A Fractal Thinker Looks at Student Ratings

© 1994-2010 Edward B. Nuhfer – Director of Faculty Development, California State University at Channel Islands

"It is the true believer's ability to shut his eyes and stop his ears to facts which in his own mind deserve never to be seen nor heard, which is the source of his unequalled fortitude and consistency."* Eric Hoffer, 1942, *The True Believer

Abstract. Fractal patterns occur in neural networks and synaptic connections. Many aspects of higher education: teaching, learning, thinking, and evaluation of all three involve fractal neural networks that manifest fractal traits. Neural networks developed during acquisition of competency in teaching are comparable to the neural networks developed during acquisition of personality types, learning styles and multiple intelligences. A trait of all fractal forms is that their characterization requires multiple measures. Distinction of personality types etc., by necessity, require multiple measures, and the tools used for diagnoses, without exception, require spectra of diagnostic responses, because types are impossible to diagnose by single response items. Likewise, summative student ratings alone cannot define "good teaching," and the employment of students' responses to single questions to characterize "good teachers" represents a single-measure attempt to characterize a fractal system. To those who understand fractals, it is obvious why such practices are flawed and doomed to failure. However, less obvious is the fact that such evaluations are not simply inept, but also destructive. All learning, including the learning of teaching competency, involves both the cognitive and affective domains of the brain. Meta-analyses have already defined self-esteem and enthusiasm as essential affective attributes of successful teachers. When student satisfaction ratings become used tyrannically, the result is an attack on the positive affective attributes of teachers' minds. The result is counterproductive to teaching, learning, and even to the personal core of the individuals caught in such hapless situations. Fear is perhaps one of the greatest destroyers of faculty collegiality, creativity, and productivity. Although the students' voice must be heard and is an essential part of faculty evaluation, using ratings to the degree that they effectively vest students with the power to retain and fire faculty is a way to make faculty afraid of their own students. It's unconscionable, and no administrator who cares about educational quality will ever allow ratings to be used to that degree.
Summative ratings result from a mix of cognitive and affective factors. Correlations reported between ratings and affective first impressions (thin slices) are even higher than those reported between ratings and learning performance. Thin-slices research confirms powerful, affective influence on ratings. Evaluative ratings are certainly honest expressions of student satisfaction and therefore represent one form of valuable information--one measure that becomes useful when informed by equally important multiple measures. However, students, unlike "customers," have responsibilities that go beyond paying for a product, so such satisfaction is not equivalent to "customer satisfaction." Affective influence is only slightly weaker on formative ratings, which use diagnostic items to deduce the pedagogical practices present in a class and the degree to which each is present. There is strong research evidence for benefits of particular instructional practices on both student learning and student satisfaction. This provides reasonable basis for including, as one essential multiple measure of a faculty member, a formative profile of pedagogical practices. Still, affect exerts strong influence even on items that would appear to solicit only an objective response.

Summative rating of faculty by college students is an evaluative challenge—at the highest level of challenge in Bloom's 1956 taxonomy. The ability of students to do evaluative thinking rests upon their ability to use evidence and to meet a high Bloom-level challenge with a high-level thinking response on the Perry (1999) scale or one another of the related research-based taxonomies of thinking, all of which seem to be generally variations that map well to Perry's model. Students' ability to handle well the evaluative challenge in the special case of rating professors is no different from that to handle other evaluative challenges. Research confirms that this ability in undergraduates (King and Kitchener, 1994) is generally marginal.

Knowledge surveys (see Case stories at http://elixr.merlot.org/) are a special kind of student ratings instrument. They constitute a direct and detailed portrayal of content learning from the viewpoint of students. They too are strongly influenced by affect, but more so by the students' feelings of their confidence to meet specific cognitive challenges addressed by particular knowledge survey items rather

than by their general satisfaction with their professors. Data yielded from knowledge surveys proves to be reliable and useful as an assessment tool. Instead of reliance on summative satisfaction, I recommend that "student ratings" of faculty proficiency should include summative ratings, formative profiles and knowledge surveys. This gives far greater voice to students, and provides revealing information about content being taught and pedagogy being employed. Although multiple measures of evaluative input from students are superior to summative surveys alone, even these together do not constitute a thorough review required for career decisions about faculty rewards and retention.

Preface

This review paper began as a reference for Boot Camp for Profs® and receives periodic updates. It evolved greatly between 1994 and 2005, as we became aware of the challenge provided by the fractal qualities of what we were trying to measure. This version has two parts. Part I is a general description of student ratings and the results of key studies, as well as a presentation of arguments based on evidence. Part II addresses student ratings through the fractal perspective of understanding education. I was surprised when a colleague noted this paper had been cited by "The New York Times" in an article March 11, 2010. There, journalist Virginia Heffernan at <http://www.nytimes.com/2010/03/14/magazine/14FOB-medium-t.html?scp=1&sq=Rate%20my%20professor&st=cse> attempted to cite this particular paper along with those authored by James Felton ("Attractiveness, Easiness and Other Issues") and Mark Edmondson ("On the Uses of a Liberal Education") as angry professors hungry for revenge "lashing out with analysis" against the commercial site, "Rate my Professor." The corporate site started in 1999, and this paper and those of other authors Heffernan cited all pre-dated the site by several years, so it wasn't possible for Rate My Professors to have inspired any of the scholarship the reporter cites. A bit of real investigative journalism would have led this New York Times reporter to "Evaluation: Rate My Professor & Professors Strike Back" (Rhem, 2009), for the article that details how the commercial site inspired negative reactions by several professors. However, neither this paper nor the others Heffernan mentions qualifies as such.

The area of student ratings differs from most research areas in that it is perhaps the paragon of emotionally charged academic topics. Emotions result from both personal experiences and conflicts of interest. As a result, the literature of student ratings contains more than the usual share of diatribes and polemics, and some take on vitriolic tones. On one side are staunch champions of student ratings, whose rhetoric seems like advocacy for student ratings being the strongest cure for any classroom ill that affects higher education. On the other side are those who hold equally unshakable belief that students have no business evaluating faculty, that student ratings are, at best, popularity votes based on attributes that have little to do with educational value. When hearing both camps arguing, one is reminded of Eric Hoffer's description of "the true believer." In the middle are the vast majority of academics (faculty and administrators) who have not read much research on student ratings, and hold opinions based mainly upon personal experiences.

Pseudo-evaluation damages the credibility of legitimate evaluation and victimizes individuals by irresponsibly publishing comments about them derived from anonymous sources. This is voyeurism passed off as "evaluation." Examples lie at http://www.pickaprof.com/ and http://ratemyprofessors.com/index.jsp. Neither site provides evaluation of faculty through criteria that might be valuable to a student seeking a professor who is conducive to their learning, thinking or intellectual growth. Both sites are transparently obvious in their advocacy that describes a "good teacher" as an easy grader. The former site proudly displays the quote: "...the most vital academic tool[s] to students seeking good grades." as a quotation from the Houston Chronicle. Presenter Phil Abrami (see Theall, 2005), rated the latter site as "The worst evaluation I've seen" during a panel discussion on student ratings at the 2005 annual AERA meeting.

Lest one think that only charlatans armed with web servers structure such abuses, an institutionalized example of irresponsible use of student ratings can be found at University of Colorado's http://www.colorado.edu/pba/fcq/. Any individual, not just prospective students, but faculty members' friends & children and associates inside or outside Colorado can snoop through what should normally be confidential personnel data. The excuse: "Collection, publication and use of student ratings is mandated by the Regents" hasn't yet been accompanied by any "Rate your Administrator" or "Evaluate your Regent" equivalents. Applying evaluative practices to others that are not applied equally to self speaks volumes about caste mentalities coupled with power abuses. In practice, most institutions show better judgment in management of their faculty and student evaluation data. Yet, the following statement accurately captures the importance administrators in general ascribe to student ratings.

> "*Student ratings of teaching* serve as an important component of many faculty evaluation systems. Either by design or default, institutions often place *great weight* on *student rating data* in making decisions that impact *faculty rewards, career progress* and *professional growth*. It is critical that student rating forms be designed and constructed in such a way as to provide *valid* and *reliable* information for these purposes." (See http://www.cedanet.com/sr_description.htm web site current as of June 10, 2009.)

Some, perhaps most, institutions fail to treat them as simply an "important component" and use student ratings as the defining measure of success. The wording from an actual evaluation report reveals such mentality: " *Dr. X's performance in the area of teaching meets expectations. However, if overall student ratings of Dr. X do not improve… teaching may not meet performance expectations*."

This quote illustrates the thinking that equates satisfactory teaching performance as synonymous with high student ratings and often specifically to tabulated agreement with a single item such as: "Overall this was an excellent course." Such examples reveal a true paradox: cultures with some awareness of the necessity for multiple measures being unable to overcome their own practices of making personnel decisions based upon single-measure convenience. Units that collect data from multiple sources often subscribe to the same mentality by failure to incorporate "other measures" in meaningful ways. Thus, they too default to "evaluation" as a product of single measures.

Part I: What's Known?

Seldin (1993) notes that "*...hundreds of studies have determined that student ratings are generally both reliable (yielding similar results consistently) and valid (measuring what the instrument is supposed to measure.)*" The volume of literature written about student ratings is indeed immense—larger than any other single topic in higher education. Cashin (1988) noted that over 1300 articles on the topic existed in 1988 and the number today is more than double that figure. Workers who survey the growing literature on the subject express favor for the usefulness of student ratings (Cashin, 1988 and 1995; Cohen, 1981; d'Apollonia and Abrami, 1997; Dunkin and Barnes, 1986; Gravestock and Gregor-Greenleaf, 2008; Greenwald, 1997; Theall, Abrami, and Mets, 2001). This is primarily because there is a very general trend for highly rated teachers to be associated with students who achieve well (McKeachie, 1986). However, practitioners often confuse "reliable" and "valid" with "highly predictive," "precise" and even "accurate."

Student ratings come in two types

In colloquial use, "student ratings" and "student evaluations" are often used interchangeably, but it is more accurate to refer to students as rating their professors than evaluating them. There are two very different kinds of student ratings instruments: "*formative*" (those that use student feedback in ways that are diagnostic and allow professors to improve their teaching) and "*summative*" (those used to evaluate professors for rank, salary and tenure purposes). Formative evaluations/ratings given during the ongoing course, usually about mid-term, ask detailed questions that provide a profile of pedagogy and strategy being employed.

Summative evaluations/ratings given at the end of a course are direct measures of student satisfaction. "Satisfaction" is the sum of complex factors that include learning, teaching traits, and affective personal reactions that are products of both what happens in a class and what an individual brings with him or her to the class in form of bias and motivation.

It is maddening when writers of papers and books about "student ratings" or "student ratings" fail to specify whether they are talking about summative or formative tools. The thorough compilation by Theall, Abrami, and Mets (2001) is somewhat damaged by lack of such specificity, because when one talks of the utility of evaluations to help to improve teaching, one cannot be talking about summative ratings.

Formative items are specific rather than general questions. Typical formative items include the following.

"*Discusses recent developments in the field*";
"*Uses examples and illustrations*";
"*Is well prepared*";
"*States objectives of each class session*";
"*Encourages class discussion/participation*";
and "*Is enthusiastic.*"

Such items reveal specific teaching practices, and averages of responses reflect the degree to which these were used in a course. Items established based on research (like that of Hildebrand *et. al.*, 1971; or

Feldman, 1986) reveal the importance of a particular trait or practice as used to promote students' success. An instructor can use the information provided to add new practices or emphasize particular ones of his/her choice. As such, formative tools yield the information required about how to improve. They seek to provide a detailed picture of facets of specific practice rather than responses to general satisfaction.

Summative items that describe general satisfaction receive the greatest (and unfortunately sometimes the only) attention by evaluative supervisors. Summative items called "global" solicit a general overview of the experience. Typical global summative items include:

> *"Overall, how do you rate this instructor's teaching ability compared to all other college instructors you have now and have had in the past?"*
>
> *"Overall, how do you rate this course compared to all other college courses you have now and have had in the past?"*
>
> *"How do you rate this course as a learning experience?"*

Note that no considerations of any specific attribute, event, or content are triggered by global questions. As a result, the responses should arise more from affective feelings than any other. Likewise, questions that ask for competitive ratings of *all* courses and *all* teachers trigger cannot possibly carry an expectation that students can recall *all*, let alone compare *all* and lead to immense disconnect between what the item wordings literally demand of students and what is possible for them to do. Scriven (1997) recommends against framing questions with the competitive wording shown here in phrases such as "*compared to all other college instructors*."

Formative rating and summative rating tools probe for vastly different kinds of information, even though both evaluations usually solicit responses based on choices on a Likert scale from "strongly agree" to "strongly disagree." At the same time, there is a remarkably strong correlation between averages of good formative items and global ratings. This indicates that the affective feelings are *informed* affective feelings and, with few exceptions, are not whimsical responses.

Brief history of formative and summative uses

The following was provided *via* email by Dr. Michael Theall, (now at Youngstown State University), who has written extensively on student ratings (see Theall & Franklin, 1990; Theall, Abrami, and Mets, 2001).

> "The earliest distinction between formative and summative uses was by Mike Scriven, who coined the terms in his (1967) 'Methodology of evaluation' in Taylor, Gagne, & Scriven's 'Perspectives of curriculum evaluation'. The earliest studies were by H. Remmers (e.g., 'Experimental data on the Purdue Rating Scale for Instructors' in 1927) and they were concerned with exploring student opinions as one way to find out more about teaching/learning for 'self-improvement of instruction' and for psychometric reasons (i.e. to validate the scale). In 1928, Remmers investigated 'student marks and student attitude toward instructors'. This time, the psychometric properties of ratings were more the focus, (perhaps due to increasing summative use and resulting validity questions?). By 1949, Remmers was referring (in 'Are student ratings of their instructors related to their grades') to students' opinions of the teacher as one of the 'Two criteria by which teachers are often evaluated...' In the 1949 study, Remmers, concluded that 'There is warrant for ascribing validity to student ratings not merely as measures of student attitude toward instructors...but also as measured by what students learn of the content of the course.'
>
> The timing of the administration of the instrument isn't mentioned in the 1927 study, for example, but in the 1949 study, the evaluations were done at the close of the term. So it looks like: 1) the earliest intent was formative; 2) summative uses developed fairly quickly; 3) psychometric properties were first a measurement issue and then a matter of establishing ratings validity due to summative use; and 4) specifics of the evaluation process gradually evolved from end-of-term administration to other timing and process changes."

Evidence for Value

Where is support for the belief that student ratings are meaningful reflections of instructional competence? The link presumed between student ratings and student learning, which Theall mentions began in 1949 with Remmers' study, remained unproven for many years. Objective support for this belief now rests upon good data and numerical analysis. The most common statistical tool used is the calculated correlation coefficient (r) between student ratings and other measures—in particular student ratings and measures of student achievement, as measured by examinations. Positive numerical coefficients can range between r = 0 (no correlation) to r = 1 (perfect correlation—see Figure 9), and negative correlations between r = 0 (no correlation) to r = -1 (perfect inverse correlation). What constitutes a positive correlation "good enough," given such data, to warrant being called "supportive?" Cashin (1988) provides some guidelines and suggests regarding student ratings validity: "Correlations between 0.20 and 0.49 are practically useful" and correlations above 0.50 "are very useful but they are rare...." Researchers generally accept these guidelines.

Reviews of the actual results from large numbers of teacher ratings show that global questions, in particular, correlate very highly with one another (Cashin, 1995). The correlations between some global questions commonly reach higher than r = 0.8 (see Figure 1). For example, a professor who is rated highly on one global question that has to do with his/her overall rating as a good professor will likely also get a high rating in an overall question about the quality of his/her course. He/she will also likely get a high rating on averages of multiple items on a formative survey. Global questions are often worded in ways that often carry a great deal of redundancy. This should not be forgotten when discussing the actual merits of the correlation coefficient or the high factor loadings in a factor analysis (see for example Marsh, 1983).

Figure 1 displays scatter plots associated with correlation coefficients within the ranges of those represented between student ratings and other measures.

The strongest argument that student ratings are related to student learning occurred about a quarter century ago. Two of the largest meta-analyses aimed to resolve the relationship (Cohen, 1981; Feldman, 1989) found consistent correlations of r = about 0.5 between student learning and student ratings. These provide the strongest basis to date that student ratings reflect cognitive gains and that high ratings generally reflect better student learning. Cohen (1981) utilized students' scores on an external exam as a measure of student learning and compared them with ratings given by students on their evaluation questionnaires. The ratings of own achievement ("rating of how much I learned" r = 0.47), "overall rating of course" (r = 0.47) and "overall rating of instructor effectiveness" (r = 0.44) all fall solidly within the upper-most regions of Cashin's values dubbed as "useful." These findings are frequently cited to support the value of student ratings. They demonstrate irrefutably that students *generally* know when they are learning and rate their teachers accordingly for providing these learning opportunities. Based upon the size and care of these studies, it is unlikely that this particular relationship will change with further study; the relationship seems irrefutable. Expectations that correlation coefficients between student ratings and achievement examinations and grades should be in the range needed for high predictability ($r > 0.8$) and criticisms of the range of r = 0.4 to 0.5 as "too low to be of value" are expressions made by those who, to put it bluntly, are abysmally ignorant of test reliability. Most faculty made tests and grades don't correlate with themselves in the range of r greater than 0.5, so the expectation that they can or should produce higher correlations with a separate measure has no grounding in reality. Even standardized tests seldom achieve reliabilities in the ranges needed for high predictability.

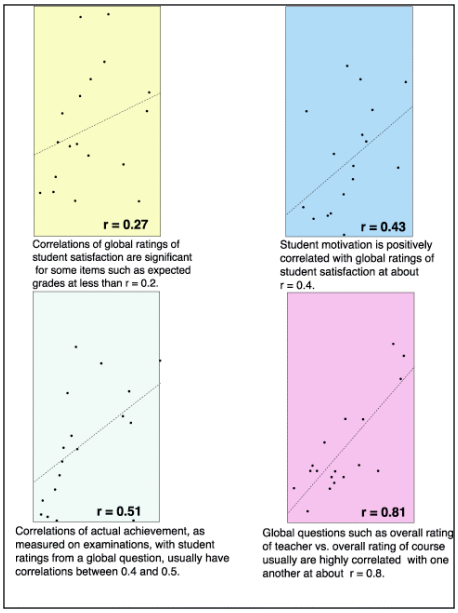

Figure 1. Scatter plots showing correlations typical of student ratings (global ratings of overall satisfaction) with other important parameters. The research that established the relationships shown in accompanying tables were done on much larger populations than shown by the points used to create these graphs. These scatter plots display "best-fit" lines to the data that show degree of prediction of Y from X which can be expected with varied correlations. A "significant correlation" does not mean "high degree of predictability." Perfect predictability (r = 1) would place all points on the best-fit line. From these graphs, it is obvious why a correlation established on a large population cannot be applied reliably to judge an individual. This is why multiple means of assessment are required; student ratings are never in themselves sufficient to judge individuals.

Cohen's (1981) landmark paper also indicated that particular instructional practices influence student ratings. He utilized students' scores on an external exam as a measure of student learning and compared them with ratings given by students on their evaluation questionnaire. He discovered that ratings correlated positively with particular teaching traits including: "explains clearly" (r = 0.50), and teacher structure "uses class time well;" (r = 0.47. These two lead us to look at whether particular practices, in addition to their documented relationship to satisfaction, are indeed related to increased learning. Thus, we now have a tie between formative evaluation data and summative ratings. As we shall detail below, research that teased apart student ratings as functions of particular practices shown by research to be useful to promoting learning, provides another strong argument for the value of student ratings. We have data to show that high global ratings in general are related to teachers' practices that indeed have a basis in research as being beneficial to learning.

Teaching Dimensions. The most critical literature to construction of most ratings forms was that which established the legitimacy of the *teaching dimensions*. The dimensions are traits and practices that tend to account for summative ratings. The reports of Theall and Feldman (2007) and Abrami, Rosenfield, and Dedic (2007) and Abrami, d'Apollonia and Rosenfield (2007) show that Feldman's original dimensions (see Feldman, 1998) remain valid. The dimensions of greatest significance established through thorough meta-analytical studies listed in order of importance (from #1 as most important) to learning (from Feldman, 1998) follow in Table 1.

"Top 5" Instructional Dimensions Based on Different Indicators
(Modified from Feldman, 1998)

Instructional Dimension	Ranked Importance with Learning Achievement	Ranked Importance with Satisfaction	Achievement % Variance Explained
Preparation and organization	1	6	30-35%
Clarity and understandableness	2	2	24 -30%
Perceived outcome or impact	3	3	15-20%
Stimulation of interest in content	4	1	10-15%
Encouragement and openness	5.5	11	<10%
Availability and helpfulness	5.5	16	<10%

Table 1. Instructional dimensions compared with their ranks of importance in producing satisfaction and producing learning. These reveal similarity but not congruence. The trait most important to develop in order to produce highest levels of student learning is attention to course organization and preparation. This importance was also confirmed by the National Study for Student Learning (Pascarella, 2001). Yet, this is only the sixth most important practice for producing high ratings of satisfaction.

Professor Kenneth Feldman studied the relationships between student ratings and practices perhaps more than any individual, and he teased out the formative components that lead to summative ratings of satisfaction and components that lead to measurably increased learning. The "Teaching Dimensions" thus unite formative measures as meaningful to summative ratings.

There is similarity in the ranking (Table 1), but there are important differences. Recent studies indicate that "the most effective" teaching practices seen by students vary across ethnic groups (Sanders and Wiseman, 1998). Further, many of the classic studies cited here draw their inferences primarily from classes dominated by lecture-discussion pedagogy. Finally, we now know that it is not safe to assume that the general relative importance of these dimensions that Feldman established from his large database has the same relative importance at one's own campus. On a campus that has small classes and significantly better student-teacher interaction, "availability and helpfulness" rises from accounting for less than 10% of variance to having parity with "clarity and organization."

Comparable studies derived from classes that use alternative pedagogies are yet to be produced, and perhaps this is one of the greatest dangers of using long-established forms: they reward great lecturers and marginalize other practices. This has led a few schools that promote active learning methods to depart from the usual purchase of established forms and to develop contemporary forms that are learner centered and pedagogically independent (Langley, 2007; Nuhfer and others, 2008a,b). A recent contribution that makes development of psychometrically sound "home-grown" instruments is Berk (2006). Particularly useful are his twenty criteria formulated to screen the development of satisfactory items. Some items on established commercial forms cannot pass Berk's criteria, and screening the average "home grown" form with his criteria usually reveals why home-grown forms have a horrid reputation; those created in the absence of significant consultation with current literature (Berk, 2006; Perry and Smart, 2007) are destined to become embarrassments.

Erdle and Murray (1986) showed that certain behaviors in class affect student satisfaction, and further, that the relative importance of these behaviors varies between disciplines (Table 2). Erdle and Murray's work indicates that the nature of what we are trying to teach influences the traits we should probably seek to address to obtain improvement.

Correlations between Ratings of Overall Teaching Effectiveness and Teaching Behavior Factors
(After Erdle and Murray, 1986)

Behavior	Perceived Importance to teaching by students of:		
	Humanities	Social Science	Physical/life science
Rapport	0.43	0.70	0.59
Interest	0.50	0.71	0.37
Disclosure	0.30	0.65	0.25
Organization	0.51	0.56	0.47
Interaction	0.48	0.51	0.34
Course Pacing	0.53	0.45	0.62
Speech Clarity	0.53	0.45	0.62
Expressiveness	0.58	0.59	0.51
Emphasis	0.61	0.58	0.51
Mannerisms	-0.53	-0.42	-0.28
Use of Graphic	0.22	0.35	0.37
Vocabulary	0.16	0.35	0.37
Presentation Rate	0.23	0.14	0.31
Media Use	0.30	0.23	0.11

Table 2. Relationship of teaching traits to summative global student ratings. The variations between how students value various traits depend upon the subject being taught. Affective factors such as rapport and expressiveness exert a powerful influence, but so do traits related to learning such as organization, clarity, and emphasis of important points.

Cashin (1988, 1995) helped promote awareness through concise summaries of the research on student ratings, and he presented this data along with correlation coefficients (Table 3). His compilations show that professors who teach classes where students are motivated (such as classes taken by choice or in one's own major) have a major advantage in their ratings over other teachers. Those who teach large classes are generally at some slight disadvantage in their ratings, and over *large populations*, ratings of students are generally consistent with those of alumni, colleagues and administrators. Those who are productive in research are generally rated more highly in classes they teach, but the relationship is so slight that research productivity is useless as a predictor of student satisfaction with teaching. The higher professorial ranks have slightly better student satisfaction, but the relationship is so weak that rank cannot be used as any predictor. The relationship between grade expectations and student satisfaction is weak.

Relationships with Student ratings: Correlations between Various Influences and Ratings of Overall Teaching Effectiveness
(NR = Not Related at any Significance)

FACTOR	CORRELATION WITH GLOBAL RATING
Sex of Instructor	NR
Sex of Student	NR
Level of Student	NR
Rank of Professor	0.10
Research productivity	0.12
Student's GPA	NR
Age of Student	NR
Age of Professor	NR
Time of day	NR
Class size	-0.18
Student Motivation	0.39
Expected Grades	0.12
Course Level	0.07
Colleagues' Ratings	0.48 to 0.69
Administrators' Ratings	0.47 to 0.62
Alumni Ratings	0.40 to 0.75

Table 3. Relationships of various factors to student ratings, come from various studies cited in Cashin, 1988, with exception of relationship to sex of instructor, which comes from Feldman, 1992, and Centra and Gaubatz, 1998. The results are all outcomes based upon studies of large populations. Student motivation (willingness to participate actively in the learning process) has the greatest positive influence on student satisfaction of any instructional factor shown. Student ratings are also consistent with those of faculty colleagues and administrators, and the ratings remain consistent, as students become alumni. Of interest is the fact that, in practice, administrators and colleagues spend little to no time in classroom from which these ratings are derived. Their only means of obtaining information are either hearsay from the students, or from seeing the results of the student ratings. Thus, colleagues' and administrators' ratings could simply be redundant with, rather than independently supportive expressions, of the student evaluation ratings.

We know that formative evaluations, done properly, are highly beneficial. Abrami, Leventhal, and Perry (1982), Dunkin and Barnes (1986), Murray, (1984), Stevens and Aleamoni (1985), and Cashin (1988) all show that formative evaluation particularly with follow-up consultation leads to improvement. Based upon a synthesis of 22 studies, Cohen (1980) showed that instructors with no student ratings rated in the 50th percentile at the end of a term; those who obtained student evaluation feedback were rated in the 58th percentile and those who received feedback with follow-up consultation were rated in the 74th percentile. It is a demonstrated fact that formative evaluation with follow-up consultation is effective in raising summative student ratings and that simply doing final evaluation, as is the common practice, is not effective. Follow-up consultations for individuals are best done in a neutral and supportive environment.

Given all this, is there any basis for arguments against the value of student ratings?

Evidence for the Contrary

The weaknesses and misuses of student ratings have been addressed by a number of writers. Representative are Nerger and others, (1995), Williams and Ceci (1997), Trout (1997), and Wilson (1998). They voice discontent about student ratings--usually because student ratings instruments solicit expressions of satisfaction or dissatisfaction based upon affective attributes rather than on learning or cognitive growth. However, objective support for such contrary beliefs also rests upon context of use and, paradoxically, on some of the same numerical analyses that support value of student ratings. The context is the fact that

student evaluation data is routinely collected and used for the purpose of judging individuals, deeming them as "good" or "bad" teachers and tying their student ratings to their careers. Critics note correctly that administrators routinely collect student ratings data that are neither gathered under research conditions nor used to address a question in educational research. Instead, the collectors of data use it to judge, reward and punish individuals. Where universities post ratings of professors on the web, one purpose for collecting the data seems to be to foster intimidation by placing students in positions of power over professors.

Critics are correct that evaluation of an individual tests a different hypothesis than deducing the general trend across a populace. Figure 1 depicted the difference clearly. Certainly, we can deduce a trend shown by the fit lines based upon many individuals' varied performances. However, in the case of a single individual's evaluation, we have a point. As Figure 1 shows, at correlation values of $r = 0.5$, almost none of the points actually fit on the line. Based on predicting learning based on student ratings, one is more likely to widely overestimate or underestimate the learning produced by an individual than to arrive at an accurate evaluation. The trends are statistically reliable, but a trend to evaluate an individual must not simply be valid and reliable; the trend also must reflect a high degree of predictability. Sufficient predictability to judge individuals responsibly simply isn't present at correlation values ranging of about $r = 0.5$ and below. A measure that yields a "significant" correlation coefficient at this level between student learning and summative evaluation scores over a large population of faculty (Cohen, 1981; d'Apollonia, and Abrami, 1997) is not something that one can reliably apply to an individual faculty member in a rank-salary-tenure decision.

A reason that Cohen and Feldman resorted to meta-analysis arose from the fact that smaller individual studies produced conflicting conclusions. Critics correctly point out that a correlative association between student ratings and learning strong enough to allow a user to accurately deduce students' learning from a faculty member's ratings, if it existed, would have been discovered without need to resort to meta-analyses. Strong associations with high predictability do not require meta-analyses to discern them. They have an inherently reasonable consistence whether one uses a large or small study group to generate a database. Such is not the case with student ratings. Application of student ratings to deduce the career success of any single professor is a different challenge from statistically describing paired relationships within large databases. The same research that supports the validity of student ratings illuminates the dangers of trying to apply results of good established associations, determined as valid on large populations, onto individuals.

Additional sources of conflict between pro and con advocates regarding use of student ratings is based on relationships between ratings and factors that are not related to learning or to any mission or goals of education. Cohen's (1981) landmark paper provided an example when it indicated that particular affective practices also contributed to higher ratings ("Teacher rapport," $r = 0.31$). Naturally, the question arose "Can a teacher promote high ratings through emphasizing affective traits without really producing the kinds of learning in accord with goals and mission of an institution?"

Feldman (1986) showed that professors' personalities affect students' ratings of overall teaching effectiveness (Table 3). One striking aspect of Table 4 is the demonstration of how teachers tend <u>not</u> to see themselves as others see them. Of the personality traits, the only two traits that peers, students, and teachers agree upon as being of significant importance are enthusiasm and self esteem, and students and peers give these much more importance than we tend to give them in ourselves.

Overall Teaching Effectiveness and Personal Attributes of Professors
(After Feldman, 1986)

PERSONALITY TRAIT	IMPORTANCE AS SEEN		
	By Self	By Students	By Peers
Self Esteem	0.38	0.51	not rated
Energy (enthusiasm)	0.27	0.62	0.51
Warmth	0.15	0.55	0.50
Cautiousness	-0.09	-0.02	-0.26
Leadership	0.07	0.56	0.48
Sensitivity	0.07	0.53	0.47
Flexibility	0.05	0.57	0.46
Emotional Stability	-0.02	0.47	0.54
Friendliness	0.04	0.42	0.49
Neuroticism	-0.04	-0.49	-0.35
Responsible/orderly	0.06	0.31	0.25
Brightness	-0.05	0.36	0.22
Independence	-0.12	0.01	0.08
Aggressiveness	0.23	0.05	0.02

Table 3. (After Feldman, 1986) reveals that professors' personal traits do affect ratings by peers and students. They further reveal that professors are apt to underestimate or misjudge the importance of the effects of these traits on others. Many of the correlations that students see as important to their satisfaction are higher than the value established in research between student satisfaction and student learning.

Any university leader who recognizes the significance of Feldman's research will do everything possible to insure that self-esteem and enthusiasm of faculty are nurtured. An evaluation system that humiliates faculty rather than strengthens them will likely damage teaching through destroying self-esteem and enthusiasm on an institutional scale.

Perhaps the most heretical of all studies concerning affective influences on student ratings was the famed "Dr. Fox experiment" (Naftulin, Ware and Donnelly, 1973) in which a hired actor posed as Doctor Fox and lectured to three groups of educators in a manner which was highly expressive but low in content. The groups consisting of professors, professionals and administrators gave satisfactory content marks to the actor, thus demonstrating a tremendously disturbing fact—even those who are above average in intelligence, trained in critical thinking and are well-educated cannot always tell when a lecture has substantive educational value. Critics noted that if professionals could not make this distinction, then how could average undergraduates make it? Could they be expected to know whether a professor was providing substantive content, if the content were current, or would they rate their professors more on expressiveness (or worse, entertainment value) rather than content value? One implication from the study was that student ratings were not valid criteria to evaluate actual teaching effectiveness by lecture. The implication was argued based on similar data both *pro* (Ware and Williams, 1975) and *con* (Marsh and Ware, 1982). Marsh and Ware (1982) used factor analysis to divide "evaluation" into several dimensions and showed, that of the two most important influences, expressiveness (number 1 in importance) was registered primarily through the rating of "Instructor Enthusiasm," whereas content coverage (number 2) was expressed through "Instructor Knowledge." The more comprehensive of the later studies (Perry, Abrami and Leventhal, 1979; Abrami, Leventhal and Perry, 1982 - see Dunkin and Barnes, 1986) respectively replicated the Dr. Fox experiment and analyzed data from their own and from 11 other studies. They found that the effect of expressiveness alone on *overall* student ratings was "significant and reasonably large" whereas the effect of content alone was sadly "inconsistent and generally much smaller." However, on overall student achievement, content became significant and expressiveness became insignificant.

Critics of the Dr Fox study speculate that such a ruse could not remain successful over an extended time and that students would eventually discover the hoax. Contradicting this speculation is an actual case study described in the first half of *Generation X Goes to College* by Peter Sacks. This autobiography of a tenure track professor in an unnamed community college discloses the teacher initially finding himself in trouble with student ratings. Sacks exploited affective factors to deliberately obtain

higher ratings, and described in detail how he did so in the chapter, “The Sandbox Experiment.” Sacks obtained subsequent high ratings through his efforts, but not through promotion of any learning outcomes. For years, he managed not only to deceive students, but also peers and administrators and eventually was awarded tenure based on higher student ratings. Sacks demonstrated that a professor can emphasize particular practices that will change student ratings but not necessarily produce the best learning outcomes. His book is a brutal disclosure about himself and his institution. The case shows clearly that (1) a teacher can manipulate satisfaction without attending to students' learning, and (2) that inept faculty peer reviewers and administrators promote the actions that Sacks chose by placing faculty careers in the hands of student raters.

The experience also made Sacks a powerful voice—not necessarily against the ratings themselves, but rather in deprecation of the tyrannical ways in which administrators can use these. Sacks describes his view in a sidebar in Trout (1997):

> "Once employed as an innocuous tool for feedback about teaching, student surveys have evolved into surveillance and control devices for decisions about tenure and promotion. Add the consumeristic and entertainment values of the culture beyond academe and the result can be ugly: pandering teachers doing what's necessary to keep their student-consumers satisfied and their jobs secure."

Sacks' quotation captures a poignant source of conflict between those who champion student ratings and those who deplore them: the vitriol in the conflict often framed as an argument over the research is likely not about the research at all. Rather, the emotional polemics (typified by Fish, 2005) are about the *use, and particularly the misuses* of ratings. In the end, misuse can lead to institution-wide erosion of self-esteem and enthusiasm, destroying, wholesale, the very traits in instructors that Feldman's work (1986) shows are most important to their success. Clearly, the practices through which student ratings are used perhaps, more often than not, are directly at odds with the research on student ratings. It is little wonder that the messy state of things led to the AERA panel in 2005 titled "Valid faculty evaluation data: are there any?” (Theall and others, 2005).

The primary reason that faculty increasingly resist student ratings instruments is because administrators increasingly misuse the information. Where administrators will not admit to the damage they do through the misuse of ratings, such defensive, blocking reactions subvert any possible solutions. Recently, The "Chronicle of Higher Education" carried a column on the research of Anthony G. Greenwald, professor of psychology at the University of Washington, and Steven D. Falkenberg, professor of psychology at Eastern Kentucky University that confirms widespread misuse (Glenn, 2007). Lack of accountability in producing good student learning is an open invitation to set institutions onto other agendas that do not support student learning. When institutions' administrators are held as fully accountable for producing student learning and performing good faculty evaluation as their faculty are held accountable for good teaching and scholarship, misuses will likely decrease.

Part II. A Fractal Thinker Looks at Student ratings

Nature is full of fractal forms: trees, clouds, blood vessels, and landforms, to name a few. A fractal form is complex, but although it has the illusion of being randomly irregular and seemingly impossible to quantify at first sight, this intimidating complexity has an order within that provides a means to understand the form in surprising ways. Order of fractal forms consists of complex forms built from recursive operations on a small unit called a generator (Figure 2). It results in forms that have the characteristics of similarity when viewed at different scales, and predictable growth of a dimension such as length in accord with decreasing length of measuring tool.

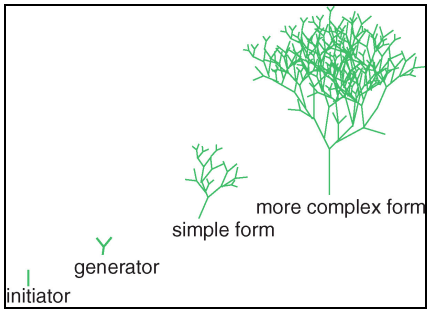

Figure 2. Concept of a fractal form, in this case a branching network built from recursive operations on a generator--each branch of the Y being replaced with subsequent Y-shapes. (From Nuhfer, 2003a).

Fractals provide important insights to understanding much about education, because learning occurs by increasing the strength and numbers of synaptic connections (Leamnson, 1999) usually through repeated use. Exceptions to repetition occur when learning is accompanied by strong emotional-affective influences that seem to establish strong permanent connections instantly. Growth of these connections, much like growth of a tree, produces immensely complex forms by recursive growth of a simple generator into branching patterns (Nuhfer, 2003a; 2003b). Education is replete with fractal characteristics in both space and time, probably because neural networks, like blood vessels, are fractal networks. In space, physical brain changes include growth of such networks in the process of becoming educated. Many natural temporal patterns in time are fractal, and learning, too, is a product of a series of events in time. Relationships between student ratings and other measures are myriad and complex. Student ratings, as well as all other educational endeavors, arise from the brain's branching neural networks. Fractal thinking proves useful to many educational endeavors (Figure 3. See also Nuhfer-2003a-c 2004a-c, 2005; Nuhfer, and Adkison, 2003; Nuhfer, Krest and Handelsman, 2004; Nuhfer, Leonard, and Akersten, 2004.) This paper concerns the specific case of student ratings.

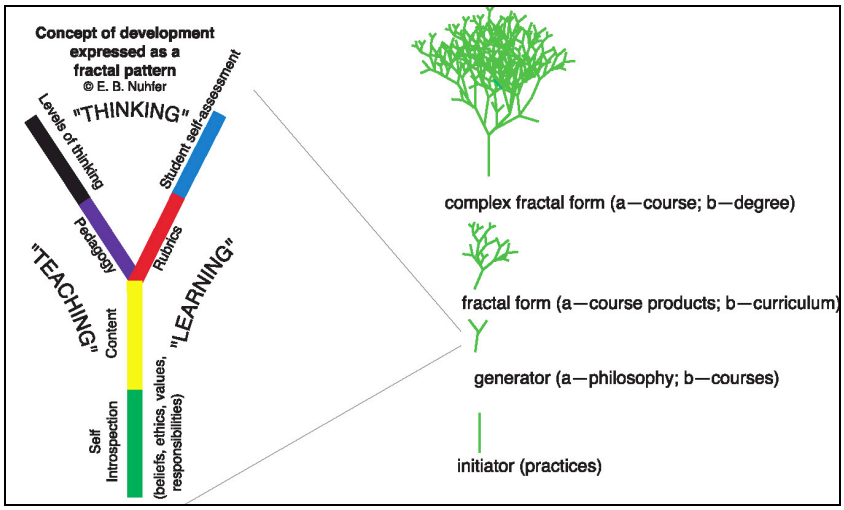

Figure 3. Generator for a professional application in college teaching (after Nuhfer, 2003a) deduced as result of years of design in faculty development. Note that the base, self-introspection, addresses primarily affective attributes. This model recognizes that all cognitive choices in selection of content, pedagogy, levels of thinking, rubrics and exercises to produce student self-assessment are rooted in and connect with affective feelings. All course products likewise manifest such feelings. Nuhfer (2007) attaches extreme importance to the generator in recognizing how earliest experiences and practices shape course outcomes and products in major ways.

The nature of learning and teaching (which is learned) involve both cognitive and affective portions of the brain that begin to become established from at least as early as birth. The cognitive is always tied, both verbally and non-verbally, to the affective regions of the brain. I believe that one of the major obstacles in resolving the student ratings controversies lies in inability to perceive the role and

magnitude of the affective components. Separate camps—one that deprecates value of student ratings and another that advocates for use of student ratings—seem unable to really hear one another.

Advocates who value student ratings too much can come across as arrogant and dismissive of others' experiences. They act as if the dominant positive trend precludes exceptions, and they label faculty disclosures of true events as “anecdotal,” “misbeliefs” or even “myths.” The expected characteristic of correlations in the ranges of r = 0.5 and less are such that many individuals WILL be exceptions—they won't "plot on the line." Some researchers (e. g. Boice, 1990) suggest that “countering” constitutes the proper response to such disclosures. "Countering" might be well suited to arguing about trends in general populations, but “countering," in the context of faculty development is an exercise unbridled by emotional intelligence. It denies the reality of individuals, and it is a humiliating response to faculty seeking help. Heavy-handed "countering," produces the undesired effect of scorning individuals’ actual hardships and experiences. The skills needed to be a researcher in education, psychology, or student ratings are not the same as those required to help an individual become a stronger teacher. Other traits, many of them affective, are far more important in administrators than the ability to argue.

There are many reasons why student ratings are intimately tied to affective domains, beyond the usual affective reaction of individual researcher's desires for his/her research to be both correct and influentially important. A faculty member who has received abusive ratings, been humiliated by inept uses of these ratings, or both, will naturally react against arguments for what will be heard as advocacy for putting her/him self into an endless cycle of abusive experiences. The ability to evaluate faculty provides a differential in which administrators have power over faculty. The tool that comes with more ease of acquisition than perhaps any other and that gives administrators more power than anything to affect faculty lives outside of the workplace are the evaluative ratings by students. Any intimation that administrators may be wielding such power badly or irresponsibly, or are not even doing real evaluation will simply be discounted and heard as threatening to both them and their positions. Experts who sell evaluation forms and workshops that promote student evaluation may have built the most impervious of all affective defenses against seeing student ratings as just one of a number of possible multiple measures. A characteristic of fractal thinking is perpetual awareness that the ratings debate is not about only an objective application of research findings.

Affective feelings largely control responses to rating items associated with the general experience with the class. The instructor is surely a major contributor to such feelings, but not nearly all. Perception of a course experience depends heavily on what individual students bring to class pre-wired within their neural networks as expectations and levels of intellectual sophistication. Bimodal distributions on ratings showing a “love-hate” division within a single class produced by students who have undergone the same educational experience are common. These reveal that a rating provided by a student is as much about the student as it is the experience. A misapplication of student ratings in using ratings alone to judge good teaching is usually based on the rash presumption that the rating is all about the teacher and reflects the instructor’s ability as a teacher to meet their assigned duties, the primary one of which is to produce beneficial cognitive growth in students. In contrast, a fractal thinker anticipates that a summative student evaluation will be a very honest expression of satisfaction, which is largely an affective trait, but which has some connection with the cognitive domain.

Fractals provide a particularly damning exposure of the practice of employment of single global questions as the over dominating basis for an evaluation. Because teaching practice is learned behavior, an evaluation of teaching is an evaluation of the neural networks a faculty member has developed to deal with those practices. The neural network is fractal, so our problem of faculty evaluation can be framed as one of understanding a complex fractal form.

The fractal dimension is one of the essential manifestations required to describe a fractal form. Derivation of a fractal dimension requires multiple measures taken at different scales. A trait of a fractal form is that its dimensions increase in a surprisingly predictable way, depending upon the length of the measuring instrument one uses to measure the form. The fact that something would change in length depending upon the length of our measuring sticks defies our common sense, but that is exactly how fractals behave. For instance, take the profile of a coastline (Figure 4) on a large map. If one measures its length with a pair of dividers that are set initially at say, four inches, and then again, at two inches, one, inch, a half-inch and so forth, each measured length of the coastline will increase, as the dividers get smaller. To some, it may seem obvious that if one takes a very crooked line (like a coastline) and measures it first crudely and then more precisely, its length will increase. However, the length of a fractal form just doesn’t increase, it increases with such regularity that one can predict accurately what length one will

measure based upon any setting of the divider (Figure 5). This growth will be so regular that it will plot as a straight line with slope produced by log(L)/log(1/r) where L= length derived from the number of divider widths required to measure the length of the feature and r = the width setting of the divider. The slope of that line defined in a single number is an expression of the fractal dimension for that landform. This provides a concise description—a name expressed in numbers—to distinguish one kind of coastline from another.

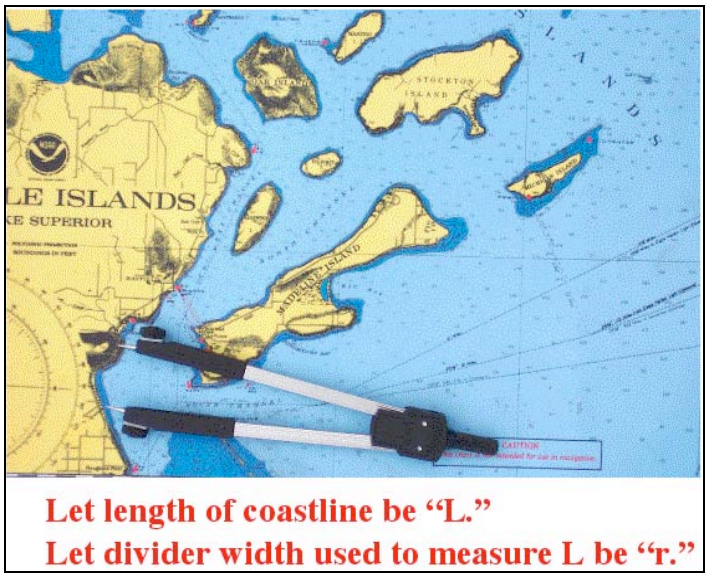

Figure 4. Measuring a fractal coastline several times by choosing first narrow and then wide divider widths, walking the divider along the coast and summing the number of divider steps and multiplying by the width to get the length of the coastline. The length of the coastline increases as divider length decreases (see Figure 4).

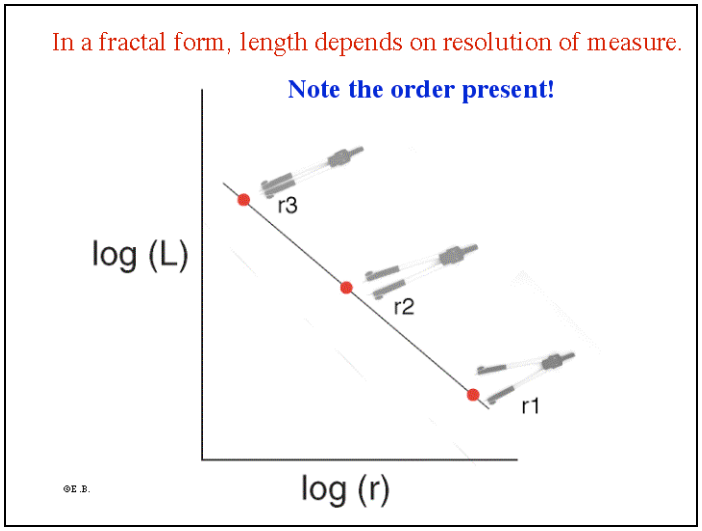

Figure 5. Understanding the coastline requires multiple measures, not just one. Note how the order of a fractal form becomes apparent as the coastal length changes almost in a perfectly predictable manner as the divider lengths are changed. This depiction of a fractal form was inspired by Benoit Mandelbrot's article in Science, 1967, "How long is the Coast of Britain?"

To measure a fractal dimension, there is no substitute for multiple measures; no single measure can capture the quality of a fractal form. Now, consider the practice of trying to measure teaching effectiveness with a single question/item. Chances of being able to capture teaching effectiveness through such an effort are much less than being able to capture the length of a coastline; the neural networks that govern teaching are much more complex than any line described by the interface of land and water.

All substantive classifications of patterns of thinking and acting have been demonstrated to require multiple measures. Consider the established patterns of learning styles, multiple intelligences and personality types. The way all originators who developed such tools were forced to classify individuals was through a battery of survey items--many measures. Not one of these researchers—David Kolb, Howard Gardner, or Isabel Myers & Katharine Briggs—was able to succeed in diagnosing an individual's respective learning style, intelligence type or personality based upon one question for each of their types. To a fractal thinker, it is obvious why all such classifications required multiple measures. Without being overtly aware of the fact, all these researchers were trying to characterize fractal forms of particular neural networks. There was no way any of them could possibly succeed without employing multiple measures.

It should be equally obvious why attempts to evaluate faculty based on student ratings alone, or worse, a single global item from a survey, is a doomed approach. The nature of what we are trying to evaluate simply cannot be captured with single measures. Capturing successful teaching is an endeavor at understanding neural networks that are far more complex than those involved in personality types, learning styles or intelligence type. Fractal thinking would show that teaching characteristics play out from all of these three and more.

Arguments over relative merits of formative and summative evaluation are somewhat like arguing for diagnosing learning styles while denying that diagnoses of multiple intelligences, or personality types serves any useful purpose. In contrast, Figure 6 is an example of how a fractal thinker perceives multiple measures.

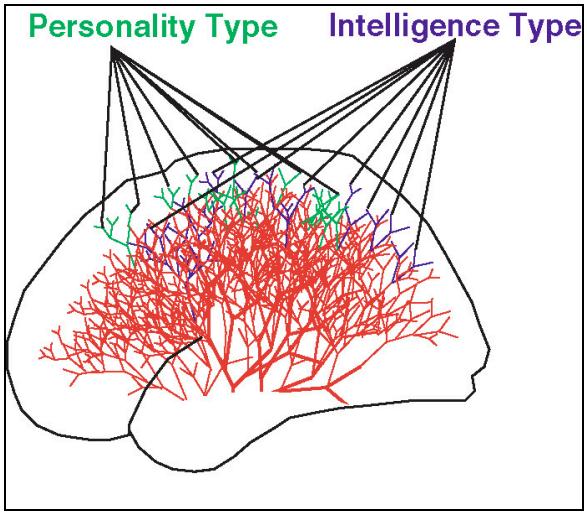

Figure 6. A model in which multiple measures deduce "type" by sampling neural networks differently. The different "types" are both useful to know, and result simply because their diagnoses employ different samplings of the affective and cognitive neural networks.

A similar conceptualization can depict the summative versus formative merits of student ratings. Figure 7 displays a single global measure (summative) *versus* a diagnostic battery of multiple practices that formative tools try to capture. It depicts formative and summative data as drawing upon different information and yielding useful measures derived through a general assessment of students' satisfaction and a pedagogical profile of practices.

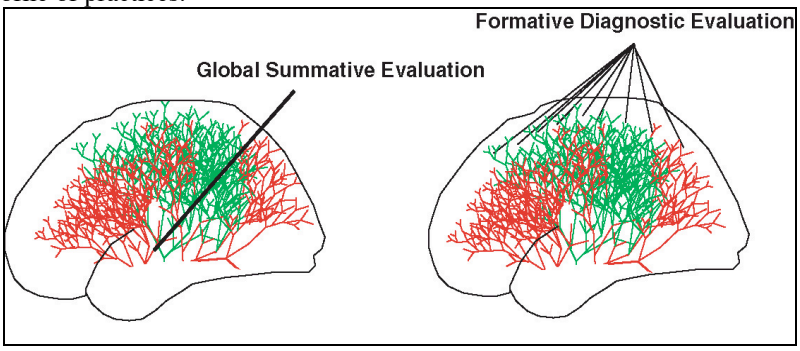

Figure 7. Comparison of sampling of a summative item: "Overall this was an excellent course;" with a battery of formative measures such as "Uses examples and illustrations," and "States objectives of

each class session." The global item is likely to trigger a response mainly from general feelings that arise in the limbic area along with some cognitive recollection. The items on the survey are more aligned to elicit a response that is more dominated by the cognitive, but nevertheless is tied to affective feelings. Of importance here is that the global summative question taps information that the formative does not, and *vice versa*. Both are useful multiple measures, but neither are sufficient in themselves for making judgmental ratings of individual professors.

Formative efforts to define, disclose and improve weaknesses along with consultations with a developer to gain improvements need to be confidential. Although advocates note that formative and summative surveys should be separated, where students confirm that good practices are in fact at work in the classes, then this "pedagogical fingerprint" is probably even more valuable in deducing the summative quality of teaching than a satisfaction rating. A fractal perspective confirms a need for full employment of multiple measures.

The assessment movement appears to have incorporated fractal thinking much better than practitioners who are caught up too much in student ratings. Assessment refuses to accept convenient single measures as adequate, and assessment is beginning to make educators aware that improving student learning requires more than evaluating the individuals who promote this learning.

Student ratings are an evaluative exercise.

The ability of students to evaluate faculty is rarely considered in terms of the research on student thinking. The research on levels of thinking is extensive and it is as solid as that on evaluative tools. The best-known model is that of Perry (1999). Perry deduced nine levels of thinking, six of which apply to undergraduates and that have been repeatedly corroborated by others' work (see Nuhfer and Pavelich, 2001, 2002). The average high school graduate reasons at about a level of 3.7 on Perry's 9-point scale. The average college graduate reasons at about a level of 4. There is scarcely little gain between high school and college in ability to reason at higher levels or to think reflectively. Reasons for this are outside this discussion. A characteristic of levels below 5 is that students may do evaluative thinking but do it poorly. When a student evaluates a class or a course, this is an exercise in evaluative thinking. The degree to which individual students can do this well does indeed depend upon the level of thinking each has reached. Those who argue that their undergraduates can evaluate faculty well may, depending upon their institution, be unwittingly arguing against a massive amount of research revealing knowledge about intellectual development.

The research on levels of thinking differs in an important way from the research on multiple intelligences, learning styles, *etc.,* where individuals create useful classifications through different tools based upon different objectives and considerations. In contrast, researchers who study levels of thinking independently and with some very differently held values nevertheless produce a remarkable concurrence about discernible differences between specific thinking levels. Leamnson (1999) captured this distinction by referring to the different classifications produced by tools as "inventions," whereas the repeatable concurrence typified by disparate studies on levels of thinking would fall more into the realm of "discovery."

Thin slices and other affective manifestations

One of the most surprising findings came from Ambady's and Rosenthal's (1993) "thin slice" studies. These researchers determined that students arrived at ratings for teachers after watching 30 seconds of silent content-free video that were highly consistent ($r = 0.76$) with end-of-semester ratings. Further, viewing of several 3-second video segments yielded only somewhat lower correlations ($r= 0.68$)—both of which are higher than the established relationship of learning to ratings. Certainly, content-free video clips observed for a few seconds cannot confer learning, and these correlations are not reasonably explained as arising from cognitive growth. An explanation is that affective reactions form neural networks quickly, stabilize early and persist to the end of the course where they manifest as a rating of the professor.

This is a find truly appealing to a fractal thinker's fancy! The first class period establishes a generator in the mind of the student--likely even unconsciously, and the character of this generator should persist in subsequent growth that the neural network produces during the course. The thins-slices finding isn't so surprising to a fractal thinker, but the strength of the generator and the incredible speed with which it forms are astounding. It shows particularly why the first day of class had better be a product that has been carefully planned in accord with one's highest aspirations and teaching philosophy.

Recent work underscores other affective influences on student ratings and observations that fit well with a fractal thinker's perspective. University of Texas economist Daniel Hamermesh (Hamermesh

and Parker, 2004) recently verified an influential relationship behind student ratings—beauty of the professor: "Instructors who are viewed as better looking receive higher instructional ratings, with the impact of a move from the 10th to the 90th percentile of beauty being substantial."

Because of affective influences attached to the very discussion of the merits of student ratings, researchers who report affective factors' strong influence on student ratings, do so at some peril. "True believers" in student ratings invariably attack such studies as "biased," "erroneous" and/or "methodologically flawed." Responses to the work of Hamermesh and Parker posted on the POD Network reveal that affective hostile responses can extend past the work to the authors themselves: "*... deprecation of student eval's. *may* be the intent of the economist-authors. (Who knows, maybe the researchers themselves are not "beautiful," received low ratings, were denied raises, etc.; I empathize/sympathize)*"

In the case cited above, Hamermesh, a well-respected teacher and researcher, had a long prior record of researching the influence of personal appearance on job success in many settings. The "deprecation" of student ratings was obviously not on the personal agendas of these researchers. Their research, though unappreciated, reported real events. However, the comment in response shows the power of the affective over both reason and civility when this particular topic arises. The topic of student ratings is accompanied with emotions that few, if any, other area of academic research carries with it.

To a fractal thinker, the discovery of powerful affective influences is neither disturbing, unanticipated, nor a detraction from sound discoveries linking ratings to cognitive factors. Because student ratings tap both affective and cognitive neural components, it would be surprising if explored relationships between affective attributes and student ratings showed no strong relationships. Statistical practices such as factor analyses and regression that focused on cognitive attributes seem to have been interpreted by some as indicative that there can be no powerful influences outside what was actually studied. The reason more affective influences were not found by early researchers whom advocates regard as orthodox is not that the influences are not real. Rather, it is because early researchers didn't look for them. We should hardly expect all meaningful data on student ratings to arise only from the cognitive domain or for student ratings to be understood completely through pedagogical practices and exam scores. Such an expectation runs counter to everything known about how the brain learns and operates.

Using Ratings More Effectively

Use multiple measures. NEVER evaluate faculty based on summative ratings alone

By default, student ratings often become the sole basis for career decisions. Although data provided by them is useful and convenient to use, ratings based on such data alone are wholly inadequate. One reason that multiple measures are not used in practice is because even when researchers on student ratings advise employment of multiple measures, no measure other than the summative student evaluation tool is specifically recommended. Here, I recommend that the student input consist of three specific measures (Figure 8): (1) summative ratings as a measure of student satisfaction, (2) formative survey information to provide a picture of pedagogical practices and (3) knowledge survey data to provide information on content, levels of challenge and a student-based report on their learning.

Knowledge surveys (Nuhfer, 1993; 1995; Nuhfer and Knipp, 2003--See link in References Cited; Knowledge surveys are detailed there) provide an additional source of detailed information gathered from students on perceptions of their learning. Knowledge surveys are an assessment tool that gathers information bridging that yielded by tests and by student ratings. Statistically, knowledge surveys prove to be highly reliable (Figure 9). They are ideal for supporting learning in any course, and the student responses yield a wealth of both formative and summative information.

Assessment of student learning, not faculty ratings, is the major outcome now demanded from evaluators and accreditors, and school administrators have been generally more engaged in rating faculty than in improving student learning. Although some advocates of student ratings argue that student learning should not be a component of faculty evaluation, I disagree. Student learning is the only outcome that, in itself, can justify the maintenance of educational institutions. Faculty are the primary group responsible for both the curriculum and for the maintenance of an environment conducive to learning. In evaluating faculty, programs and institutions, it is simply good management to monitor learning outcomes at all scales. As an outcome, student learning is more important than student satisfaction, and to act otherwise communicates that happy customers are more important than an educated populace. Such an approach subverts the goals and missions of colleges and universities. Most faculty welcome evaluation that is based

on the work they do and are asked to do, rather than review based primarily on personal attributes that they perceive are slanted toward how well they are liked by an audience.

Regarding the "student as customer model," Chemist Mike Chejlava at Lafayette College notes such "forces lead students to believe that they must get the RIGHT answer the FIRST time ...(and) any faculty member who gives work that they cannot master the first time is trying to keep them from their goals by setting standards too high." In her dissertation, "Bridging the Gap Between What is Praised and What is Practiced: Supporting the Work of Change as Anatomy & Physiology Instructors Introduce Instructional Strategies to Promote Student Active Learning in Undergraduate Classrooms," Thorn (2003& personal communication) revealed that all instructors in her study received lower student ratings while attempting to emphasize critical thinking. One, called to task by her dean, was ordered to stop that emphasis because of low satisfaction ratings. Weimer (2002) is refreshingly candid in revealing that effective learner-centered practices will not receive initial appreciation from students. She stops short of stating that this "resistance" may express itself through lowered global ratings of the faculty member who introduces them. Peter Sack's entered the "Generation X" college with the original intent of educating students, but, like the examples above, learned that it was safer to please students than to educate them. Student ratings were the lever that pressured Sacks to change his goals, and it is likely that faculty such as those in Thorn's (2003) study face similar pressure. Damage done to both education and to individuals through inept use of evaluation by forcing faculty to please "students as customers" is considerable. The assessment movement, particularly its emphasis on direct assessment of student learning, offers the most promising road out of misuse and abuse of student ratings (see Huba and Freed, 2000). The greatest reason that professors should embrace assessment (looking at the work that is done and the student learning that results) is to extricate themselves from the morass of having their livelihoods depend upon how others “feel” about them.

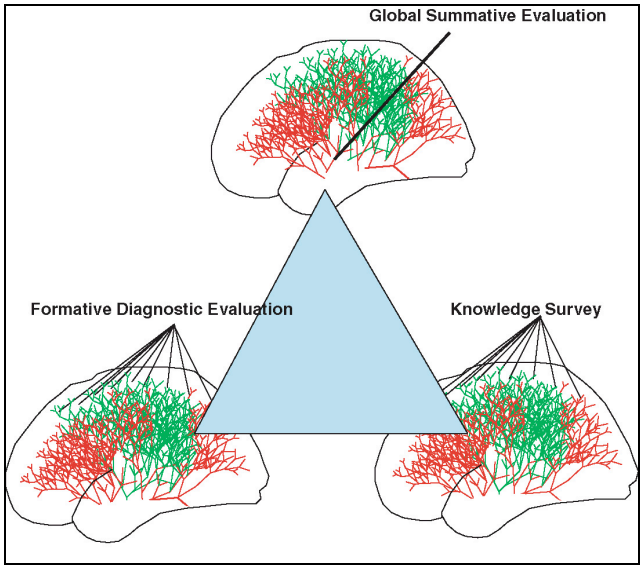

Figure 8. Suggested modern approach to evaluation of professors based on student input. Multiple measures of students responses should include (a) summative ratings interpreted as expressions of student satisfaction (not good or bad teaching), (b) formative data based on multiple measures sufficient to provide student confirmation of the profile of pedagogies teachers employ and (c) knowledge survey data that shows the detail of course content, intellectual challenge and students responses. The three: satisfaction, pedagogical practices, and content engagement provide data that bridges affective and cognitive responses. This no longer places professors' careers at the mercy of how well professors are liked.

It is better to produce evaluative tools that directly address attainment of specific educational goals (Nuhfer and Knipp, 2003; McKeachie, 1997) than to continue to rely solely on tools that produce relative ratings of professors based on an unspecific mix of outcomes and feelings. As faculty, we would never permit ourselves to grade students on the basis of how we “feel” about them, but we have come perilously close to accepting that very basis for evaluating professors.

Administer Ratings with more care.

If we simply send a student assistant into the class who passes out the forms and says little more than "Fill 'em out!" we are not administering any paper evaluation with care. In fact, we may be creating a "thin slice" perception of the evaluative process colored by a shoddy administering of the survey.

Because university - wide surveys may have questions that simply do not apply to particular classes, students really do need to hear cautions to leave questions blank that they believe may not apply to a particular class. Pitfalls arise in questions such as: "*Is the professor accessible to students outside of class?*" Only 10% of the students in many classes ever go to the professor's office for help. Students who have never been to the professor's office simply do not know whether the professor is actually available. In rating on a scale from 1 (poor) to 5 (outstanding), most students who are not cautioned about leaving responses blank unless they have first-hand knowledge of the question are prone to circle a "3" as their own expression of "I don't know." or "I really don't care much about this question." Of course, when the responses from 90% of students who don't have first-hand information are tabulated, their responses overwhelm those of the 10% who are furnishing solid information. In this way a professor who has kept all of his or her office hours and has perhaps even given out the home telephone and encouraged students to call will receive the same ratings as the person who abrogates all responsibility for being accessible. Students need to be told that if they do not have first-hand knowledge about a particular item, then they should leave the response blank.

Increase dialogue with students about the process of teaching and learning.

Global questions may evoke mainly satisfaction ratings, but satisfaction is not trivial. If students aren't satisfied with an experience, then the classroom will almost certainly be neither an inspiring nor satisfying experience for the teacher. It is possible to have both, but we cannot have both through simply evaluating teachers. Scriven (1994) provides some reasons to gather information from students: (1) The unique position of students as raters of their own learning; (2) The unique position of students as raters of changes in their motivation; (3) The unique position of students as raters of observable fact; (4) the unique position of students as raters of style indicators and (5) The position of students as raters of matters such as the face validity of tests.

The fact that student ratings are often the only dialogue about teaching with our students is a sad commentary, in general, on the nature of communication within higher education. Research by Cashin, Noma, and Hanna (1977) shows that some irreverence about such simplistic evaluation exercises is warranted. That infrequent student ratings are capable of replacing legitimate open dialogue is probably one of the most limiting concepts subscribed to by students, faculty and administrators. Ratings should be the result of interaction and dialogue throughout a course, not merely an exercise at the end. Ratings can serve as an ongoing basis to establish dialogue with student management teams (Nuhfer and others, 1990 - 2004). Such interactive discussions to improve teaching that involve students in responsibility for the quality of the classroom community are immensely rewarding.

Understand the statistics involved with evaluative ratings

Faculty frequently dismiss the low correlations between exam results and learning in the range of 0.5 or less as meaningless. Indeed, the ability to predict doesn't begin until passing r values of greater than 0.5 and is only a fair predictor at r-values of about 0.75. Prediction of one measure by another doesn't really get good until r-values achieved by correlating the two of surpasses r=0.9 (Jacobs and Chase, 1992).

An error faculty make in such judgments is a presumption tests are reliable but student ratings, knowledge surveys, etc. are not. Homegrown evaluative tools and routine tests used in classes have terrible records for reliability. The average faculty test probably has a reliability (Spearman-Brown measure; see Jacobs and Chase, 1992, p. 36) of about 0.3, which means that the average class test correlates with itself through method of split-halves only with a value of about r = 0.2. (Note: to my knowledge no one has compiled and published an average reliability measure for routine class tests, the 0.3 figure here came as result of verbal communication with Raoul Arreola on April 14, 2005. Arreola a well-respected psychometrician with many years' experience with tests and evaluations (see Arreola, 2006), kindly provided this value at my request for his estimate, after our panel discussion at the 2005 AERA Annual Meeting in Montreal. Based on his years of experiences, I believe this is a reasonable estimate from a good source.)

One cannot expect correlations between exams and other measures to be better than the internal reliability of tests with themselves. Further, obtaining good correlations depends not on there being simply a strong relationship in cause-effect but also that there be a wide scatter, ideally a normal distribution in the

data. For example, when faculty are highly successful, the entire class has high student ratings, high knowledge survey gains and high grades as result of exams and other evaluative measures. When faculty are unsuccessful, the class exhibits low student ratings, low knowledge survey gains and probably low grades as result of exams and other evaluative measures. Although the faculty influence on such outcomes is great, neither situation provides for the degree of scatter needed to deduce this; scatter plots much like those in Figure 9B are liable to result. Both situations produce correlations, but low correlations in such cases do not mean there was not a strong cause-effect relationship in what took place in the class. Thus, data from single classes is unlikely to generate meaningful insights, and a professor should not expect to get the results reported from meta analyses in which large data sets from many courses provide more scatter (and likely involved standardized tests with validated items tuned for optimum reliability) from her/his own tests and evaluations. The reader should understand that a reason for low correlations is not simply the result of lack of reliability in measures, such as student ratings, correlated with tests, but also with problems of low reliability with tests and grades themselves and with the nature of data generated in routine class practices.

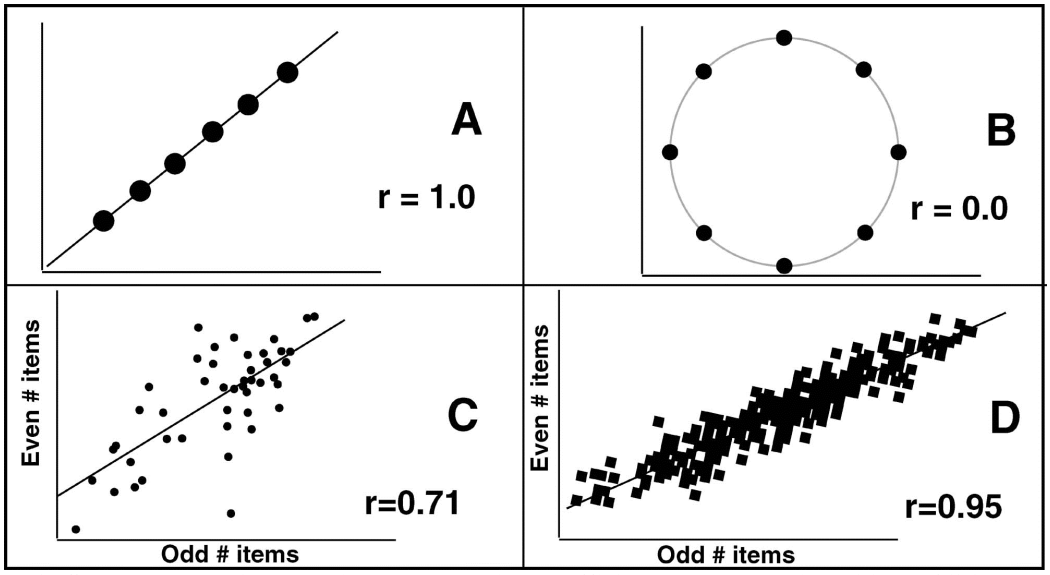

Figure 9. Scatter plots with associated correlation coefficients show predictability. "A" is a perfect correlation between two variables; r= 1.0. It has perfect predictability of one variable when the other is measured. "B" is zero correlation. "C" is from test data in a freshman course that yields r=0.71. The scatter is apparent but the pattern is elongate. One can see by picking any individual point at random that predictability is only fair at best. "D" is from a knowledge survey data from nearly 300 freshmen. The scatter pattern is very elongate and yields an r-value of 0.95. Predictability for nearly all individuals is very good

Summary

1) Summative student ratings are valid means of assessing student satisfaction. Student satisfaction is important and must be addressed. At the same time, student ratings should never be the sole criteria for rating any professor's teaching effectiveness. Ratings of student satisfaction are not measures of learning, and student ratings are not assessment tools suited to deduce student learning outcomes.
2) Summative student ratings do not look directly or cleanly at the work being done. They are mixtures of affective feelings and learning. Formative ratings look directly at the pedagogical work that is done; they reveal the practices being used and degree to which each is being used. To learn if a teacher is improving, it is better to look for increased employment of practices that the research has proven to be valuable than it is to look only at increases in higher summative ratings.
3) Correlations established on large populations, even those relationships that have been proven as "valid" and "reliable," cannot be safely applied as a tool to judge individuals. The validity and reliability exist only for large enough populations to produce them. Because validity and reliability are not discernable from studies of small populations, it is even more perilous to project such correlations onto individuals.
4) To use student ratings to evaluate individuals without some direct measure of student learning omits the most important outcome of the educational process. Knowledge surveys can help fill this gap.

Recent emphases on assessment of student learning outcomes underscores that summative rating of faculty contributes nothing to improvement of student learning.

5) Successful evaluation programs demand careful design. Helpful characteristics include:
    a. a well-designed evaluation tool
    b. a student populace that is educated about the pitfalls of such surveys
    c. campus-wide awareness of meaning, validity, and effects of ratings
    d. a mechanism that prevents abuse of evaluation results
    e. a mechanism that supports attempts by faculty to improve.
6) Self-esteem and enthusiasm are important traits for successful teaching. A university that aspires to excel in teaching can't afford policies or administrators that damage these traits in faculty. Inept evaluation will damage faculty morale institution-wide.
7) In recognizing the complexity of student-teacher relationships that occur in our classes, regular communication with our students about teaching is essential for continuous improvement. Just as Edwards Deming noted that products cannot be improved by inspections at the end of production, quality of teaching cannot be improved by final summative ratings at the end of courses. To gain improvements, formative assessment is essential and summative ratings are ineffective.
8) Professors and administrators need to become familiar with the body of literature that addresses learning, teaching, evaluation and assessment.
9) Knowing one's student audience and making public to students a clear set of learning objectives is the professor's best defense against producing substandard results (in both satisfaction and learning).
10) No professor needs to be stuck for life with poor student ratings. Means exist through which a professor can (1) become a more effective teacher and (2) without pandering, can raise his or her student ratings. The problem is primarily finding out the critical areas to focus upon given the complexity of the teaching environment. Consultation with developers, other faculty, or with student management teams can help in getting this focus.
11) The cases described in the Dr. Fox experiments and *Generation X Goes to College* are real. Summative evaluation can be manipulated or subverted by deliberate efforts to do so.
12) Fractal thinking and the neural nature of learning explains much about education and provides valuable insights about faculty evaluation and the nature of teaching and learning

References

Abrami, P. C., d'Apollonia, S., and Rosenfield, S., 2007, The dimensionality of student ratings and instruction: What we know and what we do not. in R.P. Perry and J.C. Smart, (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Approach*. Springer, 385-546.

Abrami, P. C., Leventhal, L., and Perry, R. P., 1982, Can feedback from student ratings help improve college teaching?: 5th Intnl. Conf. on Improving University Teaching, London.

Abrami, P. C., Rosenfield, S. and Dedic, H. (2007). The dimensionality of student ratings and instruction: An update on we know, do not know, and what we need to do. in R.P. Perry and J.C. Smart, (eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Approach*. Springer, 446-456.

Ambady, N. and R. Rosenthal, 1993, Half a minute: predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness: Journal of Personality and Social Psychology, v. 64, pp. 431-41.

Arreola, R., 2006, *Developing a Comprehensive Faculty Evaluation System: A Handbook for College Faculty and Administrators on Designing and Operating a Comprehensive Faculty Evaluation System, Third Edition*: Bolton, MA, Anker.

Basow, S. A., and Silberg, N. T., 1987, Student evaluations of college professors: are female and male professors rated differently?: Jour. Educ. Psychology, v. 79, pp. 308-314.

Berk, R. A., 2006, *Thirteen Strategies to Measure College Teaching*. Sterling, VA: Stylus.

Boex, L. (2000). Identifying the attributes of effective economics instructors: An analysis of student evaluations. Journal of Economic Education, 31(3), 211-227.

Boice, R., 1990, Countering common misbeliefs about student ratings of teaching: Teaching Excellence, v. 2, n. 2.

Bransford, J., Brown, A., & Cocking, R. (Eds.) (2000). How people learn: Brain, mind, experience, and school. Washington, DC: National Academy Press.

Braskamp L. A., Brandenberg, D. C., and Ory, J. C., 1984, Evaluating teaching effectiveness: a practical guide: Sage Pub., Beverly Hills, CA.

Cashin, W. E. (1988). Student ratings of teaching: a summary of the research: Kansas State Univ. Center for Faculty Evaluation and Development, Idea Paper n. 20. Available March 31, 2008 from http://www.idea.ksu.edu/resources/Papers.html.

Cashin, W. E., 1990, Student ratings of teaching: recommendations for use: Kansas State Univ. Center for Faculty Evaluation and Development, Idea Paper n. 22.
Cashin, W. E., 1995, Student ratings of teaching: the research revisited: Kansas State Univ. Center for Faculty Evaluation and Development, Idea Paper n. 32.
Cashin, W. E., and Slawson, H. M., 1977, Description of data base 1976 - 1977: IDEA Technical Report n. 2, Kansas State Univ. Center for Faculty Evaluation and Development.
Cashin, W. E., Noma, A., and Hanna, G. S., 1977, Comparative data by academic field: IDEA Technical Report n. 4, Kansas State Univ. Center for Faculty Evaluation and Development.
Centra, J. A., and Gaubatz, N. B., 1998, Is there gender bias in student ratings of instruction?: Journal of Higher Education, v 70, pp 17-33.
Cohen, 1980, Effectiveness of Student-Rating Feedback for Improving College Instruction: Research in Higher Education, v. 13, pp. 321 -341.
Cohen, P. A., 1981, Student ratings of instruction and student achievement: a meta-analysis of multisection validity studies: Review of Educ. Res., v. 51, pp. 281 - 309.
d'Apollonia, S., and Abrami, P. C., 1997, Navigating student ratings of instruction: American Psychologist, v. 52, pp. 1198-1208.
Davis, G. A., and Thomas, M. A., 1989, *Effective Schools and Effective Teachers*: Boston, MA, Allyn and Baker, 214 p.
Dunkin, M. J., and Barnes, J., 1986, Research on teaching in higher education: in *Handbook of Research on Teaching*, M. C. Wittrock, ed., pp. 754 - 777.
Eder, D. (November, 2006). Assessment vs. evaluation: Dealing with the differences. Paper presented at the 2006 Assessment Institute, Indianapolis, IN.
Erdle, S., and Murray, H. G., 1986, Interfaculty differences in classroom teaching behaviors and their relationship to student instructional ratings: Research in Higher Education, v. 24, n. 2, pp. 115 - 127
Feldman, K. A., 1983, Seniority and experience of college teachers as related to evaluations they receive from students: Research in Higher Educ., v. 18, pp. 3 - 124.
Feldman, K. A., 1986, The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: a review and synthesis: Research in Higher Educ., v. 24, pp. 129 - 213.
Feldman, K. A., 1987, Research productivity and scholarly accomplishment of college teachers as related to their instructional effectiveness: Research in Higher Educ., v. 26, pp. 227 - 298.
Feldman, K. A., 1989, Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators and external (neutral) observers: Research in Higher Educ., v. 30, pp. 137-194.
Feldman, K. A., 1992, College students' view of male and female college teachers: Part I—Evidence from the social laboratory and experiments: Research in Higher Educ., v. 33, pp. 317 - 375.
Feldman, K. A., 1998, Identifying exemplary teachers and teaching: evidence from student ratings: in *Teaching and Learning in the College Classroom* 2nd ed., K. A. Feldman and M. B. Paulsen, eds., Needham Heights, MA, Simon & Schuster, pp. 391-414.
Feldman, K. A. (2007), Identifying exemplary teachers and teaching: evidence from student ratings. in R.P. Perry and J.C. Smart, (Eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Approach. Springer, 93-129.
Fish, Stanley, 2005, Who's in Charge Here?: The Chronicle of Higher Education, February 4, 2005, available at http://chronicle.com/jobs/2005/02/2005020401c.htm.
Glenn, D., 2007, Method of Using Student Evaluations to Assess Professors Is Flawed but Fixable, 2 Scholars Say: The Chronicle of Higher Education, Tuesday, May 29.
Greenwald, A. G., 1997, Validity concerns and usefulness of student ratings of instruction: American Psychologist, v. 52, pp. 1182-1186.
Gravestock, P. and Gregor-Greenleaf, E., 2008, Student Course Evaluations: Research, Models and Trends. Toronto: Higher Education Quality Council of Ontario.
Hamermesh, D. S., and Parker, A. M., 2004, Beauty in the classroom: Professorial pulchritude and putative pedagogical productivity," Economics of Education Review, 2005 (pdf version available on 7/ 11/04 through Hamermesh's University of TX personal web site accessible through http://www.utexas.edu/.
Hildebrand, M., Wilson, R. C., and Dienst, E. R., 1971, Evaluating university teaching: Handbook published by Center for Research and Development in Higher Education, U of CA, Berkeley.
Hines, C. V., Cruickshank, D. R., & Kennedy, J. J. (1985). Teacher clarity and its relationship to student achievement and satisfaction. American Educational Research Journal, 22, 87-99.
Howard, G. S., and Maxwell, S. E., 1982, Do grades contaminate student evaluations of instruction?: Research in Higher Educ., v. 16, pp. 175 - 188.
Huba, M. E., and Freed, J. E., 2000, *Learner-Centered Assessment on College Campuses: Shifting the Focus from Teaching to Learning*: Boston, MA, Allyn and Bacon, 286 p.
Jacobs, L. C., and Chase, C. I., 1992, *Developing and Using Tests Effectively: A Guide for Faculty*: San Francisco, Jossey-Bass, 231 p.

Joint Committee: The California State University, California Faculty Association and Academic Senate CSU, (2008). Report on Student ratings of Teaching. Long Beach, CA: 13 p.
King and Kitchener, K, 1994, *Developing Reflective Judgment*: San Francisco, Jossey-Bass, 323 p.
Langley, D., and others. (2007). Recommendations on modifying the student evaluation of teaching form at the University of Minnesota. Final Report of the Ad Hoc SCEP/SCFA Subcommittee, March 1, 2007, Revised report November 8, 2007.
Leamnson, R., 1999, *Thinking About Teaching and Learning: Developing Habits of Learning with First Year College and University Student:* Sterling, VA: Stylus, 169p.
Light, R. (2001). Making the most of college: Students speak their minds. Cambridge, MA: Harvard University Press.
Marsh, H. W., 1982, The use of path analysis to estimate teacher and course effects in students' ratings of instructional effectiveness: Applied Psychological Measurement. v. 6, pp. 47 - 59.
Marsh, H. W., 1983, Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics: Jour. Educational Psychology, v. 75, pp. 150-166.
Marsh, H. W., 1984, Students' evaluations of university teaching: dimensionality, reliability, validity, potential biases and utility: Jour. Educ. Psyc., v. 76, pp. 707 - 754.
Marsh, H. W., and Ware, J. E., 1982, Effects of expressiveness, content coverage, and incentive on multidimensional student rating scales: new interpretations of the Dr. Fox effect: Jour. Educ. Psyc., v. 74, pp. 126 - 134.
McKeachie, W. J., 1997, Student ratings—the validity of use: American Psychologist, v. 52, pp. 1218-1225.
McKeachie, W. J. 2007, Good teaching makes a difference—and we know what it is. in R.P. Perry and J.C. Smart, (eds.), The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Approach. Springer, 457-474.
McKeachie, W. J., and Kaplan, M., 1996. Persistent problems in evaluating college teaching. AAHE Bulletin, February 1996, pp 5-8.
McKeachie, W. J., and others, 1994, *Teaching Tips - A Guidebook for the Beginning College Teacher (9th ed.)*: Lexington, MA, D. C. Heath, 444 p.
Murray, H. G., 1985, Classroom behaviors related to college teaching effectiveness: in Using Research to Improve Teaching, J. G. Donald and A. M. Sullivan, eds., San Francisco, Jossey-Bass.
Naftulin, D. H., Ware, J. E., and Donnelly, F. A., 1973, The Doctor Fox lecture: a paradigm of educational seduction: Jour. Medical Educ., v. 48, pp. 630 - 635.
Nerger, J. L., Volbrecht, V. J., and Ayde, C. J., 1995, Student ratings of teaching effectiveness: Use and misuse: The Midwest Quarterly, v. 38, pp. 218-233
Nuhfer, E. B., 1993, Bottom-line disclosure and assessment: Teaching Professor, v. 7, n. 7, p. 8.
Nuhfer, E. B., 1996, The place of formative evaluations in assessment and ways to reap their benefits: Jour. Geoscience Education, v. 44, n. 4, pp 385-394.
Nuhfer, E. B., 2003a, Developing in fractal patterns I: Moving beyond diagnoses, evaluations and fixes: National Teaching and Learning Forum, v. 12, n. 2, pp. 7-9.
Nuhfer, E. B., 2003b, Developing in fractal patterns II: A tour of the generator: National Teaching and Learning Forum, v. 12, n. 4, pp. 9-11.
Nuhfer, E. B., 2003c, Content coverage, courses, and controversy Part 1: Developing in Fractal Patterns V: National Teaching and Learning Forum, v. 13, n. 1, pp. 8-10.
Nuhfer, E. B., 2004a, Fractal thoughts on the forbidden affective in teaching evaluation & high level thinking: Educating in Fractal Patterns X: National Teaching and Learning Forum, v. 14, n. 1, pp. 9-11.
Nuhfer, E. B., 2004b, Why Rubrics?: Educating in Fractal Patterns IX: National Teaching and Learning Forum, v. 13, n. 6, pp. 9-11.
Nuhfer, E. B., 2004c, Student management teams: Fractals for Students Too—Developing in Fractal Patterns VII: National Teaching and Learning Forum, v. 13, n. 4, pp. 7-11.
Nuhfer, E. B., 2005a, Fractal views on good testing practices: Educating in Fractal Patterns XII: National Teaching and Learning Forum, v. 14, n. 4, pp. 9-11.
Nuhfer, E. B., 2005b, Tests as anchors that wobble: Understanding imperfect correlations in educational measurements: Educating in Fractal Patterns XI: National Teaching and Learning Forum, v. 14, n. 2, pp. 8-11.
Nuhfer, E. B., 2005c, De Bono's red hat on Krathwohl's head: Irrational means to rational ends— More fractal thoughts on the forbidden affective: Educating in Fractal Patterns XIII: National Teaching and Learning Forum, v. 14, n. 5, pp. 7-11.
Nuhfer, E. B., 2007, The ABCs of Fractal Thinking in Higher Education: To Improve the Academy, v. 25, pp. 71-89.
Nuhfer, E. B., 2008, A fractal thinker looks at student ratings. An updated version of paper delivered at the 2005 AERA Meeting in Montreal in Theall, M., Abrami, P.C., Arreola, R., Franklin, J., Nuhfer, E., and Scriven, M. (2005). Valid faculty evaluation data: Are there any? AERA Annual Meetings Program Interactive Panel Presentation, American Educational Research Association Symposium, Montreal: April 14, 2005. 240. (summaries available at http://www.cedanet.com/meta/AERA2005valid.pdf and http://profcamp.tripod.com/fractalevals07.pdf)

Nuhfer, E. B., 2008, The affective domain and the formation of the generator: Educating in fractal patterns XXIII part 1: National Teaching and Learning Forum, 18/2 8-11. and part 2, National Teaching and Learning Forum, 18/3 9-11. Available from any computer on the CSUCI campus at http://www.ntlf.com/restricted/.

Nuhfer, E. B., and Adkison, S., 2003, Developing in fractal patterns IV: Unit level development -Teaching philosophies at the unit level: National Teaching and Learning Forum, v. 12, n. 6, pp. 4-7.

Nuhfer, E. B. and Dewar, J., 2008,"Guidelines for Creating Good Items for Student Ratings of Professors." SoCal Faculty Developers Learning Community. Los Angeles, CA: http://www.lmu.edu/pagefactory.aspx?PageID=41603.

Nuhfer, E. B., and Knipp, D., 2003, The knowledge survey: a tool for all reasons**:** To Improve the Academy, V. 21, pp. 59-78 (available at http://www.isu.edu/ctl/facultydev/KnowS_files/KnowS.htm).

Nuhfer, E. B., and others, 1990-2002, *A Handbook for Student Management Teams*: Center for Teaching and Learning, Idaho State University, 60 p.

Nuhfer, E. B., and others, 1992, Involve your students in improving their teaching and learning community: 12th Annual Lilly Conference on College Teaching: The Greening of the Future: Oxford, Ohio, pp. 347 -350.

Nuhfer, E. B., and Pavelich, M., 2001, Levels of thinking and educational outcomes: National Teaching and Learning Forum, v. 11, n. 1, pp. 5-8.

Nuhfer, E. B., and Pavelich, M., 2002, Using what we know to promote high level thinking outcomes: National Teaching and Learning Forum, v. 11, n. 3, pp. 6-8.

Nuhfer, E. B., Bleicher B., Adams, V., Buchanan M., Elliott, J., Furmanski, M., Christopher, R., Smith, P., and Wood, G., (2008a), Final report of the task force to create a new student ratings form for California State University Channel Islands, 37 p.

Nuhfer, E. B., Bleicher B., Adams, V., Buchanan M., Elliott, J., Furmanski, M., Christopher, R., Smith, P. ,Wood, G., and Baker, H. (2008b). Poster: The Role of Faculty Development in Designing a Mission-based Student Ratings Instrument. North American Council for Staff, Program and Organizational Development & Professional and Organizational Development Network in Higher Education Joint Conference. Reno, NV. Oct 23, 2008.

Nuhfer, E. B., Krest, M., and Handelsman, M., 2003, Developing in fractal patterns III: A guide for composing teaching philosophies: National Teaching and Learning Forum, v. 12, n. 5, pp. 10-11.

Nuhfer, E. B., Leonard, L., and Akersten, S., 2004, Content coverage, courses, and controversy Part 2: Developing in Fractal Patterns VI: National Teaching and Learning Forum, v. 13, n. 2, pp. 8-11.

Pascarella, E. T., 2001, Cognitive growth in college: Change, v. 33, n. 1, pp. 21-27

Pascarella, E. T., Seifert, T. A., and Whitt, E. J., (2008). The role of classroom instruction in first-year persistence in college: Some new evidence. New Directions in Teaching and Learning, Jossey-Bass, in press.

Perry, R. P., Abrami, P. C., and Leventhal, L., 1979, Educational seduction: the effect of instructor expressiveness and lecture content on student ratings and achievement: Jour. Educ. Psyc., v. 71, pp. 107 - 116.

Perry, R. P., and Smart, J. C., (Eds), 2007, *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*. Secaucus, NJ, Springer, 813 p.

Perry, W. G. Jr., 1999, *Forms of Ethical and Intellectual Development in the College Years*: San Francisco, CA: Jossey-Bass (a reprint of the original 1968 work with minor updating).

Rheinberg, F., Vollmeyer, R., and Rollett, W., (2005). Motivation and action in self-regulated learning. in Handbook of Self-Regulation. Monique Boekarts, Paul Pintrich and Moshe Zeidner (eds.) Elsevier, 503-529.

Rhem, J., 2008, The affective field. National Teaching and Learning Forum, 17/2 4-5.

Rhem. J., 2009, EVALUATION: Rate My Professor & Professors Strike Back: National Teaching and Learning Forum, 18/3 5-7.

Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. Contemporary Educational Psychology, 25, 54-67.

Sacks, P., 1996*, Generation X Goes to College*: Chicago IL, Open Court Pub., 208 p.

Sanders, J. A., and Wiseman, R. L., 1998, the effects of verbal and nonverbal teacher immediacy on perceived cognitive, affective, and behavioral learning in the multicultural classroom:: in *Teaching and Learning in the College Classroom* 2nd ed., K. A. Feldman and M. B. Paulsen, eds., Needham Heights, MA, Simon & Schuster, pp. 455-466.

Schonwetter, D., Menec, V., & Perry, R. (1995). An empirical comparison of two effective college teaching behaviors: Expressiveness and organization. Paper presented at the annual meeting of the American Educational Research Association, San Francisco, CA: April, 1995.

Scriven, M., 1997, Student ratings offer useful input to teacher evaluations: ERIC/AE Digest, http://ericae.net/db/digs/ed398240.htm.

Seldin, P., 1993, The use and abuse of student ratings of professors: The Chronicle of Higher Education, v. 39, n. 46, p. A 40.

Smith, M., and Glass, G., 1980, Meta-analysis of research on class size and its relationship to attitudes and instruction: Amer. Educ. Research Jour., v. 17, pp. 419 - 433.

Stevens, J. J., and Aleamoni, L. M., 1985, The use of evaluative feedback for instructional improvement: a longitudinal perspective: Instructional Science: v. 13, pp. 285 - 304.

Svinicki, M., 2008, When does enough feedback become too much? National Teaching and Learning Forum, 17/3 12.
Theall, M., Abrami, P. C., Arreola, R., Franklin, J., Nuhfer, E., and Scriven, M., 2005,Valid faculty evaluation data, are there any?: American Educational Research Association Symposium, Montreal, April 14.
Theall, M., Abrami, P.C., and Mets, L. M., (eds.), 2001, *The student ratings debate: Are they valid? How can we best use them?*: San Francisco, Jossey-Bass, New Directions for Institutional Research, n. 109.
Theall, M., and Feldman, K. A., 2007, Commentary and update of Feldman's (1997) "Identifying exemplary teachers and teaching: evidence from student ratings." in R.P. Perry and J.C. Smart, (Eds.), *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Approach*. Springer, 130-143.
Theall, M., and Franklin, J. (eds.), 1990, *Student Ratings of Instruction: Issues for Improving Practice*: San Francisco, Jossey-Bass, New Directions for Teaching and Learning, n. 43, 135 p.
Thorn, P. M., 2003, Bridging the gap between what is praised and what is practiced: supporting the work of change as anatomy & physiology instructors introduce active learning into their undergraduate classrooms: Austin, Texas, University of Texas, PhD dissertation, 384 p.
Ware, J. E., and Williams, R. G., 1975, The Dr. Fox effect: a study of lecture effectiveness and ratings of instruction: Jour. Medical Educ., v. 50, pp. 149 - 156.
Weimer, M., 2002, *Learner-centered Teaching: Five Key Changes to Practice*: San Francisco, Jossey-Bass, 258 p.
Williams, W. M., and Ceci, S. J., 1997, How' M I doing?: Change, v. 29, n. 5, pp. 12-24.
Wilson, R., 1998, New research cast doubt on the value of student evaluations of professors: Chronicle of Higher Education, v. 44, n. 19, pp. A12-A15.
Wirth, K. R., Perkins, D., and Nuhfer, E. B. ,2005, Knowledge surveys: A tool for assessing learning, courses, and programs (abs.): 2005 Geological Society of America Annual Meetings Program, October 14-20 Salt Lake City Utah.