

CHAPTER 2

LITERATURE REVIEW

Prior to the 1970s, the assessment of teaching quality was more popular in elementary and secondary education. In colleges and universities faculty evaluation was conducted in an informal way. As Whitman and Weiss (1982) state, professors were evaluated in order to make decisions about their promotion, retention and tenure, including their capacity to "get along and not make waves" (p. 1).

Preliminary evaluations of college teaching during the 1970s were conducted with formative purposes. The evaluation was conducted with the intention of providing faculty members with valuable information for improving their teaching. This is reflected in the creation of instructional evaluation protocols and professional development offices in most colleges and universities in the U.S. during that time period. (Ory, 1998).

In the 1980s, the interest in teaching and its evaluation increased as a result of pressures for accountability, changes in enrollment trends, financial retrenchment, and government concerns about higher education (Licata, 1986; Whitman and Weiss, 1982). Different reports harshly criticized the quality of undergraduate teaching and demanded the improvement of college instruction (Boyer, 1987; Nyquist and Staton-Spicer, 1987; Cross, 1987). In addition, the interest in instructional evaluation increased as faculty evaluation became an indicator for assessing the overall performance of colleges and universities (Cave et al., 1988). The strong pressures for accountability and for improvement of college teaching during this time period changed the focus of the evaluation from improvement to decision-

making (i.e., tenure, promotion, salary increases). Evaluation results began to be used for both formative and summative purposes (Ory, 1998).

The 1990s were also characterized by increased pressures for accountability (Bok, 1992) and rising concerns about the use of student evaluations of instruction to make administrative decisions such as tenure, promotion and other rewards of academic personnel. The decrease of tenure and promotion opportunities (Marsh and Overall, 1980; Ory and Ryan, 2001) increased concerns for the validity of the evaluation specially when the assessment relies on student ratings as a single source for evaluating the quality of teaching. The use of evaluation data for formative and summative purposes also raised new concerns about the possible misuse of evaluation results and the negative consequences that could result from this misuse, such as diminished educational quality and the violation of a professor's academic freedom (Haskell, 1997).

As Ory (1998) states, most of the concerns about evaluation of teaching in higher education "are coming from today's faculty . . . (who) are making demands that teaching evaluation be more fair, more accurate --a better portrayal of the complexity of their work today" (p. 1). At the same time that administrators have become more concerned about "outcomes assessment and performance based evaluations," the use of evaluation results for administrative decisions has led to present threats of litigation with some decisions being challenged in the courts (Ory, 1998; Haskell, 1997).

As time passed, the evaluation of teaching evolved from an informal process towards a formal and systematic approach. The evaluation is no longer the sole concern of researchers but has become increasingly important to other audiences such as politicians, administrators,

faculty developers, and the public outside Academia. As new purposes arise, evaluation has brought new dilemmas and concerns about the traditional approach for evaluating teaching at the college level.

The Evaluation of Teaching Under a Positivistic Perspective

The traditional positivistic approach for evaluating teaching in higher education is characterized by a strong emphasis on objectivity in measurement that excludes attention to values behind the practice. The researcher or evaluator is an “objective” data gatherer who strongly relies on quantitative methods. According to Erickson (1986), the mainstream paradigm for research on teaching has its roots in the traditional model of the natural sciences:

"The history of the positivistic research on teaching for the past 20 years is one of analytical bootstrapping with very partial theoretical models of the teaching process, on the assumptions that what was generic across classrooms would emerge across studies, and that the subtle variations across classrooms were trivial and could be washed out of the analysis as error variance." (p. 131).

Researchers following this paradigm tend to link the idea of teaching to the idea of treatment, and evaluation to the idea of effectiveness. Teaching effectiveness, then, is “measured by looking at end-of-the-year scores or standardized achievement tests, and to particular teaching practices.” (Erickson, 196, p. 131).

A clear example of this paradigm is the process-product research that strongly supports “direct” instruction, the presentation and recitation of desired knowledge and behaviors. In this research, the effectiveness of teaching is “attributable to combinations of discrete and observable teaching performances per se, operating relatively independent of time and space.” (Shulman, 1986, p. 10).

According to Dunkin and Barnes (1986), process-product research of the 60's and early 70's is the underlying rationale for teaching in higher education today. But unlike most of this research (conducted at other levels of education), in higher education “the process part has been assumed on the basis of prescriptive definitions, or rated by untrained observers, rather than documented through careful observation.” (Dunkin & Barnes, p. 774).

Other researchers point out that the formal conception of good teaching in higher education has not resulted from process-product research but from lists of characteristics or qualities that are used as descriptors of good teaching. Some of these lists are the result of surveys to faculty members and students who have been asked to describe what constitutes "good teaching." A summary of studies identifying these characteristics is included in Table 1:

Table 1.

Characteristics of Good Teaching as Defined in Different Studies

Author	Characteristics of “Good Teaching”
Bousfield. (1940) University of Connecticut as listed in order of importance by 61 undergraduates.	Fairness, mastery of the subject, interesting presentation of the material, well organized material, clearness of exposition, interest in students, helpfulness, ability to direct discussion, sincerity, keenness of intellect.
Clinton (1930) Oregon State University as defined in order of importance by 177 junior students.	Knowledge of the subject matter, pleasing personality, neatness in appearance and work, fairness, kind and sympathetic, keen sense of humor, interest in profession, interesting presentation, alertness and broad-mindedness, knowledge of methods
Desphane, et al (1970) as defined in order	Motivation, rapport, structure, clarity,

of importance by 674 undergraduates and 32 engineering teachers	content mastery, overload (too much work), evaluation procedures, use of teaching aids, instructional skills, teaching styles.
---	--

(table continues)

Table 1 (continued)

Author	Characteristics of “Good Teaching”
Feldman (1988) review of 31 studies in which students and faculty specified the instructional characteristics they considered particularly important to good teaching and effective instruction.	Student perceived outcomes or impact of instruction, teacher-stimulation of interest in the course and its subject matter, teacher's availability and helpfulness, teacher concern and respect for students, friendliness; nature, quality and frequency of the feedback from teacher to students, teacher's sensitivity to, and concern with class level and progress, teacher preparation, organization of the course, teacher's encouragement of questions and discussion, and openness to other's opinion, clarity of course, objectives and requirements.
French (1957) as listed in order of importance by undergraduates at the University of Washington	Interprets ideas clearly, develops student interest, develops skills of thinking, broadens interests, stresses important materials, good pedagogical methods, motivates to do best work, knowledge of subject, conveys new viewpoints, clear explanations
Gadzella (1968) as listed in order of importance by 443 undergraduates at Washington State College.	Knowledge of subject matter, interest in subject, flexibility, well prepared, uses appropriate vocabulary
Hidebrandt (1971) as listed by 308 undergraduates and graduate students at the	Dynamic and energetic person, explains clearly, interesting presentation, enjoys

university of California-Davis.	teaching, has interest in students, encourages class discussion, and discusses other points of view.
Perry (1969) as listed in order of importance by 1493 students, faculty and alumni at the University of Toledo	Well prepared classes, sincere interest in subject, knowledge of subject, effective teaching methods, tests for understanding, fair in evaluation, effective communication, encourages independent thought, course organized logically, motivates students.

(table continues)

Table 1 (continued)

Author	Characteristics of "Good Teaching"
Pogue (1967) as listed in order of importance by 307 students at Philander Smith College	Knowledge of subject, fair evaluator, explains clearly.

Note. Adapted from Developing Programs for Faculty Evaluation (p. 32-33), by R. I. Miller, 1974, San Francisco CA: Jossey-Bass. Copyright 1974 by Jossey-Bass Inc., publishers.

Feldman (1988), Frey (1979) and Marsh (1991) support the use of these characteristics or behaviors for designing instruments for assessing teaching quality. They say that including multiple dimensions can produce useful information as feedback for faculty about their teaching, and for identifying faculty needs for improvement of instruction.

Another group of researchers supports a different point of view. These researchers claim that teaching should be evaluated "globally" rather than paying attention to particular characteristics or dimensions of instruction. Cashin and Downey (1992), Cohen (1986) and Abrami et al (1990; 1993), are among the researchers who support the use of global items or a "carefully weighted average of the factor scores" when the ratings are used for making

administrative decisions (Abrami et al., 1990, p. 98). Abrami says that although "good teaching" is a construct of multiple components, it is more appropriate to evaluate teaching "globally" for comparing instructors across courses, departments and settings. He expresses concern about construct validity problems that could result if the evaluation is unable to include all relevant dimensions and characteristics of good teaching.

Assessing Teaching Effectiveness

The evaluation of teaching under the traditional positivistic paradigm puts strong emphasis on generalization, and on the establishment of cause and effect linkages. In most cases, the evaluation is conducted by using rating scales, semi-structured interviews, personality tests or questionnaires (Feldman, 1986, 1989; Falk, 1971). Questionnaires, however, are the most used instruments in the evaluation of college instruction by researchers following a positivistic orientation. In most cases, the instruments include global items and/or pre-ordinate standardized sets of items about teaching characteristics and dimensions. The administration of the evaluation forms is standardized. Findings of these surveys are commonly analyzed in a way that reduces the results to a rating or score. Then, results obtained from the evaluation are compared with those obtained by other faculty members or against a predetermined criterion or standard. When classroom observations are conducted, the tendency again is towards quantification.

Reviews of the research on the evaluation of teaching in higher education using a positivistic paradigm have been conducted with an emphasis on measurement issues, looking especially to the strengths and limitations of different evaluation sources and to issues of reliability, and internal or external validity of the evaluation ratings.

Evaluation Sources

In the literature on evaluating college teaching, different studies have been conducted to identify the advantages and limitations of different data sources in the evaluation of teaching in higher education: students, colleagues, administrators, teachers themselves, former students and external observers.

Students Ratings. The main source used in the evaluation of college teaching is current students (Seldin, 1993). Only a few studies have used other sources for evaluating teaching and even fewer have used more than one source simultaneously to assess quality of instruction. As Bukalski and Zirpola (1993) mention, students' appraisal of instruction is a prevailing "part of the academic life at almost all colleges and universities in the United States" (p. 23)

Findings reported in major literature reviews during the last twenty years support the reliability of the evaluation results obtained from student ratings of instruction (Marsh, 1987; Centra, 1993; Costin, Greenough, and Menges, 1971; McKiechie, 1979; Cashin, 1988, 1995). According to Marsh and Bailey (1993, student evaluations of teaching effectiveness are "multidimensional, reliable, stable and relatively valid against a variety of indicators of effective teaching." Marsh (1987) also states that the ratings are "relatively unaffected by multiple variables identified as potential biases to the evaluation, and are useful for faculty as feedback about their teaching" (p. 255). Other researchers, such as Ryan and Johnson, (1998), although sharing the belief that the research can be defensible from a logical and psychometric viewpoint, express concern when the evaluation is used for satisfying more than one purpose.

This concern is also shared by other researchers in the field of faculty evaluation, such as Hawley (1977), who says

If the purpose of the evaluation is to improve the quality of the instruction, faculty members will rightly feel sabotaged when the data are used also in making decisions about tenured salaries. In the first case, evaluation can be seen as helpful; in the second case, it takes an adversary tone. (p. 10)

In addition, although many researchers see the evaluation as potentially useful for improving the quality of teaching, this is an area where there is no conclusive evidence, because as Dunkin & Barnes (1986) said:

Much more research is needed to demonstrate ways in which... (the) evaluation can be put to effective use in improving teaching. In particular, research on the effects of feedback from student ratings upon change in teaching processes is needed.

Peer review. The phrase "peer review" in faculty evaluation traditionally refers to the evaluation of a professor for another faculty member or a panel of colleagues. Some authors, such as Wicks (1992) maintain that peer review is a flexible system capable of generating creditable judgments based on wide ranges of evidence. It is adaptable, can reflect and respect the traditions of those being evaluated, and can grow and change to satisfy new needs. In addition, Menges (1987) indicates that faculty members seem to report higher satisfaction, more interactions with other faculty members, increased motivation, and renewal when peer reviews are used, particularly for formative purposes. Braskamp, Brandenburg and Ory (1984) also state that colleagues can be a major source of information for faculty evaluation. They maintain that one of the advantages of colleagues as sources is that they have the necessary expertise in the discipline in which the faculty member is being evaluated.

Braskamp, Brandenburg and Ory believe that colleagues are in an excellent position to judge:

(a) instructor's knowledge and expertise in major field as reflected by the course syllabus and the reading list; (b) the instructor selection of realistic course objectives; (c) instructor assignments, group projects, and examinations; (d) student achievement as indicated by performance on exams and projects, and (e) instructor involvement on instructional research.

However, some studies have identified problems with using colleagues as a source of evaluation. One of the problems is that faculty do not study each other's classes systematically, so it is very difficult for them to perceive the instructional effectiveness of their colleagues (Andrews, 1985; Moffett, 1997). In addition, there is controversy about the reliability of colleagues as raters of college instruction. Some studies have reported that colleagues tend to give very high ratings to their peers (Andrews, 1985). In fact, various researchers such as Centra (1979) have found that colleagues tend to give higher ratings to their peers than do students. On the other hand, other studies have found that faculty members tend to give lower ratings to their colleagues when evaluating college instruction. Longman (1978) has reported that faculty members tend to perceive that their colleagues teaching is below average and that their peers need help in improving the quality of their instruction.

One of the limitations of peer review, according to the researchers with a positivistic orientation, is the lack of statistical reliability of peer ratings. Centra (1979) reports a correlation of only $r = 0.26$ among different colleagues who were evaluating the same instructor. However, this author considers that training faculty on observation techniques can help to improve the reliability of faculty colleague ratings. More research is also needed because the literature provides only a few examples of research-based activities where

instructional performance was completely and systematically evaluated by colleagues for formative or summative purposes (Menges, 1991).

Administrators as evaluators. Centra (1993) and Genova et al. (1976) have considered the use of administrators, such as department chairs for evaluating college instruction, although they agree that the use of this source is controversial. One of the main concerns about using administrators is that many administrators seem to base their evaluation on considerations other than instructional effectiveness. Feldman (1989) reports that administrators seem to base their evaluation on the reputation of the faculty, and their perceived participation in university activities, such as faculty committees. Most administrators do not visit the classrooms systematically, so they do not have many opportunities to appreciate the instructional quality of their faculty. In addition, when administrators participate in the process of evaluating faculty they play a dual role as decision-makers and suppliers of evaluative information. This dual role could interfere with the validity of their evaluation of faculty instruction (Genova, et al, 1976). More research is needed about the use and validity of this source that has been practically unstudied.

Self-evaluation. The ratings that each instructor gives to his own teaching, also called self-evaluation or self-assessment, have played a prominent part in many faculty and administrative evaluations (Miller, 1974). Dressel (1970) asserts that self-evaluation is essential to faculty improvement. Moreover, Washton (1988) and Dressel (1970) believe that this kind of evaluation can give individuals an opportunity to reflect about their work and to confront their professional weaknesses. In addition, Braskamp and Ory (1994) think that self-evaluation is the most important source of assessment because only faculty themselves

can provide information about "the thinking behind their work, career goals, strengths and weaknesses, plans for meeting perceived instructional goals, changes in their work based on assessment, their evaluation plans and their implementation" (p. 103). Moreover, the use of course and teaching portfolios for self-evaluation and reflection can be particularly useful for engaging faculty in both individual and institutional improvement (Braskamp and Ory, 1994).

One of the limitations of self-evaluation identified by the researchers with a positivistic orientation is in relation to the very modest relationship between student ratings and faculty ratings of the instruction (See Feldman's review of the literature, 1989). Least similarity has also been found between faculty self-ratings and those of their colleagues (Feldman, 1989). In general, the use of self-evaluation deserves more attention because the research on this evaluation source is practically unexamined.

Multiple sources. Feldman (1989) and Braskamp, Brandenburg and Ory (1984) suggest the use of records¹, alumni and external observers as additional sources for evaluating a faculty member's instructional effectiveness. Feldman (1989) synthesized the research by comparing overall ratings of college teachers' effectiveness that were collected from six different sources: (a) current students (b) former students, (c) colleagues, (d) administrators, (e) external observers, and (f) teachers themselves. After Feldman reviewed the similarities and differences among the ratings from the six sources, he found the highest relative similarities between: (a) current students and former students, (b) current students and colleagues, and between (c) current students and external observers. Feldman found that the

¹ According to Braskamp and Ory (1994), records includes "pieces of factual information, summaries, and faculty materials, such as listings of courses, committee assignments, advising loads, consultation activities, grant applications and rewards, clinical charts, progress reports and daily logs." (p. 226).

three lowest average correlations were: (a) teachers as self-raters and current students, (b) teachers as self-raters and colleagues and (c) teachers as self-raters and administrators.

Feldman believes that more research is needed because of the possibility that there could be an interaction between students and colleagues ratings. Colleagues may base their evaluations on rumors from students, the teacher's reputation, and perhaps even on "the teacher's own discussion with colleagues about his or her student evaluations." (p. 165). In addition, Feldman's work presents some methodological problems. He did not analyze all the sources at the same time. He only reviewed the statistical data of the studies included in his summary.

Reliability Studies

Reliability studies examine the extent to which the evaluation procedure measures teaching consistently for the purposes it serves (Seldin, 1984). Studies on the reliability of student ratings have focused mainly on item analysis (internal consistency) and agreement over time (stability). Internal consistency refers to the degree there is an agreement among students within a class rating their instructor and course. Stability refers to the extent in which a single instructor and course are rated similarly by the same students at two different times (Braskamp and Ory, 1994).

The findings of the research on the reliability of the ratings are summarized by Braskamp and Ory (1994):

1. Student agreement on global ratings of the instruction is high if the class has more than fifteen students (Crooks and Kane, 1981; Feldman, 1977, 1978; Marsh and Overall, 1981; Marsh, Overall, and Kesler, 1979).
2. Students are consistent in their global ratings of the same instructor at different times in the course (Centra, 1977).

3. The overall instructor teaching effectiveness can be reliably generalized from student ratings from five or more classes taught by the instructor when the class has at least fifteen students (Crooks and Kane, 1981).
4. Similar global ratings from each section are obtained by instructors who teach different sections of the same course (Overall and Marsh, 1979; Shingles, 1977).

According to Cashin and Perrin (1978) in internal consistency studies, the average item reliabilities tend to increase as the number of raters increases. Cautionary statements, however, have come from Marsh (1984) who states that while internal consistency of the items tends to be high, we need to be careful in our interpretations because the research on internal consistency provide "an inflated estimate of reliability due to the fact that it ignores the substantial portion of error resulting from lack of agreement among the different students." (p. 716)

Studies on the stability of student ratings have focused on the agreement of the ratings over time. Findings of these studies are reported from Overall and Marsh (1980) who found a high correlation of 0.83 when comparing the ratings of the same students at the end of the course and several years later. In general, Murray and others (1990) who summarize the findings of the research on the reliability of the ratings, state "Although findings are sometimes contradictory, the weight of the evidence suggests that student ratings of a given instructor are stable across items, raters, and time periods" (p. 250).

Validity and Evaluations of College Teaching

Validity is concerned with the questions: Are we measuring what we think we measure? Are our inferences and actions about the evaluand² supported by evidence? Because validity is linked to the meaning, value and the appropriateness of interpretation, validity is the most critical consideration in evaluation. Evaluations are not valid when they have "a weak representation of value" (Stake, 1996, p. 1). To be valid, evaluations must represent the quality of the evaluation object (Stake, 1999) and "take into account all relevant factors, given the whole context of the evaluation..." (Scriven, 1991, p. 373).

Since evaluation is "a search for goodness and badness, for merit and shortcoming, for quality" (Stake, 1999, p. 1), a valid evaluation of teaching has to represent the quality of the teaching within its context and complexity and "to convey the sense of quality to others" (Stake, 1999, p. 1). The evaluation will not be valid if it ends with an assessment of teaching that lacks relevance and utility, and its value implications and social consequences are not addressed because as Messick (1989; 1995) argues, meaning and consequences are essential to validity.

² Evaluation object.

Studies on the validity of the evaluation of teaching in higher education have centered on the study of the validity of student ratings. According to Ory and Ryan (2001) five approaches have been used on studies validating students evaluations of teaching in higher education: (a) multi-section; (b) multi-trait; (c) "bias"; (d) laboratory designs and (e) dimensionality of the ratings. Among the five, multi-section and multi-trait are the most used approaches to determine the validity of student ratings of instruction.

Multi-section Studies. In these studies, multiple sections of the same college course taught by different instructor are examined. Then, a measure of student ratings is correlated with a measure of student achievement. Then, the researchers calculate a correlation between "the section mean of student ratings with the section mean of student achievement scores on a common examination" (Ory and Ryan, 2001, p. 6). The higher the correlation between both measures, the higher the validity of the ratings (El-Hassan, 1995).

Some researchers such as Marsh (1984), who have conducted research on the generalizability of student ratings across sections, found a correlation of $r = 0.61$ for the same instructor teaching different courses. He also found a higher correlation ($r = 0.72$) when the same instructor teaches different sections of the same course. Marsh concluded that the instructor and not the course was the main determinant of student rating results. Other researchers such as Gillmore, Kane and Naccarato (1978) and Hogan (1973) have supported Marsh's findings. In addition, Cohen (1981, 1983) found a stronger correlation in multi-section studies ($r = 0.38$) and used these results to support his argument for the validity of student ratings as measures of instructional quality.

Unpersuaded, Dowell and Neal (1982) found that while the multi-section studies support the validity of student ratings, “this validity is not consistent across various situations” (p. 59). The evidence in Dowell's study suggested that validity of student ratings “is modest at best (maximum $r= 0.26$) and quite variable” when comparing the ratings across different sections of a course (p. 59). El-Hassan (1995) believes that part of the differences among these studies may be attributed to the differences in methods and controls used by the researchers. A review of several dozen multi-section studies conducted by Abrami, D' Apollonia, and Cohen (1990) showed that although consistent, more research needs to be conducted to understand the limits on generalizability of rating validity across rating dimensions, effectiveness criteria, and conditions of instruction. As is well known, correlational research findings need to be taken carefully because group homogeneity and other factors vary so. In addition, many of these studies were conducted in low learning, introductory courses taught primarily to freshmen and sophomores.

Multi-trait Studies. In these studies, multiple traits are evaluated using multiple methods such as student ratings, peer evaluation, observations, etc. Although results of these studies have "shown evidence for both discriminant³ and convergent validity⁴" (Howards, Conway, and Maxwell, 1985; Marsh, 1982), these studies have been mostly meta-analyses which have compared the findings of different studies. These meta-analyses have ignored the complexities of the context within which teaching is evaluated as well as the differences

³ Discriminant validity addresses the question: Are ratings influenced by variables unrelated to teaching?

⁴ Convergent validity addresses the question: How well are ratings measures correlated with other indicators of effective teaching?

among the type of institutions, time periods of data collection, and types of instruments used.

Moreover, the data of these studies may not be equivalent for comparison or may be too limited in number for comparison. As Feldman (1989) mentions, " for the... comparison pairs (used in the meta-analyses), many had either no studies with data pertinent to them or only one or two studies" (p. 167). Abrami, D' Apollonia and Cohen (1990) who reviewed several multi-trait studies conclude that these "designs provide weak evidence for the validity of student ratings as measures for instructional effectiveness" (p. 221).

"Bias Studies". These studies are conducted with the intention of identifying "extraneous influences on student ratings" (Ory and Ryan, 2001). In these studies student ratings are correlated with other variables that may influence the ratings, such as characteristics of the instructor (i.e. age, sex, teaching experience, personality), students (i.e. age, sex, level of the student) and the course (i.e. class size, time of the day, etc.). A significant number of studies on the biasing factors affecting students ratings of teaching have been conducted over the past twenty years. According to El-Hassan (1995), some of the factors that seem to differently influence student ratings of instruction are:

1. Faculty rank (i.e., junior versus senior), Braskamp, Brandenburg and Ory (1984)
2. Student motivation (Feldman, 1978; Marsh, 1984)
3. Expected grades (Feldman, 1976; Marsh, 1984; Greenwald and Gillmore, 1976)
4. Course type (elective versus required) (Aleamoni and Hexner, 1980; Braskamp and Ory, 1994)
5. Academic discipline (Feldman 1978; Marsh, 1984; Miller, 1987; Braskamp and Ory, 1994; Cashin, 1990), and
6. Workload/difficulty (Marsh, 1984, Greenwald and Guilmore, 1996).

Braskamp and Ory (1994) review of the literature on biasing factors influencing student ratings found similar relationships between the ratings and some biasing factors like those mentioned by El-Hassan. But, Braskamp and Ory were more attentive to the elective/required nature of the course as a "biasing factor" than other variables such as grades or academic discipline. Ory and Ryan (2001) added:

At the present time, the elective/required nature of a course is the only context variable that we account for in our student rating system. However, we are considering making some changes based on some old and new research revealing differences in the ratings collected in courses of different disciplines.

Ory and Ryan (2001) also state that there is less certainty about the possible influence of other variables in student ratings, such as student grade expectations and the gender of the instructor. This is in part because of the mixed results obtained in studies focused on the influence of these variables in student ratings.

The research on the influence of grades on student ratings has found evidence about the possible relationship between student ratings and grade inflation. While some researchers such as Feldman (1976) and Marsh (1984) affirm that the relationship between these variables is generally small (around .20), recent research by Brodie (1998) found that prior studies on the relationship between grade inflation and student ratings "have underestimated the biasing effect of grading leniency" (p. 17). After analyzing 1,939 student evaluations from 75 classes, Brodie found that "when grades varied markedly across sections of the same course, the professor assigning highest grades with least studying received highest evaluations" (p. 17).

The research on gender differences and their influence on student perceptions of teaching has produced some controversial results (Goodwin and Stevens, 1993). The studies have reported mixed evidence about the influence of gender in student ratings of the instruction. Most of the studies have focused on three main aspects: (a) teacher gender and its influence on instructional effectiveness, (b) influence of student's gender on the ratings, and (c) teacher gender and perceptions of good teaching.

Following the first aspect, Feldman (1993) found that while many studies have reported no significant differences between male and female faculty evaluations due to gender differences, other have found that female professors received somehow higher ratings than males on a variety of specific or global aspects. Felman's findings indicate that students tend to give higher scores to their female teachers in the aspects:

1. the nature and value of supplementary teaching aids,
2. pursuing and/or meeting the course objectives,
3. teacher sensitivity and concern with class level and progress,
4. teacher availability and helpfulness,
5. teacher encouragement of student questions,
6. teacher openness to other's opinions, and
7. use of more varied and valuable course material.

According to Goodwin and Stevens (1993) female professors tended to receive higher student ratings in traditionally female disciplines (nursing, education) compared to female professors in traditionally male disciplines (engineering and science). On the other hand, male professors seem to receive higher student rates in the dimensions: (a) teacher's clarity and

understandability, (b) teacher knowledge of subject matter and (c) personality (Feldman, 1993; Goodwin and Stevens, 1993).

Another group of studies found that the gender of the students could influence the way that they rate their instructors. The findings of Goodwin & Stevens, (1993) study indicate that while male professors seem to be rated similarly by male and female students, female students seem to rate higher their female professors than their male professors. In addition to this, Goodwin & Stevens found that the gender of the instructors could influence their perceptions of good teaching and the activities that they implement in the classroom. These authors found a few gender differences between men and women instructors when they were questioned about what is the meaning of good teaching. According to their findings both men and female women said that good teaching is concerned with higher-order thinking skills. But, women were also concerned about the self-esteem of the students, teacher-student interactions via small group activities, and the development of a variety of learning levels via exams and discussions. In addition, female teachers seemed to use a combination of audiovisual aids and were more open to consult with their colleagues about how to improve their teaching. Men differed from women in that they said to place greater value on student evaluations than did women. According to the authors, the preference of female teacher for active methods of instruction could influence negatively their evaluation ratings. Some studies have found that there is evidence of a relationship between lower student ratings to their teachers in classes with higher levels of student. One reason for these findings can be the presence of gender stereotypes and performance evaluation bias. Another reason can be that some researchers have analyzed student rating data at the item level whereas others have

looked at differences at the global level (Feldman, 1993). According to Feldman (1993) More research is needed on the area of gender and communication styles between male and female instructors.

There are also mixed findings in relation to the effects of class size on student ratings. For example, Feldman (1978; 1984) found that while some studies discovered no relationship between class size and ratings, other studies showed some evidence that students tend to give low ratings to their teachers' effectiveness in large classes. Feldman's findings (1984) indicated that an average correlation between student ratings and class size was of $r = -0.9$ (52 studies). Recent studies by Greenwald and Guilmore (1996) clarify the influence of class size on the ratings but more research is needed in this area to better understand the complex relationship between class size and teaching and learning. In general, educators are concerned with the influence of class size, especially in cases when evaluation results are used for making administrative decisions. (McKeachie, 1997).

Other studies have found inter-relations among different variables influencing student ratings. For example, a correlation seems to take place between the course discipline and the age of the instructor. These studies found that when students evaluate their college teachers, they tend to give higher ratings to some academic disciplines than to others (Barnes and Barnes, 1993, Biglan, 1973; Newman and Newman, 1983, 1985; Cashin, 1990; Feldman, 1978, and Marsh, 1984). According to these authors, professors of humanities receive a higher score than do faculty of social sciences or physical sciences, mathematics and engineering. This relationship may also be influenced by the age of the instructor, with elders in social sciences rated higher. Some authors have reported a significant but small impact of

the instructor's age on student evaluations that may influence the relationship between ratings and academic discipline. Centra (1979) reports that instructors with over twenty years of experience tend to receive lower ratings than instructors with up to twelve years of experience. Kinney and Smith (1992) found that students tend to give higher rates to professors in the social sciences and humanities as they approach the mandatory retirement age. At the same time, these authors found that students were giving lower ratings to professors in the physical and natural sciences as the instructor advanced toward retirement age (Kinney and Smith, 1992). Miller (1987) commented that more than age, burnout, boredom, physical and diminished energy, may affect views of teaching competence.

Laboratory Studies. As Ory and Ryan (2001) mention, these studies "examine the relationship between student ratings and experimenter-controlled variables in a non naturalistic setting, e.g., videotaped lessons, lab-delivered lectures." (p. 7). Findings of the research using laboratory designs have shown that these studies are not adequate for studying how well the instructor influences student learning in the actual classroom. In addition, as Abrami, D' Apollonia, and Cohen, 1990, mention, this research approach "lacks comprehensiveness because it fails to represent the many instructor characteristics that could affect validity... and the actual differences among instructors in the field." (p. 222)

Dimensionality Studies. Kulik and McKeachie 1975), Feldman (1987) are among the researchers who have tried to identify the underlying construct measured by the evaluation. As mentioned before, these studies focused on identifying the "conceptual structure of the ratings" (Ory and Ryan, 2001, p. 8). Using meta-analyses, the researchers have tried and failed to identify a "common set of factors underlying the construct being measured by

student ratings of instruction." (Ory and Ryan, 2001, p. 8). Although there is consistency across different studies, the researchers have not been able to identify a single set of characteristics and behaviors, as those essential for defining the construct teaching quality. As Ory and Ryan state, "Without a clearly defined target domain of effective instructional characteristics", it is unclear how institutions select the content of their evaluation forms, and more importantly, what do these institutions infer as the meaning of their ratings." (Ory and Ryan, 2001, p. 11).

New Validity Framework and Current Evaluations of Teaching

In the late 80's and early 90's, a shift took place in the assessment literature that resulted in a new conceptualization of validity. Samuel Messick was responsible for this shift with his famous chapter on validity (1989), followed by Shepard (1993), and other authors, such as Lane, Park, and Stone, (1998); Moss, (1992, 1998); Reckase, (1998); Yen, (1998); Cronbach, (1989), and Moss (1992, 1996).

The new framework moves away from a fragmented to a unified concept of validity. Under this new framework, all validity is about construct validity. As Messick states, the new framework "integrates considerations of content, criteria, and consequences into a construct framework for the empirical testing of rational hypotheses about score meaning and theoretically relevant relationships including those of an applied and a scientific nature" (Messick, 1995, p. 751).

In addition, validity is not a property of a test but "an overall judgment of the extent of which empirical evidence and theory support the adequacy and appropriateness of the

interpretations based on the assessment" (Messick, 1995, p. 741). Moreover, validity refers not only to meanings and interpretation of assessment scores, but also to the inferences and social consequences that result from the evaluation. Indeed, meaning and consequences are essential to validity (Messick, 1989, 1995).

Aspects of Construct Validity

Messick (1989, 1995) identified six important aspects of construct validity to be used for all educational assessments to identify sources of invalidity: construct, substantive, structural, external, generalizability, and consequential. The issues and sources of evidence emphasized by each of the aspects are⁵:

1. Content aspect: Includes evidence of content relevance, representativeness, and technical quality (Lennon, 1956; Messick, 1989)
2. Substantive aspect: Refers to theoretical rationales for the observed consistencies in test responses, including process models of task performance (Ebreton, 1983), along with empirical evidence that the theoretical processes are actually engaged by respondents in the assessment tasks.
3. Structural aspect: Appraises the fidelity of the scoring structure to the structure of the construct domain at issue (Loevinger, 1957)
4. External aspect: Includes convergent and discriminant evidence from multitrait-method comparisons (Campbell & Fiske, 1959), as well as evidence of criterion relevance and applied utility (Crombach & Glesser, 1965)
5. Generalizability aspect: Examines the extent to which score properties and interpretations generalize to and across population groups, and tasks (Cook & Campbell, 1979; Shulman, 1970), including validity generalization of test-criterion relationships (Hunter, Schmidt, & Jackson, 1982).

⁵ Messick, 1994, p. 11-12.

6. Consequential aspect: Appraises the value implications of score interpretation as a basis for action as well as the actual potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice (Messick, 1980, 1989)

In their analysis of how current evaluations of teaching in higher education match the new framework of validity, Ory and Ryan (2001) found that some research has been conducted on some aspects of validity but that other important aspects have not been addressed by the research.

Content Validity. One of the most important aspects of validity is the capacity of the evaluation to reflect the content of the construct that it is intended to measure. This aspect of validity addresses the question: Is there a relationship between the content of the evaluation and the construct intended to be measured?

Two main sources of invalidity can be associated with content aspects: construct under-representation and construct-irrelevant validity. Construct under-representation takes place when the assessment is too narrow in representing the construct being measured. Construct-irrelevant variance takes place when the assessment includes elements that are irrelevant to those of the construct being measured.

In the evaluation of teaching in higher education, content validity refers to the capacity of the evaluation to measure teaching quality. Consequently, there is construct under-representation when the evaluation is not broad enough broad to measure all the components of good teaching. There is construct irrelevant variance when the assessment includes variables other than teaching quality.

According to Ory and Ryan (2001), because most evaluation forms “are developed without too much thought of theory or construct domains” (p. 11), there is a raising concern about the validity of the interpretations based on evaluation scores.

In addition, the use of standardized procedures for evaluating college raises issues of construct under-representation if the assessment fails in representing the construct been measured either because it lacks important elements of the construct or because the measurement includes variables non relevant to this construct (Stake, 1999).

In addition, validity problems increase because current evaluations of teaching remain focus on narrow definitions of teaching that are not consistent with current theories of teaching and learning. Indeed, while the research on teaching has evolved from simplistic to more complex conceptualizations, these changes have not taken place in the evaluation.

Unless there is certainty about the meaning of the scores, it can not be said for certain that the evaluation results are valid representations of instructional quality. As Ory and Ryan (2001) state, construct under-representation will occur if the scores are interpreted as good teaching when in fact they have failed to include all the important elements of the construct. Having different conditions of learning taking place in different classrooms and implementing course content in different ways can also result in a threat to validity if as a result of this, a sub-group of instructors is given an unfair advantage in the evaluation.

It is also important to know if the assessment is encouraging a particular type of teaching and learning and also if the assessment result in punishing alternative approaches that stress non-traditional views of teaching and learning.

Substantive Validity. For this aspect of construct validity it is important to analyze response processes of those taking the test and completing the evaluation forms in order to see if there is a fit between the process used to answer and the process for which the assessment was developed. Evidence of substantial validity can be found when there is a fit between what is been tested and the construct measured. As Ory and Ryan (2001) illustrate, "When an examinee uses critical thinking to answer items on a test of critical thinking there is evidence for the substantial validity of the test scores." (p. 14).

Studies on the substantive validity aspects of construct validity focus on questions such as: What accounts for score differences? What do we know about the response processes in different situations? If students respond more positively in a given situation, are they responding more or less truthfully? Does the nature of the evaluation process match the construct being measured?

It is not enough to know that the scores change in different situations, it is necessary to know why the change takes place. We also need to understand how students use the rating scales to respond, and if there is a fit between the intended meaning of the scale and the meaning of the scale for students. In addition, it is important to determine if all students follow similar processes when responding to the tests. Do some subgroups of students respond differently than others? Is the assessment appropriate for different groups of students of diverse ethnic and cultural backgrounds?

According to Ory and Ryan, several studies (Marlin, 1987; Dwinell and Higben, 1993; Ballantyne, 1998) have been conducted about student attitudes about the evaluation, specifically towards the student ratings. But, little is still known about "the actual process

followed by students to respond to rating forms.” (Ory and Ryan, p. 26) According to these authors “past research efforts have indicated how ratings change in different situations but they do little to help us understand why the change occurs” (p. 15). In addition, more research is needed to understand how students use the rating scales to respond. As Ory and Ryan (2001) state, “If items are presented with a five point Likert scale, how do students interpret and use the middle category? Do students mark a “3” to indicate an inability to respond, a middle response, or a lack of interest? If only the endpoints are labeled how do students interpret and use the other scale points? Are some students more reluctant than others to use the extreme ends of the scale? Do some students believe that a “perfect six” is unobtainable? To make valid inferences from student ratings we need to determine if there is proper fit between what the meaning of the scale was for students and the intended meaning of the scale.” (p. 15)

The problem with standardized evaluations of teaching on campus, as identified by critical scholars, is that the scores do not necessarily reflect real differences among people, and they often do not adequately eliminate underlying biased cultural assumptions built into the test as a whole.

There is a need for conducting research on the substantive validity of the evaluation of teaching in higher education. Examining the substantive aspects of construct validity is important because “the response consistencies or performance regularities are reflective of the domain processes.” (Messick, 1994, p. 13). The interpretation and use of evaluation results can be improved if the student response pattern and the differences in response patterns among different students are understood.

Structural Aspect of Validity. This aspect of construct validity stresses that "the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria and rubrics." (Messick, 1994, p. 15). This aspect of validity addresses the question: To what extent do the relationship among different components of the evaluation procedures correspond with the construct being evaluated? In this way, the evaluation needs to provide evidence that the relationship among the different components of the assessment instrument correspond with the structure of the construct domain. It is also important to know how well the scoring structure is consistent with the construct domain.

As mentioned before, a number of studies have been conducted in order to determine the characteristics or behaviors that constitute good teaching (See Table 1). Many instruments for evaluating teaching are based on those characteristics. Some researchers have found correlations between these and other sets of characteristics and behaviors and the ratings of instruction. For example, Centra (1993) and Feldman (1976) found common dimensions after analyzing several evaluation forms. As Ory and Ryan said,

items are included on many forms because students appear to respond similarly to particular ones not because they come from a known domain of targeted characteristics. It is somewhat like analyzing student responses to hundreds of math items, grouping the items into response-based clusters, and then identifying the clusters as essential skills necessary to solve math problems. (p. 18)

There is no empirical evidence that the items selected are indeed elements of good teaching.

External Aspect of Validity. This aspect of validity analyzes the relationship of the evaluation to other variables, external to the assessment in order to provide source of validity evidence. In this way, "the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning" (Messick, 1994, p. 16). In addition, "special importance among external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings" (Messick, 1994, p. 17).

As mentioned before, prior research on the validity of student ratings has been conducted to address this important aspect of validity. Some of these studies have been conducted to determine if there is a relationship between student ratings and student achievement.⁶ The multisection studies mentioned early are an example of this kind of research that determines the validity of the evaluation by analyzing the correlation of evaluation results of a single course that is taught by different instructors with "the section mean student achievement" (Ory and Ryan, 2001). In addition to Cohen correlation studies (1981) on the relationship between student ratings and student achievement, other researchers such as Murray (1983) who used trained observers to determine teaching differences between instructors who obtained high and low ratings.

Other studies have compared the results obtained from student ratings with other data sources, such as peers, alumni, self-ratings, etc. Researchers have studied different evaluation

⁶ Defined as grades.

sources to determine the consistency among different data sources in evaluating teaching. Researchers have found high positive correlations between student ratings and alumni ratings.

In addition, another group of studies have studied the correlation between different forms of data collection, such as “student overall ratings of instructor competence as measured rating items, written comments to open-ended items, and group interviews (Ory, Braskamp, and Pieper, 1980; Ory and Ryan, 2001).

Generalizability Aspect of Validity. This aspect of validity examines if there is correlation "of the assessed tasks with other tasks representing the construct or aspects of the construct" (Messick, 1994, p. 15). Generalizability refers to the "boundaries of score meaning" (Messick, 194, 15).

Generalizability as an aspect of construct validity addresses questions such as, Can we make comparable inferences about the meaning of the scores across subjects, settings, and time. Can we make the same inference about ratings collected in different settings? Can we make valid comparisons between scores used for one purpose versus another purpose? Are assessment scores collected in different settings comparable? Generalizability studies focus on determining differences, understanding why they occur, and learning how to account for them in reporting assessment results to enhance the validity of the assessment process.

Although some researchers support the generalizability of student ratings across different sections, other researchers such as Abrami, d’Apollonia, and Cohen (1990) have questioned the generalizability of the evaluations of teaching using student ratings “because so many of the studies were conducted in lower-learning, introductory courses taught

primarily to freshmen and sophomores” (Ory and Ryan, 2001, p. 20). More research is needed on this important aspect of construct validity

Consequential Validity. This aspect of validity refers to the short and long-term consequences of evaluation use and the consequences associated with interpretations of evaluation scores (Wilson, 1999). Both intended and unintended consequences are important. Consequential validity also implies the need for appraising the value implications of the theory underlying evaluation scores, as well as the ideology in which the theory is embedded. When collecting evidence about the consequential aspect of validity, especial emphasis is given to consequences "associated with bias in scoring and interpretation or with unfairness in test use" (Messick, 1994, 17).

According to Ory and Ryan (2001), the consequential validity of the evaluation of teaching in higher education is an area that has received little attention by the researchers. More research is needed on the value implications of the evaluation results, the intended and unintended consequences of using certain criteria for defining and assessing good teaching, “the ideology within which the theory is imbedded, (and) the potential or actual problems that could result for the institution as a result of the consequences” (p. 26).

Limitations of the Positivistic Paradigm

A positivistic orientation to the evaluation of teaching has the benefit of organization and simplicity. But many see the limitations and problems surpassing its benefits. As with standardized student assessment, this orientation could result in some serious negative consequences.

First, exclusive use of characteristics or behavioral attributes is “limiting to a certain kind of knowledge about teaching and learning.” (Dunkin and Barnes, 1986, p. 774). The evaluation usually centers on a kind of teaching that is teacher-centered. In other words, a kind of teaching in which the instructor’s task “is to cover a well defined set of topics for a course systematically and precisely, while the student’s task is to master the course content through traditional assignments and study methods,” (Centra & Boneshell, 1990). Instructors using a different teaching approach or style may be at a disadvantage. There can be a mismatch between the evaluation and any teaching consistent with constructivist learning theory, as well as with theories of human and cognitive development (Mabry, 1999).

The use of a list of characteristics and behaviors from the process-product research, although correlated with student performance in exams or tests, raises validity questions. Their validity has not been determined empirically. These behaviors and characteristics are the outcome of synthesis from aggregate data, but there is “little evidence that any observed teacher had ever performed in the classroom congruent with the collective pattern of the composite.” (Shulman, p. 12). The use of characteristics, styles, traits and behaviors identified from surveys to faculty and students also presents a problem. There is little empirical evidence that any of them constitute good teaching, or that they are related to student learning (Miller, 1974; Genova et al, 1986).

There are problems when not all relevant characteristics and behaviors used as criteria are included in the assessment, problems of content validity. As said earlier, using a general and narrow definition of teaching is problematic because it may not be appropriate for all teaching situations (Stake and Cisneros-Cohernour, 2000). Doyle (1982) says, "... it seems

most unlikely that any one set of characteristics will apply with equal force to teaching all kinds of materials to all kinds of students under all kinds of circumstances... To prepare such a list entails a substantial risk" (p. 27). When a number of characteristics are adopted as indicators of teaching quality, their prescribed use can result in limiting instructional creativity, and can become a barrier for professional development. The use of traits and/or teaching styles as criteria for evaluating teaching constrain diversity in instructors, penalizing those who do not "fall within the norm" (Stake and Cisneros-Cohernour, 2000).

The tendency to summarize teaching quality in a numerical index can lead to the unintended consequence of people focusing more on improving the scores than on improving their teaching (Cisneros-Cohernour, 1997). Comparisons made without a rigorous control of variables influencing the teaching and learning process, can lead to unfairness (Stake and Cisneros-Cohernour, 2000).

It is important to review the claims of objectivity made by those conducting research on the evaluation of teaching. In the U.S., most of those conducting research on the validity of the evaluations of teaching in higher education have a dual role, as scholars and as those who develop and implement the evaluation. Their scholarly work directly or indirectly more often than not supports the validity of their work as administrators in the institution.

The publications of teaching research in the US contain few studies that contradict the main findings of this research community. An exception is the work of Brodie, a Canadian researcher, who in 1988 found that prior studies on the relationship between grade inflation and student ratings "have underestimated the biasing effect of grading leniency" (p. 17). In that study, Brodie found evidence that "when grades varied markedly across sections of the

same course, the professors assigning highest grades with less studying received highest evaluations" (p. 17). In addition, Brodie's (1999) review of the research on the correlation between student evaluations of teaching and student learning raised new issues. He found discrepancy between the results in research reports and the published articles of the same study. He encountered evidence that correlations between certain teaching characteristics and the ratings as reported in several journals have been inflated and that "some researchers have deleted low and/or negative correlations, but also created positive correlations by reversing the rating scale" (p. 1). Although no research has been conducted to confirm the findings of Brodie's research, or about the influence of the dual role of the researcher as scholar and as administrator of the evaluation system, these important questions deserve more attention.

Other critics of the positivistic paradigm perceive that the emphasis put on measurement by the researchers is so strong that it has replaced "the concept with the formula, and causation with rule and probability." (Horkheimer and Adorno (1948, p.11).

Magunsson (2000), in her discussion of the appropriateness and validity of the evaluation for assessing the quality of teaching of instructors of diverse cultural or ethnical background, adds:

The problem with an analysis that equates 'minority' with small systemic variance, or other such measurement concepts, is that it constructs the issue once again within the technical discourse of psychometrics. The problem is that if there is racism, this is systemic to the entire organization and can't be reflected merely as systemic variance related to measurement.

Menges (1998), also claims that more research is needed with still little known about how the information from the evaluation is interpreted and used, and about how teachers "use the evaluation in planning, implementing, and appraising their own teaching." (p. 3) He adds

that the main shortcoming of the research is "its lack of recognition of the context of teaching ... (ignoring) the perspectives of different participants, and their personal, organizational, and political contexts" (p. 4).

What is promising is that among the researchers supporting the positivistic paradigm is a growing interest for testing the assumptions held by the research, and for examining the validity of the construct being evaluated. (Menges (1998); Theall & Franklin, (1990, 2000); Ryan & Johnson (1998); and Ory and Ryan (2001).

Summary

The evaluation of teaching in higher education has evolved from informal to systematic approaches as pressures for accountability increased in this level of education. In addition, as administrators began to worry about measuring outcomes, concerns have increased among the faculty about the fairness and use of evaluation results for making administrative decisions, such as tenure, promotion and salary increases.

As the research on teaching and learning have evolved to more complex understanding of these the teaching and learning processes, new questions are raised about the validity of the traditional positivistic approach for evaluating teaching in higher education. Under this positivistic approach, teaching is linked to the idea of treatment and evaluation to the idea of effectiveness. Good teaching is defined as a set of ideal characteristics or behaviors expected from the instructor. Although, some researchers define teaching as a global construct as opposite to a set of characteristics, behaviors or dimensions. Studies under the positivistic approach have centered on the reliability and stability of student ratings of instruction, as

well as on the study of several variables that could negatively influence the evaluation. Other studies have been conducted about the strengths and limitations of different evaluative sources (i.e. student ratings, peers, external observers, administrators, self-evaluation, etc.), and on the relationship between student ratings and some variables, such as student achievement. Reliability studies have been conducted about the internal consistency of the evaluation forms (item analysis), and stability of the ratings over time.

Supporters of the positivistic approach for evaluating teaching claim that student ratings of instruction are reliable sources for evaluating teaching. The researchers have found correlations between student ratings and student achievement,⁷ and stability of the ratings when comparing the ratings of the same instructor in different sections. Multi-trait studies have also found some evidence of discriminant and convergent validity of the ratings. In addition, researchers studying the possible variables that could negatively influence the ratings have found evidence of some biasing influence, primarily by course type (required versus elective) and course discipline as biasing factors influencing the ratings.

Critics of the positivistic approach state that the strong emphasis on generalization and the establishment of causal and effect linkages have been overstressed by the research. Studies on the dimensionality of the student ratings have failed to identify the essential elements of the construct “good teaching.” The over reliance on meta analyses of students on

⁷ Defined as grades.

student ratings of instruction has also been questioned, especially when evaluation data is used for making administrative decisions that can affect faculty careers. In addition, prior research on the validity of the evaluations of teaching in higher education has been conducted using a traditional approach to validity.

But the main limitation of the research is about the validity of the evaluation in representing the construct been evaluated. Although some research on the validity of student ratings have been conducted on the generalizability of the ratings and their external validity, No research has been conducted on the conceptual, substantive and consequential validity aspects of the evaluation. There is also a need for understanding if the evaluation fairly represents the quality of teaching within its context, and how decision makers and teachers use evaluation results for professional development and for making administrative decisions.