Education Policy Analysis Archives

Volume 5 Number 18 **August 25, 1997** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal.

Editor: Gene V Glass Glass@ASU.EDU.

College of Education

Arizona State University,Tempe AZ 85287-2411

Academic Freedom, Promotion, Reappointment, Tenure And The Administrative Use of Student Evaluation of Faculty (SEF): *(Part III)* Analysis And Implications of Views From The Court in Relation to Accuracy and Psychometric Validity

Robert E. Haskell [1]

University of New England

This is the third of four articles by Haskell on this subject. The other articles can be found at

- Volume 5 Number 6
- Volume 5 Number 17
- Volume 5 Number 21

Abstract: In two previous papers, it was noted that while a controversial history of research on the reliability and validity of student evaluation of faculty (SEF) exists, it has not been typically viewed as an infringement on academic freedom, promotion, reappointment, and tenure rights. As a consequence, legal aspects of SEF are neither readily apparent, nor available. Legal rulings, their implications and assumptions in relation to their accuracy and psychometric validity where SEF are integral to the denial of academic freedom, tenure, promotion, and reappointment are reviewed along with the legal principles of Disparate

Treatment and Disparate Impact, and the scientific Precautionary Principle in policy decisions.

Table of Contents

.....As I indicated in previous papers on SEF (Haskell, 1997a, 1997b), the history of legal rights demonstrates that issues not considered to have legal standing only come to have legal standing after a long process of advocacy. The evolution of a policy or legal principle requires the accumulation of data, coalescing judgements and arguments. To this end, this paper, will continue to examine court reasoning and rulings on SEF in cases involving the denial of academic freedom, tenure, promotion, and reappointment (AFTPR) decisions in relation to its implications and assumptions regarding accuracy and psychometric validity.[2]

.....In a second paper on SEF, (Haskell, 1997b), I abstracted from the text of located legal cases views from the court pertinent to SEF. The appendix of the second paper provided a verbatim abstracting of the text of each case relative to its SEF content. As a consequence, in summarizing the pertinent findings of that paper for the present one, for convenient referencing the specific case textual material for each section will be in placed in a footnote indicated at the beginning of each section heading and the indented "Summaries" are carried over from Part II.

For convenience, I will use these abstracted legal views and rulings to examine their implications for the courts use of SEF in relation to their accuracy and psychometric validity. A final paper will address the implications of court rulings for academic freedom and instruction. As I noted in my second paper, not only are legal cases prima facially complex, but when specific legal definitions (e.g., disparate treatment and impact) and other special Congressional Acts (e.g., EEOC) are superimposed on them, they become logically unwieldy, not just to the non legal scholar, but apparently to the Courts as well.[3]

.....Finally, I would like to point out that the issues examined in this series of papers are not primarily concerned with individual faculty rights but with the implications of SEF when used for administrative purposes on academic freedom, educational quality, standards, and ultimately on the competence of graduates.[4]

Brief Overview of the Validity of SEF

.....As shown in Haskell (1977b) and reiterated below, views from the court on the appropriate use of SEF vary so greatly that the concept of variation might more descriptively be replaced by the concept of "randomness" were it not for the fact that there has been a consistent trend by the courts to accept SEF data as it is presented to them by institutions. In presenting an analysis and implications of these views from the court in relation to their validity, the detailed research literature on the validity of SEF will largely have to be bracketed. To do otherwise would take this article to far afield. Nevertheless, because the issue of validity is so central to this paper, an overview of the SEF validity literature is a necessary foundation for the following analyses.

.....There is a long and controversial research history on SEF, with most early reviews and extant opinion---though certainly not all---suggesting their general validity, with validity referring to the accuracy of SEF measuring teaching effectiveness. More recently, however, sophisticated statistical reviews of this past literature strongly suggest that earlier reviews of SEF literature were not rigorously analyzed and controlled methodologically, thus casting serious doubt on their validity. As Barnett (1996), Greenwald (1997), Greenwald and Gillmore (1996) demonstrate, past reviews have tended to not be sophisticated critiques. Positions, suggesting cautious support for validity of SEF while at the same time expressing concerns about the adequacy of their support, include, Abrami, Dickens, Perry, & Leventhal (1980). Reviews and empirical critiques that are critical of the validity of SEF include, Chacko (1983), Dowell and Neal (1982), Powell (1977), Snyder and Clair (1976), Vasta and Sarmiento (1979), and Worthington and Wong (1979). Some of the past reviews that have categorized the significant research that have found SEF to be essentially a valid measure of quality of instruction are: Cashin (1995), Cohen (1981), Franklin & Theall (1990), Holmes (1972), Howard, Conway, and Maxwell (1985), Howard and Maxwell (1980, 1982), Marsh (1980, 1982, 1984), Marsh and Dunkin (1992), and McKeachie (1979).

.....Cahn (1987) suggests that student ratings do not measure the instructional effectiveness or the intellectual achievement of students. SEF measure student satisfaction, attitudes toward instructors course, student personality, and their psychosocial needs. Cahn further suggests, students know if instructors are likeable, not if they are knowledgeable; they know if lectures are enjoyable, not if they are reliable. In a meta-analysis Cohen (1983)---who basically accepts the validity of SEF---concludes from his study, "While the magnitude of the average rating/achievement correlation for the thirty-three multisection courses is not overwhelming [14.4% of shared variance between ratings and the criteria], the relationship is

certainly stronger and more consistent than we were led to believe..." (p. 455). And Dowell & Neal (1982) conclude that

> "The research literature can be seen as yielding unimpressive estimates of the validity of student ratings. At their most valid, then, validity of SEF refers to only 14% of the total variance. The literature does not, therefore, support claims that the validity of student ratings is a consistent quantity across situations. Rather, the evidence suggests that the validity of student ratings is modest at best and quite variable...The variability in obtained validity coefficients even in studies with reasonable methodological requirement lead us to suspect that the validity of student ratings is influenced by situational factors to such an extent that a meaningful, generalizable estimate of their validity does not exist. In general . . .no meaningful estimate of the validity of student ratings can be provided with confidence that is generalizable enough to be useful..." (59-61).

For example, studies demonstrate the following confounding variables: (1) Age, (2) gender, (3) class size, (4) year of student, (5) level of student, (6) instructor style, (7) subject matter, (8) major or elective course, (9) student interest in subject matter, (10) instructor grading difficulty, (11) anonymous v.s signed ratings, (12) whether students are informed of their use, (13) instructor present v.s instructor absent while completing the evaluation (see for example, Divoky and Rothermel, 1988), (14) length of class period, and a host of other variables. .....Finally, the philosopher of science, Michael Scriven who has conducted rigorous work on evaluation procedures, (1995, 1993, 1991, 1988), particularly on the justification of inferring from ratings to conclusions about the merit of teaching on the basis of statistical correlations between ratings and student learning gains. He suggests that such inferences are invalid, unless a number of stringent conditions are met on the design, administration, and use of such ratings. He further suggests of faculty evaluation in general that, "All are face-invalid and certainly provide a worse basis for adverse personnel action than the polygraph in criminal cases. Based on examination of some hundreds of forms that are or have been used for personnel decisions (as well as professional development), the previous considerations entail that not more than one or two could stand up in a serious hearing." Given this highly questionable state of affairs on the validity of SEF, the question is how do courts view validity in relations to its use for administrative purposes?

.

The Courts' Approach to the General Accuracy and Psychometric Validity of SEF [5]

.....An issue directly related to the reliance on SEF for administrative purposes is its validity. Presumably the more valid SEF data in a given case, the more justifiable is the reliance on it for administrative purposes. From the legal cases reviewed (in Haskell, 1997b), it is clear that the courts tend to accept SEF data as presented to them by institutions.

> *Summary:* With regard to requiring the general and statistical accuracy of SEF, legal reasoning and rulings can be summarized (see Haskell, 1997b) as ranging from: (1) accepting statistical analyses as a part of a plaintiff's effort to establish discriminatory treatment if it reaches proportions comparable to those in cases establishing a prima facie racial discrimination, (2) cautioning that statistics are not irrefutable, with their usefulness depending on surrounding facts and circumstances of a case, (3) maintaining that the court need not consider validity and is under no obligation to establish the accuracy of administrative interpretations of SEF, (4) that tenure criteria are not drawn with "mathematical nicety," (5) administrator's failure to perform statistical comparisons is not

> arbitrary and is reasonable, (6) especially if such is not required by a Faculty Association Contract, (7) nearly any use made of SEF, regardless of its validity, is acceptable if it followed the standard practice of the university, (8) that creativity, rapport with students and colleagues, teaching ability, and other qualities are intangibles which cannot be measured by objective standards.

Some courts (e.g., *Fields V. Clark University*, 1987) have noted even when SEF are not gathered and evaluated according to accepted standards of scientific polling procedures it is nevertheless acceptable if the process followed standard practice involved in other tenure decisions at the university (p.671).

.....While there does exist a "substantial evidence" standard which gauges whether an institution's decision-making body carefully considered the evidence and had a substantial body of evidence on which to base its decision, and an "arbitrary and capricious" standard which gauges whether a deciding body acted without reason or irrationally, (See Kaplin,1995, section 1.4.3.6. Standards of Judicial Review and Burdens of Proof 35), it appears these standards are frequently ignored in relation decisions based on SEF.

.....In general, the exception to the courts almost total disregard for the validity of SEF has been in cases involving EEOC issues. In such cases, the courts require precise accuracy. I will address this issue in more detail in the section of disparate treatment and impact below.

Historical Overview of the Courts' Approach to the Validity of Faculty Evaluation Data

.....As noted previously, unlike general performance evaluations of faculty, SEF does not have a categorical legal history. Since SEF is but a subset of faculty performance evaluation in general, it is appropriate to briefly review the history of this more general area. Given SEF as a subset of faculty performance in general, it is accordingly not surprising to see that the view from the courts on the validity of SEF parallels that of the courts view of faculty performance evaluation.

.....Historically, in terms of faculty evaluation instruments in general, (on both secondary and postsecondary levels) it is widely agreed by legal scholars (Baez, Benjamin, and Centra, 1995) that "Despite the subjectivity of measuring the quality of a faculty member's scholarship, service and teaching accomplishments, courts will rarely, if ever, question the appropriateness of an institution's criteria (or how they measure them) for granting reappointment, promotion, or tenure....they will rarely substitute their judgments for those of peer review committees....Although juries may have less deference" (p.139). It might also be added that courts will seldom question administrative judgements of evaluations. It seems that faculty who challenge institutional evaluation tools very rarely succeed. Although the legal "competent and substantial evidence" standard places a significant burden of proof on the educational organization, it has not generally required that faculty assessment instruments are professionally validated (Rebell, 1990; Kaplin and Lee, 1995). Such rulings do, however, appear to vary by state or federal jurisdiction.

.....Psychometric standards of validity, reliability, and specific evaluation techniques, are rarely incorporated in state laws, regulations, or common-law standards. Accordingly, cases that involve evaluation have tended to focus on adherence to specific procedural requirements as set forth in state law or on general common-law notions of fairness and due process, not on expert psychometric standards. Although state courts will require strict adherence to the procedural aspects of these requirements and will strike down an arbitrary failure to use any apparent evaluative criteria, the state courts tend not to probe the substance of evaluation criteria or methods (Rebell, 1990; Kaplin and Lee, 1995). As Copeland and Murry (1996) have put it, "the judiciary has generally behaved as though it believed that

evaluations were made only after careful deliberation and with procedural due process protections. In short, the judiciary has tended to act as if colleges and universities could be trusted to act in good faith" (p.246).

.....Rebell (1990) outlines what he describes as a "striking example of the courts' traditional deferential attitude toward teacher evaluation" data (p.337). The decision of the United States Court of Appeals for the Eighth Circuit in Scheelhaase v. Woodbury Central Community School District (1973), involved the dismissal of a teacher whose contract had previously been renewed over a ten-year period. The reason for her termination was that she was incompetent as indicated by the low scores of her students on the Iowa Test of Basic Skills (ITBS) and the Iowa Test of Educational Development (ITED). Despite the a number of expert witnesses testifying that it was inappropriate to use such test scores as a basis for evaluating a teacher performance, the court dismissed Scheelhaase's claim. The claims were considered basically irrelevant by the court because "such matters as the competence of teachers and the standards of its measurement" are not matters of constitutional dimension.

.....This early case involving a public school teacher is significant both because (a) of the Court's apparent lack of concern with the serious psychometric issues raised by a reliance on student achievement scores as a sole stated basis for termination and (b) because of the Court's almost total reliance on a school administrator's psychometrically unsubstantiated, and quite possibly equally erroneous evaluation. One of the concurring Scheelhaase case judges bluntly stated:

> The Board was entitled to rely upon the recommendation of conclusions of its superintendent, not-withstanding the existence of strong opinions contrary to his regarding the use of the ITBS or ITED tests as a tool of Leacher evaluation...Thus, its decision, even though premised upon an apparently erroneous 'expert opinion 'cannot be faulted as arbitrary and capricious. The Board's mere mistake in judgment or in weighing the evidence does not demonstrate any violation of substantive due process. (Emphasis added).

Thus, even when states use student achievement scores as an index of faculty proficiency, [6] courts have had an "apparent lack of concern with serious psychometric issues raised by reliance on student achievement scores as a sole stated basis for termination," again, relying on administrator's unsubstantiated evaluations (Rebell, 1990). Thus, courts have historically adopted the position that they are not qualified to second guess peer-review committees, at least as long as committees do not act arbitrarily and instruments are consistently and fairly applied (Baez, Benjamin, and Centra, 1995; Kaplin and Lee, 1995; Rebell, 1990). Traditionally, notes Rebell, most other courts have tended to take a similar deferential stance in teacher evaluation cases.

.....There seems to be two exception to this. The first is in discrimination cases. In general, courts have tended to only require precise accuracy in cases where EEOC issues are involved (See below). The second, is in claims of unfair treatment because of exercise of First Amendment free speech rights, including union-organizing activities, or allegations of denial of Fourteenth Amendment rights to due process by tenured teachers or others with a reasonable expectation of continued employment will trigger federal court jurisdiction with greater scrutiny of data (Rebell, 1990).

.

Acceptance of Administrative Subjective and Untrained Evaluator Judgements Of SEF Data [7]

.....An issue directly related to both the reliance on and statistical accuracy of SEF are views of the court regarding accepting or not accepting subjective administrative judgements of faculty teaching effectiveness.

> *Summary:* With regard to accepting the subjective judgements of administrators evaluation of SEF, the legal reasoning and rulings can be summarized as ranging from: (1) accepting administrative subjective judgements if (2) they are deemed sincere (3) grounded on some evidentiary basis (4) if made on the "vigor and variety of student criticisms" (5) "not arbitrary or capricious and were exercised honestly upon due consideration," (6) based upon "much experience in reviewing student evaluations, (31) reasonably draw on that experience (7) and have ruled that Presidents are not bound by factual findings made by majority members of a faculty.

Not only have the courts not traditionally examined faculty evaluations rigorously, they have tended not to require that evaluators be trained in the use, analysis, and interpretation of evaluation instruments. In general, state courts reviewing teacher evaluation practices will not analyze directly the substantive criteria used to evaluate teachers, nor the or qualifications of the raters. (Rebell, 1990). There are exceptions, however.
.....Some states, like Florida and Pennsylvania now mandate such training. Florida specifically mandates school boards to provide training programs to "ensure that all individuals with evaluation responsibilities understand the proper use of the assessment criteria and procedures" (Fla. Educ. Code, /sec 231.29(2). In Pennsylvania (Rebell, 1990), employees must be evaluated "by an approved rating system which shall give due consideration to personality, preparation, technique and pupil reaction in accordance with standards and regulations for such scoring as determined by rating cards to be prepared by the Department of Public Education...." (p.345-6).

SEF as Social Judgement and Diagnosis

.....Given the courts assumptions regarding validity and the untrained judgement of those making decisions based on SEF, a part of influencing the courts is demonstrating relevant research. In the research on social judgement and clinical diagnosis, it is clear that the manner in which nearly all SEF data are analyzed is but a subset of the social judgement and clinical diagnosis literature, involving the same logical and cognitive bias and distortions that result in the pervasive inaccuracy of social judgement in general and clinical diagnosis in specific. The findings of the judgement research literature applies to students making such judgements in evaluating faculty and to those interpreting the results; they are in fact making diagnoses.
.....Psychological research has recognized the severe cognitive problems and limitations of "intuitive," and "experience-informed" everyday judgements for over thirty years, (Dawes, Faust, and Meehl, 1989; Faust, Guilmette, Hart, Arkes, Fishburne and Davey, 1988; Garb, H. N. 1989; Hayes, 1991; Larkin, McDermott, Simon, and Simon, 1980; Rabinowitz, 1993) yet the mistakes continue in everyday practice situations. Interpretation of SEF are no different. As two authors who consider SEF literature valid (Franklin & Theall, 1990)---point out:

> Even given the inherently less than perfect nature of ratings data and the analytical inclinations of academics, the problem of unskilled users, making decisions based on invalid interpretations of ambiguous or frankly bad data, deserves attention. According to Thompson (1988, p. 217) "Bayes Theorem shows that anything close to an accurate interpretation of the results of imperfect predictors is very elusive at the intuitive level. Indeed, empirical studies have shown that persons unfamiliar with conditional probability

> are quite poor at doing so (that is, interpreting ratings results) unless the situation is quite simple." It seems likely that the combination of less than perfect data with less than perfect users could quickly yield completely unacceptable practices, unless safeguards were in place to insure that users knew how to recognize problems of validity and reliability, understood the inherent limitations of rating data and knew valid procedures for using ratings data in the contexts of summative and formative evaluation. (79-80).

The authors conclude by noting, "It is hard to ignore the mounting anecdotal evidence of abuse. Our findings, and the evidence that ratings use is on the increase, taken together, suggest that ratings malpractice, causing harm to individual careers and undermining institutional goals, deserves our attention." (p.79-80). Recognizing such problems is not methodological nit-picking; they are pragmatic, paradigmatic, and scientifically fundamental.

.

Variables Affecting Validity Not Taken Into Account When Assessing SEF [8]

.....In conducting any research, it is a given there are a host of variables that affect outcomes. Put in experimental terms, there are a host of independent variables that affect the dependent variable (here teaching effectiveness). The question is, how have courts addressed this crucial issue that impacts so centrally on validity of SEF data?

Instructional Variables

.....Legal cases concerned with the validity of SEF occasionally note various instructional factors that were not controlled in the faculty evaluation process.

> *Summary:* The variables noted in the legal cases reviewed include, (55) not controlling for class size, i.e., those obtained in small seminars from those obtained in large lecture classes, (56) those obtained from tenured faculty from those obtained from non tenured junior faculty, (57) not performing appropriate comparisons of SEF with other faculty, (58) noting SEF in all courses, not just to problem courses, (59) not mistaking student 'response' figures for actual student enrolment figures when using them to determine student attraction to a course, (60) using all courses taught, (61) taking into consideration faculty teaching a wide range of courses, versus those with lighter teaching loads, (62) number of new courses taught in a year, (63) whether graduate courses were taught at the same time as teaching undergraduate courses, (64) selectively mentioning only negative student comments, or (65) overly weighting negative comments, and (66) different procedures for gathering student opinion.

Courts sometimes weigh these variables heavily, in most cases, however, the courts either ignore them or do not weigh them very heavily in the total context of a particular case. [9]

.

Student Biases Variables [10]

.....A significant issue is how courts view student biases in assessing the reliability and validity of SEF.

Summary: Student bias variables include reactions to (48) academically demanding faculty, that (49) thus thwart student expectations, (50) difficult examinations (51) tough grading policy, (52) heavy workload in a course. (53) While most courts ignore these student biases in SEF, (54) occasionally a court will recognize that difficult courses have to be given to the students and that such material is difficult for even the best teacher to get the material across.

In general, however, it is overwhelmingly clear that courts seldom take these variable into account, despite the fact that such reactions often function as generalized affective overlays on SEF (see below).

.

Popularity Variables and Effectiveness [11]

.....A related student variable issue is the extent to which SEF measures popularity, not teaching effectiveness. Accordingly, it is instructive to see how courts view this issue.

Summary: Court rulings range from saying that (9) in cases of exceptional research faculty that popularity should not play a role in termination due to teaching, to (10) in normal cases that a measure of popularity is related to teaching effectiveness.

While not noted frequently, popularity appears to be generally assumed to be involved in teaching effectiveness. But again, the courts are mixed on this issue as well. In terms of the research literature there is little to no support for popularity being a measure of teaching effectiveness in higher education. [12]

.

The Courts' Reliance on Both Quantitative Data and Qualitative Comments in SEF [13]

Reliance on SEF v. Peer Evaluation

.....Is it considered acceptable, for example, to rely heavily or even solely on SEF, or must they be used in conjunction with other evaluative methods?

Summary: From the cases analyzed, it can be seen that court rulings range from saying that (1) relying primarily or solely on student evaluations is acceptable, to (2) placing little exclusive reliance on SEF, (3) in rare cases SEF can not be permitted to stand in the way of promoting or retaining professors who are excellent in non teaching areas, (4) tenure decisions can not be based solely on SEF by students who have not been made aware of the ramifications of their evaluations, (5) anonymous documents or those "based on hearsay" should not be included in a faculty member's file, (6) students should be made aware of the purpose and ramifications of their evaluations of faculty, (7) anonymous student evaluations should not be used, (8) peer evaluations must also be a part of evaluating teaching.

Again, courts range widely on the exclusiveness or non exclusiveness of SEF, even though

books on how to conduct faculty evaluation (by authors who basically accept the validity of SEF, e.g., Seldin, 1984; Theall, and Franklin, 1990) for some time now have consistently emphasized that SEF should not be used as the only and/or primary method for assessing teaching effectiveness.

.

Numerical Ranking of Faculty [14]

.....An important issue is how the courts view the relative weighting of SEF in administrative decisions of teaching competence. It seems to be common practice to ordinally rank and compare faculty to each other according to average SEF numerical scores.

> *Summary:* From the cases reviewed, numerical scores from SEF often result in faculty (22) being compared relative to other faculty, (23) being ranked relative other faculty, (24) with distinctions often being made on the basis of tenths of a decimal, (25) with most courts accepting these fine decimal distinctions.

Despite the above overview of the research on the highly questionable validity of SEF, institution administrators and the courts continue to make and accept fine numerical distinctions in faculty scores from student evaluation questionnaires to ordinally rank faculty. Even given that SEF is valid to a level accounting for 14% of the variance, it is not psychometrically appropriate to accept such ordinal rankings.
.....It should be noted that SEF rate the majority of faculty as above average---whatever this means.
.....Ordinal scales do not tell us if a faculty half way down the scale is only half as good as the top ranked member. Thus without a criterion referenced standard, we have no way of knowing if everyone on the scale is an effective teacher, or conversely an ineffective teacher. Moreover, should all faculty who fall below the statistical "average" be eliminated? And if so, using the same logic, should we rank order and thereby eliminate all Olympic team members who fall below the team average? If the answer is 'yes,' then (a) we eliminate highly functioning athletes, and (b) it leads to an infinite regress where we end with only one or two on any given team. Currently, we have no idea if "statistical average" means good, bad, or indifferent teaching in terms of instructional effectiveness.

.

Use of Qualitative Written Student Comments [15]

.....Over and above quantitative data, the use of written comments, often *single instances*, by students on their SEF forms seems wide spread by both educational administrators, faculty evaluation committees, and the courts.

> *Summary:* For the use of student comments, court views ranges from (33) placing importance on a single comment (34) to several comments as significant information, (35) maintaining that statistical analyses of SEF need to be bolstered by individual comments, (36) maintaining that while some very negative---e.g., racist, sexist---comments may be found, the court may find that they do not render SEF unreliable, (18) that such instances or "*impressions*" may be validated after the fact, (37) negative comments often seem to outweigh positive ones, and (38) may often outweigh numerical data to the contrary, (39) negative comments need not be verified before acting on them, to (40) that negative

comments can not be used to undermine otherwise generally favorable comments received in an annual performance review.

Clearly the views from the court suggest the legitimacy of not only using what is in fact anecdotal data, but often to raise it above more systematic (averaged) data.

.

Mixed Student Comments [16]

.....Just as quantitative SEF data may be bimodal, so too written student comments may also be bimodal or mixed. How do courts (indeed, educational administrators, and faculty evaluation committees) view and pronounce on such data?

> *Summary:* With regard to non numerically assessed written student comments, they are often qualitatively characterized as (41) a few were ambivalent, (42) a considerable number, (43) of mixed result, and selectively recognized: (44) it would only be fair to add that there were a number of comments in favor, (45) there were also some negative comments, (45) sometimes placing the greater weight on past evaluations of teaching over current comments, (47) sometimes placing greater weight on current comments over past positive evaluation of teaching.

Again, with regard to single and mixed comments on SEF, the courts (administration, and faculty evaluation committees, See Appendix) tend to weigh them far above their non representative and anecdotal-data value.
.....It seems to be generally assumed by most faculty and administrators that SEF are used by virtually all schools in the U.S. It is further assumed by many that SEF is necessary for both faculty evaluation of teaching effectiveness and thus for quality control of student learning. While its use is clearly wide spread (see Seldin, 1984; Crumbley, and Fliedner, 1995) in the U.S., and is increasing in Europe (Husbands, and Fosh, 1993), what is not generally recognized is that there are schools that preclude its use in salary, promotion and tenure decision either totally, or in part, by precluding the use of qualitative students comments.[17]

.

Transcendent Value of a Professor Over Teaching Quality [18]

.....Despite the importance placed on teaching, there is precedent for both school policy and the courts---under certain conditions---to ignore poor teaching as indicted by SEF.

> *Summary:* (11) The courts and educational administrations can not allow low SEF to stand in the way of promoting or retaining professors who may be world renowned scientists, (12) deemed nationally or internationally exceptional as a researcher, courts may nevertheless disregard SEF, (13) at least in these two cases the courts did not find the faculty exceptional. It would be interesting to see if what the court seems to accept in principle exists in fact.

The above collective categories abstracted from court cases are illustrated by a denial of tenure case described in the Appendix below, by a (non litigated) case that contains an interesting difference from most of the cases reviewed here.

Procedural, Burden of Proof, and Policy-Decision Criteria in Assessment of SEF

.....Other overlooked issues involving SEF and its validity are the problems of (a) content versus process, i.e., whether the assessment of SEF data constitutes a process or procedural issue or (b) is simply a content issue.

Validity Assessment of SEF as Procedural or Process Issue

.....An exemplar of the content and the procedural/process distinction is often exhibited between trial and appellate courts. The latter often only judge if correct procedural/due process was followed by a lower court. The content v procedural/due process distinction is typically used by college campus grievance committees. When a tenure committee, for example, renders an unacceptable decision, a faculty member may challenge the decision. A grievance or appeal committee then may review the decision only in terms of if the correct process or procedures by which the decision was made was followed. The point here is that many such appeals committees do not define looking at the procedures by which SEF data were gathered and analyzed by a tenure committee or administrative evaluator as procedure/due process (e.g., whether the tenure committee just 'eye balled' the data and student comments, whether they compared the data to other similar faculty SEF, etc.), but as content and therefore not within its purview. Grievance committees often therefore will not review the substantive content of SEF data on grounds that it is not a procedural or process issue.

.....In general, given the courts tendency to accept the validity of SEF data, at least by default, how SEF data are assessed and used is often considered to not be a process/procedural issue. At least one court has, however, considered how SEF data is assessed and used as procedural. This is evidenced in *Christopher Turner v. The President of the University of British Columbia* (1993), where it was stated that

> the Dean said, "there were few students in undergraduate literature courses since 1986/7---(3,8, and 6 respectively," thus mistaking student 'response' figures for actual student enrolment. The Board concluded that (5) "This misunderstanding is in our opinion sufficient in itself for a reconsideration, since teaching was the focus..." (p.3), and (7) "we think that the comments and emphasis on the size of Dr. Turner's classes as evidence of poor teaching are open to objection and *constitute errors of procedure and/or evidence*" (p.6). [italics added]

As noted above, however, it appears that most courts, and indeed, perhaps most faculty grievance committees (See Appendix below) have not considered how SEF data is analyzed as a procedure/due process issue. The issue of the validity of SEF, then, would appear to have legal "due process" implications.

Decision Criteria and the Scientific Precautionary Principle

.....Since SEF has haphazardly evolved along with a general acceptance of its validity as an appropriate measure of faculty teaching effectiveness, the burden of proof somehow has been placed on faculty-as-challengers of such data to scientifically prove that SEF data is *not*

valid--- a strange state of affairs, at least in science. And the standard of proof required has been typically high. In effect, faculty are guilty until proven innocent. So the process that exists is:

1. Either (a) a legal abdication of the assessment of SEF by the court, relying on the good faith evaluation of SEF data by the institution, or (b) the court simply assuming its validly.
2. Placing the burden of proof on faculty who challenge the data of demonstrating with scientific levels of certainty (statistical significance or confidence level) that the data is not valid.

.....Given---at the very least---the controversial assessment of the validity level of SEF in measuring teaching effectiveness, in terms of decisions and policy perhaps we should err on the side of caution in applying such data for administrative purposes. In the field of environmental science, Lemons (1996) and Lemons, Shrader-Frechette and Cranor (in press) have suggested a Precautionary Principle when making policy decisions. In essence, this principle says that when making policy decisions about environmental harm, given (a) a certain level of possible harm, (b) the complexity/uncertainty of data, and (c) the high level of proof (typically a 95 per cent confidence level) required for a scientific finding to be accepted by scientists, setting policy should not be based on this level of scientific proof. The reason is this: To wait for such a confidence level may be too risky given the level of harm that may be indicated (by the existence of data with a lesser confidence level suggests). In short, using scientific criteria that have been adopted for doing science may often not be appropriate criteria for making policy decisions.

.....The reasoning surrounding the Precautionary Principle is too complex to fully delineate here. The reader is referred to the citations. In the meantime consider the following analogy that in broad outline exemplifies the spirit of the Precautionary Principle: A dangerous tiger has escaped from a local zoo a few miles from your house. In the back of your house is a wooded area. Your child wants to go out and play in the woods. No one has actually seen the tiger in the woods or anywhere else around the neighborhood. In other words, there is no scientific level of evidence that the tiger is anywhere around, or that your child would be in immediate danger by playing in the woods. Do you let your child out to play in the woods? In most areas of science, the rule is to avoid type-I error---asserting there is an effect when there is none, and therefore place the burden on those who postulate an effect rather than on those who postulate no effect---and not so much concerned with avoiding type-II error---asserting no effect when there is one. In adopting SEF data as indicating teaching effectiveness administrators, faculty evaluation committees and the court have engaged in type-I error---given both the level and burden of proof.

.....Now there are two implications for the Precautionary Principle as applied to SEF in relation to faculty and instructional quality. First, given (a) the haphazard way SEF have been introduced and accepted by the courts (b) the level of possible harm of accepting SEF for administrative purposes of salary, promotion, denial of tenure or non reappointment, to that faculty and more importantly (c) the effects of SEF used for such purposes has on the quality standards of higher education (see Haskell, 1997a) should such a burden of proof be demanded by the court of faculty challenging SEF data? Certainly, as shown below in disparate treatment and disparate impact cases, a kind of Precautionary Principle is already in effect. Second, given the at least clearly conflicting evidence of whether SEF demonstrates teaching effectiveness of a faculty, should not administrators and faculty evaluation committees apply, for the same reasons, a similar Precautionary Principle stance?

.

The Court's Approach to Validity of SEF in Relation to the Principles of Disparate Treatment and Disparate Impact [19]

.....Given the above findings on how the courts have tended to treat SEF validity issues, I would now like to further look at the implications. Federal courts---and to a lesser degree state courts---have adopted a more stringent approach to testing teacher evaluation cases, at least regarding primary and secondary teachers. According to Rebell, the four main reasons for this change are (1) the wider use of more stringent evaluation techniques by institutions, which largely stem from legislative reform initiatives that have led to an increased number of denials of teacher certification and terminations, (2) a disproportionate number of these certifications and termination involve members of minority groups, (3) legal developments have broadened the jurisdiction of the federal courts to consider issues of social reform, and (4) judges' own increased experience in assessing psychometric techniques in employment discrimination cases. It is perhaps 2 and 3, however, that have had the most impact on the courts (Kaplin and Lee, 1995; Rebell, 1990).
.....Educational reform issues from desegregation, special education, and other school-based litigations, has made the courts more experienced and more inclined to scrutinize educational testing requirements. As the consequence of federal Equal Employment Opportunity Commission (EEOC) criteria, in today's civil rights climate, courts are more likely to scrutinize the validity of the faculty evaluation instrument, especially in terms of racial, gender, and age discrimination.

Disparate Impact

.....In regard to teacher evaluation in general in cases involving claims of discrimination under the equal protection clause of the Fourteenth Amendment, or under the anti discrimination statutes enacted to protect members of racial and ethnic minorities, women, handicap conditions, age, and other protected groups scrutiny of the case tends to be more probing and stringent. Such cases are of two basic kinds: (1) those involving discriminatory intent, called disparate *treatment* claims, and (2) those involving no intent, called disparate *impact* claims (see Kaplin and Lee, 1995, section 3.3.2.1.).
.....Disparate impact claims in personnel evaluation is the use of assessment procedures that are facially (on their surface, or methodologically) neutral in their treatment of different groups, but which produce evaluation outcomes that inadvertently fall more harshly on one group than on another. Thus, proof of a discriminatory motive is not necessary to establish a disparate impact claim. To establish a *prima facie* case of such adverse impact, a minority need only show a causal connection between the facially neutral employment practice and the disproportionate negative or adverse effects on him or herself as a member of a protected group. For example, a university tenure process may be found to discriminate against females because the evaluation process or evaluation criteria favors male faculty more than female. In such cases, rigorous statistical analysis is typically used to establish disparate impact.
.....Discriminatory treatment and disparate impact claims has made courts more inclined to specifically analyze educational testing instruments for validity, and this increased involvement by the courts is predicted to increase. As of 1990, 41 states have mandated some form of standardized testing requirements as part of their teacher certification process Because many of these exams are claimed to have a disproportionate negative impact on minority candidates, competency tests have triggered a number of large scale federal class

suits. Again judge Rebell (1990) notes,

> In June 1988, the United States Supreme Court issued a ruling which is likely to accelerate the trend toward increased judicial involvement in teacher evaluation matters. That case, *Watson v. Fort Worth Bank and Trust* (1988), extended to judgmental employment practices the Court's 1971 holding in Criggs v. Duke Power Company (1971) that standardized employment tests having a **disparate impact** on minorities must be shown to be job-related. Although the Court's ruling in Watson was unanimous, there was substantial disagreement among the Justices as to how closely courts should scrutinize particular practices and **validation techniques**. Whatever the precise standard of review ultimately implemented, there is little doubt that the federal courts will be more likely to scrutinize nonobjective evaluation procedures as a result of Watson (p.339).

Thus any instrument or evaluation criteria that *in effect* places an unfair burden on those being evaluated has been judged to exhibit what is legally termed *disparate impact*. .....The present point is that while the courts have not, and continue to not rigorously scrutinize SEF, they have for sometime now applied fairly rigorous standards to evaluations both in the workplace and in academia to cases involving discrimination of protected groups, whether the discrimination is purposeful, or by disparate impact. Not every indication of racism, however, may be considered by a court to be proof of discrimination. .....For example, In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), the Labor Relations Board said, "with respect to comments that while some students had written that she was a "slant eyed bitch," and that she should "go back to China....We also are not persuaded that the racism evident in the student evaluations of Grievant made student evaluation results unreliable. The percentage of evaluations in which racism by students was evident was approximately one percent of the total evaluations" (p.306). Assuming some level of covert racism, how does one disentangle the generalized affective racist and sexist overlay of students evaluation on a total questionnaire?[20]

The Disparate Treatment and Impact Principles Generalized

.....In the evolution of any legal policy or principle its extension often occurs by generalization or analogical transfer, extending a principle thought to apply to only one area to other areas (See for example, Anderson and Schadewald, 1991; Golding, 1984; Levi, 1949; Marchant, Robinson, Sunstein, 1993). Currently both corporate and academic cases of straight forward discrimination and the more inadvertent discrimination cases based on disparate impact often trigger the courts to rigorously scrutinize the methodology and statistical data of such evaluations not typically accorded to non discrimination cases. Presumably, in discrimination cases the court's interest is in establishing validity and using rigorous statistical methods in ascertaining the "truth." If this is the case, then by clear logical implication and inference---as we have seen---in generic cases of evaluation the court could be said not to be in the truth business. As documented above, in non discriminatory cases, courts have assumed the "truth" lay in the appropriateness of an institution's criteria and rarely would substitute their judgments for those of peer review committees, adopting the position that they are not qualified to second guess peer-review committees, at least as long as committees do not act arbitrarily and instruments are consistently and fairly applied. The burden of proof is on the faculty challenging an institutional decision. Some courts have only been concerned with consistency and fairness of application, even if the methods of evaluation are clearly defective.[21] Generally, however, the courts have acted as though they believed that institutional evaluations were made only after careful deliberation and with

procedural due process protections.

.....The question is: why not make the same assumptions regarding discrimination? The answer is that, understandably, the courts have accepted that there has existed a widespread conscious and non conscious ethnic, gender, age , religious belief, sexual orientation, and handicap bias in society, such that they can not simply rely on the "truth" or good faith behavior of an institution or its data. Given this, the argument is made that herein lies the distinction, and reason, for treating non discriminatory cases differently from cases where either discrimination has been charged (treatment claim) or where discrimination has be inadvertent (disparate impact claim).

.....Thus the courts have tended to accept the judgement and "good faith" motivations of organizations. Unlike in the past, however, just as the data are in regarding discrimination of protected groups in academia, so too the data are now sufficiently in to cast serious doubt on the courts assumption of "truth" residing in corporate and academic data on discrimination, so too is it in on (a) the questionable validity of SEF, (b) the internal politics of administration and faculty relations which can revolve around student retention and unpopular ideas, and (c) the economic pressures on institution to not tenure faculty and to sometimes terminate tenured faculty, all of which can have serious contaminating consequences institutional decisions.

.....The importance of this is that while courts have scrutinized SEF for evidence when civil rights discrimination has been questioned or suspected, they have not applied the same rigor to the validity of evaluation instruments or have held as suspect other institutional biasing variables. The courts continue to assume a kind of pre 1960s academic Camelot. If such a round table of academic knights ever did historically exist or was merely mythical, it certainly now exist only in myth.

.....One compendium of legal findings in higher education specifically notes SEF and recognizes the accepted application of the principles of disparate treatment and impact along racial and gender lines in SEF. It should noted that the disparate treatment and impact issues when applied to SEF, is of course, no different than any other disparate impact case, except that the student evaluation data in such cases will be scrutinized by the courts. The authors (Baez and Centra, 1995) suggests that the SEF research in the area of race and gender discrimination, has been inconsistent, and suggest that while deserving of more attention, the inconsistency of the research makes it unlikely that the courts will sustain such a claim. Some courts, however, have found in favor of faculty in such cases. For example,

> In *Cynthia J. Fisher v. Vassar College* (1995), after a bench trial, the district court found that, in denying Fisher tenure, Vassar had discriminated against her by reason of (a) her sex in violation of Title VII of the Civil Rights Act of 1964, (b) her age in violation of the Age Discrimination in Employment Act. The court found that the termination of Fisher's employment resulted not from any inadequacy of her performance, qualifications, or service, but rather from pretextual and bad faith evaluation of her qualifications. Scrutinizing Vassar's report on Fishers teaching ability which included reviews of her student evaluations that were said to reflect "consistent problems with clarity and her ability to illuminate difficult material" but which were otherwise generally positive. The district court found that the Vassar's biology department had distorted her teaching recommendations by "selectively exclud[ing] favorable ratings," by selectively "focus[ing] on the two courses in which she had difficulties" and by "applying different standards to her than were applied to other tenure candidates" (*Id.* at 1209). The court further observed that "the males tenured while Dr. Fisher was on the faculty were praised for their fine teaching while Dr. Fisher was criticized, although the facts on which the Committee's determinations were based (student evaluations, Biology Majors Reports and [Student Advisory Committee] reports) revealed that Dr. Fisher's evaluations were superior to theirs" (*Id.* at 1211). The court noted that statistical analysis may be a part of a plaintiff's

effort to establish discrimination under a theory of disparate treatment.

The point here is that if this had not been a disparate treatment discrimination case the biases and distortions of data about her teaching student evaluations would likely have gone un examined. [22]

.....It would seem, then, that this discrepancy in the discrimination-based search for "truth" should be used as---and provide justification for---a kind of generalized disparate impact principle to legally invoke or generalize a fairness principle that the same rigor be applied to non civil rights cases such as SEF. As Kaplin points out, however, current law generally prohibits courts from such generalization. [23] So the issue of change apparently becomes not so much one for the court as it is a policy issue for both higher educational administration to use as a guideline and for legislatures to legislate change.[24]

.....Age discrimination in SEF is another possible bridge in this potential extension of disparate impact. The Age Discrimination in Employment Act of 1967 (ADEA) requires employers to evaluate persons on their qualifications or ability to competently perform their job, and not on the basis of age. Like any other employer, colleges and universities are likewise prohibited from considering a faculty member's age in making decisions about employment, salary increases, promotion, tenure, and retention. Yet, there is evidence that SEF do discriminate on the basis of age, with older faculty receiving lower student ratings (Feldman, 1983). There are a host of other variables like class size, or teaching a courses within a student's major as opposed to elective course, or teaching freshman v.s upper level students, that also make a kind of default "disparate impact" if such variables are not controlled in the analysis of SEF data.

.....What is being suggested here is that in the interest of justice, equity, truth, and in "fact finding," the courts and institutions should scrutinize all SEF data as rigorously as they do disparate treatment and disparate impact cases. Currently data and conclusions from SEF are seldom scrutinized (as indeed are other issues in the denial of tenure or promotion not equally scrutinized) as they are in discrimination and disparate impact cases. Justice, however, is not only blind to ethnic, gender, age, sexual orientation, religious belief, and handicap status, it is blind to institutional economic pressures and other biasing variables within academic institutions. Thus, biases and distortions of the SEF data are not revealed in non discrimination cases as they are in disparate treatment and disparate impact cases. As a consequence, in terms of revealing unfair attributions based on SEF data, those covered under EEOC guidelines have a "truth finding" advantage over those who are not covered.

Beyond Statistical Significance of SEF Research

.....Having reviewed SEF cases and examined the significance of validity, I would now like t to turn the issue of validity on its head. Underlying statistical research on SEF that attempts to establish its validity is a complex of contextual variables and assumptions seldom addressed.[25] In this section, I will address some of these contextual variables and assumptions that I suggest cut through and render the best of statistical research on SEF showing teaching effectiveness nearly irrelevant. Understanding is not acquired by statistical significance alone. Certainly showing statistical validity of SEF is a necessary condition, but it is not a sufficient condition for understanding their meaning and for its use in administrative decisions. It is an understanding of these contexts and assumptions that underlie statistical validity research on SEF that educational policy-makers and the courts need to think long and hard about accepting SEF for assessing instructional competence and using it for promotion, tenure, and reappointment decisions.[26]

Assumption # 1: Statistical Significance of Indicators of Teaching Effectiveness

.....An assumption underlying statistical analyses of SEF is that we know what the indicators of effective teaching are. To my knowledge, the research does not support this assumption. What makes us so sure that many of the questions we ask on SEF questionnaires are all that related to effective student learning. Consider, for example, the typical question "Was your instructor organized?" This question in turn entails a myriad of assumptions and conditions about effective teaching. Would Socrates, for example, be *perceived* as organized by most students---being peripatetic and just asking a lot of questions? And what makes us think-----at least for some students and some kinds of subject matter---that just going into class, being Socratic, asking a lot of provocative questions, and confronting students by challenging their belief systems may not be the most effective instructional and learning method in the long run to get students engaged and to think critically?[27] What evidence is there that either being perceived as organized or actually being organized is a necessary condition for effective instruction? I know of no rigorous supporting evidence. Indeed, many of my friends in the humanities, much to the dismay of my behaviorist and cognitive colleagues---and sometimes myself---would suggest that systematic and sequentially structured teaching methods are simply structural analogues of our technological society (see, for example the classic by Jacques Ellul (1964).

.....Consider, too, a question that, while it is not directly asked on SEF questionnaires is implied in other questions in various forms, inquiring "Does your instructor mainly lecture?" Though there is precious little rigorous evidence showing that lecturing is inherently an ineffective teaching method, it is clearly *persona non grata* among many educational theorist. Lecturing is "out" while collaborative learning is "in"---but apparently not so considered by many faculty (for both valid and invalid reasons).

.....While I happen to agree that being organized is generally good, and that collaborative learning is perhaps good for certain student populations, subject matters and desired outcomes, the question is: are they appropriate indicators of effective teaching applied to individual faculty as claimed? The answer to this question is they are not appropriate indicators of effective teaching applied to individual faculty---and this applies even if the statistical research strongly supported the claim. This is an important point that, as I recall, is addressed in the faculty evaluation literature only by Scriven (1988). I will quote Scriven at some length.

.....Scriven observes that in the attempt to render teacher evaluation more scientific the field rushed into focusing on research-based indicators, teaching indicators which sound research supposedly demonstrated are positively correlated with successful student learning. These indicators or

> Popular envies are structured presentations, active involvement, emphasis on positive reinforcement, high eye contact, high frequency of question asking, provision of learning objectives, frequent feedback, use of multi-media (p.4)....the provision of a brief outline of topics to be covered in a day's lesson can be justified on administrative grounds, since substitute teachers must get some guidance; but the requirement that anything like that be provided to students, for pedagogical reasons--a claim often said to be supported by research---cannot be justified. The use of instructional objectives or any other kind of advance organizer is simply a characteristic of ones style of teaching, not a duty of the teacher. Nor can such an outline be required as evidence of preparation (arguably a duty), since a teacher using a textbook--or for that matter, memory--may do as well or better than one with lengthy lesson plans listing activities and testing procedures (p.7-8).

The presence or absence of these factors, says Scriven, defines a style of teaching. He maintains that any reference to a 'teaching style' in teacher evaluation is not valid, regardless of whether there exists a research basis for thinking the style is correlated with teaching effectiveness.[28] He goes on to explain:

> A major source of confusion in discussing the use of indicators is that the research is often presented as showing that 'the best way to teach' is by using high eye contact (or whatever), whereas all it really shows is that there a slight tendency for better teachers to exhibit this characteristic, for reasons which might include the fact that they were taught to use it, although in fact it's not a help at all. The reader is seduced by the relative plausibility of the style recommendations, whereas you'd never buy the idea of using eye color or skin color. But plausibility isn't necessity, and absent necessity, you're just a stylist Our kids don't need stylists, they need good teachers; and if you can't distinguish the two, you're in the wrong business (p.7).[29]

Scriven is not denying the validity of statistical inference. Useful information is contained in a statistical correlation, and there are circumstances in which that information can be put to good use. It can even be put to good use in making decisions about people---but only when no better data is available because of limitations on time or resources. [30] Scriven maintains that such teaching effectiveness indicators are invalid

> for essentially the same kind of reason that the evaluation of personnel by the color of their skin or their church affiliation is necessarily invalid. While it is true that much racial prejudice, sexism, etc., is based on false beliefs about the groups discriminated against, the essential flaw in it goes deeper than that. The essential flaw is that even if women in general are less strong than men, you shouldn't. use gender to discriminate against a particular candidate for a position as a luggage-handler, but only a job-related strength test or series of observations in a trial period on the job. And this is nor just for ethical/legal reasons, but also for scientific reasons and reasons of efficiency (p.4)....Which means you can't discriminate against a teacher on the grounds that s/he exhibits some approach to teaching that research has shown is less likely to be successful. Whites are statistically less likely to be good basketball players than blacks, but you can't kick the whites off the squad the day you discover that the statistics are worse than you thought. nor would you be any good as a coach is you used skin color as a criterion for selection. You have to look at the individual's success, not at the success of groups to which the individual belongs (p.4).

Finally, Scriven suggests a reason for the almost total disregard of the validity of SEF by the courts documented in this paper (and my previous paper, Haskell, 1977b). He understands the implications for courts recognizing the fallacy of such indicators: He says, The current fallacy of using such statistical-indicators are,"as certain to crash in the courts---eventually---as the most blatantly racist hiring practices. We may have only a short breathing space before the courts and defense attorneys begin to see the underlying similarity of these two approaches....The consequences for states and districts will be chaotic; old decisions may be reversed on appeal, huge damages may be awarded, those hearings will clog the system, and there will be no legitimate process to take the place of the illicit one (it is because of this potentiality for disaster that we are giving a longer than-usual treatment of the issue here) (p.5).

Assumption # 2: Statistical Significance of SEF of Teaching Effectiveness Measures Appropriate Learning

.....An assumption that is virtually unnoted in the literature is that given SEF is eventually found to measure teaching effectiveness---and this "given" is only for the sake of the current argument---it is assumed that what is thereby being measured is appropriate learning. This assumption is arguably incorrect for at least two reasons. I say it is arguably incorrect, as whether the assumption is correct or not depends on other differing assumptions about higher education.

.....First, let us not fool ourselves into thinking that we know what effective teaching is for all populations of students and subject matters. There is no shortage of possible indicators of effective instruction and learning, but most are not articulated within an adequate theory of effective instruction or learning. At the very least, "effective" is relative to a given student population. And when referring to teaching effectiveness are we referring to measuring short term or long term learning?[31]

.....In addition, as Abrami (1989) and others (see Cohen, 1983) have suggested, most studies on the relationships between student ratings and instructor-generated student learning have been done with learning outcomes collected largely from freshman classes, and---more importantly---learning at the lowest level of Bloom's taxonomy. Similarly, the literature on transfer of learning shows that when student transfer of learning is found, it reflects the lowest level of concrete transfer. So even if we are effective in achieving this level of effectiveness, what have we achieved? This brings me my main point.

.....I suggest that teaching effectiveness and appropriate learning in higher education are two different logical and empirical entities. I shall now address these two differing assumptions together. If the data showing (a) student level of unpreparedness, (b) student ability level as measured by most national tests, (c) unrealistic student expectations about learning, (d) grading, (e) feeling of entitlement, (f) motivation level, (g) good faith motivation for evaluating faculty, (h) maturity level, and (i) hours spent studying have been either in decline for years, or have become increasingly inappropriate is accepted, then effectiveness in teaching most of these students does not necessarily---and most likely does not---mean appropriate learning. For purposes of clarity (and at some risk of seeming not only insensitive, but as a right wing radical, which I assure the reader I am not), let me demonstrate why teaching effectiveness is separate from appropriate learning by using what may be considered an extreme scenario as an example: Suppose that the American Disabilities Act as applied to higher education is amended to include having to admit the mentally retarded, thus requiring making whatever instructional adjustments need to accommodate their disability.

..... Now assume that such adjustments are made, e.g., speaking slower, simplifying and otherwise decreasing the amount of content to be mastered, along with the depth of understanding and critical thinking. In addition, assume that if such adjustments and other classroom behaviors that were once apporpriate for a previous level of student are not accommodated and that this is reflected in low SEF score. Now assume that because of pressures such adjustments have been made and that SEF findings for those teaching the disabled students unequivocally shows teaching-effectiveness. The question then becomes: is this appropriate learning for a higher education course? [32]

.....Most will likely respond to this question with a resounding "no." Some, on the first assumption noted above may say "yes." Some will respond by maintaining the above scenario is extreme and inappropriate. The fact is, however, that this scenario is simply a quantitative extension not qualitatively different from what has been occurring in the lowering of admission and course requirement standards that has been occurring for some time. So the question now becomes, not simply teaching effectiveness but teaching effectiveness at what level of learning, and by implication, academic standards. This is an issue that needs to be

addressed nationally by faculty. Being well versed in logic as well as statistics, Scriven (1988), of course, understands this. In a similar context he notes,

> It's not even true that 'it all boils down to how much the students learn from the teacher': if it did, the teachers of mentally-retarded students would automatically be the worst teachers. In fact. they are often much better teachers than those teaching smart students, because smart students survive bad teaching better. (How many of the research studies naively treated "amount learned" as the criterion against which they "validated" the indicators?) p.7

And herein lies the ghost in the machine of most statistical validation studies of SEF---at their very best: There is nothing wrong with the statistics only with the meaning of what they are purportedly measuring. Thus the problem is not a flaw in the data or the measurement instrument, but a flaw in the measurer.

.....Finally, to conclude this section, the implications for SEF in general and for the issue of validity seem clear. The issue of validity of SEF, then, is not the primary issue it appears to be, and serves inadvertently to hide the significant issue of academic standards.[33] I will address this issue in relation to academic freedom and academic standards in more detail in my final paper. [34]

Conclusion

.....From most of the above cases---even given that, as challengers, the burden of proof has been on faculty --- it seems clear that the courts have not been kind to faculty with regard to student evaluations.[35] Some clearly see the courts various involvements in academic matters as detrimental to academic freedom. Arguably, rulings do often seem to shape it in inappropriate---and not so arguably---inconsistent and contradictory ways. "It is not clear, however," suggests Rebell (1990b), "that increased judicial involvement will have such a detrimental impact. In some measurement situations, courts have exhibited a sophisticated understanding of the complex judgmental factors at stake, and their insistence on thorough-going implementation of improved, fairer assessment devices has enhanced, rather than impeded, the development of professional standards" (p.340). He goes on to point out that, "because the state of the art concerning teacher-evaluation practices is at a sensitive developmental stage, extensive court intervention at this point can substantially influence---for better or worse---the future direction of basic practice in the field" (Rebell, 1990b, p.344). Thus whether increased judicial intervention in faculty matters will have a positive or a negative impact on professional evaluation practice depends on providing the courts with appropriate psychometric data and other scientific procedures.

.....Given the above rulings and the courts propensity to accept faculty/institutional agreements, it would seem as Kaplin and Lee advise, regarding academic freedom that "it is especially crucial for institutions to develop their own guidelines on academic freedom and to have internal systems for protecting academic freedom in accordance with institutional policy" (p. 192) would be especially true for a detailed SEF policy, especially including how the data is to be assessed.

.....The fourth and final paper will address the implications of court reasoning and rulings for academic freedom, standards, and instructional decisions.

Notes

1. Address correspondence to: Robert E. Haskell, Ph.D., Professor of Psychology, Department of Social and Behavioral Sciences, University of New England, Biddeford, Me. 04005. Email: rhaskel1@maine.rr.com. I would like to thank Professor John Damron, of Douglas College for continually providing me with sources, support, and advice, and especially Professor William A. Kaplin, School of Law, Catholic University of America for his invaluable legal counsel and for reading a draft of this paper. Interpretive liberties with the legal material and any other problems and omitted legal nuances are my responsibilities.
[BACK to document]

2. As with my second paper (Haskell, 1997b), the focus here will be delimited to how the courts reviewed have addressed SEF issues within various legal challenges to the denial of academic freedom, tenure, promotion, and reappointment by institutions of higher education. There are multiple legal variables that define an action or influence an outcome in a particular case. Among them are the statutes or other sources of law being applied, the cause of action being asserted, the prescribed prima facie case, the allocation of burdens of proof, and the standards of judicial review (see, e.g., Kaplin and Lee, 1995, section 1.3 & section 1.4.3.6). For my purposes here, I will not be concerned with these variables. Accordingly, this paper will neither be concerned with the outcome of the legal rulings, nor with the complex legal reasoning on which the rulings were based. My purpose is to review the general reasoning of the courts on SEF from a "reasonable man" standard and from a policy point of view.
[BACK to document]

3. To the layman, legal rulings regarding SEF are a veritable thicket, often seeming that the use of context to differentiate one apparently similar case from another functions as a kind of ad hoc carte blanche to justify preconceptions and positions.
[BACK to document]

4. A largely neglected---or ignored---important function of education is its social function. Education is not just for the benefit of the individual but for the benefit of society. Like it or not, we in higher education have accepted the social function of certifying competence of our students entering into an increasingly complex world. The certifying function has become especially important since the introduction of vocational programs into university curricula.
[BACK to document]

5.....In *Johnson v. University of Pittsburgh* (1977), the court said (7) "We have repeatedly approved the use of statistical proof where it reached proportions comparable to those in this case to establish a prima facie case of racial discrimination in jury selection cases . . Statistics are equally competent in proving employment discrimination. We caution only that statistics are not irrefutable. They come in an infinite variety and, like any other kind of evidence they may be rebutted. In short, their usefulness depends on all of the surrounding facts and circumstances" (8) The court further said in Footnote # 20: "Considerations such as small sample size may of course detract from the value of such evidence" " (p.1361).
5.....In *Peters v. Middlebury College* (1977), it was maintained that (5) "A professor's value depends upon his creativity, his rapport with students and colleagues, his teaching ability, and numerous other intangible qualities which cannot be measured by objective standards" (p.860).
5.....In *Fields V. Clark University* (1987), the court noted that (10) Fields' "attacks" the university's use of her student evaluations because they were not gathered and evaluated

according to accepted standards of scientific polling procedures. In response, the court agreed, saying, "She is probably correct. The use made of the student evaluations in her case, however, followed the practice at the defendant's university in other tenure decisions" (p.671).
5.....In *Cynthia J. Fisher v. Vassar College* (1995), the court noted that (7) "statistical analyses may be a part of a plaintiff's effort to establish discriminatory treatment" (p.1209).
5.....In *Yu Chuen Wei and the Vermont State Colleges Faculty Federation* (1995), the court ruled that (4) "The Court need not consider the accuracy of these administrative determinations, and that (24) tenure criteria "are not drawn with mathematical nicety." The board further ruled that (25) "the Dean and the President, both reviewed Grievant's student evaluations carefully. Their failure to take it a step further, and perform a statistical comparison of Grievant's student evaluations with those of other faculty members who have been granted tenure was not arbitrary and was reasonable; (26) Such a comparison is nowhere required by the Contract, [and] (27) we decline to hold such an involved comparison is necessary before a reasonable tenure determination can be made" (p.311).
5.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), the court concluded (38) "that the instrument was not perfect, that it had flaws, and that the very limited number of samples (because of the very limited number of courses and students surveyed over the period) impaired its reliability. (p.30). (39) "However, we accept the evidence of Dr. [X] that the instrument has some value, directed toward the specified factors. The court noted that (28) "One problem with the questionnaire is that it solicits bad points as well as good points. Despite that caveat, we conclude that the inclusion of the qualitative comments was not a significant error" (p.32).
5.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the Board said (19) Given certain Departmental procedures, "there is a danger that some negative class commentary will dominate the discussion and will not be the 'independent' opinion of all of the students. (20) This is especially true in the context of the direction to assess "effectiveness" versus "popularity" (p.10). They further noted, (18) Given that "There was no peer review at all; no member of the Department audited any of Dr. Kramer's lectures. There was, therefore, nothing to guide the Department but the student comments," and "no way to test the accuracy or fairness of the undoubtedly disturbing comments in Asian Studies" (p.10).
5.....In *University of Regina Faculty Association v. University of Regina* (1993), The Board argued (6) that "the University was under an obligation to verify negative comments before acting on them" (p.4).
5.....In *Christopher Turner v. The President of the University of British Columbia* (1993), the Board said, (7) "while not ignoring some student unhappiness with Dr. Turner's teaching style, we think that the comments and emphasis on the size of Dr. Turner's classes as evidence of poor teaching are open to objection and constitute errors of procedure and/or evidence" (p.6).
[BACK to document]

6. This is an important area but will not be dealt with here because student achievement scores as a measure of teaching effectiveness is almost exclusively used on the secondary level of education.
[BACK to document]

7.....In *Dyson v. Lavery* (1976), the court found that despite questionable errors it concluded that administrative judgements were acceptable because, "they were sincere and grounded on some evidentiary basis" (p.111); and (5) "In the absence of a finding that same were sexually motivated, the administration's professional judgment must be respected"

(p.111 all italics added).
7.....In *William Sypher v. Vermont State Colleges Faculty Federation* (1982), (7) sufficient evidence exists from which the Dean and President could have reasonably concluded Sypher was not above average in his teaching effectiveness; (8) the Board went on to say that if they adopted the Colleges' view that Sypher was not reappointed because of his teaching effectiveness, no argument advanced by him defending his teaching was likely to persuade the President because his decision was made on the "vigor and variety of student criticisms" (p.135).
7.....In *Carley v. Arizona Board of Regents* (1987), The court ruled (18) the University president was free to consider factual findings made by minority members of the academic freedom and tenure committee and any other evidence which he found relevant in determining whether to deny renewal of teaching contract to non tenured instructor. The president was not bound by factual findings made by majority members of committee (P.1103).
7.....In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), it was noted that (28) The Dean and the President obviously had much experience in reviewing student evaluations, and could reasonably draw on that experience in each tenure review. (p.311); judgements "were not arbitrary or capricious and were exercised honestly upon due consideration,"....that Deans and Presidents have "much experience in reviewing student evaluations, and could reasonably draw on that experience" (p.311).
7.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), the court said, (40) The relevance and quality of the scores are "a matter of weight for the various decision-makers, and we assume that they were reasonably aware of the limitations of student evaluations and gave them the weight they deserve" (p.30).
7.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the board concluded, "In the final analysis, we feel that this review of the Head's comments on teaching, which would be the sole evidence upon which the Dean and the President could rely, shows that it was incomplete and might have been misleading" (p.12-14).
7.....In *University of Regina Faculty Association v. University of Regina* (1993), he Board said teaching was wrongfully evaluated, but upheld denial of tenure on grounds of inadequate scholarship.
7.....In *Christopher Turner v. The President of the University of British Columbia* (1993), The board concluded that (11) "there were sufficient errors of procedure and/or evidence to return the case for reconsideration" (p.11).
[BACK to document]

8. In *Lieberman v. Grant* (1979), Lieberman attempted to introduce approximately ten personnel files concerning the tenure proceedings of other faculty in the English department for comparison. (6) Recognizing that such evidence would have had some minimal probative value, the Court, exercised its discretion under Fed. R.Ev. 403, and excluded it on the ground that "such probative value would be substantially outweighed by the delay and waste of time, which introduction of such evidence would have necessarily entailed....The plaintiffs case without such evidence seemed almost interminable, consuming 52 trial days over a two-year period. That is long enough" (p.873).
8.....*In Fields V. Clark University* (1987) notes but does not admonish the non separation of student remarks from small seminar courses and those from large lecture classes.
8.....In *Cynthia J. Fisher v. Vassar College* (1995), the district court found (2) that the biology department distorted Fisher's teaching recommendations by (3) "selectively exclud[ing] favorable ratings," by "focus[ing] on the two courses in which Dr. Fisher had difficulties" and (4) by "applying different standards to her than were applied to other tenure candidates" (p.1209).

8.....In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), it was noted that (19) "The statistical comparison demonstrates that Grievant was evaluated higher by students than her [male colleague] with respect to upper level classes, but that (20) [male colleague] was evaluated higher than Grievant in lower level classes. Given (21) this "mixed" result, the statistical comparison of evaluations does not demonstrate by a preponderance of the evidence that Grievant's students rated her the same, or better, than [male colleague]" (p.305). Wei maintained that (16) her students rated her the same or higher than the male colleague's students rated him. The Board disagreed, saying, (19) "We note that the comparison offered by Grievant is somewhat weak since [male colleague] was tenured in 1988, and those student evaluations of his which were compared with Grievant post-dated his tenure review by a number of years...further saying, "we decline to hold such an involved comparison is necessary before a reasonable tenure determination can be made" (p.305).
8.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), the Board noted that (19) the reviewing faculty held in-class discussions about his teaching.
8.....In *Robert Kramer v. The President of the University of British Columbia* (1992), Kramer argued that the most significant mistake was the failure to consider all aspects of his teaching. For example, only his teaching in 1989-90 was considered, whereas (9) he had taught a wide range of courses over the previous three years (10) had three new courses that year, (11) plus a graduate course. Moreover, (17) The department head indicated that his teaching was not up to the departmental "standard." The standard appeared to be the performance of the tenure-track faculty, though Kramer was one of the most junior faculty members (p.8). (15) Only one of the more than thirty numerically rated questions was used: "Rate instructor bad to good." (16) While a number of negative student comments were quoted in the department Head's letter, there were a number of very positive comments, and these were not mentioned at all.
8.....In *Christopher Turner v. The President of the University of British Columbia* (1993), the Dean said, "there were few students in undergraduate literature courses since 1986/7---(3,8, and 6 respectively," thus mistaking student 'response' figures for actual student enrolment. The Board concluded that (5) "This misunderstanding is in our opinion sufficient in itself for a reconsideration, since teaching was the focus..." (p.3), and (7) "we think that the comments and emphasis on the size of Dr. Turner's classes as evidence of poor teaching are open to objection and constitute errors of procedure and/or evidence" (p.6).
[BACK to document]

9. Given the extensive variation of rulings on SEF cases, from the perspective of a non legal professional it seem that legal reasoning carries the use of contextual analysis and variables to an extreme, making it possible---and justifiably legally---to rule just about anyway a court wants to rule. The logical extension of such reasoning would lead to each case being unique and nonsignificantly related to any other case.
[BACK to document]

10.....In *Johnson v. University of Pittsburgh* (1977), the court noted that (10) "It has also been pointed out that in some cases difficult courses have to be given to the students and the material is such that it is difficult for even the best teacher to get it across.
10.....In *Carley v. Arizona Board of Regents* (1987), he (7) characterized his professional style as being a "demanding teacher contrary to some student expectations," (8) Because of this, he maintained his popularity suffered and resulted in low student evaluations, (9) examination of his student comments indicated that Carley was correct in his assessment as 61% (49 out of 80) negative student comments focused on these values. The court ignored these findings.
10.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), it

was noted that (21) While the knowledge, interest and enthusiasm of Dr. MacLean were acknowledged, "the problem appeared to be one of style or personality."
10.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the Board noted that (26) It was obvious that almost all of the classes were upset about an examination which was considered more geography than Asian Studies, and (27) they didn't like the marking. (28) They also felt the workload was far too heavy for an "introductory" course. The Board apparently only noted this variable.
[BACK to document]

11.....In *Johnson v. University of Pittsburgh* (1977), the court said, "It is also obvious that the court and the administration of universities cannot permit students to exercise a veto over professors who may be world renowned scientists and yet if the students rate them unfavorably can be terminated at any time because of unpopularity" (p.1366-7).
11.....In *Carley v. Arizona Board of Regents* (1987), he (8) he maintained his popularity suffered as reflected in his low student evaluations
11.....In *Robert Kramer v. The President of the University of British Columbia* (1992), he maintained that (14) Student evaluations were considered from the standpoint of his popularity, not his effectiveness.
11.....In *Brian Maclean v. President of The University of British Columbia* (1991), (35) The Faculty Agreement specified that "Evaluation of teaching shall be based on the effectiveness rather than the popularity of the instructor." Courts have ruled in various directions on this issue.
11.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the board noted (21) "As for the 'popularity vs. effectiveness' debate, a discouraging or hostile attitude is a part of effectiveness as much as it is of popularity" (p.8).
11.....In *Christopher Turner v. The President of the University of British Columbia* (1993), the Board ruled, (8) while popularity is not competence nor effectiveness, to the extent that it encourages students it has some relation to both" (p.7).
[BACK to document]

12. There may well be research showing that being a popular teacher affects learning on elementary and secondary levels of education, I know of no such rigorous research on the post secondary level. In my view, one of the problems is that all too often we automatically transfer findings from elementary and secondary levels to higher education.
[BACK to document]

13.....In *Johnson v. University of Pittsburgh* (1977), the court noted that it (5) "has placed little reliance on students' surveys....students in a given course rating a teacher, or professor, some of them as excellent, others as terrible and in between, many who say passable, mediocre etc.... we cannot say it was unreasonable for the tenured faculty to consider this along with other matters" (p.1359). (8) "It is also obvious that the court and the administration of universities cannot permit students to exercise a veto over professors who may be world renowned scientists" (p.1366-7). A similar view was expressed in *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995).
13.....In *Peters v. Middlebury College* (1977), the court gave some weight to an administrative devaluing of a set of positive student evaluations of a faculty that said (2) "The department chair sent a letter to the president of the college, saying, " The course of action I recommend is not likely to be popular with students who, though they in part recognize her intellectual limitation, are warmly responsive to her enthusiasm, energy, openness and ready human concern" (p.860).
13.....In *Carley v. Arizona Board of Regents* (1987), the court said, (23) "Carley has cited

no authority that relying primarily or solely on student evaluations would be impermissible. We have found none" (p.1105, italics added).
13.....In *Guam Federation of Teachers v. The University of Guam* (1990), the Guam Federation of Teachers challenged the use of SEF in tenure and promotion decisions (Blum, 1990). The Board (1) ruled to remove anonymous student evaluations from professors' tenure files, (2) The union said the use of SEF violated the union's contract with the university, (3) which provides that anonymous documents or those "based on hearsay" should not be included in a faculty member's file, (4) The court further ruled that (5) students should be made aware of the purpose and ramifications of their evaluations, and (6) anonymous student evaluations should not be used.
13.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the Board noted that (18) "The most important perceived error in the teaching evaluation, in the opinion of the Board, is the reliance solely upon the student evaluations and written comments for the 1989 course evaluations. There was no peer review at all; no member of the Department audited any of Dr. Kramer's lectures" (p.10).
13.....In *University of Regina Faculty Association v. University of Regina* (1993) a Canadian Arbitration Board ruled that (3) "With respect to teaching, it is our opinion that the evidence of unsatisfactory performance is very weak indeed ...It is important to note that the basis of the comments, particularly the negative ones in the fall of 1992, were written student assessments... [and] Although these assessments are expressly recognized in Art. 17.19 of the collective agreement, to base important career decisions on them only does not seem justified" (p.4). The Board further ruled (4) that tenure decisions could not be based solely on assessments which were completed by students who had never been made aware of the ramifications of their statements. (5) [I]f evaluations are to be used for serious career development purposes those completing them should be aware of the potential consequences of their participation" (p.4) (8) "To base serious career decisions narrowly on student evaluations is not to be encourage... (9) If teaching is to be seriously evaluated for career purposes, whether for positive or negative purposes, it seems incumbent upon Faculties not to rely only on classroom administered evaluations but to broaden the base of assessment" (p.4).
13.....In *Christopher Turner v. The President of the University of British Columbia* (1993), the Board ruled, (9) while the [Faculty Association] Agreement permits, but does not mandate either student reviews or peer reviews, and the methods of assessment 'may vary', we do conclude that the reliance placed on these very limited student reviews must have been great, since there was no other evaluation referred to. Where there is no other evidence sought, student comments will have an apparent importance and credibility that they may not deserve... (10) We would strongly recommend peer review in the reconsideration which we are requiring" (p.7). The board further noted that (8) "This board has been asked on a number of occasions to pass judgment on the relevance of student evaluations to the [Faculty Association] Agreement criteria for good teaching. Good teaching is an elusive concept. Students may not be good judges during a course; their judgment might be quite different several years later in life. (p.7).
[BACK to document]

14.....In *Dyson v. Lavery* (1976), a student evaluation ranked her 46th of 48 teachers.
14.....In *Lieberman v. Grant* (1979), the court noted (4) a compilation of student ratings showed that the cumulative ratings for members of the department ranged from a low of 4.09 to a high of 8.95. She had a cumulative rating of 7.06, which ranked her 12th out of the 15 junior faculty members. The 7.06 figure included the ratings from a previous semester in which the plaintiff received a rating of 8.18. Prior to this rating in the spring of 1972, the plaintiff's cumulative rating was 6.7.
14.....In *Carley v. Arizona Board of Regents* (1987), it was noted that (1) of the 13 faculty

in his department of art, he was ranked fifth, (2) by his chairman he was ranked 7th, (3) student evaluations, however, ranked him last: 13th of 13 (p.1105).
14.....In *Robert Kramer v. The President of the University of British Columbia* (1992), the court noted (24) scores in the other two courses were higher---3.45 in one, 3.91 in another, against a "faculty average" of 4.22. The board further noted, "In the result, one got a 2.82 and one got a 3.07...the difference is statistically invalid in any event" (p.10).
[BACK to document]

15.15.....In *Dyson v. Lavery* (1976), the course said (1) "A number of students apparently had voiced displeasure over the quality of her class preparation and presentation" (p. 111 (3) "These *impressions*" said the court, "were largely confirmed after the initial decision to not rehire her had been made, by a student evaluation that ranked her 46th of 48 teachers in the Business Department" (p.111, italics added).
15.....In *Johnson v. University of Pittsburgh* (1977), the court said, (3) "we have *the instance* referred to in Finding 27 (p.1359, italics added).
15.....In *Lieberman v. Grant* (1979), the court noted (3) based on complaints received from "*several students*," to the effect that Lieberman's interest in feminism caused her to ignore other themes in literature (p.873, italics added).
15.....In *William Sypher v. Vermont State Colleges Faculty Federation* (1982), (1) some of the student comments noted that, "When students try to disagree he shoots you down and tries to degrade you in front of the class," (p.115), while others said, "encourages student participation as much as possible... encourages student to express their ideas freely and not worrying how 'dumb' it may sound...always wants you point of view." (P.115) (2) With regard to the numerical ratings, the Board's opinion was that (3) "regardless of a strong majority of students' rating his teaching as above average, (4) the existence of a significant minority of students feeling degraded, humiliated, and embarrassed can reasonably lead an evaluator to question a teacher's effectiveness" (p.115).
15.....In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), the Board said, (22) "the statistical comparison does not take account of the comments made by students on the evaluation forms. Grievant's student evaluations are striking in how often mention is made of Grievant's communication difficulties, particularly language difficulties (p.304-5). The board further noted with respect to comments that while some students had written that she was a "slant eyed bitch," and that she should "go back to China," (30) "We also are not persuaded that the racism evident in the student evaluations of Grievant made student evaluation results unreliable. The percentage of evaluations in which racism by students was evident was approximately one percent of the total evaluations" (p.306).
15.....In *Robert Kramer v. The President of the University of British Columbia* (1992), (2). The department Head viewed Kramer's 1989-90 course evaluations "with some alarm"....(4) Even more disturbing to the department Head was that a considerable number of students in their written comments stated that Dr. Kramer was biased, sarcastic, and hostile to the material and that a number of students had stated that Dr. Kramer's teaching would cause them to stay away from the Asian Studies department. (5) There were also some diametrically apposed positive comments" (p.10).
15.....In *University of Regina Faculty Association v. University of Regina* (1993), The Board argued (6) that the University was under an obligation to verify negative comments before acting on them. Consequently, (7) the fact that Dr Jalan had received some negative evaluations from students could not be used to undermine the otherwise generally favorable comments he had received in his annual performance reviews" (p.4).
15.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), the court noted that (25) "With respect to the "qualitative" scores---i.e., the "comments," there was a clear error. The qualitative comments from a number of courses were read and

commented on, and conclusions were drawn from them which went into the "file." Both Reviewing faculty read and commented on them, as did the Department Chair in her letter to the Dean. Yet the Dean had clearly stated in a departmental memo that the qualitative comments were not to be used for administrative or promotion purposes. (26) While in the abstract there is no reason why such comments would not be relevant, if the Department had a rule against their use, or in other words if they were "for the professor's eyes only," then it was a significant breach of Departmental rules to use them" (p.31). (27) In the opinion of the Board, so long as the comments were fairly presented, they offered the PAT [Promotion and Tenure Committee] and others a better balanced view of the teaching qualities and problems of Dr. MacLean than the quantitative statements alone" (p.31). (28) The court noted that "One problem with the questionnaire is that it solicits bad points as well as good points. Despite that caveat, we conclude that the inclusion of the qualitative comments was not a significant error" (p.32).
[BACK to document]

16.16.....In *Johnson v. University of Pittsburgh* (1977), the court noted (2) they "approached this question of teaching ability with considerable doubt, in view of the fact that in prior years there does not appear to have been any criticism of her teaching and also in view of the fact that...there was evidence that the department chairman, had informed her after one of her lectures in 1971 what a great lecture it had been;" On the other hand, the court said (3) "we have the *instance* referred to in Finding 27 (p.1359, italics added).
16.....In *Fields V. Clark University* (1987, it was observed (3) a few of which, from students in Fields' seminars, were "wildly enthusiastic" about her enthusiasm, commitment and presentations; (4) a few were ambivalent; (5) with a considerable number being extremely negative, particularly (6) with regard to her large lecture classes in basic courses in sociology.
16.....In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), moreover, they said, (19) "The statistical comparison demonstrates that Grievant was evaluated higher by students than [her male colleague] with respect to upper level classes, but that (20) [the male colleague] was evaluated higher than Grievant in lower level classes. Given (21) this "mixed" result, the statistical comparison of evaluations does not demonstrate by a preponderance of the evidence that Grievant's students rated her the same, or better, than [male colleague]" (p.305).
16.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), it was noted that (20) In general, the in-class peer reports were mixed but favourable. The in-class discussions were more problematic. (p.30). (21) While the knowledge, interest and enthusiasm of Dr. MacLean were acknowledged, "the problem appeared to be one of style or personality." It was further noted that (29) "As against the low figures, they disclosed a number of good qualities in Dr. MacLean---enthusiasm for his subject, wide knowledge of the literature, much out of class assistance to students, and a commitment to seeking good work from students. (p.31). (30) The reviewing faculty report noted the comments about Dr. MacLean's "derogatory manner, biased opinion, unwillingness to listen," were matched by "clear, stimulating, very helpful after class." And, (31) "some students have told us that the comments made were not representative of the class as a whole and were unduly influenced by the process" (p.41). (32) "A number of students, both from earlier years and from his current classes, furnished letters of support, and in preparation for the appeal, some furnished affidavits with respect to particular matters such as the 'intimidation' discussion in Soc. 250 and events in Soc. 490 and 520 in the fall of 1989." (p.33)
16.....In *Robert Kramer v. The President of the University of British Columbia* (1992). (16) While a number of negative student comments were quoted in the department Head's letter, there were a number of very positive comments, and these were not mentioned at all.

(25) "We have examined all of these written comments. There was a very wide range of comments. There were not 29 comments saying sarcastic and biased comments; but there were certainly 29 comments which included either cynical, sarcastic, biased, insulting, negative, condescending, belittling, opinionated, arrogant, nihilist, and destructive.... (29) However, it would only be fair to add that there were a number of comments in favour of Dr. Kramer, stating that the student "liked the course immensely," "now interested in Asian Studies;" "helps create a relaxed atmosphere," "really enjoyed him," "very approachable and knowledgeable," "very enthusiastic," "captivates audiences with his humour," "very effective" (p.12). (30) "In the other two courses, both small, both Japanese language, there were also some negative comments" (p.12).

16.....In *Christopher Turner v. The President of the University of British Columbia* (1993), the board noted that (6) "While there is no question of Dr. Turner's competence as a teacher at all levels, teaching evaluations for the last several years show that his effectiveness is marred by what students perceive as excessive formality, lack of enthusiasm and dullness....In a previous promotion attempt, his teaching was briefly described as "very competent" but student evaluations indicate further improvement to be "better than adequate" (p.2)

[BACK to document]

17.....I wish to thank to Patrick B. Shaw, Attorney for AAUP for referring me to Ms. Linda Lott, Administrative Coordinator, Hofstra Univeristy Chapter, AAUP, who conducted a search for me of a faculty collective bargaining contract database being developed there. Ms. Lott searched the database with "several key words that relate to academic freedom, teaching methodology and student evaluations. The only word that was identified in some of the contract provisions was 'student evaluation'"(Personal communication, March 21, 1997). It should be noted that very few explicit references in the contracts to the use of signed/unsigned SEF or the use/nonuse of comments were found in this developing database. Some of the instances found are:

17.....At Rider University, the agreement stated "The College may not use course evaluations for purposes of discipline, promotion, or tenure, unless introduced for such purposes by the faculty member."

17.....At Western Michigan University, the agreement stated "Only the ratings shall be included in all promotion, reappointment, merit, and tenure recommendations, together with such other evaluations of teaching competence as may be employed by faculty members and made available. Western agrees to consider all the evidence of teaching competence that is presented in evaluating teaching faculty and shall not use unsubstantiated structured comments in personnel decisions." I have already noted the ruling at the University of Guam (Blum D. E. (1990, October 3). which stated that (1) students not being made aware of the purpose and ramifications of their evaluations, (2) the anonymous nature of student evaluations, (3) the invalid analysis of SEF, and therefore, (4) SEF in effect being anecdotal and hearsay data. Since most SEF results are prepared anonymously, an instructor has no recourse to confront his/her evaluators. As will be addressed below, the anonymous nature of SEF is beginning to also be questioned by arbitration boards.

17.....I am informed from a colleague at St. John's University (New York) that, though SEF are mandated, they are not used administratively. I suspect there are many more schools (likely those who have union contracts) that do not use SEF administratively or who limit its use. I might note here for those who maintain that without SEF used administrative that there is no quality control over instruction and that therefore student learning will suffer, to check with the schools who do not use SEF administratively for a reality check on their assumption.

[BACK to document]

18.....In *Johnson v. University of Pittsburgh* (1977), the court said, "It is also obvious that the court and the administration of universities cannot permit students to exercise a veto over professors who may be world renowned scientists" (p.1366-7), noting, "It is obvious that a professor may be possessed of excellent qualifications as a research scientist and not necessarily be able to prove his or her worth as a teacher, concluding that, (9) "in cases where one has an outstanding scientist of national or international reputation, one may decide to promote and give tenure notwithstanding inability to come across as a teacher, this however is not one of those cases" (p.1366-7).
18.....In *Yu Chuen Wei v. Vermont State Colleges Faculty Federation* (1995), (31) Wei's last claim charged that the College violated the Contract by denying her a promotion, even though both her scholarly performance and professional activities were exceptional. Article 22(E) of the College provides for otherwise granting promotion if the President decides that "performance in one of three areas has been exceptional" (p.314). The Board concluded that "Although Grievant had a significant publication record, most of it was developed before coming to Castleton" (p.315). (33) In terms of exceptional scholarship, Dr. Wei maintained she had solved a significant mathematical problem (apparently published). The Board's response was, (34) "although Grievant claimed to have solved the Erdos conjecture, [the]Dean reasonably concluded that she had not established that she actually had solved the conjecture. Under these circumstances, and given our consideration of the discrimination issue previously discussed, we conclude that (35) Grievant has not established discrimination. The Colleges reasonably, and based on legitimate reasons, concluded that Grievant had met the tenure standards in this performance area but that her performance was not exceptional" (p.315).
18.....In *Dr. Brian Maclean v. President of The University of British Columbia* (1991), the court said, (34) "while a superior research and publication record cannot overcome a poor teaching record, it might tip the scales where the teaching record was 'on the edge'" (p.10).
[BACK to document]

19. I would like to again acknowledge the invaluable assistance of Professor William Kaplin (Personal conversation, May 28, 1997). In constructing this section on the generalization of the principles disparate treatment and impact, I have gone where wiser and more skilled sailors would perhaps have elected not to sail. Without Bill's counsel I would clearly have sailed off the edge of this legal world. As it is, I may be dangerously close. But since my intent in not a strictly legal one I can perhaps be given some latitude.
[BACK to document]

20. Statistically, a case can be made that the racist comments of one percent represent a larger (but unknown) number of racist attitudes that simply were not overtly stated. Thus even one percent overtly racist comments should be a red-flag to look deeper into such a situation. And given a commitment to eradicating racial and gender discrimination, should not the overt one percent (plus) evidence of racism on a SEF be treated more seriously?
[BACK to document]

21. For example, in *Fields V. Clark University* (1987), the court noted that (10) Fields' "attacks" the university's use of her student evaluations because they were not gathered and evaluated according to accepted standards of scientific polling procedures. In response, the court agreed, saying, "She is probably correct. The use made of the student evaluations in her case, however, followed the practice at the defendant's university in other tenure decisions" (p.671).
[BACK to document]

22. Given the current attitude of the courts in non EEOC cases toward accepting the institutional interpretation of the SEF data, for those covered as members of a protected group ---assuming sufficient *prima facie* evidence to do so, and possible illegalities notwithstanding---using one's protected status could be used as a strategy for insuring a rigorous analysis of one's SEF data.
[BACK to document]

23. Once again, numerous legal variables prohibit such generalization. See Kaplin and Lee, 1995:
23.....***1.4.3.6.*** *Standards of judicial review and burdens of proof.* Postsecondary institutions have numerous processes for making internal decisions regarding the status of faculty, students, and staff, and for internally resolving disputes among members of the campus community. Whenever a disappointed party seeks judicial review of an institution's internal decision, the reviewing court must determine what "standard of review" it will apply in deciding the case. This standard of review establishes the degree of scrutiny the court will give to the institution's decision, the reasons behind it, and the evidence supporting it. Put another way, the standard of review helps establish the extent to which the court will defer to the institution's decision and the value and fact judgments undergirding it. The more deference the court is willing to accord the decision, the less scrutiny it will give to the decision and the greater is the likelihood the court will uphold it. Issues regarding standards of review are thus crucial in most litigation.
23.....In turn, standards of review are related to the "burdens of proof" for the litigation. After a court determines which party is responsible for demonstrating that the institution's decision does or does not meet the standard of review, the court allocates the burden of proof to that party. This burden can shift during the course of the litigation (see, for example, Section 3.3.2.1). Burdens of proof also elucidate the elements or type of proof each party must submit to meet its burden on each claim.
23.....***1.4.3.6.*** *Standards of Judicial Review and Burdens of Proof 35* or defense presented. Such issues are also critical to the outcome of litigation and can become very complicated (see, for example, Section 3.3.2.1).
23.....There are many possible standards of review (and likewise many variations of burdens of proof). The standard that applies in any particular litigation will depend on numerous factors: the type of institution subject to the review (whether public or private); the type of claim that the plaintiff makes; the institution's internal rules for reviewing decisions of the type being challenged; the character of the contractual relationship between the institution and the party seeking court review; and the common law and statutory administrative law of the particular state (see this volume, Section 1.3.1), insofar as it prescribes standards of review for particular situations. At a subtler level, the court's selection of a standard of review may also depend on comparative competence-the court's sense of its own competence, compared with that of the institution, to explore and resolve the types of issues presented by the case.
23.....If a court is reviewing the **substance of a decision (whether the institution is right or wrong on the merits)**, it may be more deferential than it would be if it were reviewing the adequacy of the procedures the institution followed in making its decision-the difference being attributable to the court's expertise regarding procedural matters and relative lack of expertise regarding substantive judgments (such as whether a faculty member's credentials are sufficient to warrant a grant of tenure).
[BACK to document]

24. Again I am indebted to Bill Kaplin for this important point.

[BACK to document]

25. An assumption that students are qualified evaluators of teaching effectiveness I will deal with in the following paper.
[BACK to document]

26. I am not against using SEF for feedback to faculty. Some method of assessing faculty teaching effectiveness needs to be developed. No profession can be completely self policing. In terms of using SEF to assess teaching effectiveness and its use in tenure, promotion, reappointment, and merit salary increases, however, we need to proceed much more carefully than we have. As an initial general resolution to the problem of their use, I suggest that it be used as a "red flag" that can then set in motion a systematic faculty inquiry into the situation.
[BACK to document]

27. One teaches in such a classic mode today at one's litigious peril. See Pinsker (1989).
[BACK to document]

28. Scriven further suggests that "Unfortunately, most of what you see in a classroom are the features of teaching style, and you can't use any of them, because no amount of research can justify you in counting off brownie points for style against demonstrably badly or well-performed duties. You might as well try subtracting 10% of the purse from a professional golfer's pay-out on the grounds that he or she has an inelegant swing. You must get data on all duties; and when you have that, why would you need anything else?" (p.6).
[BACK to document]

29. SEF has become politicized. At a recent faculty development meeting on campus a well known "consultant" from a business school who has published books and articles of faculty evaluation was brought in by administration to work with faculty on a SEF and teaching portfolios. I mentioned Scriven's work to him. The consultants flip response was to call Scriven's views "fringe." I let it be known that this kind of non scholarly and ad hominem response was not acceptable. He has been critiquing and developing methods of assessment for some time. He takes a rigorous no-nonsense, yet practical approach to evaluation. For readers who may not know of Scriven's background his work, he is a philosopher of science of international reputation who was co-editor of the foundational series, *The Minnesota Studies in the Philosophy of Science*. In 1967 he coined the terms *formative* and *summative* evaluations that are coin of the realm today in evaluation research (Scriven 1981); he founded and was the first president of what is now the American Evaluation Association. Thus to call Scriven's views "fringe" is at best the epitome of arrogance, at worse, its more likely due to ignorance.
[BACK to document]

30. Most rigorous statisticians, however, would as a matter of course agree with Scriven. One of the first things we learn in psychological research is that statistical correlations can not be automatically applied to individuals. How this is generally transferred even by those who understand it in one domain to a different domain is quite another---yet important educational---question in itself.
[BACK to document]

31. Research into transfer of learning suggests many traditional methods that appear to produce effective learning may in fact be counterproductive relative to long term transfer of learning. For example, it's generally the case that immediate feedback during learning results in more efficient learning. It therefore seems to logically follow that immediate feedback during learning would result in more efficient transfer of learning. Recent findings, however,

indicate that under certain conditions *delayed* feedback is more efficient. Other examples point out the difference between understanding typical learning principles on the one hand, and principles of transfer of learning (i.e., generalization of learning and long term application) on the other. Schmidt and Bjork (1992), note "we have repeatedly encountered research findings that seem to violate some basic assumptions about how to optimize learning in real-world settings." For example, increasing the frequency of information about errors to learners during practice improves their performance. The fact is, that increasing the frequency about errors can work in just the opposite manner for long term retention and for transfer. Further counterintuitive effects come from research showing that increasing the variability of a task during practice depresses performance during training, but may increase transfer of performance after training when conditions are altered from the original training situation. Still other data show that performance on solving a puzzle is virtually perfect with no delay between instruction and application, but rapidly declines as the delay is increased (e.g., where periods of delay are two weeks and one, two, three, and four months). In contrast, performance on a similar puzzle was worse than performance on the same puzzle at first, but stayed relatively constant over a delay of four-months. In other words, the transfer effect was much more persistent than the specific effects of learning a particular puzzle. Singley and Anderson (1989). And so it goes.
[BACK to document]

32. As I was writing this section (5/10/97), I heard on the NBC national news about Jon Westling, (1995) the new President and former Provost of Boston University in trouble for suggesting that disability laws requiring special standards for students are contributing to lowering academic standards. I then went on the internet and found the following: "The disability laws are sacred cows, but they must at the very least be tethered so that they cannot be used to force universities to lower academic and other standards."
[BACK to document]

33. This section should avert the frequent "he-said-she said, you-show-me-yours-I'll-show-you-mine" approach to assessment of the SEF validity literature by administrators and some faculty, where the apparent conflicts in the literature are rationalized away by simply saying "well, some studies show that SEF are largely not valid, but other studies show it is valid."
[BACK to document]

34. In my first paper (Haskell, 1997a), I suggested that one of the reasons that SEF has not been viewed as a threat to academic freedom and generated more interest is because many do not consider it high status research and do not see its encompassing implications for quality education---even after reading some of the findings on SEF. Many of my colleagues, including those who have basically supported my efforts and work on the issue have said to me "Well, you've made your point about SEF by making us aware of it and by publishing articles, why don't you now put it aside and get back to your *real* scholarship."
[BACK to document]

35. As one scholar (Damron, personal communication, April, 1997) who read a draft of this paper observed: from "the legal decisions you review in your paper it is clear that untenured and/or politically incorrect faculty are often considered to be "fair game" by administrators, with literally _any_ superficially plausibly excuse serving as a rationale for dismissal. Use of such strategies reveal that faculty are often regarded as little more than term employees who are as disposable (and replaceable) as tissues. Clearly, there is a very serious ethical issue

here, and a hugely hostile attitude toward academic freedom and faculty in general....the great variety of decisions you've reviewed and their assorted implications for the coherence and ethics of the legal processes that gave rise to them...it seems to me that many judges and arbitration panelists have little sense of how to proceed in hearing[s] involving academics." [BACK to document]

References

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

Abrami, P.C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education*, *30*(2). 221-27.

Anderson, U., & Schadewald, M., (1991). Analogical transfer and expertise in legal reasoning. *Organizational Behavior and Human Decision Processes*, *48*, 272-290.

Baez, B., & Centra, J. (1995). Tenure, promotion, and reappointment: Legal and administrative implications. *Ashe-eric Higher Education Report No. 1*.

Barnett, L.D. (1996). Are teaching evaluation questionnaires valid? Assessing the evidence. *Journal of Collective Negotiations in the Public Sector*, *25*(4). 335-349.

Bauer, H.H. (1996). The New Generations: Students who don't study (prepared for the symposium, "The Technological Society at Risk" Annual Meeting, AOAC International, Orlando (FL), 10 September 1996). Email hhbauer@vt.edu for a copy of the article.

Blum, D.E. (1990, October 3). U. of Guam removes evaluations from files. *The Chronicle of Higher Education*, Section: Personal & Professional, p. A21

Brian Maclean v. President of The University of British Columbia (1991). Held at Vancouver, B.C. January 28--June 20.

Brimelow, Peter (1996). Devalued diplomas. *Forbes*, April 22.

Carley V. Arizona Board of Regents, 737 P.2d 1099 (Ariz. App. 1987).

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

Chacko, T.I. (1983). Student ratings of instruction: A function of grading standards. *Educational Research Quarterly*, *8*(2), 19-25.

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

Cynthia J. Fisher v. Vassar College (1995). United States Court of Appeals for the Second Circuit Nos. 1179, 1303, 2275 Docket Nos. 94-7737, 94-7785, 94-9125

Christopher Turner v. The President of the University of British Columbia (1993).

Arbitration Appeals Board Case No.: AI-256 Vancouver, B.C. July 19,20 and 21.

Cohen, P.A. (1983). Comment on 'A Selective review of the validity of student ratings of instruction.' *Journal of Higher Education*, *54*, (no) 4.

Copeland, J.D., & Murry, J.W., Jr. (1996). Getting tossed from the Ivory Tower. *Missouri Law Review*, *61*, 233-327.

Divoky, J.J., & Rothermel, A. (1988). Student perception of the relative importance of dimensions of teaching performance across type of class. *Educational Research Quarterly*, *12*, 40-45

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

Dyson V. Lavery, 417 F.supp. 103 (E.d. Va. 1976).

Dawes, R.M., Faust, D., & Meehl, P.E. (1989). Clinical versus actuarial judgment. *Science*, *243*, 1668-1674

Dilts, D.A., Samavati, H., Moghadam, M.R., & Haber, L.J. (1994). Student evaluation of instruction: Objective evidence and decision making. *Journal of Individual Employment Rights*, *2*, 73-86.

Ellul J. (1964). *The technological society*. New York: Vintage Books.

Faust, D., Guilmette, T.J., Hart, K., Arkes, H.R., Fishburne, F.J., & Davey, L. (1988). Neuropsychologists' training, experience, and judgement accuracy. *Archives of Clinical Neuropsychology*, *3*, 145-163.

Feldman, K.A. (1983). Seniority and experience of college teachers as related to evaluations they receive from students. *Research in Higher Education*, *18*, 3-124

Fields V. Clark University (1986). Civil Action No. 80-1011-s, United States District Court for the District of Massachusetts, 40 Fair Empl. Prac. Cas. (Bna) 670, March 14,; Vacated and Remanded May 8, 1987.

Fighting grade inflation. (1994) *Science*, *264* (27 May), 1255.

Franklin, J., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. In Theall, M. & Franklin J. (Eds.), *Student Ratings of Instruction: Issues For Improving Practice*. San Francisco: Jossey-Bass.

Garb, H.N. (1989). Clinical judgment, clinical training, and professional experience. *Psychological Bulletin*, *105*, 387-392.

Golding, M.P. (1984). *Legal reasoning*. New York: Alfred A. Knopf.

Greenwald, A.G. (1996). *Applying social psychology to reveal a major (but correctable) flaw in student evaluations of teaching*. University of Washington, Draft Manuscript, March 1, submitted for publication.

Greenwald, A., & Gillmore, G. (1997). No pain, no gain? The importance of measuring

course workload in student ratings of instructions. *Journal of Educational Psychology* (forthcoming).

Haskell, R.E. (1997a). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century. *Education Policy Analysis Archives*, *5*(6) [Refereed Online Journal]. Available: http://olam.ed.asu.edu/epaa/v5n6.html

Haskell, R.E. (1997b). Abridgement of academic freedom, promotion, reappointment and tenure rights by the administrative use of student evaluation of faculty: (Part II) Views from the court. *Education Policy Analysis Archives*, *5*(6), [Refereed Online Journal]. Available: http://olam.ed.asu.edu/epaa

Hayes, S.C. (1991). The emperor's clothes: Examining the 'delusions' of professional psychology: The healthy skepticism of David Faust. *Science*, *1*, 22-25.

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

Howard, G.S., & Maxwell, S.E. (1982). Do grades contaminate student evaluations of instruction? *Research in Higher Education*, *16*, 175-188.

Howard, G.S., & Maxwell, S.E. (1980). Correlation between student satisfaction and grades: A case of mistaken causation? *Journal of Educational Psychology*, *72*, 810-820.

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

Johnson V. University of Pittsburgh (1977). 435 F.supp. 1328 (W.d. Pa.).

Kaplin, W.A., & Lee, B. (1995). *The law of higher education: A comprehensive guide to legal implications of administrative decision making*. San Francisco: Jossey Bass.

Larkin, J., McDermott, J., Simon, & Simon, H. (1980). Expert and novice performance in solving physics problems. *Science*, *208*, 1335-1342.

Lasch, C. (1979). Schooling and the new illiteracy (Chapter Six), in *The culture of narcissism: American life in an age of diminishing expectations*. New York: Warner Books.

Lemons, J. (ed.) (1996). *Scientific uncertainty and environmental problem-solving*. Cambridge, MA: Blackwell Science

Lemons, J., Shrader-Frechette K., & Cranor C. (in press). *The precautionary principle: Scientific uncertainty and type-I and type-II errors*. Foundations of Science.

Leo, J. (1996). No books, please; We're students. *U.S. News and World Report*, September 16.

Levi, E.H. (1949). *An introduction to legal reasoning*. Chicago: University of Chicago Press.

Lieberman V. Grant (1979). 474 F.supp. 848 (D.conn.).

Lovelace v. Southeastern Massachusetts University (1986). 793 F.2d 419 (1st Cir.).

Marchant, G., Robinson, J., Anderson, U., & Schadewald, M. (1991). Analogical transfer and expertise in legal reasoning. *Organizational Behavior and Human Decision Processes*, *48*, 272-290.

Marsh, H.W., & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. *Higher Education: Handbook of Theory and Research*, *8*, 143-233.

Marsh, H.W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology*, *76*, 707-754.

Marsh, H.W. (1982). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology*, *74*, 264-279.

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

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

Neath, I. (1996). How to improve your teaching evaluations without improving your teaching. *Psychological Reports*, *78*, 1363-1372.

Peters V Middlebury College (1977). 409 F.supp. 857 (D. Vt.).

Pinsker, S. (1989). Teaching in a litigious age. *Change*, *21*, (July/August), 50-54.

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

Rabinowitz, J. (1993). Diagnostic reasoning and reliability: A review of the literature and a model of decision-making. *The Journal of Mind and Behavior*, *14*, 297-316

Rebell, M.A. (1990). Legal issues concerning teacher evaluation. In Millman, J., & Darling-Hammond, L., (Eds.), (pp. 337-355) *The new handbook of teacher evaluation*. Beverly Hills, CA: SAGE Publications.

Robert Kramer v. The President of the University of British Columbia (1992). Arbitration Appeals Board Case No. AI-245. Held at the University April 22 and 23.

Sacks, Peter Generation X Goes to College, Chicago & LaSalle (IL): Open Court, 1996.

Singley, M.K., & Anderson, J.R. (1989). *The transfer of cognitive skill*. Cambridge, MA: Harvard University Press (p.199).

Scheelhaase v. Woodbury Central Community School District (1973), In Rebell, M, A. (1990). Legal issues concerning teacher evaluation. In Millman, Jason& Darling-Hammond, Linda, (Eds.), (pp. 337-355) The new handbook of teacher evaluation. Beverly Hil ls: CA. Sage Pub.

Schmidt, R. A., & Bjork, R.A.(1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*. 207-217.

Scriven, M. (1988). The new crisis in teacher evaluation: The improper use of research-based' indicators. *Professional Personnel Evaluation News*. (January) p.4.

Scriven, M. (1987). Validity in personnel evaluation. *Journal of Personnel Evaluation in Education*, *1*, 923.

Scriven, M. (1991). *Evaluation Thesaurus*. (4th Ed.) Newbury Park, CA: Sage Publications

Scriven, M. (1993). The validity of student ratings: In Teacher evaluation. *Evaluation & Development Group*.

Scriven, M. (1995). A unified theory approach to teacher evaluation. *Studies in Educational Evaluation*, *21*, 111-129

Scriven, M. (1981). Summative Teacher Evaluation. In J. Millman (Ed.) *Handbook of teacher evaluation* (pp. 244271). National Council on Measurement in Education. Beverly Hills, CS: Sage Publications.

Simon, W.E. (1996). The dumbing down of higher education. *Wall Street Journal*, March 19.

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

Stern, J., & Flynn, P.D. (1995). Students propose a course of action for grade inflation. *The Bucknellian*. (Feb, 20).[Online] Available:

Sunstein, C.R. (1993). On analogical reasoning. *Harvard Law Review*, *106*, 741-79.

University of Regina V. University of Regina Faculty Association and Dr. Pradeep Jalan. (1993). Case No.: AI-298 Regina, Saskatchewan, July 9.

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

Westling, J. (1995). Getting Government Out of Higher Education. Boston University, delivered May 3. Available Online: http://www.heritage.org/heritage/library/categories/education/lect533.html

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

William Sypher v. Vermont State Colleges Faculty Federation. (1982). 5 VLRB 102.

Yu Chuen Wei and the Vermont State Colleges Faculty Federation. (1995) 18 VLRB 261.

Appendix: A Non Litigated Case of SEF Used in the Denial of Tenure and Reappointment

.....The following is a case of the denial of tenure and reappointment primarily on grounds of apparent non outstanding teaching as measured by SEF. The case illustrates most of the

issues discussed in this paper. As in most such denials, Dr. Tichenor did not go to litigation as litigation is costly given the low odds of the courts finding in favor of faculty. The case is an example of the likely thousands of such cases (given that there are over 3,500 colleges in the U.S.) that do not go to litigation. In most respects this case is probably typical relative to the inappropriate use of SEF in faculty dismissals.

.....The case involves Dr. Linda L. Tichenor, assistant professor of biology, who was denied reappointment during the 1996/97 academic year at a small private, student tuition-dependent university. The SEF material presented here was given to me for use here by Dr. Tichenor. It is from a draft paper by her that will appear in a special monograph series to be published by *The Society of College Science Teaching*, edited by Mario W. Caprio.

.....The unique aspect of this case of non reappointment due to SEF, ostensibly reflecting teaching ineffectiveness, is that Dr. Tichenor graduated from Idaho State University, Doctorate of Arts program in biology. The program, established by the Carnegie-Mellon Foundation in the late 1960's, specifically prepares future faculty for college science teaching pedagogy as well as for teaching a broad range of courses within the discipline of life sciences. The program requires a breadth of background in the biological sciences, knowledge of learning processes and pedagogy, awareness of the objectives of an undergraduate education, and the development of a sound educational philosophy. "As a result of my training," says Dr. Tichenor, "my classroom was to become my 'research bench.'" This is fairly unique in higher education. Few higher education faculty have any formal training in teaching (with the exception of faculty in the discipline of Education). Granted, while a degree in teaching does not guarantee effective teaching, it does attest to skills that most other faculty have not formally acquired. Dr. Tichenor has been awarded research grants and has published her pedagogical views and experiences in college science teaching journals.[1]

.....As Dr. Tichenor points out, "several national reports have called for reform in college science teaching (Michael 1989; AAAS Report on the National Science Foundation Disciplinary Workshops on Undergraduate Education 1990; Moore 1993; Sigma Xi 1990). The AAAS Report (1990), suggests that conventional science courses do not reflect the practice of science "at its best." The report recommends that pedagogical techniques be directed toward open-ended and investigatory laboratories in order that the teaching of science be driven by real problems rather than contrived textbook exercises and bring the spirit of scientific inquiry to undergraduate studies. The idea of student-designed laboratories supports the pedagogy suggested to foster student inquiry. I implemented these types of laboratories into my physiology course (Tichenor, 1996)."

.....At her university, it is required that teaching be assessed as "outstanding" for tenure to be granted. A faculty tenure committee rates teaching efficacy from (a) peer reviews, (b) student teaching evaluations, (c) yearly departmental chair reports, and (d) other evidence submitted by the candidate. Dr. Tichenor notes that, "Although the committee's final report in my case was positive overall, tenure was denied because my teaching was considered not 'outstanding.' Its decision was supported by the dean of my college." The letter from the faculty Reappointment, Promotion and Tenure committee reads:

> Dr. Tichenor has an unusual doctorate (my emphasis) focused on college teaching in the sciences and is very interested in non-traditional teaching methods. In spite of, or maybe due to this, her teaching evaluations are mixed. The review process is exacerbated when consistently mixed student evaluations are at odds with supportive statements made by peer reviewers. Often, the latter reflect peer approval of the philosophy and pedagogical approach of the instructor, while student comments address the instructor's relative success in facilitating the students' learning. Such dichotomy of opinion is especially problematic when attempting to evaluate a candidate for tenure. Contradictory statements

are quite evident in this candidate's teaching evaluations. Her dean described her teaching effort as energetic, innovative and valuable to the University, but concluded that she had not distinguished herself in this preparation.

.....In reaching their decision, the review committee reportedly did not do a statistical analysis and comparison of her SEF with other faculty and selectively picked out certain negative comments by some students.
.....From the four criteria of (a) peer reviews, (b) student teaching evaluations, (c) yearly departmental chair reports, and (d) other evidence submitted by the candidate, only SEF was apparently used for the denial decision. Nearly the entire department faculty supported her and protested as a collective body the denial of tenure decision to the Dean and to the President, to no avail. A subsequent grievance committee failed to reverse the negative reappointment decision.
.....In commenting on her SEF responses, Dr. Tichenor points out that here evaluations in general physiology over a four-year period reflect evidence of difficulties that students have with innovative teaching:

Q: What do you suggest to improve this course?
A: 'Rely more on traditional teaching concepts than on "progressive" new teaching ideas . . '
A. 'In my opinion, we could have learned a lot more had we stuck to a conventional lecture format.'
A. 'Shifting to a more traditional style would be an improvement, the way it is set up now is a nice idea, but doesn't work as well as its supposed to.'
A. 'More and better notes would be good.' and finally,
A. 'I would suggest that there be more lecturing and less time sitting in a circle and wasting time staring at our neighbors and then when questions are asked we are told to look it up for the exam.'

.....On the other hand, some students who have grasped the active learning style more readily may say:

Q: What part of the course was most beneficial to you?
A. 'The fact that we had to learn through our own research, interests, etc. were most beneficial to me. I learned a lot more on my own through independent presentations than I would have merely listening to a lecture on the subjects.'
A. 'Learning groups require people to be able to communicate with one another and therefore teach communication skills not usually taught in school. Learning groups also help people to learn problem solving skills and ways of approaching problems. Learning these skills are very important. When students leave college they are usually filled with a lot of book knowledge but not much knowledge of how to approach solving a problem. Learning groups can do this because the student is responsible for assimilating information and helping other students understand it. Also, group learning is important because it helps students learn to work in a team and take responsibility for themselves and others.'
A. 'The part of the course most beneficial to me was the group discussions in class about the material. We are all able to relate to terms and events that enables the material to be better understood.'

.....Dr. Tichenor, perhaps typically, observes: "The association of mixed reviews and innovative teaching is probably familiar to many teachers who take the risk of implementing these methods. By innovative, I mean something different from the "teacher-as-textbook" model of teaching." She also, perhaps typically, that "I have found that numerical ratings for me are more positive than the open ended questions; and as it turns out, my [numerical]

teaching evaluations are not at all 'mixed' but above average."
.....In terms of preparing students for a non lecture teaching format she has engaged in-class discussions "on primary literature, oral presentations, cases studies presented both orally and written formally, student-designed laboratories, peer evaluation, and other active learning strategies. Experiences students have disliked the most have been class periods built upon a discussion of previously assigned readings. Even though I give them a list of discussion questions prior to the class, they sometimes fail to read the discussion material."
.....Dr. Tichenor concludes: "there are some difficulties with the use of active learning models. Since most students have grown up with the traditional lecture method, many assume that lecturing is the only way to conduct a class. Students may feel cheated by a new approach especially if they are asked to generate their own material for class. This innovative style places new and unfamiliar demands on students. Students may feel a lack of self confidence and feel overwhelmed with the type of work expected of them."

Commentary

.....This case demonstrates most of the findings outlined above in this paper: (1) assumed validity of SEF, (2) reliance on SEF for administrative assessment of teaching effectiveness, (3) the subjective interpretation of SEF data by (4) untrained evaluators, both administrative and faculty, (5) reliance on SEF over peer evaluation, (6) ignoring the many variables in the implicit comparison of SEF data involved in the final decision that Dr. Tichenor was not outstanding, including (7) student bias variables, (8) selective use of qualitative written student comments, (9) over superior numerical averages on the SEF instrument, (10) justifying the interpretation of unacceptable teaching effectiveness by selective and subjective emphasis on the negative student comments in a mixed series of comments that include positive ones, (11) assumes the metaphor of student as consumer, (12) that students should have "vote" in what is appropriate teaching methods, which in term assumes, (13) students are qualified to do so, and (14) assumes that SEF validly measures teaching effectiveness.
.....There is an interesting set of interrelated ironies involved in this case. The university, which prides itself on being a teaching institution in apparent contradistinction to research oriented universities, denied Dr. Tichenor tenure and reappointment, a faculty who not only (a) has a rare doctoral degree that specifically prepared her for college science teaching, but (b) was engaged in what the literature suggests is one of the most effective teaching methods---collaborative and student-centered teaching----,[2] (c) was creatively trying to improve on those methods, (d) had been awarded a related grant, and (e) had published articles on her teaching in professional journals. A final irony is that Dr. Tichenor has accepted a new appointment at large state university where part of her duties will be to assess teaching.
.....Given the teaching orientation of the university, it would seem reasonable to expect support and encouragement for such teaching activity, especially in the light her numerical SEF score being above average. The question remains as to why this irony exists. While the answer is certainly complex, one of the answers is to retain student tuition dollars. I will deal in more detail with this issue in my final paper.

Appendix Endnotes

1. Tichenor, Linda L. (1996). Student-designed physiology laboratories: Creative instructional alternatives at a resource-poor New England university. *Journal of College Science Teaching*, *26*(3), 175-181; Tichenor, Linda L. and Joseph Kakareka. (1995). An

interdisciplinary teaching approach by integrating cell biology and biochemistry: A scientific learning community at the University of New England. *Journal of College Science Teaching*, *25*(2), 144-149.
[BACK to document]

2. I might note that I am not a supporter of collaborative, student-centered teaching methods for certain populations of students such as are currently found on many colleges.
[BACK to document]

About the Author

Robert E. Haskell

Robert E. Haskell has been teaching college and university level courses for over twenty years. He earned his Ph.D. from the Pennsylvania State University in Psychology and Social Relations, his M.A., and B.A. from San Francisco State University. His areas of research and teaching include: transfer of learning, analogical reasoning, small group dynamics. Major publications include: four books, the latest of which is, *The Future of Education and Transfer of Learning: A Cognitive Theory of Learning and Instruction For The 21st Century* (forthcoming), and numerous presentations, chapters, and research articles in national and international journals. He also serves on several editorial review boards, and is Associate Editor of The Journal of Mind and Behavior. He is former Chair, and currently Professor of Psychology, Department of Social and Behavioral Sciences, University of New England.

Professor of Psychology
University of New England
Biddeford, Maine 04005

UNE Home Page: http://home.maine.rr.com/une/

E-mail: haskellr@maine.rr.com

The World Wide Web address for the *Education Policy Analysis Archives* is
http://olam.ed.asu.edu/epaa

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, Glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-2411. (602-965-2692)

EPAA Editorial Board