera distinta a los mismos docentes. Ha sido comprobado en las evaluaciones y retrospectivas realizadas por algunas universidades de habla inglesa, que el grupo estudiantil antiguo o graduado, continúa evaluando de manera semejante al cuerpo docente, que cuando asistían a cursos regulares; los resultados de estas nuevas evaluaciones mantienen la misma visión de los mejores y los peores docentes, al igual que lo hicieron en el pasado.

Evidentemente, no se trata de verdades absolutas y sin matices, pero en la práctica, las evaluaciones longitudinales nos han demostrado, con resultados más o menos satisfactorios, evaluaciones completamente estables a lo largo del tiempo, y sugieren, como una perspectiva añadida, que no altera las puntuaciones dadas al final de la asignatura. Drucker y Remmers (1950), Centra (1979) y Marsh y Overall (1981) concluyen que las puntuaciones que los docentes reciben están generalmente correlacionadas en un período de tiempo.

Hemos visto que las valoraciones de estudiantes dadas en los cuestionarios de opinión elaborados, sistemáticamente, son *unánimes y estables*, lo que demuestra fiabilidad en aquellos casos reportados en las universidades de habla inglesa; valdría la pena comenzar a revisar los análisis de los instrumentos utilizados en las Universidades centroamericanos y sus reportes.

En este devenir a veces se entiende por fiabilidad la coherencia entre diversos evaluadores, ya sean colegas, directores o estudiantes, pero esta coherencia es una prueba

de validez convergente, no de fiabilidad. Ciertamente *la validez* de este tipo de cuestionarios es más difícil de comprobar que su fiabilidad, debido a la existencia de múltiples indicadores de validez, aunque para Cronbach y Meehl (1955) la validez de constructo reúne todos los diferentes tipos existentes y es la única a tomar en cuenta.

## 3. LA VALIDEZ EN LOS CUESTIONARIOS UNA OPERACIÓN COMPLICADA

La literatura sobre *la validez* de los cuestionarios de opinión para evaluar la competencia docente ha originado un cuerpo impresionante de literatura. Feldman (1997), Marsh y Dunkin (1992) y Marsh (2001, 1987) demuestran en sus estudios que estos son multidimensionales, fiables, estables y razonablemente válidos con respecto a una gran variedad de indicadores de eficacia docente.

Por el contrario, Shadish (1998) y Weinbach (1988) cuestionan la validez indicando que los cuestionarios presentan muy poca validez o que no tienen del todo, por tanto, no deben ser utilizadas para tomar decisiones sobre el empleo del cuerpo docente. Su discurso asevera que, aunque la literatura sobre evaluación del profesorado es extensa, en muchas ocasiones el nivel conceptual y metodológico de estos instrumentos es muy mediocre. Otros autores mantienen una posición intermedia como Greenwald y Gillmore (1997) y Franklin y Theall (1996) expresando que las encuestas de opinión del grupo estudiantil son generalmente fiables y los indicadores de efectividad docente son válidos, aunque es necesaria una evaluación adicional independiente.

### 3.1. Los años 70 y la nota final

Esta época se destaca por una amplia variedad de temas de estudio, sin embargo, el que más preocupa es el sesgo producido por la nota final. Snyder y Clare (1976) manifiestan que las valoraciones de estudiantes están sesgadas por la nota final del curso. La propuesta expresa que si el profesorado asigna buenas notas, el estudiantado valorará mejor al cuerpo docente. Refiriéndose a lo anterior Worthington y Wong (1979) advierten que los hallazgos reportados sugieren que la validez del cuestionario de opinión está cuestionada seriamente. Si este fuera el caso, un instrumento de evaluación afectado por esta variable, invalida totalmente la evaluación docente realizada, no se podría considerar para nada la información aportada.

Durante estos años se realizaron una serie de trabajos experimentales, tratando de demostrar el sesgo presentado por las encuestas de estudiantes al introducir la variable de nota final del curso. Según Chacko (1983), Holmes (1972), Powell (1977), Vasta y Sarmiento (1979) y Worthington y Wong (1979), la metodología consistía en manipular la nota del estudiantado hacia arriba o hacia abajo, según el objetivo del experimento, para determinar el grado de influencia ejercida sobre la opinión del grupo estudiantil; al rellenar el cuestionario de evaluación. Indudablemente, estos investigadores encontraron algunos efectos en sus grupos de trabajo y publicaron sus resultados.

No obstante, al contrastar los resultados obtenidos en estos experimentos, utilizando los mismos datos, investigadores de la talla de Abrami, Dickens, Perry y Leventahl (1980), Marsh (1987) y Marsh y Dunkin (1992) se encontraron con deficiencias en el diseño y en la metodología de investigación, lo cual a su vez invalidaba enfáticamente los resultados obtenidos. Por lo tanto, estos recomiendan no tomar en cuenta sus resultados, porque metodológicamente estaban defectuosos. Pese a todo, la línea de los años 70 trabajó mucho sobre la hipótesis de que la nota del curso afectaba los cuestionarios de opinión y fue totalmente apoyada por las investigaciones experimentales realizadas en esos años.

### 3.2. De los 80 hasta el presente; el uso de otras técnicas

En los 80 los estudios e investigaciones trataron de determinar la validez de constructo, suministrando tres tipos de evidencia: los estudios multisección, los path análisis y los estudios multirasgo - multimétodo.

El primero, los estudios *multisección,* es quizás el mejor y mayor grupo de todos los estudios de validez de constructo y se refieren a la multisección de un mismo curso en el que participaron varios docentes, utilizando como criterio de rendimiento un examen cualquiera aplicado al grupo estudiantil. Los investigadores correlacionan la media de la valoración del estudiantado y la media de logro de un curso Universitario. Una correlación positiva y significativa era tomada como evidencia de la validez de las encuestas de estudiantes. Los estudios demostraron que las diferencias en el rendimiento del grupo estudiantil, que fue atendido en un mismo curso por diferentes docentes, se reflejaron en las evaluaciones hechas por la clase.

Estos resultados han sido muy analizados en varios meta-análisis, no obstante, no se llega a un acuerdo relacionado con la validez de los cuestionarios de opinión, aunque ciertamente se ha logrado una modesta validez convergente. Al respecto, Abrami, Cohen y

d'Apollonia (1988) y finalmente D'Apollonia y Abrami (1997) destacan que las correlaciones entre los exámenes, como medidas del logro del estudiantado, tienen una media de r = 0.40, valor a tomar con precaución. Marsh y Dunkin (1992), refiriéndose a este índice advierten el resultado puede estar afectado por la variable motivación, en las diferentes secciones o por el grado de satisfacción con la nota y es un elemento por controlar, antes de realizar los análisis de datos de la evaluación docente.

Si revisamos la literatura sobre el tema, podemos ver la existencia de varias publicaciones[4] sobre los estudios multisección, los cuales han marcado una diferencia clara en cuanto a la validez de los cuestionarios de evaluación. Por un lado, Cohen (1981) señala: "*existe un fuerte apoyo a la validez de las valoraciones de estudiantes como medidas de la eficacia docente*" (p. 300), por otro lado, Dowell y Neal (1982) tienen sus dudas al expresar que: "*los estimados de validez de las valoraciones de estudiantes no convencen*" (p. 52).

Sin embargo, McCallun (1984) concluye que los resultados de los cuestionarios son homogéneos. Abrami (1990), no apoya esta afirmación y menciona que Rodin y Rodin (1972) encontraron una fuerte y negativa correlación (-0.75) entre las valoraciones y el logro; mientras Centra (1977) reportó una fuerte y positiva correlación (0.92) en sus estudios. De todo este cuerpo de literatura se puede discernir que las conclusiones sobre el método multisección no cuenta con claros resultados en torno a la validez de estas evaluaciones.

Greenwald (1997b) destaca que los estudios de validez multisección a favor de la validez de constructo de las valoraciones de estudiantes, se apoyan en la interpretación de las correlaciones observadas entre las notas y estas valoraciones en términos de efectos paralelos de terceras variables. Para el autor, si la nota correlaciona con la evaluación es fundamentalmente porque los buenos docentes generan altas notas y altas valoraciones, lo cual es válido.

Otro grupo de estudio son los *Path análisis* y exploran la validez de constructo sobre la idea de una tercera variable, explicando la correlación entre la nota esperada y las evaluaciones, pero considerando terceras variables. Howard y Maxwell (1980) midieron el grado de motivación del estudiantado y concluyeron: "*la relación entre la nota y la satisfacción del estudiante puede ser vista como un resultado de causa importante en las relaciones entre otras variables, por ejemplo, como evidencia de contaminación del grado de indulgencia del profesor*" (p. 810). En otro ejemplo similar Marsh (1980), destaca

[4] Ver: Abrami, 1989; Cohen, 1981; 1982, 1983; Dowell y Neal, 1982 y McCallum, 1984.

> los path análisis han demostrado que el estudiantado como sujeto principal de interés tiene un fuerte impacto en las evaluaciones y (...) también cuentan como una tercera parte de la relación entre la esperanza de nota y estas evaluaciones (…) por su parte la expectativa de nota es vista como sesgo – aunque pequeño- en las evaluaciones y esta interpretación está abierta a las interpretaciones alternativas (p.196).

Con el desarrollo de los modernos métodos de análisis estadístico para estudiar las relaciones causales entre las variables (LISREL, EQS y AMOS), se abre toda una gama de posibilidades para explorar con profundidad la competencia docente universitaria. Ejemplos de este método se pueden observar en los trabajos de García (1996, 1997, 1998) y Acevedo (2004), quienes elaboraron propuestas a partir de los modelos de ecuaciones estructurales en la validez de constructo utilizando Lisrel y Amos, respectivamente.

El tercer grupo para estudiar la validez de los cuestionarios de opinión ha sido el de los estudios *multirasgo - multimétodo.* Esta técnica fue desarrollada por Campbell y Fiske (1959) como un medio para obtener índices acerca de la validez convergente y validez discriminante. El modelo se basa, según López Feal (1986), en la diferencia entre rasgo como atributo, característica o propiedad mensurable y método como forma de aproximación a la medida del rasgo.

Los estudios multirasgo – multimétodo en la investigación sobre la competencia docente se han utilizado para demostrar que la valoración del grupo estudiantil tiene validez convergente y discriminante. En este sentido, los autores como Marsh (1982) y García Ramos (1999a) han tratado de demostrar que las encuestas correlacionan: a) bien con las medidas basadas en otros métodos de evaluación del constructo de la calidad docente y b) relativamente menos bien con medidas de otros constructos. Estos estudios han demostrado la validez convergente y discriminante de los cuestionarios de opinión. No obstante, Marsh (1987) señala que el diseño de estas investigaciones requieren de un gran control por las amenazas a la validez interna, a la externa y a ambas. Específicamente, el diseño puede presentar amenazas a la validez interna, pero controlando las características externas se pueden conocer lo que significan las diferencias de clase. Lógicamente estos estudios no se adaptna a las características extrañas o ajenas de la actividad docente como las variables del cuerpo estudiantil, los cursos y otras.

Por su parte, Greenwald (1997b) destaca que ninguno de los estudios ha considerado la expectativa de nota como un recurso de contaminación en la evaluación de la

competencia docente. Acerca de este tema, McKeachie (1979) afirma que la mayoría de los factores que pueden invalidar las evaluaciones de estudiantes tienen efectos muy pequeños, los cuales no pueden ser considerados como contaminantes del proceso de evaluación. De la misma forma se expresa Marsh (1984, 1997, 2001), quien manifiesta que, posiblemente, el grado de indulgencia produzca un efecto de sesgo, pero el apoyo empírico estadístico a esta idea es muy débil y pequeño como para ser tomado en consideración. Las evaluaciones de estudiantes tienden a ser estadísticamente fiables y válidas y relativamente libres de sesgo.

### 3.3. Nuevas perspectivas de estudio y análisis

Actualmente se toma en cuenta más de un tipo de validez y cada una tiene un fuerte enfoque que marca la diferencia. Por ejemplo Marsh y Roche (1997) centran su trabajo en la estructura conceptual de las valoraciones. Su principal punto de vista es analizar la eficacia docente como un constructo multidimensional. Por lo tanto, los autores tienden a utilizar la medida de las valoraciones para capturar la amplitud de esas dimensiones. D'Apollonia y Abrami (1997) destacan la validez convergente como el foco principal de su trabajo. De su revisión de la literatura sobre validez multisección, los autores concluyen que las valoraciones presentan correlaciones substanciales entre el logro de los y las estudiantes y el examen como medida de rendimiento. Greenwald y Gillmore (1997) concentran su trabajo en la validez discriminante, analizando las regularidades observadas entre la correlación de la nota esperada y las valoraciones desde múltiples teorías y perspectivas estadísticas. Los autores concluyen que la más fuerte contribución de la correlación entre la nota y la valoración de estudiantes es imperceptible (aunque estadísticamente correcta). McKeachie (1997) está a favor de la validez convergente y discriminante de las valoraciones de estudiantes, pero su idea es que no deben ser empleadas en muchos grupos.

Un resumen de la posición de los diferentes autores se puede observar en la siguiente tabla.

**TABLA N°1 AUTORES Y POSICIONES TEÓRICAS RELACIONADA CON LA VALIDEZ**

| Autor | Estructura conceptual | Validez convergente | Validez discriminante | Validez consiguiente [5] |
|---|---|---|---|---|
| **Marsh y Roche (1997)** | La eficacia docente es conceptual y empíricamente multidimensional. Su validez, y particularmente su uso como feedback, son socavados por la ignorancia de su multidimensionalidad. | Las diferentes dimensiones de estas encuestas están consistentemente relacionadas con los criterios conocidos de eficacia docente. Ello apoya la validez de constructo. | Estas valoraciones no se ven afectadas por el sesgo. El sesgo ha sido mal interpretado. | La multidimensionalidad se ve fortalecida por la consulta. La mejora de la eficacia docente es el más importante propósito de las encuestas. Su uso, para decisiones personales, podría ser más informativo y sistemático |
| **D'Apollonnia y Abrami (1997)** | Aunque la enseñanza es multidimensional las evaluaciones de estudiantes contienen un gran factor global. | Las valoraciones globales de los estudiantes están moderadamente correlacionadas con lo que el profesor produce y el aprendizaje de los estudiantes. | Existe una pequeña evidencia de sesgo, aunque pocas de esas características conocidas afectan las valoraciones. | Las valoraciones proveen una información válida de la eficacia docente. Sin embargo, no pueden ser el único recurso de información. |
| **Greewald y Gillmore (1997)** | Las evaluaciones de estudiantes están dominadas por un factor global evaluativo y muchos de los ítems detectan sólo este factor. | Las medidas de las valoraciones presentan moderada correlación con el logro en los diseños multisección | El mismo instructor puede tener altas valoraciones si brinda altas notas o enseña en clases con pocos alumnos. Los valoraciones aumentan con un estilo entusiasta del (la) profesor (a) | La búsqueda de altas valoraciones está inducida sutilmente por la nota o por la reducción de los contenidos académicos del curso. |
| **McKeachie (1997)** | Hay un factor "g" en las valoraciones, pero este es también discriminante de viejos y bajos factores | Las valoraciones de estudiantes son validas, aunque son medidas imperfectas de la eficacia docente. | Se ven influenciadas por otras variables que están relacionada con el contexto. | Éstas contribuyen a juzgar la eficacia docente, pero su uso debe estar dirigido hacia la mejora. |

**Fuente**: Greenwald (1997), pág. 1185.

Como se ha visto, la investigación sobre la evaluación de estudiantes ha sido conducida en muchas perspectivas. Algunos estudios, como los de Barnes y Barnes (1993), Hativa (1996) y Marsh y Overall, (1981), han evaluado su fiabilidad y su validez a través del tiempo, de los cursos y de los instructores.

Otros estudios han evaluado la validez concurrente de las valoraciones de estudiantes, correlacionándolas con otros supuestos criterios de medida, con el fin de contabilizar la

[5] Se refiere al uso que se le da a las evaluaciones y si éstas benefician el sistema educativo.

varianza con el constructo multidimensional del rendimiento de clase. Los criterios encontrados que correlacionan substancialmente con las valoraciones de estudiantes de su profesor incluyen: estudiantes antiguos, colegas, evaluadores externos y administradores. Feldman (1989) halló que el criterio menos correlacionado con estas evaluaciones es el de "productividad investigadora" del docente (Noser, 1996) y el logro del grupo estuidantil (Koon y Murray, 1995). Indudablemente, la explicación podría encontrarse en el hecho de que los estudiantes están menos preocupados por la productividad investigadora de sus docentes y más por la calidad de sus explicaciones, imparcialidad de sus procesos evaluativos, etc.

Más allá de los elementos puramente psicométricos relacionados con la fiabilidad y la validez de este tipo de evaluación, los investigadores han examinado también el nivel y el diseño de estas evaluaciones. Los estudios relacionados con este tema sugieren que la evaluación varía según el tamaño de la clase, el formato de la clase y el nivel de estudios. Broder y Dorfman (1994) estudian el diseño e indican que los mismos estudiantes pueden exhibir diferentes patrones de evaluaciones de un mismo profesor, dependiendo de sus necesidades.

Para Braskamp y Ory (1984), los diversos estudios sobre la validez de los cuestionarios de opinión del estudiantado, de la competencia docente universitaria apuntan hacia *dos rumbos* muy delimitados:

1.- Analizar la medida en que factores extraños pueden *sesgar* los cuestionarios de evaluación.
2.- Estudios correlacionales entre las valoraciones del grupo estudiantil y otras medidas consideradas indicadores *reales* de la efectividad docente.

Estas dos líneas de investigación han sido objeto de muchos trabajos, crítica y resultados dispares en el mundo de la evaluación docente universitaria, debido a que unos buscan los factores asociados al "buen" docente, a aquellas competencias que se deben tener para manejar con sabiduría la clase. Después, un cuerpo extenso de investigadores ha tratado de identificar aquellas características externas del estudiante, profesor y la clase que pueden atentar contra la validez de este tipo de evaluaciones.

### 3.4. Fuentes de sesgo

Respecto a la primera línea de investigación, existe una serie de trabajos relacionados con el contexto de los involucrados y la evaluación del estudiantado. Wachtel (1998) considera que, en los últimos años, el foco de atención de las investigaciones se ha desplazado hacia: "*el interés metodológico de las características específicas de contexto que puedan dañar la validez*" (p. 192) de estas evaluaciones. Esto se refiere a la posibilidad de que las características de contexto, factores ajenos a las competencias del profesorado universitario, puedan sesgar las evaluaciones hechas por el grupo estudiantil.

De manera semejante expresa Worthington (2002), al considerar que la validez y fiabilidad de estas evaluaciones se ve afectada por las variables del contexto, tanto del cuerpo docente, como del estudiantado. Ting (2001), por su parte, reconoce que existen tres determinantes fundamentales que afectan estas valoraciones: las características contextuales del curso, del grupo estudiantil y del profesorado. Acevedo (2006), propone que estas diferencias se encuentran alrededor de las características del cuerpo de profesores y el estudiantado.

Relacionado con las *características de contexto,* Feldman (1978) Cheng y Hoshower (1998) y Wachtel (1998) han estudiado y analizado las particularidades del proceso de *administración de la evaluación*. Feldman (1977, 1978, 1979) observa que el tiempo, el propósito de la evaluación, el anonimato y la presencia del instructor en clase, pueden posiblemente influenciar los resultados de la evaluación y, en alguna manera, crear sesgo. Marsh y Dunkin (1992), Braskamp y Ory (1994) postulan que las influencias en la evaluación del y de la docente están relacionadas con las características del curso. En la actualidad, está muy difundido y se reconoce el impacto que tiene de la electividad del curso (si es obligatorio o no), el nivel, el área de conocimiento y la cantidad de trabajo, entre otros elementos.

Anderson y Siegfried (1997) y Wachtel (1998), entre muchos otros[6], estipulan que la validez de estas evaluaciones tiene relación con las *características del instructor*. Esas características incluyen entre otras: el rango, la experiencia, la reputación, las habilidades de investigación, el género y la apariencia física.

[6] Ver por ejemplo los textos clásicos de: Centra y Creech, 1976; Feldman, 1983; Hamilton, 1980; Mash, 1980, 1987; Perry, Abrami y Leventhal 1979; Erdle, Murray y Rushton, 1985; Greenwald y Gillmore, 1997; Powell, 1978, 1977; Elmore y Pohlmann, 1978

Koermer y Petelle (1991), Tatro (1995), Cheng y Hoshower (1998) han analizado los factores relacionados con las características del grupo estudiantil[7] e hipotetizan que el interés del estudiantado, el género, su expectativa de nota y su edad tienen una influencia que puede sesgar la evaluación de la competencia docente universitaria.

Cuando examinamos la investigación existente asociada al *sesgo* en la evaluación del grupo estudiantil, acerca de la docencia universitaria, se observa que existe una gran cantidad de puntos emergentes. Si bien algunas de esas características, como la administración de la evaluación, el instructor y el curso, han sido extensivamente estudiadas, muy poca atención se ha puesto en determinar cuál es la variable de sesgo más importante, de todas las mencionadas.

Evidentemente, aunque algunos estudios se han empezado a realizar en otros países, como el Reino Unido, España, Australia o Hong Kong, el grueso de las investigaciones realizadas han sido efectuadas exclusivamente en Estados Unidos y Canadá. Un ejemplo de esos trabajos incluye a Tatro (1995), Anderson y Siegfried (1997), Cheng y Hoshower (1998) y Ting (2001). Algunos otros estudios, como los de Casey (1997) y Timpson y Anderson (1997), sobre el proceso de evaluación en Australia, no han puesto su atención en el sesgo que se presenta en estas evaluaciones. Aunque ciertamente pueden existir algunas diferencias significativas entre el cuerpo docente de otros países y los Estados Unidos, actualmente no se cuenta con estándares de comparación. A pesar de esto, Marsh (1994, 1992, 1988, 1981, 1985, 1997), con la colaboración de otros investigadores, han intentado consolidar la aplicación del instrumento SEEQ en diferentes países con resultados muy halagadores, lo cual pone en evidencia que existen ciertas características docentes universales.

En otro sentido, la mayoría de los trabajos existentes no han centrado el problema de investigación, en las características del *contexto de la docencia* universitaria y se han enfocado, solamente, en tomar elementos particulares o micro variables, como si estos no fueran componentes de un conjunto. Este tipo de investigación es un problema, porque no muestra la docencia universitaria en toda su complejidad, pues deja al margen de análisis una serie de factores que pueden o no estar enlazados con otros.

---

[7] A manera de ejemplo ver también: Marsh 2001; Zoller, 1992; Perry, 1990; Kember, 1994; Marsh y Ware, 1982; Abrami, Leventhal y Perry, 1982; Braskamp, Ory y Pieper, 1981.

Aunado a ello, otras variables de contexto, como la edad del estudiantado, el nivel del curso, tamaño de la clase, el interés previo del grupo estudiantil, aún no han sido suficientemente investigadas como un conjunto que integra un fenómeno de estudio. Por ello, su estudio aislado no es del todo recomendable, sino más bien utilizando los modelos jerárquicos lineales se podría indagar con mayor profundidad como lo han hecho los estudios de Ting (2002) y Acevedo (2006).

A nuestro juicio, consideramos particularmente importante la aplicación de rigurosos métodos de análisis empíricos, porque con ello se podrían facilitar el estatus de las características de contexto en la evaluación de la competencia docente universitaria. En este sentido, McKeachie (1997) advierte sobre todo a los investigadores relacionados con el tema de sesgo en estas evaluaciones, que es necesaria mayor precaución en la interpretación de los datos para ofrecer resultados coherentes con la realidad del fenómeno estudiado, porque se están tomando para análisis las variables de manera aislada.

Si bien los estudios de sesgo en estas valoraciones tratan de determinar hasta dónde la diferentes variables afectan a los resultados, otros estudios tratan de determinar más bien la relación positiva con los criterios identificados de competencia docente.

## 3.5 Correlaciones positivas entre las dimensiones e indicadores de competencia docente

La mayor evidencia sobre la validez de los cuestionarios de opinión de estudiantes, procede de los estudios en los cuales existe correlación positiva en las dimensiones de estos instrumentos, reunidas alrededor de lo considerado como un "buen profesor", por ejemplo: organización, evaluación, interacción, comunicación y apoyo, entre otras. Además, se ha observado que *correlacionan positivamente* con otros indicadores como: evaluación por colegas, autoevaluación, valoración por expertos, evaluación por alumnos graduados y por el aprendizaje del grupo estudiantil. Al respecto, Abrami (1990), manifiesta que los investigadores han validado las dimensiones consideradas propias del "buen profesor", como medidas del proceso instruccional, cuyos hallazgos han determinado que estas dimensiones correlacionan con el juicio total del estudiantado sobre la competencia docente y con el juicio que hacen los colegas, antiguos estudiantes, administradores y observadores externos[8].

[8] Un pequeño ejemplo de los investigadores que han trabajado en esta línea se pueden ver en: Abrami, d'Apollonia y Cohen (1990), Cohen (1989), Dickinson (1990), Drews, Burroughs y Nokovich

Para validar los cuestionarios de opinión se correlacionan con: (1) evaluación de los colegas, (2) juicio de expertos, (3) valoraciones de los y las estudiantes y graduados, y (4) con medidas de aprendizaje. Todas las asociaciones indican la existencia de una moderada o alta correlación positiva, lo cual viene a ser considerado como una evidencia adicional de la validez de los cuestionarios de opinión del estudiantado. Todo lo anterior se opone a los estudios realizados por Bendig (1953) y Rodin y Rodin (1972), quienes encontraron correlación negativa entre el logro del estudiantado y la evaluación del cuerpo docente. No obstante, escritos posteriores de Centra (1973a), Frey (1973), y Menges (1991) han criticado, fuertemente, la metodología utilizada por estos investigadores, pues ha objetado los resultados obtenidos en los estos estudios.

Por su parte, en la validación de estas investigaciones se han utilizado herramientas estadísticas tales como: análisis correlacionales, análisis factoriales (exploratorios y confirmatorios) y, en menor grado, los modelos jerárquicos lineales[9]. Pese a esto, McKeachie (1997) destaca que un problema adicional, relacionado con la validez de las conclusiones de estas evaluaciones, se fundamenta en el propio uso de los datos debido a la falta de sofisticación estadística en los comités o los encargados de utilizar esta información, parece ser que estos carecen de nivel o formación estadística y pueden expresar o explicar resultados de forma errónea o falsa. Por ello, se recomienda contar con personal que tenga un alto grado de especialización en la materia,

De esta forma, hemos observado la presencia de una moderada correlación del criterio de competencia docente con el logro del grupo estudiantil, en algunos estudios que investigan esta relación con los predictores de competencia docente, como: d´Appolona y Abrami (1997), Greenwald (1997), Marsh y Roche (1997), McKeachie (1997) y Saroyan y Amundsen (2001). Además, cuando las evaluaciones se asocian a variables no relacionadas con la eficacia docente, los resultados han sido muy diversos y hasta contradictorios. Mientras algunos estudios, como d´Apollonia y Abrami (1997) y Marsh y Roche (1997), sugieren que existe una pequeña evidencia de sesgo de estas valoraciones, otros, como

(1987), Gigliotti y Buchtel (1990), Harrison, Ryan y Moore (1996), Koon y Murray (1995), Nimmer y Stone (1991), O' Connell y Dickenson (1993), Prave y Baril (1993), Prosser y Trigwell (1990), Ryan y Harrinson (1995), y Vu, Marriot, Skeff, Stratos y Litzelman (1997).

[9] Para más información ver: Aleamori y Hexner, 1980; Burdsal y Bardo, 1986; Greenwald y Gillmore, 1997; Marsh, 1984, 2001; Vandewalle, 1997; Marsh y Roche, 1997; García Ramos, 1996; Tejedor, 1990; Ting, 2001; Goddard, Hoy y Woolfolk, 2001; Acevedo, 2006).

Greenwald y Gillmore (1997) proponen que el grado de benevolencia, el tamaño de la clase y el entusiasmo del instructor son elementos potenciales de sesgo.

## 4. DISCUSIÓN

El estudio de la competencia del personal docente universitario es motivo de preocupación desde los inicios del siglo XX, pasando por la década de los 50 y se intensifica entre los años 70 y 80. Obviamente, en la actualidad esta área es impostergable debido al rol del funcionario universitario, como actor social, quien debe enseñar a aprender a sus estudiantes y, a su vez, orientarlos a tomar iniciativas innovadoras en el mundo cambiante del cual es partícipe.

Indudablemente, en los últimos años la preocupación está centrada en la evaluación institucional como un todo y no en un solo componente como es la acción docente, que es una parte de ella. Sin embargo, la evaluación docente es la de mayor preocupación, mayor estudio e incuestionable actualidad. El hecho de evaluar al profesorado, permanecerá en el tapete de las inquietudes universitarias; conocer la actividad es una necesidad imperiosa en aquellas universidades preocupadas por la calidad de su enseñanza.

La coyuntura sociocultural del presente exige revisiones científicas sobre la competencia del docente universitario, por ello, la evaluación permanente del docente es labor ineludible, que debe asumirse mediante la aplicación de técnicas e instrumentos analizados rigurosamente, en donde la improvisación y el miedo no tengan entrada. Pero para llegar a este nivel, han de utilizarse las más modernas técnicas y métodos de análisis estadístico, el establecimiento de criterios de evaluación consensuados y, sobre todo, funciones claras de evaluación. Aunque, ciertamente, debemos tener presente los primeros elementos de ésta: ¿Qué evaluar?, ¿Para qué evaluar? ¿Cómo hacerlo? y ¿Con qué propósito se hace?

Respecto al apoyo y a la oposición de los implicados en este tipo de evaluación, existen diferentes puntos de vista, que señalan vértices opuestas, dado que aparecen personas a favor y en contra. Aleamori (1981) destaca que en la opinión del profesorado, sobre la utilidad de este tipo de evaluación, no hay indicios claros, pues los docentes se mantienen en los extremos, unos señalan que es fiable, válida y útil y otros dicen todo lo contrario.

Sin embargo, después de casi 70 décadas de investigación en el uso de la evaluación de la competencia docente universitaria, se puede manifestar, con seguridad, que los principales investigadores confían en que las encuestas son válidas y fiables y que vale la pena realizarlas. De hecho, uno de los más destacados autores Marsh (1984) manifiesta que la evaluación del grupo estudiantil es solamente un indicador de la eficacia docente, que tiene una validez establecida a conciencia y rigurosamente. Pero es solamente uno de los muchos indicadores necesarios para valorar la actividad docente y, por lo tanto, es deber de la administración basar sus conclusiones en los numerosos indicadores e instrumentos a fin de conocer, realmente, el grado de calidad del profesorado universitario y no apoyar sus conclusiones únicamente en un instrumento.

Esta recomendación debe permanecer constantemente en el escritorio de los administradores universitarios, porque hacerlo de otro modo sería ignorar todos los procedimientos de evaluación docente que se han utilizado para conocer su actividad en la Universidad.

Claro está que los cuestionarios de evaluación han sido muy utilizados y han sido fuertemente criticados en cuanto a su validez y a su fiabilidad; sin embargo, son una técnica fiable que ha sido estudiada por muchos investigadores, quienes han reportando coeficientes muy altos. De esa manera lo confirma Feldman (1977, 1984, 1997), quien considera que estos instrumentos son realmente fiables donde se han reportado coeficientes de fiabilidad en un rango superior de 0.90[10].

Cabe añadir que si un instrumento no está bien construido, como es señalado por los investigadores con frecuencia, obviamente su fiabilidad será muy baja y su validez nula. Por lo tanto, los resultados de la evaluación serían inútiles, así como los juicios que de ella se deriven. Millman (1981), refiriéndose a lo anterior, destaca que si el instrumento no ha sido adecuadamente construido con la ayuda de profesionales, la fiabilidad será muy baja. De igual forma, Craton y Smith (1990), concluyen que un instrumento bien desarrollado y procesado, puede brindar una fiabilidad interna muy alta.

[10] Por su parte Aleamori (1978a) encuentra también rangos de 0.81 a 0.94 para ítem y de .88 a .98 para las subescalas del CIEQ. A su vez Coffey y Gibbs (2001), en el Reino Unido, y Rindermann y Schofield (2001), en Alemania, obtienen coeficientes de fiabilidad de 0.80 a 0.97. Similares índices han sido localizados en España por investigadores tan relevantes como Tejedor y Montero (1989), Salvador, (1990), Muñíz (1991), Jornet (1995) y Abalde y otros (1995), los cuales han reportado coeficientes de fiabilidad que oscilan entre 0.93 y 0.97, puntaje que es considerado muy alto.

Ahora bien, hacemos un juicio de valro, al manifestar que toda la evidencia apunta directamente a que las encuestas de opinión del estudiantado son relativamente fiables, unánimes y estables. Su fiabilidad es más elevada que cualquiera de los otros procedimientos empleados para evaluar la docencia universitaria.

Entonces, si el concepto de validez es todavía objeto de mucha controversia y debate, también lo es *la validación* de los cuestionarios de opinión del estudiantado, dado que no existe un criterio específico sobre lo que es instrucción eficaz. En consecuencia, muchos investigadores utilizan un enfoque de validación criterial, relacionando estas encuestas con otras medidas que se asumen como indicadores de eficacia docente. Desde este enfoque, es necesario que exista relación entre estas y los diferentes indicadores de eficacia docente. Al respecto, Hilton (1993) subraya que estas valoraciones son pobres medidas de la eficacia docente y el enfoque de la validación de constructo se ve: "*disminuido por la falta de un modelo universal aceptable de "buena enseñanza*" (p. 567).

Asimismo, algunos investigadores consideran que las puntuaciones en estas evaluaciones están influenciadas por las metas y las estrategias de enseñanza de los instructores, con lo cual, en palabras de Kolitch y Dean (1999), no podrían ajustarse completamente a la concepción de enseñanza y aprendizaje descrita en un instrumento típico de evaluación. Sobre este pensamiento, Trigwell y Prosser (1996) demuestran cómo la concepción de enseñanza que se asuma tiene influencia directa en el tipo de acciones que se ejecuten para desarrollarla.

Desde este enfoque, la competencia docente obedecería al pensamiento educativo y a las actividades que el profesorado desarrolle para el aprendizaje del grupo estudiantil. McKeachie (1997) cree: "*que su eficacia dependería entonces de una definición de sus metas de enseñanza*" (p. 1219) y que: "*la mayoría de los formularios de valoración de estudiantes del profesor (...) focalizan casi completamente la enseñanza convencional de clase*" (p. 1220).

Por su parte, D'Apollonia y Abrami (1997) admiten que la definición de eficacia instruccional está ligada a la didáctica tradicional de enseñanza y: "...*no necesariamente generaliza a otros contextos instruccionales*..." (p. 1199). De la misma forma se expresa Centra (1993) quien mantiene que: "*el típico formulario de opinión de estudiantes es concebido para reflejar la eficacia de la charla, la clase y su discusión, y otros métodos centrados en el profesor*" (p. 47).

Al respecto, Wilson (1988) argumenta que esos formularios suponen una: "*pedagogía conservadora" que representan siempre a: "un estudiante pasivo y un profesor activo*" (p.90). Indudablemente, estos son elementos por considerar en la elaboración de nuevos instrumentos de evaluación. Obviamente, los existentes han comprobado, constantemente, su validez, la cual ha sido identificada, comentada y alcanzada por medio de múltiples métodos y en diferentes continentes. Marsh (1987), Marsh y Roche (1994) y Ramsden (1991), en sus estudios, deducen que la evaluación de los y las estudiantes es más útil, precisa y válida que otras medidas del rendimiento docente y tiene beneficios añadidos por ser una medida directa de la satisfacción del consumidor. Fundamentado en sus investigaciones, Marsh (1987) concluye que los SET's, es decir, los estudios evaluativos que realiza el estudiantado al cuerpo docente, conlleva probablemente: "*el más grande estudio de todas las formas de evaluación personal, y una de las mejores en términos de apoyo para la investigación empírica*" (p. 369).

Para ir finalizando, podemos retomar algunas advertencias que han sido destacadas en distintos puntos de nuestra exposición retomándolas, porque nos brindan una visión interrogante que debe ser constantemente replanteada. La idea de que las dimensiones de evaluación docente presumen ser *un consenso* no existe. Es obvio, ¿cómo podemos evaluar la eficacia docente adecuadamente si no estamos de acuerdo con lo que constituye la eficacia docente?. En ese sentido, la enseñanza es *un arte* y un sentimiento, que involucra fomentar cualidades similares, las que no son fácilmente evaluables por instrumentos de evaluación.

De este modo, el cuerpo docente puede sentir la pérdida del tiempo de clase en la administración de los formularios de evaluación y eso les puede desmotivar para experimentar con sus métodos de enseñanza, por eso, es importante revisar el tiempo de la aplicación (Centra 1993, p.93).

Por tal motivo, los docentes y los administradores de la educación tienen *poco conocimiento* de la investigación existente en el área y, por lo tanto, suelen administrar la evaluación indebidamente (Franklin y Theall, 1989).

Así, los resultados de la valoración de la eficacia del docente constituyen un reto para las autoridades universitarias, quienes deben velar por la inserción de programas de actualización y mejoramiento de las competencias del docente universitario en pro de garantizar profesionales competitivos acordes con las demandas de la sociedad actual y en

respuesta al contexto que caracteriza la universidad como semillero de vocaciones profesionales en diversos campos.

## 5. A MANERA DE CONCLUSIÓN Y CIERRE

En respuesta a las interrogantes planteadas en el presente artículo se puede decir que:

1) La confiabilidad trata de los coeficientes de correlación denominados intra – clase, relacionados con el análisis de varianza y con el coeficiente de fiabilidad Alfa de Cronbach utilizados comúnmente en los test. Estos coeficientes se valoran de 0 a 1, siendo el mejor valor el que está cercano a la unidad. Una gran cantidad de estudios dividen la muestra en dos mitades y le estiman el coeficiente de fiabilidad Alfa, para ver si obtienen los mismos resultados que con la muestra total.
2) Hablar de validez es más complicado, debido a su relación con una estructura teórica, pues ésta se refiere a si realmente estoy midiendo lo que pretendo medir y no estoy midiendo otra cosa. Ello tiene que partir de una sólida estructura teórica, al que, en el caso de la evaluación docente, ya existe una gran aproximación al docente universitario que queremos.
3) Metodológicamente hablando se han utilizado diversas técnicas estadísticas para estudiar la validez de estos cuestionarios, por ejemplo: correlaciones entre factores, estudios multisección, multirasgo – multimétodo, modelos de ecuaciones estructurales (path análisis) y modelos jerárquicos lineales. Estos dos últimos son los más sofisticados y profundos, permitiéndonos indagar, con mayor precisión, otras técnicas de análisis de datos. En los modelos de ecuaciones estructurales encontramos paquetes como: Lisrel, EQS y Amos. En los jerárquicos: HML y MlwiN, todos cuentan con óptimas herramientas de trabajo que superan fuertemente los estudios estadísticos tradicionales.
4) Desde nuestra visión, la fiabilidad y la validez de los cuestionarios de opinión ha sido ampliamente demostrada; desde luego, si un instrumento de evaluación docente está mal elaborado, aportará datos erróneos e inválidos y podrá ser cuestionado por el profesorado. Por lo tanto, lo importante será buscar especialistas y fundamentos teóricos que permitan desarrollar un instrumento robusto y consistente.

5) El cuerpo docente universitario puede confiar en los resultados de la valoración de su desempeño docente, siempre y cuando apliquen cuestionarios debidamente confeccionados por especialistas y con demostrados índices de fiabilidad y validez.

6) Finalmente, no todos los cuestionarios, que se han utilizado en el transcurso de la historia, cumplen con los índices de calidad y ajuste estadístico como para ser utilizados en la evaluación del profesorado. En este sentido, las autoridades universitarias deben velar por la confección objetiva de instrumentos de evaluación, que permitan obtener resultados asertivos para la toma de decisiones, ya sea en la contratación o ascenso de profesionales al servicio de la docencia en este nivel de formación.

## REFERENCIAS

Abalde, E.; De Salvador, X; González Carbanach, R. Y Muñóz C.; J.M. (1995). Análisis de la Evaluación de la Docencia Universitaria por los (as) alumnos (as) en la Universidad de la Coruña (1993-1994). En **Estudios de Investigación Educativa en Intervención Pedagógica**, (pp. 289-292). Valencia: AIDIPE.

Abrami, P. C., Dickens, W. J., Perry, R. P., & Leventhal, L. (1980). Do teacher standards for assigning grades affect student evaluations of instruction**. Journal of Educational Psychology, 72**, 107-118.

Abrami, P.C. (1989a). How Should We Use Student Ratings to Evaluate Teaching?. **Research in Higher Education, 30** (2), 221-227.

Abrami, P.C.; Cohen, P.A. & D'apollonia, S. (1988). Implementation Problems in Meta-Analysis. **Review of Educational Research, 58**, 151-179.

Abrami, P.C.; D'apollonia, S. & Cohen, P. (1990). Validity of Student Ratings of Instruction. What We Know and What We Do Not. **Journal of Educational Psychology, 82** (2), 219-231.

Abrami.C. y D'apollonia, S. (1990b). .The dimensionality of ratings and their use in personnel decisions. M. Theall Y J. Franklins (Eds.), In **Student Ratings of Instruction. Issues for Improving Practice** (pp. 97-111). New Directions for Teaching and Learning.

Abrami, P.S.; D'apollonia, S. & Rosenfield. (1997). .The Dimensionality Of Student Ratings of Instruction. What We Know and What We do Not. R.P. PRRY & J.C. SMART (Eds.) In **Effective Teaching in Higher Education. Research and Practice** (pp. 321-367). New York. Agathon Press.

Acevedo Alvarez, R. & Rodríguez, N. M. (2006). Factores de sesgo asociados a la validez de la evaluación docente universitaria: un modelo jerárquico lineal. **Archivos Analíticos de Políticas Educativas, 14** (34). Recuperado el 27 de marzo de 2007 de http://epaa.asu.edu/epaa/v14n34/

Acevedo, R. y Fernández, M. J. (2004). La percepción de los estudiantes universitarios en la medida de la competencia docente: validación de una escala. **Educación: Revista de la Universidad de Costa Rica, 28** (2), 145-166.

Albanese, M.A. (1991). .The validity of lecturer ratings by student and trained observers. **Academic Medicine, 66** (1), 26-28.

Aleamori, L. (1978). .Development and factorial validation of the Arizona Course/ Instructor Evaluation Questionnaire. **Educational and Psychological Measurement, 38**, 1063-1067.

Aleamori, L. (1981). **Student ratings of instruction**.. In J. MILLAMN (ed.), **Handbook of Teacher Evaluation,** pp.110-145. Newbury Park, CA. Sage.

Aleamori, L.M. (1999). Student Rating myths versus research facts from 1924 to 1998. **Journal of Personnel Evaluation in Education, 13** (2), 153-166.

Aleamori, L.M. y Hexner, P.Z. (1980). .A review of the research on student evaluation and a report on the effect of different sets of instructions on student course and instructor evaluation. **Instructional Science, 9**, 67-84.

Aleamori, L.M. y Yimer, M. (1974). **Graduating Senior Ratings' Relationship to Colleague Rating Student Rating.** Research Productivity and Academic Rank in Rating Instructional Effectiveness (Research Report N°352). Urbana. University of Illinois, Office of Instructional Resources, Measurement and Research Division.

Anderson, K.H. y Siegfried, J.J. (1997). Gender differences in rating the teaching economics. **Eastern Economic Journal, 23** (3), 347-357.

Arubayi, E. (1987). Improvement of Instruction and Teacher Effectiveness. Are Student Ratings Reliable and Valid?. **Higher Education, 16**, 267-288.

Barnes, L.B. y Barnes, M.W. (1993). Academic discipline and generalizability of student evaluations of instruction. **Research in Higher Education, 34** (2), 135-149.

Bendig, A.W. (1953). .Relation of level of course achievement of student, instructor and course ratings in introductory psychology. **Educational and Psychological Measurement, 13**, 437-488.

Braskamp, L.A. & Ory, J.C. (1994). **Assessing Faculty Work. hancing Individual and Institutional Performance**. San Francisco. Jossey-Bass.

Braskamp, L.A.; Ory, J.C. & Pieper, D.M. (1981). Student Written Comments. Dimensions of Instructional Quality. **Journal of Educational Psychology, 73** (1), 75-70.

Broder, J.M. & Dorfman, J.H. (1994). Determinants of Teaching Quality; What's Important to Students? **Research in Higher Education, 35** (2), 235-249.

Burdsal, C.A. & Bardo, J.W. (1986). Measuring Student's Perceptions of Teaching. Dimensions of Evaluation. **Educational and Psychological Measurement, 56**, 63-79.

Campbell, D.T. y Fiske, D.W. (1959).Convergent and discriminant validation by the multitrait – multimethod matrix. **Psychological Bulletin, 56**, 81-105.

Carson, B.H. (1999). Bad News in the Service of Good Teaching. Students Remember Ineffective Proffessors. **Journal On Excellence In College Teaching, 19** (1), 91-105.

Casey, R.J.; Gentile, P. y Bigger, S.W. (1997). Teaching appraisal in higher education. an Australian perspective. **Higher Education, 34** (3), 459-482.

Cashin, W. E. (1995). **Student Ratings Of Teaching**. The Research Revisited. IDEA Paper No. 32. Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W.E. (1988). **Student Ratings of teaching**. A summary of the research. IDEA. *Paper No 20.* Manhattan, KS. Kansas State University, Center Faculty Evaluation and Development.

CHACKO, T.I. (1983). Student ratings of instruction: a function of grading standards. **Educational Research Quaterly, 83** (1), 19-25.

Centra, J. (1977). Student Ratings of Instruction and Their Relationship to Student Learning. **American Educational Research Journal, 14**, 17-24.

Centra, J.A. (1972). **The Utility of Student Ratings for Instructional Improvement**. Pricenton, NJ. Educational Testing Services.

Centra, J.A. (1979). **Determining Faculty Effectiveness**. San Franciso: Jossey – Bass.

Centra, J.A. (1993). **Reflective Faculty Evaluation. hancing Teaching and Determining Faculty Effectiveness**. San Francisco: Jossey-Bass.

Cheng, Y. y Hoshower, L.B. (1998). Assessing student motivations to participate Teaching evaluations. an application of expectancy theory. **Issues in Accounting Education, 13** (3), 531-549.

Coffey, M. & Gibbs, G. (2001). The Evaluation of the Student Evaluation of Educational Quality Questionnarie (SEEG) in U.K. Higher Education. **Assessment & Evaluation in Higher Education, 26**, (1), 89-93.

Cohen, P. A. (1981). Student ratings of instruction and student achievement. A meta-analysis of multisection validity studies. **Review of Educational Research, 51**, 281-309.

Cohen, P.A. (1983). A selective review of the validity of student ratings of teaching. **Journal of Higher Education, 54**, 448-458.

Cohen, P.A. (1989). Do grades influence students' evaluation of clinical courses? **Journal of Dental Education, 53** (4), 238-240.

Costin, F. (1968). **Survey of Opinions About Lectures.** University of Illinois: Department of Psychology.

Costin, F.; Greenough, W.T. y Menges, R.J. (1971). Student Ratings Of College Teaching. Reliability, Validity and Usefulness. **Review of Educational Research, 41**, 511-535.

Craton, P. y Smith, R. A. (1990). Reconsidering the Unit of Analysis. A Model of Student Ratings of Instruction. **Journal of Educational Psychology, 82** (2), 207-212.

Cruse, D. B. (1987). Student Evaluation and the University Professor. **Higher Education, 15** (6), 723-737.

D'apollonia, S. & Abrami, P. C. (1997). Navigating Student Ratings of Instruction. **American Psychologist, 51** (11), 1198- Ratings 1208

De Neve, H. M. F. y Janssen, P. J. (1982). Validity of Studen t Evaluation of Instruction. **Higher Education, 11** (5), 543-552.

Dickinson, D. J. (1990). The relationship between ratings of teacher performance and student learning. **Contemporary Educational Psychology, 15**, 142-151.

Dowell, D. A. , & Neal, J. A. (1982). A selective view of the validity of student ratings of teaching. **Journal of Higher Education, 53**, 51-62.

Doyle, K. O. (1975). **Student Evaluation of Instruction A.** Lexington, MA: Lexington Books.

Drews, D. R. ; Burroughs, W. J. Y Nokovich, D. A. (1987). Teacher self ratings as a validity criterion for student evaluation. **Teaching of Psychology, 14** (1), 129-143.

Drucker, A. J. Y Remmers, H. H. (1950). Do Alumni and Students Differ in Their Attitudes Toward Instructors? **Purdue University Studies in Higher Education, 70**, 62-64.

Drucker, A. J. Y Remmers, H. H. (1951). Do Alumni and Students Differ in Their Attitudes Toward Instructors? **Journal of Educational Psychology, 42**, 129-143.

Erdle, S.; Murray, H.G. & Rushton, J.P (1985). Personality, Classroom Behavior and Student Ratings of College Teaching Effectiveness: A Path Analysis. **Journal of Educational Psychology, 77** (4), 394-407.

Feldman, K. A. (1977). Consistency and Variability among College Students in Rating Their Teachers and Courses. **Research in Higher Education, 6** (2), 223-274.

Feldman, K. A. (1978). Course characteristics and college students´ ratings of their teachers. What are know and what we don´t know. **Research in Higher Education, 9** (2), 199-242.

Feldman, K. A. (1979). The Significance of Cirunstances for College Students ratings of Their Teachers and Courses. **Research in Higher Education, 10** (2), 149-172.

Feldman, K. A. (1984). Class size and college students´ evaluation of teachers and courses. a closer look. **Research in Higher Education, 21** (11), 45-116.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and formed student, colleagues, administrators and external (neutral) observers. **Research in Higher Education, 30** (2), 137-194.

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

Feldman, K. A. (1997). Indentifying Exemplary Teachers and Teaching. Evidence from Student Ratings. R. P. En PERRY & J. C. SMART (eds.), **Effective teaching in Higher Education. Research and Practice**, (pp. 368-395). Bronx, N. Y: Agathon.

Fernández, J.; Mateo, M. & Muñíz, J. (1998). Is There Relationship Between Class Size and Student Ratings of Teacher Quality?. **Educational and Psychological Measurement, 58** (August), 596-604

Franklin, J. & Theall, M. (1989). **Who read ratings. Knowledge, attitude, and practice of users of student ratings of instruction.** Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Franklin, J. & Theall, M. (1989). **Who read ratings. Knowledge, attitude, and practice of users of student ratings of instruction.** Paper presented at the annual meeting of the American Educational Research Association, San Francisco.

Frey, P. W. (1973). Student Ratings of Teaching. Validity of Several Rating Factors. **Science, 182**, 83-85.

García Ramos, J. M. (1999a). Análisis multirasgo- multimétodo en la validación de instrumentos de medida para la evaluación de la calidad docente en instituciones universitarias. **Revista Española de Pedagogía, 214**, 417-444.

García Ramos, J. M. (1997). Valoración De La Competencia Docente Del Profesor Universitario. Una Aproximación Empírica. **Revista Complutense De Educación, 8** (2), 81-108

García Ramos, J. M. y Congosto Luna, E. (1996). **Un Modelo de Evaluación Institucional en la Universidad.** Salamanca**:** Studia Pedagógica

Gigliotti, R. Y Buchtel, F. (1990). Attributional Bias and Course Evaluations. **Journal of Educational Psychology, 82** (2), 341-351.

Gillmore, G. (1973). **Estimates of Reliability Coefficients for Items and Subscales of the Illinois Courses Evaluation Questionnaire.** (Research Report N°341). Urbana: University of Illinois, Office of Instructional Resources, Measurement, and Research Division.

Gillmore, G. M. ; Kane, M. T. & Maccarato, R. W. (1978). The Generalizability of Student Ratings of Instruction. Estimation of the Teacher And Course Components. **Journal of Educational Measurement, 15**, 1-13.

Gilmore, G. (1984). Student ratings as a factor in faculty employment decisions and periodic review. **Journal of College and University Law, 10**, 557-576.

Goddard, R; Hoy, W & Woolfolk, A. (2001). **Its Meaning, Measure, and Impact on Student Archievement.** Collective Teacher Efficacy. Manuscrito sometido para publicación. 1-40.

Goldman, L. (1993). On erosion of education and the eroding foundation of teacher education (or why we should nor take student evaluation of faculty seriously). **Teacher Quarterly, 20** (2), 57-64.

Greenwald, A. G. (1997b). Validity Concern and Usefulness of Student Ratings Of Instruction. **American Psychologist**, 51 (11), 1182-1186.

Greenwald, A. G. Y Gillmore, G. M. (1997). Grading Leniency is a Removable Contaminant of Student Ratings. **American Psychologist, 51** (11), 1209-1217.

Guthrie, E. R. (1954). **The Evaluation of Teaching**. A Progress Report. Seatle: University of Washington.

Harrison, P. D.; Ryan, J. M. Y Moore, P. S. (1996). College studen's self-insight and common implicit theories in the ratings of teaching effectiveness. **Journal of Educational Psychology, 88** (4), 775-782.

Hativa, N. (1996). University instructors' rating profiles. Stability over time, and disciplinary differences. **Research in Higher Education, 37** (3), 341-365.

Hativa, N. y Raviv, A. (1993). Using a single score for summative teacher evaluation by students. **Research in Higher Education, 34** (5), 625-646.

Hepworth, D. Y Oviatt, B. E. (1985). Using student course evaluations. findings, issues and recommendations. **Journal of Social Work Education, 21** (3), 105-112.

Hilton, H. (1993). Realibility and Validity of Student Evaluation. Testing Models versus Survey Research Model, p. s. **Political Science & Politics, 26**, 562-569.

Hogan, T. P. (1973). Similarity of student ratings across instructors, courses and time. **Research in Higher Education, 1**, 149-154.

Holmes, D. S. (1972). Effects of grades and disconfirmed grade expectancies on students' evaluations of their instructor. **Journal of Educational Psychology, 63,** 130-133.

Howard, G. S. , Conway, C. G, & Maxwell, S. E. (1985). Construct validity of measures of college teaching effectiveness. **Journal of Educational Psychology**, **77**, 187-196.

Howard, G. S. Y Maxwell, S. E. (1980). Correlation between student satisfaction and grades. a case of mistaken causation?. **Journal of Educational Philosophy, 72** (4), 810-820.

Howell, A. J. Y Symbaluk, D. (2001). Published Student Ratings of Instruction. Revealing and Reconciling the Views of Student and Faculty. **Journal of Educational Psychology**, (4), 790-796

Jornet, J. M.; González Such, J. Y Pérez Carbonell, A. (1995) . **Evaluación De la Actividad Universitaria.** G. Quintás (eds.), En **Reforma y Evaluación de la Universidad. Valencia** (pp. 189-244). Valencia: Servei de Publicacions de la Universitat de Valencia.

Koblitz, . N. (1990). Are student ratings unfair to women?. **Newsletter of the Association for Women in Mathematics, 20**, 17-19.

Koermer, C. D. Y Petelle, J. L. (1991). Expectancy violation and student rating of instruction. **Communication Quarterly, 39** (4), 341-350.

Kolitch, E. & Dean, A. V. (1999). Student Ratings of Instruction en the U. S. A. Hidden Assumptions and Missing Conceptions About . Good. Teaching. **Studies in Higher Education, 24**, (1), 27-42.

Koon, J. Y Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. **Journal of Higher Education. 66** (1), 61-81.

Lin, W. Y.; Watkins, D. Y Meng, Q. M. (1995). Student's Evaluation of University teaching. A China perspective. **Higher Education and Research y Development, 14** (1), 61-74.

López Feal, R. (1986). **Construcción de Instrumentos de Medida en Ciencias Conductuales y Sociales**. Barcelona: Alamex.

Mahmound, M. M. (1991). Descriptive models of student decision behaviour in evaluation of higher education. **Assessment and Evaluation in Higher Education, 16** (2), 133-148.

Marques, T. E.; Lane, D. M. Y Dorfman, P. W. (1979). Toward the development of a sustem for instructional evaluation. Is there consensus regarding what constitutes effective teaching?. **Journal of Educational Psychology, 71**, 840-849.

Marsh, H. & Roche, L. R. (1997). Making Students´ Evaluation of Teaching Effectiveness Effective. The Critical Issues of Validity, Bias, and Utility. **American Psychologist, 52**, 11, 1187-1197.

Marsh, H. (1982). SEEQ. A Realiable, Valid, and Useful Instrument for Collecting Students Evaluation of University Teaching. **British Journal of Psychology, 52**, 77-95.

Marsh H. (1984). Student's evaluation of university teaching; dimensionality, reliability, validity, potential biases ad utility. **Journal of Educational Psychology, 76** (5), 707-754.

Marsh, H. (1987a). Students´ evaluation of university teaching; Research findings, methodological issues, and directions for future research. **International Journal of Educational Research, 11**, 253-288.

Marsh, H. (1987b). Student Evaluations of Teaching. M. J. DUNKINS (eds.), **The International Encyclopedia of Teaching and Teacher Evaluation** (pp. 181-187) Oxford: Pergamon Press.

Marsh, H. y Bailey, M. (1993). Multidimensional Students' Evaluations of Teaching Effectiveness. **Journal of Higher Education, 64** (1), 1-18.

Marsh, H. A. y Overall, J. U. (1979b). **Validity of student's evaluation of teaching. A comparison with self evaluations by teaching assistants, undergraduate faculty,**

**and graduate faculty**. Paper presented Annual Meeting of the American Educational Research Association, San Francisco.

Marsh, H. W. (1977). The validity of students' evaluations of instructors independently nominated as best and worst teacher by graduating senior. **American Educational Research Journal, 14**, 441-447.

Marsh, H. W. (1980). The influence of student, course, and instructor characteristics on evaluations of university teaching. **American Educational Research Journal**, **17**, 219-237.

Marsh, H. W. (1992b). **A Longitudinal Perspective of Student's Evaluations of University Teaching. Ratings of The Same Teacher over a 13 Year Period**. Documento presentado en Annual Meeting of the American Educational Research Association (p. 18) San Francisco, Ca. Abril.

Marsh, H. W. (2001). Distinguishing between good (useful) and bad workloads on Students' Evaluations of teaching. **American Educational Research, 38** (1), 183-212.

Marsh, H. W. y Dunkin, M. J. (1992). Students' Evaluation of University Teaching. A Multidimensional Perspective. J.SMART (ed.) **Higher Education. Handbook of Theory and Research**. (pp. 143-223). New York: Agathon.

Marsh, H. W. y Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept. First and higher order factor models and their invariance across groups. **Psychological Bulletin, 97**, 562-582.

Marsh, H. W. y Overall, J. U. (1981). The relative influence of course level, course type, and instructor on students' evaluations of college teaching. **American Educational Research, 18**, 103-112.

Marsh, H. W. y Roche, L. (1992). **The Use of Student's Evaluations of University Teaching To Improve Teaching Effectiveness.** Canberra: Australian Government Publishing Services.

Marsh, H. W. y Roche, L. A. (1994). The Use of Student Evaluations of University Teaching in Different Settings. The Applicability Paradigm. **Australian Journal of Education, 36** (3), 278-300.

Marsh, H. W. y Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching. popular myth, bias, validity or innocent bystanders?. **Journal of Educational Psychology, 92** (1), 202-228.

Marsh, H. W.; Balla, J. R. y Mcdonald, R. P. (1988). Goodness of Fit Indices in Confirmatory Factor Analysis. The Effects of Sample Size. **Psychological Bulletin, 103**, 391-410.

Marsh, H. W. ; Hau, K. T. ; Chung, C. M. & Siu, T. L. P. (1997). Students' s Evaluations of University Teaching. Chinese Version Student ratings of faculty of The Student's

Evaluations of Educational Quality Instrument. **Journal of Educational Psychology, 89** (3), 568-572.

Marsh, H. & Roche, L. R. (1997b). Making Students´s Evaluation of Teaching Effectiveness Effective. The Critical Issues of Validity, Bias, and Utility. **American Psychologist, 52** (11), 1187-1197.

Marsh, H. W.; Touron, J. y Wheeler, B. (1985). Students' s Evaluations of University Instructor. The Applicability of American Instrument in a Spanish Setting. **Teaching and Teacher Education, 1**, 123-138.

Mccallum, L. W. (1984). A meta-analysis of course evaluation data and its use in the tenure decision. **Research in Higher Education**, **21**, 150-158.

McKeachie, W. J. (1979). Student ratings of faculty. A reprise. **Academe, 65**, 384-397.

McKeachie, W. J. (1987). Instructional evaluation. current issues and possible improvement. **Journal of Higher Education, 58** (3), 344-350.

McKeachie, W. J. (1990). Research on College Teaching. The Historical Background. **Journal of Educational Psychology, 82** (2), 189-200.

McKeachie, W. J. (1997). Student Ratings. The Validity of Use. **American Psychologist, 52** (11), 1218-1225.

McKeachie, W. J.; LIN, Y. G. y MENDELSON, C. N. (1978). A small study assessing teacher effectiveness. Does learning last?. **Contemporary Educational Psychology, 3**, 352-357.

Meeth, L. R. (1976). The stateless art of teaching evaluation. Report on teaching. **Change, 8**, 3-5.

Menges, R. J. (1991). The real world of teaching improvement. A faculty perspective. M THEALL & FRANKLIN (Eds.) **Effective Practices for Improving Teaching, New Directions for Teaching and Learning**, (Vol. 48, pp. 21-37). San Francisco: Jossey-Bass.

Miller, R. I. (1987). **Evaluation Faculty for Promotion and Tenure**. San Francisco: Jossey-Bass.

Miller, S. (1984). Student rating scales for tenure and promotion. **Improving College and University Teaching, 32** (2), 87-90

Millman, J. (1981). **Handbook of Teacher Evaluation**. Berverly Hills, CA: Sage.

Monroe, C. y Borzi, M. G. (1989). Methodological issues regarding student evaluation of teacher. A pilot study. **ACA Bulletin, 70**, 73-79.

Moses, I. (1986). Self and Student evaluation of academic staff. **Assessment and Evaluation in Higher Education**, 76-78.

Muñíz, J. ; García, A. y Virgos, J. M. (1991). Escala de la Universidad de Oviedo para la evaluación del profesorado. **Psicothema, 3** (2), 269-281.

Murray, H. G.; Rushton, P. & Paunonen, S. V. (1990). Teacher Personality Traits and Student Instruccional ratings in Six Types of University Courses. **Journal of Educational Psychology, 82** (2), 250-261.

Nimmer, J. G. y Stone, E. F. (1991). Effects of grading practices and time of rating on student ratings of faculty performance and student learning. **Research in Higher Education, 32** (2), 195-215.

Noser, T. C. ; Manakyan, H. y Tanner, J. R. (1996). Research productivity and perceived teaching effectiveness. A survey of economic faculty. **Research in Higher Education, 37** (3), 299-321.

O'connell, D. Q. y Dickenson, D. J. (1993). Student ratings of instruction as a function of testing conditions and perceptions amount learned**. Journal of Research and Development in Education, 27** (1), 18-23.

Overall, J. U. & Marsh, H. W. (1980). Students' s Evaluations of Instruction. A Longitudinal Study of Their Stability. **Journal of Educational Psychology, 72**, 321-325.

Palchik, N. S. (1988). Student assessment of teaching effectiveness in a multi-instructor course for multidisciplinary health professional student. **Evaluation and the Health Professions, 11** (1), 55-73.

Powell, R. W. (1977). Grades, learning, and student evaluation of instruction. **Research in Higher Education**, **7**, 193-205.

Prave, R. S. y Bavril, G. L. (1993). Instructor rating. controlling for bias from initial student interest. **Journal of Educational for Business, 68** (2), 362-366.

Prosser, M. & Trigwell, K. (1990). **How will Future Academic be Evaluated?** Using Student Study Strategies to Check the Validity of Student Evaluations of Teaching Courses. G. MULLIS (Ed)

Ramsden, P. (1991). A performance indicator of teaching quality in higher education. the course experience Questionnarie. **Studies in Higher Education, 16**, 129-150.

Rindermann, H. & Schofield, N. (2001). Generalizability of Multidimensional Student Rating of University Instruction Across Courses and Teacher. **Research in Higher Education, 42**, (4), 377-400.

Rodin, M. & Rodin, B (1972). Student Evaluation of Teachers. **Science, 177**, 1164-1166.

Rutland, P. (1990). Some considerations regarding teaching evaluations. **Political Science Teacher, 3**, 1-2.

Ryan, J. M. & Harrison, P. D. (1995). The Relationship Between Individual Instructional Characteristic and the Overall Assessment of Teaching Effectiveness Across Different Instructional Context. **Research In Higher Education, 36**, (5), 577-594

Ryans, D. G. (1960) **Characteristics of Teachers**. Washington, D. C.: American Council on Education.

Salvador, L. (1990). **Los Docentes Universitarios Exitosos Desde La Perspectiva del Alumno. Su Caracterización Psicopedagógica**. Tesis Doctoral. Universidad de Salamanca. España.

Saroyan, A. & Amundsen,Ch. (2001). Evaluating University teaching. Time to Take Stock. **Assessment & Evaluation in Higher Education, 26**, (4), 341-353

Seldin, P. (1984). **Changing Practices in Faculty Evaluation**. A Critical assessment and Recommendations for Improvement. San Francisco: Josser – Bass.

Seldin, P. (1993a). The use ad abuse of students ratings of professors. **The Chronicle of Higher Education**, 40.

Seldin, P. (1993b). **Successful Use Of Teaching Portafolios**. Bolton, MA: Anker Publishing Co.

Shadish, W. (1998). Some Evaluation Questions. **Practical Assessment, Research & Evaluation, 6** (3). Recuperado el 14 de enero de 2001 de http. //ericae. net/pare/getvn. asp?v=6&n=3].

Snyder, C. R. , & Clair, M. (1976). Effects of expected and obtained grades on teacher evaluation and attribution of performance. **Journal of Educational Psychology**, **68**, 75-82.

Spencer, P. A. Y Flyr, M. L. (1992). **The Formal Evaluation As Ac Impetus To Classroom Change. Myths or Reality?** Research/ Technical Report, Riverside, CA.

Tagomori, H. & Bishop, L. (1995). Student Evaluation of Teaching. Flaw Instruments. Thought and Action. **The National Education Association Higher Education Journal, 11**, 63-78.

Tatro, C. N. (1995). Gender effect on student evaluations of faculty. **Journal of Research and Development in Education, 28** (3), 169-173.

Tejedor, F. y Montero, L. (1990). Indicadores de la Calidad Docente para La Evaluación del Profesor Universitario. **Revista Española de Pedagogía, año XLVIII**, N° 186, mayo-agosto, 259-279.

Timpson, W. W. Y Andrew, D. (1997). Rethinking student evaluation and the improvement of teaching. Instrument for change at the University of Queensland. **Studies in Higher Education, 22** (1), 55-65.

Ting, K. F. (2001). A Multilevel Perspective On Student Ratings of Instruction. Lessons From the Chinese Experience. **Research in Higher Education. 41**, 5, 637-653.

Trigwell, K. & Prosser, M. (1996). Changing Approaches to Teaching. **A Relational Perspective. Studies in Higher Education**, **21**, 275-284.

Vandewalle (1997). Development and validation of a work domain goal orientation instrument. **Educational and Psychological Measurement, 57** (6), 995-1015.

Vasta, R., & Sarmiento, R. F. (1979). Liberal grading improves evaluations but not performance. **Journal of Educational Psychology, 71**, 207-211.

Villa, A. Y Morales, P. (1993). **La Evaluación del Profesor. Una Visión De Los Principales Problemas y Enfoques De Diversos Contextos**. Vitoria: Departamento de Educación, Universidades e Investigación. Gobierno Vasco.

Vu, T. R. ; Marrito, D. J. ; Stratos, G. A. Y Litzelman, D. K. (1997). Prioritzing areas for faculty development of clinical teachers by using student evaluations for evidence-based decisions. **Academic Medicine, 72** (10), 57-59.

Wachtel, H. K. (1998). Student Evaluation of College Teaching Effectiveness. A Brief Review. **Assessment & Evaluation In Higher Education, 23** (June), 191-211.

Weinbach, R. W. (1988). Manipulation of student evaluations. No laughing Mater. **Journal of Social Work Education, 24** (1), 37-34.

Wigington, H.; Tollefson, N. Y Rodríguez, E. (1989). Student's ratings of instructor visited. Interactions among class and instructor variables**. Research in Higher Education, 30** (3), 331-334.

Wilkerson, D.; Rogers, M. A. Y Maughan, R. (2000). Validation of Student, Principal, and Self Ratings in 360 Feedback for Teacher Evaluation**. Journal of Personnel Evaluation in Education, 14** (2), 179-192

Wilson, K. L.; Lizzio, A. & Ramsden, P. (1997). The Development , Validation, and Application of the Course Experience Questionnaire. **Studies in Higher Education, 22** (1), 33-52.

Worthington, A. G. , & Wong, P. T. P. (1979). Effects of earned and assigned grades on student evaluations of an instructor. **Journal of Educational Psychology**, **71**, 764-775.

Worthington, A. C. (2002). The impact of student perception and characteristics on teaching evaluation. A case study in finance education. **Assessment & Evaluation in Higher Education, 27** (1), 49-64.

Zoller, U. (1992). Faculty Teaching Performance Evaluation in Higher Science Education. Issues and Implications. **Science Education, 76** (6), 673-684.