EVIDENTIARY AND CONSTITUTIONAL DUE PROCESS CONSTRAINTS ON THE USES BY COLLEGES AND UNIVERSITIES OF STUDENT EVALUATIONS

ROGER W. REINSCH[*]
SUSAN M. DES ROSIERS[**]
AMY B. HIETAPELTO[***]

"I continue to believe that before the decision is made to terminate an employee's wages, the employee is entitled to an opportunity to test the strength of the evidence 'by confronting and cross-examining adverse witnesses and by presenting witnesses on his own behalf, whenever there are substantial disputes in testimonial evidence.'"[1]

INTRODUCTION

> We do not know if Plato was invited by Socrates to evaluate his discourse; nor do we know whether or not Aristotle insisted on giving Plato teaching feedback! What is certain, however, is that if he were teaching at a modern business school, Plato would have little alternative but to confront ratings from all of his students.[2]

Faculty and teaching professionals at most (perhaps all) colleges and universities today in the United States are subject to summative student evaluations.[3] Summative student evaluations[4] use numerical scores to regularly establish and

[*]Associate Professor, College of Business and Management, Northeastern Illinois University; B.S. Southwest Missouri State, 1973; J.D. University of Missouri-Columbia, 1981.

[**] Partner, Attorney Des Rosiers of Wisconsin; B.F.A. University of Wisconsin-Eau Claire, 1971; J.D. University of Wisconsin Law School, 1978.

[***]Associate Dean, College of Business and Management, Northeastern Illinois University; B.S. Michigan State University, 1978; M.B.A. Michigan State University, 1980; Ph.D. University of Minnesota, 1997.

1. Cleveland Bd. of Educ. v. Loudermill, 470 U.S. 532, 548 (1985) (Marshall, J., concurring) (emphasis omitted) (quoting Arnett v. Kennedy, 416 U.S. 134, 214 (1974) (Marshall, J., dissenting)).

2. Thomas E. Barry & Rex Thompson, *Some Intriguing Relationships In Business Teaching Evaluations*, 72 J. EDUC. FOR BUS. 303, 303 (1997).

3. *See*, Peter Seldin, *How Colleges Evaluate Professors: 1983 v. 1993*, AAHE BULL., Oct. 1993, at 6.

4. Hereinafter "student evaluations."

compare median scores of faculty in order to make employment and personnel decisions, including those regarding hiring, merit, promotion, retention, and tenure, often with life-changing consequences.[5] Student evaluations are widely used in both public and private colleges and universities, including unionized and non-unionized environments, as a standard technique for assessing the teaching effectiveness of faculty. Given the life-changing nature of employment and personnel decisions based upon student evaluations, and given "the evaluation process is important to the teacher, the student, the educational institution, and society itself,"[6] it is critically important that their meaning, use, and constitutionality be well understood when such crucial administrative decisions are made.[7] Given today's heightened emphasis on accountability and assessment, an increase in importance of the usage of student evaluations for assessment can only be expected. Yet, as shall be seen, the use by administrators of student evaluations with questionable validity, reliability, and evidentiary usefulness raises a substantive due process issue.

I. THE PROCESS AND USE OF STUDENT EVALUATIONS

At most colleges and universities, students are given evaluation question and answer sheets during a regularly-scheduled class period in one of the last few weeks of the semester. An answer sheet typically uses a Likert scale[8] rating system—often from one through five—that students utilize to rate a professor. The answer sheets are then processed to provide a mean and median for each faculty member for each class he or she teaches. Students' feedback remains anonymous supposedly in order to promote honest evaluations and to alleviate

5. *See* Philip C. Abrami, *Improving Judgments About Teaching Effectiveness Using Teacher Rating Forms*, New Directions for Institutional Research, Spring 2001, at 59–60 ("Anecdotal reports suggest that there is wide variability in how promotion and tenure committees use the results of [Teacher Rating Forms]. At one extreme are reports of discrimination between faculty and judgments about teaching based on decimal-point differences in ratings. Experts in the area are often shocked to learn of such decisions but do not have sufficient means to prevent such abuses. At the other extreme are reports that discrimination between faculty and judgments about teaching fail to take into account evidence of teaching effectiveness (in other words, instructors are assumed to teach adequately), meaning that the importance of instructional quality is substantially reduced when assessing faculty performance.").

6. Deborah C. Haynes & Holly Hunts, *Using Teaching Evaluations as a Measurement of Consumer Satisfaction*, 46 CONSUMER INTERESTS ANN. 134, 134 (2000). *See also*, Cathy King Pike, *A Validation Study of an Instrument Designed to Measure Teaching Effectiveness*, 34 J. SOC. WORK EDUC. 261 (1998); Mark Clayton, *Give Me an 'A' Professor—I'll Give You One Too*, CHRISTIAN SCI. MONITOR, Mar. 17, 1998 at B6; and Susan S. Lang, *Student Ratings Soar When Professor Uses Enthusiasm*, 25 HUMAN ECOLOGY F., Fall 1997, at 24.

7. *See* Janice L. Nerger, Wayne Viney, & Robert G. Riedel II, *Student Ratings of Teaching Effectiveness: Use and Misuse*, 38 MIDWEST Q. 218 (1997) (discussing the various issues with student evaluations).

8. A type of survey question where respondents are asked to rate the level at which they agree or disagree with a given statement. A Likert scale is used to measure attitudes, preferences, and subjective reactions. *See* Usability Glossary: Likert Scale, http://www.usabilityfirst.com /glossary/main.cgi?function=display_term&term_id=968 (last visited Nov. 6, 2005).

students' and administrators' concerns regarding the potential of "retaliation" by a faculty member receiving unfavorable evaluations. This concern is especially acute in small classes where such identification might be made more easily and in graduate-level classes where faculty members exert considerable control over important outcomes other than grades (e.g., fellowships, theses, dissertations). The median and mean generated by the electronic analysis is then compared to all of the other faculty scores in a department, college, or entire campus in order to quantify and rate the professional characteristics of a particular faculty member or group of faculty members. College and university administrators use the evaluation scores and numerical referencing as factors to determine the "quality" of teaching and/or the "qualifications" of the teacher, often without giving any substantial weight to alternative assessment mechanisms, such as peer and self evaluations.

Even though at some institutions other sources of information are gathered, in practice, the most important pieces are the student evaluations.[9] The reason for this reliance on student evaluations is that they provide quantitative evidence that appears to be "black and white."[10] On the surface, at least, student evaluations produce numbers, which seem not to lie.[11] One author calls this the "micrometer

9. *See, e.g.,* Mary Gray & Barbara R. Bergmann, *Student Teaching Evaluations: Inaccurate, Demeaning, Misused*, ACADEME, Sept.-Oct. 2003, at 44; Kathryn M. Obenchain, Tammy V. Abernathy, & Lynda R. Wiest, *The Reliability of Students' Ratings of Faculty Teaching Effectiveness*, 49 C. TEACHING 100 (2001) (noting that student evaluations are critical in tenure and promotion decisions and much emphasis is placed on student evaluations and faculty concerns regarding the use of student-completed evaluation forms as the sole or most important assessment of teaching quality have been well documented) (citing Robert E. Haskell, *Academic Freedom, Tenure, and Student Evaluations of Faculty: Galloping Polls in the 21st Century*, 5 EDUC. POL'Y ANALYSIS ARCHIVES (1997), *available at:* http://epaa.asu.edu/epaa/v5n6.html; Herbert W. Marsh, *Students' Evaluations of University Teaching: Research Findings, Methodological issues, and Directions for Future Research*, 11 INT'L J. EDUC. RES. 253 (1987); William E. Cashin, *Concerns About Using Student Ratings in Community Colleges*, NEW DIRECTIONS COMMUNITY COLLEGES, Mar. 1983, at 57; S.F. Mark, *Faculty Evaluation in Community College*, 6 Community Junior College Research Quarterly 167 (1982)); William J. Read, Dasartha V. Rama, & K. Raghunandan, *The Relationship Between Student Evaluations of Teaching and Faculty Evaluations*, 76 J. EDUC. FOR BUS. 189, 192 (2001) ("The results from our study show that SEs continue to be the tool most used by administrators of accounting departments for evaluating teaching."); Ronald L. Jirovec, Chathapuram S. Ramanathan, & Ann Rosengrant Alvarez, *Course Evaluations: What are Social Work Students Telling Us About Teaching Effectiveness?*, 34 J. SOC. WORK EDUC. 229, 229 (1998) ("Because student evaluations often receive paramount consideration when assessing teaching effectiveness, they contribute greatly to perceptions of a faculty member's competence among colleagues and administrators."); Nerger, *supra* note 7, at 218 ("In recent years, data from student ratings of instructional activities of faculty have occupied an increasingly conspicuous role in tenure, promotion, and salary exercises."); George W. Carey, *Thoughts on the Lesser Evil: Student Evaluations*, 22 PERSP. ON POL. SCI. 17, 17 (1993) ("[At Georgetown University,] student evaluations are by far the most important factor in determining how many teaching points an individual will receive.").

10. *See*, *e.g.*, David A. Dowell & James A. Neal, *The Validity and Accuracy of Student Ratings of Instruction: A Reply to Peter A Cohen*, 54 J. HIGHER EDUC. 459 (1983).

11. *See*, *e.g.*, Richard L. Abel, *Evaluating Evaluations: How Should Law Schools Judge Teaching?* 40 J. LEGAL EDUC. 407 (1990).

fallacy" because one attributes meanings to numbers simply because they can be calculated.[12] "Administrators may understand that the ratings imperfectly reflect actual teaching but continue to use them anyway because they are inexpensive to administer and provide data that can be interpreted in various ways to suit the administration's purpose."[13] In fact, colleges and universities rely heavily on student evaluations as measures of teaching effectiveness primarily because they are inexpensive, quantifiable, and easy to acquire, not because evaluations tell them much about teaching effectiveness. Well-conceived and -executed teaching assessment programs would take considerable time, energy, and money.[14]

The practice of solely or primarily using student evaluations to make these decisions goes counter to what most writers in the field recommend, because the almost universal recommendation is the use of multiple sources and types of data. As one author states, "Viewing student ratings as data rather than as evaluations [of teaching quality] may also help to put them in proper perspective. . . . No single source of data, including student rating data, provides sufficient information to make a valid judgment about teaching effectiveness."[15] Student evaluations should only be utilized as one source of data; their use must be considered both in the context of who provided the data and under what circumstances the data was provided. Effective assessment of teaching requires triangulation of multiple methods, including both direct and indirect assessment measures. The typical usage, however, means that student evaluations carry an inordinate amount of weight in making life-changing decisions regarding faculty teaching competence. In most situations, administrators' use of student evaluations is to determine whether or not a faculty member has the skill and ability to help students learn the material. If a faculty member is above the median then that faculty member is considered to be a "good" teacher. The higher a given faculty score above the median, the better that faculty member is perceived as a teacher. The presumption is that the higher the faculty member scores above the median, the more the students will learn. However, few would contend that college and university teaching and learning have improved as the use of ratings has increased.[16] Just as challenging as the absence of reliable and valid measures of both teaching and learning is the resistance to developing them within academia itself; yet such assessment is increasingly needed.[17]

A. Effective Teaching—What Is It?

Because of the inordinate amount of weight given to student evaluations and the

12. *Id*. at 428–29 (citations omitted).

13. Judith D. Fischer, *The Use and Effects of Student Ratings in Legal Writing Courses: A Plea for Holistic Evaluation of Teaching*, 10 LEGAL WRITING 111, 121–22 (2004).

14. Richard H. Hersch, *What Does College Teach?*, Atlantic Monthly, Nov. 2005, at 140.

15. WILLIAM E. CASHIN, CTR. FOR FACULTY EVALUATIONS & DEVELOPMENT, IDEA PAPER NO. 20, STUDENT RATINGS OF TEACHING: A SUMMARY OF THE RESEARCH (Kan. St. Univ. Div. Continuing Educ. 1988), http://www.idea.ksu.edu/index.html.

16. Fischer, *supra* note 13, at 112.

17. Hersch, *supra* note 14.

fact that they are used to purportedly measure effective teaching, there are a number of issues that need to be addressed. One such issue is how those involved in the process define "learning." Faculty members, students, and administrators each have different views as to what should occur in the classroom to help students "learn."

In fact, not surprisingly, research has found that teaching effectiveness is defined differently by students and the institution.[18] Research on intellectual development has shown that "knowing" evolves "from absolute knowing, through transitional and independent knowing, to contextual knowing."[19] This research has shown that most students enter the college or university at the very first stage of knowing and do not reach the last stage until after they have graduated.[20] At the start of their college or university career, students are only at the absolute knowing stage and may evolve to the transitional knowing stage, and, therefore, expect the material to fit those states of knowing. At both of these stages they want to be passive recipients of information, not active participants in the learning process.

Many faculty understand that students developmentally progress from a stage of

18. *See* Obenchain, *supra* note 9 ("Whereas previous studies looked for reliability across a group of students and over time, the comparison in this study examined the reliability of the individual student. However, if, as this study found, individuals are not consistent in their evaluations, then aggregated reliability measures are giving faculty a false sense of security. . . . The reliability of student evaluators may be further confounded by research indicating that student-completed evaluations measure 'popularity of the instructor' rather than 'teaching effectiveness'. . . . Teachers perceived as enthusiastic, good-humored, and warm consistently fare better on student evaluations. Although these characteristics are pleasant, they do not equate with teaching effectiveness. As stated earlier, student-completed evaluations are more about the instructor than about the actual course. The results, then, could be not just an issue of unreliable student evaluations or an invalid instrument. Rather, the system for evaluation itself may be inconsistent. This system requires that students use an instrument corresponding with the institution's definition of teaching effectiveness, rather than the students' definitions of teaching effectiveness. This inconsistency only becomes evident when students complete multiple measures, including measures that reflect the institution's definition and ones that reflect the students' definitions.") *Id.* at 102–03.

19. JANET DONALD, LEARNING TO THINK: DISCIPLINARY PERSPECTIVES 3 (2002), *available at* http://s11.Stanford.edu/projects/tomprof/newtomprof/postings/405.html.

20. *Id.* at 3–4 ("In Baxter Magolda's longitudinal study, most students—68 percent—entered university in a stage of absolute knowing, considering knowledge to be certain or absolute and conceiving their role as learners to be limited to obtaining knowledge from the instructor. The remaining 32 percent of entering students were in a stage of transitional knowing, considering knowledge to be partially certain and partially uncertain; their role was to understand knowledge. In both stages, students depict themselves as passive recipients of their professors' wisdom. During their senior year, some students—16 percent—displayed independent knowing; that is, they considered knowledge to be uncertain. In this stage, everyone has his or her own beliefs, and students are expected to think for themselves, share views with others, and create their own perspective. Independent knowing increased to 57 percent the year following graduation. Only in the year following graduation did a small number of students—12 percent—reach the stage of contextual knowing, where knowledge is judged on the basis of evidence in context, and the student's role is to think through problems and to integrate and apply knowledge. These findings suggest that two-thirds of entering students limit their role as learner to obtaining knowledge, and most will not be actively constructing meaning (independent knowing) until after they have graduated.") *Id.*

"transitional knowing" (acquisition of facts) to a stage of realizing that knowledge is not absolute and that understanding is more crucial. This stage of "independent knowing" heralds the perception that knowledge is uncertain and the creation of one's own perspective is paramount. This paves the way for development to the final stage, "contextual knowing," where independent thinking remains vital but is now contextualized so that certainty is dependent on context.[21] Much of learning in colleges and universities, especially at the undergraduate level, occurs at the first two stages; yet significantly more ought to occur in the latter stage. Faculty who believe that transitional knowing is not adequate education thus teach with the purpose of trying to achieve, at a bare minimum, the independent knowing stage, and, ideally, the contextual knowing stage. Hence, their approach to teaching is very different from the students' expectations of what teaching ought to be, and often much more demanding of students' efforts than students expect or believe teaching ought to be.

Many faculty members believe that their responsibility is to teach material that will be useful in the students' future professional career. Faculty members are making professional decisions as to pedagogy and the substantive content of the class. In Uniform Commercial Code terms, faculty want to teach material that is "merchantable"—fit for the purpose of the students' careers and for the reasonable future into each student's career.[22] This requires that students learn the material, understand the material, and be able to apply the information—think critically—in their future career decision-making situations. The goal of faculty is to get students to the independent knowing stage.[23] This is a difficult and time-consuming process, and one many students view as unnecessary and painful.

Students, essentially want class material to "fit for their particular purpose"—meaning their immediate and current level of understanding, which is to be able to use the information to pass all of the exams with very good grades, ideally with minimal effort. If the information is not going to be on the test, it is not relevant in their minds, and they do not want to learn it. Part of the reason for this expectation is the students' level of intellectual development.

Based on this research, it is fairly clear why students' expectations in the classroom are very different from faculty expectations. Most students' intellectual development has not progressed to the point where they care about the future use of educational material. In addition they are focused on grades, not learning. With a grade orientation, rather than a learning orientation, "students . . . expect[] . . . knowledge [to be] 'neatly packaged' and arranged for ease of access."[24] Therefore, when most students complete the student evaluations, they evaluate the "teaching effectiveness" of the professor based on their belief of what knowledge

21. This is not unlike legal reasoning at its best.

22. This Uniform Commercial Code term was chosen by the authors, however, *see,* Lynn Clouder, *Getting the 'Right Answers': Student Evaluation as a Reflection of Intellectual Development?*, 3 TEACHING IN HIGHER EDUC. 185 (1998), where the author also uses a Uniform Commercial Code analogy.

23. *See* DONALD, *supra* note 19, at 2–6 (2002), *available at* http://sll.stanford.edu/projects/tomprof/newtomprof/postings/405.html.

24. Clouder, *supra* note 22, at 190.

means and whether there was enough information provided to easily pass the exams with a minimal amount of effort.

Administrators also want the classroom material to be "fit for a particular purpose,"[25] i.e., quelling students' fears by allowing them to pass coursework more easily, thereby increasing the likelihood they remain happy and do not drop out of the college or university, taking their money with them. In addition, administrators want the evaluation process to be cheap, efficient, and require little time; therefore easily quantifiable student evaluations are often the most effective way to accomplish those multiple purposes. Note that accurately assessing learning is not generally one of those purposes. In fact, even as external demand for comprehensive educational assessment builds, at the college and university level, current measures of college and university quality and student learning are typically inexpensive, readily available measures that do not actually tell the institutions much.[26]

Essentially students and administrators use a "McDonald's Happy Meal" educational philosophy, i.e., making material readily available with minimal input or thought by and/or for the "consumer" students. If these "consumers" are kept happy, then they will return with their money and buy more prepackaged, easy-to-digest "educational meals." Whether or not such a *McDonaldized*[27] education nourishes students intellectually and professionally or otherwise provides sustenance for all involved is irrelevant to most students and administrators. It is, instead, easily digested and satisfies the immediate needs of both, though its long term value to these constituents or to society as a whole is highly suspect; research has shown that education is significantly more complex than making people happy. Education requires much more effort than many students are willing to invest in the process. Under *McDonaldized* educational circumstances, student evaluations do not address whether or not real teaching and actual learning outcomes take place, even though that is the stated purpose of the student evaluations. The evaluations simply measure the "happiness index"—positive affect—of both students and administrators relative to the ease of the educational experience and the ease of calculating the results based on the time and money each is willing to commit to the process.

Because, as seen above, the students' expectations of education is typically very different from the faculty members' definitions, those differing expectations will have an impact on evaluating the professors' ability to teach. "For example, students who adopt a surface approach [absolute learning] to learning and focus more on rote recall will typically prefer teachers who provide information and design assessment around a specifically defined set of criteria."[28] The students'

25. *Id*. (Clouder used the UCC analogy in her article, and the use of this phrase by the authors continues the Uniform Commercial Code analogy.).

26. Hersh, *supra* note 14, at 140.

27. In 1996, George Ritzer developed the term *McDonaldization*. *See* George Ritzer, THE MCDONALDIZATION THESIS: EXPLORATIONS AND EXTENSIONS (1998).

28. William W. Timpson & Desley Andrew, *Rethinking Student Evaluations and the Improvement of Teaching: Instruments for Change at the University of Queensland*, 22 STUDIES IN HIGHER EDUC. 55, 58 (1997).

expectations are that they will be passive recipients; if the professor expects them to be active participants they will not consider this "proper teaching." On the other hand, "[s]tudents who adopt a deep approach and focus more on understanding . . . will generally prefer teaching which is intellectually challenging,"[29] and, as noted above, leads to "contextual learning" which will be important in their future. Teaching evaluations may thus reflect the degree of mismatch between student-faculty expectations and behaviors, rather than accurately evaluating teaching effectiveness.

Though relationships between teaching and learning do exist, those relationships, as well as the relationships among the expectations of those involved, are highly complex and both improperly and inadequately measured by the summative evaluations. Critically important differences occur between and among students, faculty, and the teaching environment itself, yet neither evaluation documents themselves, nor the use of the raw data, account even minimally for any of these complex differences.

B. Additional Factors That Need to Be Considered

In addition to the lack of a common understanding of what teaching effectiveness actually means, there are other factors, discussed briefly below, which bias student evaluation results.

"Student ratings of teaching effectiveness [are not determined solely by the quality of the teacher, but rather are] driven more strongly by a student characteristic than they [are] by a teaching condition or a teacher characteristic."[30] Researchers "using data from several sources . . . support the view that teaching is multidimensional. Specifically, they identified nine dimensions of teaching: learning/value, instructor enthusiasm, group interaction, individual rapport, organization/clarity, breadth of coverage, examinations/grading, assignments/readings, and workload/difficulty."[31] It should be recognized that "[t]he implications [of such differing approaches, expectations and characteristics] for students' evaluations of teaching are substantial."[32]

The Center for Faculty Evaluation & Development at Kansas State University and others have found that the research suggests that there are factors that may bias student rating data, such as:

(1) Courses in the students' major fields versus elective courses,[33]

29. *Id.*

30. Kelly W. Crader & John K. Butler, Jr., *Validity Of Students' Teaching Evaluation Scores: The Wimberly-Faulkner-Moxley Questionnaire*, 56 EDUC. & PSYCHOL. MEASUREMENT 304, 304 (1996).

31. Eugene P. Sheehan & Tara DuPrey, *Student Evaluations Of University Teaching*, 26 JOURNAL OF INSTRUCTIONAL PSYCHOL. 188, 189 (1999).

32. Timpson & Andrew, *supra* note 28, at 58.

33. "Students tend to rate courses in their major fields and elective courses higher than required courses outside their majors." Barbara Gross-Davis, *Tools for Teaching—Student Rating Forms* (1993), http://teaching.berkeley.edu/bgd/ratingforms.html. *See also*, William E. Cashin, *IDEA Paper No 32: Student Ratings of Teaching: The Research Revisited* (1995), http://www.idea.ksu.edu/papers/Idea_Paper_32.pdf; John C. Ory and Katherine Ryan, *How Do*

(2) Level of the course,[34]
(3) Academic field[35] and/or specific discipline,[36]
(4) Faculty rank,[37]
(5) Gender of an instructor,[38]
(6) Workload/difficulty,[39]
(7) Class size,[40]
(8) Lecture versus discussion type of class format,[41]
(9) Expressiveness,[42]
(10) Student expectations,[43]
(11) Student motivation,[44]
(12) Expected grades,[45]

Student Ratings Measure up to a New Validity Framework?, 109 NEW DIRECTIONS FOR INSTITUTIONAL RES. 27 (2001).

34. "[H]igher level courses, especially graduate courses, tend to receive higher ratings." Cashin, *supra* note 33.

35. "[S]ome studies [suggest] that humanities and arts type courses receive higher ratings than social science type courses, which in turn receive higher ratings than math-science type courses." *Id.*

36. Davis, *supra* note 33.

37. "[R]egular faculty tend to receive higher ratings than graduate teaching assistants." Cashin, *supra* note 33 (citation omitted).

38. Davis, *supra* note 33; Kathleen S. Bean, *The Gender Gap in the Law School Classroom—Beyond Survival*, 14 VT. L. REV. 23, 25, 29 (1989); Joan M. Krauskopf, *Touching the Elephant: Perceptions of Gender in Nine Law Schools*, 44 J. LEGAL EDUC. 311, 326–327 (1994); Kristi Andersen & Elizabeth D. Miller, *Gender and Student Evaluations of Teaching*, 30 POLITICAL SCI. & POL. 216, 217 (1997); Kathryn M. Stanchi & Jan M. Levine, *Gender and Legal Writing: Law Schools Dirty Little Secret*, 16 BERKELEY WOMEN'S L.J. 3, 4 (2001).

39. Richard John Stapleton & Gene Murkison, *Optimizing the Fairness of Student Evlauations: A Study of Correlations Between Instructor Excellence, Study Production, Learning Production, and Expected Grades*, 25 J. MGMT. EDUC. 269, 280–81 (2001) (reporting that teachers who assigned more work received lower student ratings). *But see*, Cashin, *supra* note 33 (stating that students give higher ratings in difficult courses where they have to work hard).

40. Davis, *supra* note 33.

41. *Id*.

42. Cashin, *supra* note 33 (stating that student ratings may be more influenced by an instructor's style of presentation than by the substance of the content). *See also*, W. Neil Widmeyer & John W. Loy, *When You're Hot, You're Hot! Warm-Cold Effects in First Impressions of Persons and Teaching Effectiveness*, 80 J. EDUC. PSYCHOL. 118, 119 (1988). For a comprehensive review of instructor personality issues, *see* John C. Damron, *Instructor Personality and the Politics of the Classroom*, ftp://ftp.csd.uwm.edu/pub/Psychology/Behavior Analysis/educational/politics-of-instructor-evaluation-damron.

43. "[S]tudents who expect a course or teacher to be good generally find their expectations confirmed." Davis, *supra* note 33.

44. "[I]nstructors are more likely to obtain higher ratings in classes where students had a prior interest in the subject matter, or were taking the course as an elective." Cashin, *supra* note 33.

45. *Id.* (reporting a positive, but low correlation between students' ratings and expected grades); *see also*, David S. Holmes, *Effects of Grades and Disconfirmed Grade Expectancies on Students' Evaluations of Their Instructor*, 63 J. EDUC. PSYCHOL. 130 (1972); Richard Gigliotti & Foster Buchtel, *Attributional Bias and Course Evaluations*, 82 J. EDUC. PSYCHOL. 341 (1990).

(13) Non-anonymous ratings,[46]

(14) Instructor present while students complete ratings,[47] and

(15) Purpose of the ratings.[48]

Furthermore, while the research shows many examples of biases, it is also clear that administrators are well aware of potential biases.[49] There are examples of bias (due to a faculty member's personal characteristics,[50] personal opinions of students as to what should occur,[51] unexplainable reasons,[52] or affect of the professor,[53])

46. "[S]igned ratings tend to be higher." Cashin, *supra* note 33.

47. *Id*. These tend to be higher.

48. *Id*. (citations omitted).

49. During a meeting between a faculty member and the dean to review student evaluations the following occurred:

> Dean: Well, how did everything go last semester?
>
> Faculty: Not badly, but there was one unpleasant incident.
>
> Dean: Oh?
>
> Faculty: I had three sections of course X and there was a conspicuous instance of cheating in one. It involved five or six students, and I gave those people zeros on that exam.
>
> Dean (perusing the evaluation summaries): That was Section 3, I see.

Larry E. Stanfel, *An Experiment with Student Evaluations of Teaching*, 18 J. OF INSTRUCTIONAL PSYCHOL. 23, 24 (1991). This example demonstrates that administrators really know that making students unhappy creates bias and, therefore, lowers student evaluation scores.

50. "Two students disclosed to me in passing . . . that when a self-identified gay, black instructor of a course on racism left the room for students to evaluate him, two white male students joked about how they were going to 'slam the faggot.'" Heidi J. Nast, *'Sex', 'Race' And Multiculturalism: Critical Consumption And The Politics Of Course Evaluations*, 23 J. OF GEOGRAPHY IN HIGHER EDUC. 102, 106 (1999). "This study supports other research that suggests personality traits are robust predictors of students' evaluations of teaching effectiveness. However, it is difficult to determine a cause and effect relationship between instructor personality and student evaluations of faculty." Sally A. Radmacher & David J. Martin, *Identifying Significant Predictors of Student Evaluations of Faculty Through Hierarchical Regression Analysis*, 135 J. OF PSYCHOL. 259, 267 (2001).

> But a couple of factors are making it harder for professors to 'do the right thing.' First, the number of students who resent tough course loads and high grading standards seems to be growing as high schools continue to pump them out under-prepared and disengaged. And professors are encountering more and more of these students who resent, and in some cases actively resist, efforts to educate them. Some instructors, after enduring days, months, and years of scowls and pleas, eventually capitulate and make students happy 'consumers' by dumbing down their courses. . . . This increasingly means pleasing those students who don't like to read, write, think, or work hard. Even when in the minority, these disengaged students are feared, because they can drastically lower a professor's numbers. Conversely, professors have little to fear from engaged students, who tend to grade them generously because they're happy to have more study time for really challenging courses.

Paul Trout, *Evaluating the Evaluators*, Christian Science Monitor, Dec. 8, 1998 at 15; *See also*, Obenchain, Abernathy and Wiest, s*upra* note 18.

51.

> "It is unfair to drop someones (sic) grade because he/she missed too many days." "We were bombarded with information about authors that was boring with fact." "He had a tendency to be critical on objective manners (sic) such as word choice." "It is really hard to come to class when every day the material is being shoved down your throat."

> "The instructor needs to lower her standards." "I also think 2 novels to read outside of class is (sic) a bit too much. It's hard enough to get through." "She should have more concern for her students, their stress levels, and their GPAs!"

Trout, *supra* note 50, at 15.

> Taken as a whole, opinions [on student evaluations] were often contradictory, as is often the case. For some, there was too much work; for others, too little. The course was at once too demanding and not challenging enough. I was too tough or too easy, too patient or too impatient. Some praised, others criticized the textbook. . . . The outcome is not simply the result of what the professor plans, but what everyone brings to the class.

Douglas Hilt, *What Students Can Teach Professors: Reading Between The Lines of Evaluations*, CHRON. OF HIGHER EDUC., March 16, 2001, at B5.

> Utilising (sic) student evaluations to make course and/or tenure and promotion decisions is institutionally problematic for at least two reasons. First, it assumes that students have not judged what they have consumed based on whether or not they 'liked' the topic covered in the course. Liking may have to do with students' personal predilections or with the degree of emotional comfort they feel in the classroom. Non-majors commonly give lower evaluations to courses they are compelled to take, either as part of general liberal studies series, or as one of the only available elective slots that fits their schedule; for whatever reason, the course charts student anxieties and dislikes about taking something outside the desired or disciplinary field—anxieties over which an instructor has little control but which nevertheless register in teaching evaluations. Similarly, if more systemically, problematic are cases where faculty curricularly address issues of homophobia, racism, classism, misogyny or heterosexism—any or all of which may cause student discomfort. Like the evaluative impulses of non-majors in introductory classes, discomfort may result in negative evaluations, the directness with which difficult issues are broached producing different degrees of resistance.

Nast, *supra* note 50, at 104 (citations omitted).

52. Professor Stanfel conducted research that involved specific questions as to whether or not students had a clear understanding of the grading process. He passed out a memo that bore this request:

> Please sign this and return [it] to me if you have read my note on grade computation in QBA 4020 for fall [of] '87 and if it is perfectly clear to you. [O]therwise, make an appointment with me so that I can explain it to you again. Thanks. signed." Each student did sign and return that sheet. Thus did the author, confident of straight 1s for item 2, distribute his evaluation forms. The average response of that class to item 2 was 2.2. The departmental average response for item 2 was 1.628 and for full-time faculty there, 1.59.

Stanfel, *supra* note 49, at 36–37. This result demonstrates that after making a very serious effort to inform students about the grade computation process, including a signature that each student understood the process, the professor performed worse than those who made no such effort. Professors Nerger and Viney detailed Viney's experience.

> The second author (Viney) has always included an item on his questionnaire that asks student to rate his availability. As an administrator he continued to teach his course and he made it clear that students could come to his office any time between 8 a.m. and 5 p.m. five days per week. If he could not see the student immediately, a mutually convenient appointment would be worked out immediately by the secretary. Objectively, the instructor was available to students for a very large portion of each day. The students apparently did not see it that way because the mean score on the availability item on the student questionnaire was 3.12. Upon relinquishing administrative duties and assuming full-time professorial duties, the author kept three regular office hours per week. Objectively, the professor was available far less than he had been as an administrator, yet student ratings improved dramatically to a mean of

that measure "popularity of the instructor" rather than "teaching effectiveness,"[54] or a number of other reasons.[55]

Most professors have also sensed the effects of the size of the classroom and how close they can get to students (proxemics) on their effectiveness. The larger lecture halls increase the physical space between the professor and the students, which results in greater psychological space and creates various communication problems. Since some professors teach in rooms that are small or have round

> 3.64. . . . Thus availability ratings appear to have been affected by the students' perceptions of the professor's approachability [instead of the actual availability]. Student evaluations, in such a case, said nothing about actual availability and were, in that sense, not valid.

Janice L. Nerger and Wayne Viney, *Student Ratings of Teaching Effectiveness: Use and Misuse,* 38 MIDWEST Q. 218, 229–30 (Winter 1997). In a similar research project, the researcher provided the students with the information about course evaluation method and then tested them on their knowledge. On the test, they all demonstrated that they fully understood the method for course evaluations. However, even though

> all students had proven themselves aware of how they would be evaluated, but again, at course evaluation time, only one decided this continued to be true. Over twenty-eight percent were uncertain about what previously they had shown to be true, and upwards of forty-six percent then disagreed with the evidence they themselves provided earlier.

Larry E. Stanfel, *Measuring The Accuracy of Student Evaluations of Teaching*, 22 J. OF INSTRUCTIONAL PSYCHOL. 117, 120 (1995). Another portion of this same research project involved the question on the evaluation document regarding prompt return of all graded material.

> Insofar as reasonably prompt return of graded documents was concerned [when all were returned the next class period], only five persons could strongly agree that the earliest possible moment qualifies as reasonably prompt, and over one quarter of the group was either uncertain or in disagreement that this proven policy could be so regarded.

Id. at 120. Even though all of the graded documents had been returned with the utmost promptness, very few of the students' answers reflected the actual facts of the situation.

53. In a research project the researchers had the same professor teach the same class using the same syllabus, same exams and same lectures, etc. The only change between the two semesters was that the professor was trained in how to deliver the lectures more "enthusiastically." The results on the student evaluations for the two semesters were significant. The professor was rated on such factors as Knowledgeable, Tolerant, Enthusiastic, Accessible, and Organized. The mean scores increased from .69 to .95 after the "enthusiastic" training. Everything, except the style of delivery remained the same, so there was no logical reason for those scores to go up that much. Wendy M. Williams & Stephen J. Ceci, *How'm I Doing?, Problems With Student Rating of Instructors and Courses*, 29 CHANGE 13 (1997). These authors went on to say, "our modest study nevertheless shows that student ratings are far from the bias-free indicators of instructor effectiveness that many have touted them to be. Moreover, student ratings can make or break the careers of instructors on grounds unrelated to objective measures of student learning . . . ." *Id*. at 21.

54. *See generally*, Obenchain, *supra* note 9.

55.

> [R]atings are higher if the instructor is present while the forms are being filled out; non-anonymous ratings are higher; and ratings are higher in classes that meet with more intensive time schedules. Ratings are also higher on items custom-designed by the instructor, as compared to items on standardized forms. Again, if instructors are to be compared with each other, these factors must somehow be taken into account.

Nerger and Viney, *supra* note 7, at 220–21 (citations omitted).

tables in them and others teach in the large informal lecture halls, the effects of proxemics are not equally distributed across all professors.[56] Rarely are these biasing factors considered in the use of student evaluations for making the life-changing decisions in regard to faculty members.

Since all these variables could impact ratings, control of all substantially meaningful variables among the multidimensional aspects of teaching must be considered in the student evaluations, rather than being based on mere student comfort and grade satisfaction as typically occurs. Yet, administrator analysis and interpretation of student evaluations makes no meaningful attempt to do that. In practice, little or no attempt is made to use instruments which meaningfully evidence and incorporate the multitude of substantial, yet variable factors, potentially impacting the scores students choose to evaluate professors and/or other college and university teachers, because every teaching situation and every course is treated identically.

> Although the standardized questions alone could yield basic information as to whether students liked a particular instructor, his exams, or his grading, they could provide no meaningful information as to why this was the case and, therefore, easily confounded similar results arising from vastly different circumstances . . . . They, therefore, provided the department with inadequate information for any kind of meaningful evaluation and the teachers with inadequate information to help them improve their performances . . . . In short, . . . the computerized answers literally produced academic junk.[57]

II. THE LEGAL ISSUES

Due to the above issues and unknowns, there are several legal issues regarding student evaluations and their use, which arise when student evaluations are used at either public or some private institutions[58] to make employment, retention, and tenure decisions for tenure-track or probationary faculty and to make employment and retention decisions for other constitutionally protected teaching professionals.[59] These legal issues consist of the constitutional issues of

56. "Thus, any fair comparison of one instructor with another must factor out the effects of this potentially important variable." *Id*. at 219.

57. Robert Justin Godstein, *Some Thoughts About Standardized Teaching Evaluations*, 22 PERSP. ON POL. SCI. 8, 10 (1993).

58. Constitutional rights attach when the private institutions receive enough government funds and/or other governmental aid to allow a court to say that there is enough "state action" to require them to comply with the Constitution. However, *see* Steven K. Berenson, *What Should Law School Student Conduct Codes Do?* 38 AKRON L. REV. 803, 837 (2005) "[A] number of theories have been applied to impose upon private schools similar procedural due process requirements to those that apply to public schools. First, it has been argued that because many private universities receive federal financial assistance, are heavily regulated, and engage in a variety of projects with government entities, such universities are 'state actors' for purposes of due process analysis. However, such arguments have been rejected."

59. For simplicity the term "protected faculty" shall be used in the rest of this article to mean tenured, tenure track, and other faculty who have a protected interest because of their

fundamental rights, substantive due process, and a related issue when students are being treated as *de facto* "experts" in pedagogy.

A. Fundamental Right

"[P]eople who seek to challenge governmental action under the due process clause must first demonstrate to the court they have a constitutionally protected liberty or property interest. If they do, and only if they do, does the court then take the next step and determine what process is due them."[60] Therefore, not all college and university faculty members may be constitutionally protected, but for some faculty members this protected liberty or property interest does exist. In January of 1972, the U.S. Supreme Court heard two cases involving college and university faculty members' or teaching professionals' rights in regard to continued employment.[61] In *Roth*, the plaintiff had a contract for a fixed term of one academic year, which was not renewed. He was simply informed he would not be hired for the following academic year. Roth challenged the non-renewal as a denial of his constitutional right to due process. The Court reasoned that because Wisconsin law and regulations do not grant Roth a legal right to an "expectation" of renewal, no due process rights attached to his claim. However, the Court recognized "'Liberty' and 'property' are broad and majestic terms,"[62] and as such, by definition would include more than the merely common understanding of property. The *Roth* Court recognized there could be a "property" right in the faculty position when there is some expectation created by some "understanding or tacit agreement"[63] the job will continue.

In *Sindermann*, the U.S. Supreme Court affirmed the Court of Appeals, which held "that, despite the respondent's lack of tenure, the failure to allow him an opportunity for a hearing would violate the constitutional guarantee of procedural due process if the respondent could show he had an 'expectancy' of re-employment."[64] The Court agreed in this factual situation there was an evidentiary issue as to whether or not he had a legitimate "expectancy" of continued employment. The college had certain rules and practices that could be construed as giving one the expectancy of continued employment. Thus, if Sindermann could prove he had such expectancy, then at least procedural due process would attach to that right.[65] The Court went on to say:

> We have made clear in *Roth* that "property" interests subject to procedural due process protection are not limited by a few rigid, technical forms. Rather, "property" denotes a broad range of interests

expectation of continued employment for other reasons.

60. William P. Quigley, *Due Process Rights of Grade School Students Subjected to High-Stakes Testing*, 10 B.U. PUB. INT. L. J. 284, 290 (2001).

61. *See* Bd. of Regents of State Colleges v. Roth, 408 U.S. 564 (1972); Perry v. Sindermann, 408 U.S. 593 (1972).

62. *Roth*, 408 U.S. at 571.

63. *See id.*

64. *Sindermann*, 408 U.S. at 596.

65. *See id.*

> that are secured by "existing rules or understandings." A person's interest in a benefit is a "property" interest for due process purposes if there are such rules or mutually explicit understandings that support his claim of entitlement to the benefit and that he may invoke at a hearing.[66]

The court in *Regents of the University of Michigan v. Ewing* reiterated this by saying, "We recognize, of course, that 'mutually explicit understandings' may operate to create property interests."[67]

At most colleges and universities there is likely a combination of tenured faculty, probationary faculty, and academic teaching staff. Generally, all of these faculty members participate in an annual review to determine whether or not each will receive the next year's contract, and for those who are tenure-track, whether they are progressing according to the tenure guidelines, so that they may receive tenure at the end of the six-year period. The process that is used for this review and the guidelines will be found either in the system-wide rules at the state level or at the local level. Generally the state level provides the broad outlines for annual reviews of faculty members, while the local (college/university and/or department) rules fill in the details for both retention and, ultimately, for tenure. If the review rules that apply to the various categories of faculty include either an explicit or implicit expectation of employment, then those faculty members are protected by the Due Process Clause of the Fourteenth Amendment, and shall be considered protected faculty. When, for some categories of faculty, there is no expectation of continued employment, either explicit or implicit, those are unprotected faculty and shall not be part of this discussion.

At all colleges and universities the review rules include reviewing an individual's teaching, research, and service. The value or "weight" given to each of these three categories will vary at each college or university, but all review these three categories to some degree. The faculty members' property interests and the dimensions that are created depend upon the existing rules and/or understandings that come from an independent source such as state law, college and university rules, or understandings that secure these benefits and then support a claim of entitlement to those benefits.[68] The standards for each of these categories will contain specific criteria for the renewal of the contract. To meet each of the criteria requires highly detailed information. Everyone involved in the process clearly understands if the faculty member provides the requisite information that meets the specified criteria, his/her contract must be renewed. No similar expectation and process routinely applies to teaching professionals whose status may not otherwise be constitutionally protected.

As seen in the U. S. Supreme Court and other courts, decisions have stated that an expectation of continued employment by a faculty member, created by the

66. *Id.* at 601 (citing *Roth*, 408 U.S. at 571–72, 577).

67. Regents of the Univ. of Mich. v. Ewing, 474 U.S. 214, 222 n.9 (1985) (citing *Sindermann*, 408 U.S. at 601).

68. Bd. of Regents of State Colleges v. Roth, 408 U.S. 564, 571 (1972).

applicable rules, is a fundamental right under the U.S. Constitution,[69] and due process attaches. College and university rules provide some guidelines for the process that must be followed, but,

> [o]nce a claimant establishes a right protected by due process, a court must decide what process is "due." The existence of mandatory procedures may help establish a due process entitlement, but the Constitution neither gives an individual the right to have those procedures followed nor does it restrict an individual's rights only to those procedures. The constitutional requirements of due process are independent.[70]

B. Substantive Due Process

The Fourteenth Amendment contains three things in addition to the Equal Protection Clause and procedural due process, "it contains a substantive component, sometimes referred to as 'substantive due process,' which bars certain arbitrary government actions regardless of the fairness of the procedures used to implement them."[71] This means that due process consists of both procedural and substantive due process.

> On the substantive side, the law holds that some rights are so profoundly inherent in the American system of justice that they cannot be limited or deprived arbitrarily, even if the procedures afforded the individual are fair. Substantive due process challenges strike at the fairness of the state action itself, not the method by which it is achieved.[72]
>
> The substantive due process doctrine turns due process from a mechanism ensuring procedural fairness when the government attempts to deny life, liberty, or property, into a fourth protected entity that determines whether or not fundamental rights exist that are not enumerated within the Constitution. Under the doctrine, due process has some "substantive" quality that forms and then falls under the liberty provision.[73]
>
> [The] Due Process Clause protects "the substantive aspects of liberty against impermissible government restrictions." Courts have determined that the Due Process Clause requires that the government avoid taking action that is arbitrary, capricious, does not achieve a legitimate state

69. *See Roth*, 408 U.S. 564, *and* Perry v. Sindermann, 408 U.S. 593 (1972).

70. Tim Searchinger, *The Procedural Due Process Approach to Administrative Discretion: The Courts' Inverted Analysis*, 95 YALE L.J. 1017, 1023 (1986).

71. Daniels v. Williams, 474 U.S. 327, 337 (1986); *see also*, Debra P. v. Turlington, 644 F.2d 397 (5th Cir. 1981); Crump v. Gilmer Indep. Sch. Dist., 797 F. Supp. 552, 555 (E.D. Tex. 1992).

72. Quigley, *supra* note 60, at 305.

73. Christopher J. Schmidt, *Revitalizing the Quiet Ninth Amendment: Determining Unenumerated Rights and Eliminating Substantive Due Process*, 32 U. BALT. L. REV. 169, 169 (2003).

> interest, or is fundamentally unfair. A substantive due process violation is deemed to occur where such state action "encroaches upon concepts of justice lying at the basis of our civil and political institutions."[74]

The "fundamental right" of some faculty members has already been established, therefore both procedural and substantive due process attaches to that right when the state is trying to take it away.

Lochner v. New York[75] "effectively immortalized the substantive due process mechanism that is still the standard for analyzing claims regarding unenumerated constitutional rights today."[76]

> The "fundamental liberty interest" or "unenumerated right" branch of substantive due process . . . has gained a remarkable degree of at least formal acceptance by the current Supreme Court. The doctrine was put on the most solid doctrinal footing in its history by its explication in the Court's 1992 decision in *Planned Parenthood v. Casey*.[77]

The Clause "provides heightened protection against government interference with certain fundamental rights and liberty interests."[78] Therefore, the substantive due process clause has

> become [not only the] bulwark . . . against arbitrary legislation; but, [also against other arbitrary action] . . . as it would be incongruous to measure and restrict [it to only process] . . . [as it] . . . must be held to guarantee not [only] particular forms of procedure, but the very substance of individual rights to life, liberty, and property.[79]

Since actions by the government "can be arbitrary in more than one sense[,] . . . the Due Process Clause has been construed to provide protection against more than one type of arbitrary government action."[80]

> The categories of substance and procedure are distinct. Were the rule otherwise, the Clause would be reduced to a mere tautology. "Property" cannot be defined by the procedures provided for its deprivation any more than can life or liberty. The right to due process "is conferred, not by legislative grace, but by constitutional guarantee. While the legislature may elect not to confer a property interest in

74. Paul T. O'Neill, *High Stakes Testing Law and Litigation*, 2003 BYU EDUC. & L.J. 623, 641 (2003) (citing 16B Am. Jur. 2d Constitutional Law 901 (2002); Debra P. v. Turlington, 644 F. 2d 397, 404 (5th Cir. 1981)) (footnotes omitted).

75. 198 U.S. 45 (1905).

76. Schmidt, *supra* note 73 at 172 (footnote omitted).

77. Peter J. Rubin, *Square Pegs and Round Holes: Substantive Due Process, Procedural Due Process, and the Bill of Rights*, 103 COLUM. L. REV. 833, 836 (2003). *See also*, Planned Parenthood v. Casey, 505 U.S. 833 (1992).

78. Rubin, *supra* note 77 at 836 (quoting Washington v. Glucksberg, 521 U.S. 702, 719–20 (1997)).

79. Hurtado v. California, 110 U.S. 516, 532 (1884).

80. Rubin, s*upra* note 77 at 841; *see also*, Marc C. Niles, *Ninth Amendment Adjudication: An Alternative to Substantive Due Process Analysis of Personal Autonomy Rights*, 48 UCLA L. REV. 85, 144 (2000); Bruce N. Morton, John Locke, & Robert Bork, *Natural Rights and the Interpretation of the Constitution*, 22 SETON HALL L. REV. 709 (1992).

> [public] employment, it may not constitutionally authorize the deprivation of such an interest, once conferred, without appropriate procedural safeguards."[81]

A substantive due process right deals with the ability of a person to defend/explain, in substance, what is being done—essentially a "fairness" issue. For example, the court in *Debra P. v. Turlington*[82] held that for a test that was required to be taken prior to graduation to be valid, the state must be able to show that the test fairly assessed what was actually taught in the school, because the students had a protected property interest (a fundament right) in the expectation of receiving a diploma.[83] Another court citing *Debra P.* stated that "fundamental fairness requires that the state be put to test on the issue of whether the students were tested on material they were or were not taught."[84] The concept of "fundamental fairness" is part of the substantive due process doctrine. The analogy to the current use of student evaluations should be quite obvious—it is "fundamentally unfair" to use data, the numbers resulting from the student evaluations, that may have little or no relationship to what is actually being "tested"—the quality of teaching.

III. RELATED LEGAL ISSUES WITH STUDENT EVALUATIONS AS "EVIDENCE"

When a college or university uses student evaluations as the sole or primary criterion for personnel decisions and views these student evaluations in a context similar to that described herein, in effect, the students' evaluations become "testimony" or evidence relative to the pedagogical ability and the substantive knowledge of the professor. Students thus serve in the de facto role of expert witnesses in this process. Yet presently an institution does not have to show that there is any evidentiary validity to student evaluation "testimony," nor must it first qualify the students as "experts."

A. Validity

Validity addresses the issue of whether what is supposed to be measured is what is actually measured.[85] Student evaluations are intended to measure students' objective perception of the teaching process (pedagogy) and teaching effectiveness[86] (substantive knowledge) of the individual professor being assessed.

81. Cleveland Bd. of Educ. v. Loudermill, 470 U.S. 532, 541 (1985) (footnotes omitted).

82. 644 F.2d 397 (5th Cir. 1981).

83. *Id.* at 404–05.

84. Crump v. Gilmer Indep. Sch. Dist., 797 F. Supp. 552, 555 (E.D. Tex. 1992).

85. Judith D. Fischer, *The Use and Effects of Student Ratings in Legal Writing Courses: A Plea for Holistic Evaluation of Teaching*, 10 LEGAL WRITING 111, 117 (2004) ("Yet there is a continuing lack of consensus among scholars about a number of points, including the important issue of the ratings' validity, that is, whether they actually do measure teaching quality."). *See also*, Philip Abrami, et al., *Validity of Student Ratings of Instruction: What We Know and What We Do Not*, 82 J. EDUC. PSYCHOL. 219 (1990).

86. *See* James A. Kulik, *Student Ratings: Validity, Utility, and Controversy*, 109 NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH 9, 10 (2001) ("To say that student ratings are valid

To be valid, teaching effectiveness ought to be correlated to effective course design and effective delivery, ultimately resulting in increased learning of the subject area. "If ratings are valid, students will give good ratings to effective teachers and poor ratings to ineffective ones."[87] Therefore, if the questions on the student evaluations are intended to measure effective teaching—their claimed purpose—which results in increased learning, then the questions must be valid in regard to measuring the correlations between teaching and learning, with irrelevant factors eliminated and relevant factors controlled. Since, as we saw above, there are innumerable other factors that come into play when students answer the questions on the instrument, there is no certainty that the instrument actually measures what it is purported to measure, and its use for that purpose is therefore invalid,[88] or in legal terms, arbitrary, capricious, or "fundamentally unfair." Because of this invalidity most student evaluations probably measure affect, instead of effect.

There are four different types of validity—conclusion validity, internal validity, construct validity, and external validity.[89] All of these are problematic in the usage of the typical teaching evaluation.

The conclusion validity is the relationship between the two variables—the questions and what the questions are intended to measure (effective teaching). This means that the questions should be framed so that one can determine whether learning resulted from the classroom experience. However, most questions on the student evaluations actually address how much the students liked the process of learning,[90] instead of how much was actually learned. When invalid evaluations are used to show that a professor is not a good teacher, there is no conclusion validity.

Internal validity addresses whether the relationship is a causal one. Just because a professor does all of the things that a good professor does, are those necessarily the cause of student learning? For example, speaking clearly, knowing the material well, and other such questions are presumed to contribute to learning, but speaking clearly and knowing the material well may have no effect if the student has not prepared, does not study, or does not come to class very often. The underlying presumption as to cause and effect could be invalid. There may have been any number of other causes for either the increased learning or poor learning. Each student brings his or her unique background and work habits to the classroom. Could those factors have been the cause of the increased learning,

is to say that they reflect teaching effectiveness."). *See also,* Fischer, *supra* note 85, at 116 ("Student achievement is often proposed as the appropriate indicator of effective teaching, but there is no universally accepted means of measuring it.") (citations omitted).

87. Kulik, *supra* note 86, at 10.

88. *See* Herbert W. Marsh, *Student Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Baiases, and Utility*, 76 J. EDUC. PSYCHOL. 707, 749 (1984).

89. William Trochim, Introduction to Validity, http://trochim.human.cornell.edu/kb/introval.htm (last visited Oct. 16, 2005).

90. Fischer, *supra* note 85, at 118–19 ("Other scholars have stated that student ratings measure not teaching effectiveness but student *perceptions* of teaching effectiveness or feelings that are not directly related to good teaching and learning.") (citations omitted).

instead of what the professor did or did not do? The student evaluations do not even consider most of those factors. Therefore, there is little, if any, internal validity.

Construct validity addresses whether or not student evaluations ask the critical questions that would actually measure the outcome we wanted to assess—increased learning.

> [The] [s]ubstantive aspects [of construct validity] involve evidence supporting the theoretical and empirical analysis of the processes, strategies, and knowledge proposed to account for respondents' item or task performance on the assessment (or both). Sources of evidence include analysis of individual responses or response processes through think-aloud protocols *or simply asking respondents about their responses*."[91]

Because those who respond to these items on student evaluations are kept anonymous, they can never be quizzed as to the reasons for their choices on the document. There is no way to establish construct validity.

External validity assesses if there is a causal relationship as to cause and effect that can be generalized to other teaching situations. If this professor taught another group of students, would that new group of students have a similar level of learning? Clearly, the evaluations don't do that either, since there is no evidence that they even measure learning. They probably do measure affect, and in that sense there may be external validity, but that is not what they are used for—they are used as a surrogate for effective teaching. There is no external validity.

Based on the four different types of validity, student evaluations meet none of these standards and are invalid instruments. Yet they are used to make life-changing decisions without any ability by the person against whom they are being used to show that they are invalid and unfair.[92] From the discussion it is clear that there is no validity to student evaluations, because for "any inference or conclusion, there are always possible threats to validity—reasons the conclusions or inference might be wrong. Ideally, one tries to reduce the plausibility of the most likely threats to validity, thereby leaving as most plausible the conclusions reached by the study."[93] In regard to the conclusions drawn from the questions on student evaluations—that these questions provide the conclusion that one who scores high is a good teacher—there are far too many threats to validity that have not been effectively controlled.

91. John C. Ory & Katherine Ryan, *How Do Student Ratings Measure Up to a New Validity Framework*, 109 NEW DIRECTIONS FOR INSTITUTIONAL RESEARCH 27, 29 (2001) (emphasis added).

92. Fischer, *supra* note 85, at 119 ("McKeachie declared that 'for personnel purposes, faculty and administrators rightfully have great concerns about the validity and reliability of evaluation data.' Others have bluntly called the ratings 'risky business,' 'pernicious,' or 'an unqualified failure' with a 'dysfunctional' impact.") (citations omitted).

93. William Trochim, Introduction to Validity, http://trochim.human.cornell.edu/kb/introval.htm (last visited Oct. 16, 2005).

B. Standards for Expert Testimony

Even though the student evaluations aren't being used in a trial, they are being used as evidence in critical life-changing decisions that could result in a right to a hearing and ultimately an appeal. Once that procedural due process right to a hearing is exercised by the faculty member, that faculty member must have the substantive due process right to get at the basis of the evidence used against him/her, because that is what fundamental fairness requires in this situation. Because the student evaluations are supposedly evidence of effective teaching, and that evidence is treated as if an expert provided it, by analogy the same or similar standards that apply to expert witnesses and expert testimony should apply to the right to use this "testimony" when a faculty member exercises his/her due process rights.

In a federal courtroom the Federal Rules of Evidence would apply, and Rule 702 states: "If scientific, technical, or other specialized knowledge will assist the finder of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert by knowledge, skill, experience, training, or education, may testify in the form of an opinion or otherwise."[94] The students' opinions expressed in the evaluations are used by the finder of fact, the faculty committee, and others who use this information to make their recommendations, to help them determine a fact issue—whether or not this faculty member is a good teacher. Rule 702 states that only a "qualified . . . expert . . . may testify . . . in the form of an opinion."[95] Since students can't qualify as experts,[96] their opinion as to the quality of teaching cannot be used because the witness providing his or her opinion must be a person who is an expert on pedagogy. Such a qualified expert would testify as to whether the methods used by the faculty member should be effective, based on the available scientific research in pedagogy. Expert testimony would be necessary to prove causation that poor pedagogy caused the result of not much learning, instead of any number of other variables that are not considered in student evaluations. The problem is not that "opinions" are used, rather that the "witness"—each student—is not an expert whose opinion can help clarify or help the trier of facts to understand a fact issue. Rule 702 clearly states that laypersons' opinions, such as these student opinions, could not be used for the ultimate issue of whether the

94. FED. R. EVID. 702.

95. *Id.*

96. Neither undergraduate students nor graduate students could qualify as experts in pedagogy, because even graduate students are not taught much about "how to teach." For a discussion of some of the issues with graduate students and their understanding of "how to teach," *see*, *e.g.*, Colleen Conway, Erin Hansen, Andrew Schulz, Jeff Stimson, & Jill Wozniak-Reese, *Becoming a Teacher: Stories of the First Few Years*, 91 MUSIC EDUCATORS J. 45 (2004); Rose Mary Carroll-Johnson, *Learning to Teach 32* ONCOLOGY NURSING F. 889 (2005); Carol Anderson Darling & Eileen M. Earhart, *A Model For Preparing Graduate Students As Educators*, 39 FAM. REL. 341 (1990); Stephen F. Davis & Jason P. Kring, *A Model for Training and Evaluating Graduate Teaching Assistants*, 35 C. Student J. 45 (2001); Elizabeth H. Morrison & Janet Palmer Hafler, *Yesterday a Learner, Today a Teacher Too: Residents as Teachers in 2000*, 105 PEDIATRICS 238 (2000); Wayne Wanta, Paul Parsons, Sharon Dunwoody, William C Christ, Richard L. Barton, & Beth Barnes, *Preparing Graduate Students to Teach: Obligation and Practice,* 58 JOURNALISM & MASS COMM. EDUCATOR 209 (2003).

faculty member is a good teacher, since they are not experts at determining what pedagogical elements are necessary to make one a good teacher. These students, as laypersons, are clearly able to "testify" as to their experiences in the classroom—the facts that they are privy to—but not the ultimate issue of whether that makes one a good teacher or not. Furthermore, graduate students, both at the masters and doctoral level, rarely receive any sort of training in teaching techniques or pedagogy, even if they are expected to teach in their respective disciplines once they receive their doctorate, and thus would not qualify as expert witnesses. Additionally, graduate students also experience a severe conflict of interest: even if they were qualified as expert witnesses, which few would be, they are also far less likely to be honest because critical rewards beyond merely receiving grades (e.g., theses, preliminary exams, progress-towards-degree assessments, dissertation defenses) are highly contingent upon rating their professors highly, a small number of faculty teach the same graduate students over and over, especially at the doctoral level, and class sizes are so small as to make anonymity unlikely. Yet, student evaluations are, in fact, used to help make the determination of whether someone is a good teacher, and in some colleges and universities, they are the most critical piece of "evidence" in that process.

C. Recent Supreme Court Cases on the Use of Expert Witnesses

The Supreme Court has created a "gate keeping"[97] function in regard to what an expert witness could testify about. Four foundation levels are relevant to the admissibility of testimony by an expert witness: competency, theory, technique, and application. The first level, competency, establishes the expertise of the witness and the "competency" of that person's testimony based on Rule 702. In the first step a judge's "gate keeping" function is to determine whether the witness is an expert—in this situation the student who is offering the opinion on the professor's ability to teach him or her. "In exercising the trial judge's gate keeping responsibility under Rule 702, the trial court has broad discretion in not only determining the general competency issue, but also whether a particular subject matter is beyond the scope of the expert's expertise."[98] However, even with that broad discretion, none of the students whose "expert" testimony is used are qualified by any knowledge, skill, experience, training or education as to the subject matter their testimony is used for—pedagogy. Such testimony is clearly beyond the scope of any "expertise" students may have. Therefore, students could not be used to testify as to the faculty members' professional teaching skills, since they lack any knowledge of pedagogy and have no expertise in the faculty members' subject matter knowledge.

Even if a judge, based on the above factors, could determine that students are experts, the next step is for the judge to function as the gatekeeper as to the

97. *See* Daubert v. Merrell Dow Pharm. Inc., 509 U.S. 579, 597 (1993).

98. Richard Collin Mangrum, *Interpreting Nebraska Rule of Evidence 702 After the Nebraska Supreme Court Adopted the Federal* Daubert *Standard for the Admissibility of Expert Testimony in* Schafersman v. Agland Corp.*,* 35 CREIGHTON L. REV. 31, 81 (2001) (citations omitted).

reliability[99] of expert testimony.[100] This process must follow Supreme Court standards for admitting scientific and non-scientific testimony.

> [The] United States Supreme Court embarked on a journey to create standards for admitting both scientific and nonscientific expert testimony. The evolution of this journey, as demonstrated by *Daubert v. Merrell Dow Pharmaceuticals Inc.*, *General Electric Co. v. Joiner*, and *Kumho Tire Co., Ltd. v. Carmichael*, illustrated the Court's recognition that all admissible expert testimony must achieve a certain level of reliability and relevance.[101]

In *Daubert*,[102]

> the Court identified four non-exclusive factors to aid in determining the admissibility of scientific evidence: (1) whether the theory or scientific technique has been tested; (2) whether it has been subjected to peer review or publication; (3) the known or potential rate of error; and (4) whether the principle was generally accepted in the relevant scientific community.[103]

These factors are non-exclusive, and others may be considered as to the reliability of the proffered testimony.[104] Regardless of which, or how many, factors, are used, the testimony by the expert cannot be "couched in terms of mere possibility, as compared with probability or certainty, [because that] provides an insufficient basis for admitting expert testimony."[105]

Until *Kumho Tire*,[106] these were some of the factors to be considered as to reliability of scientific evidence. *Kumho Tire* expanded these factors to the use of nonscientific evidence.[107] These two cases made it clear that the judge is the gatekeeper as to expert testimony in both scientific and nonscientific testimony as to the reliability of the testimony by considering several factors. "Therefore, the proper application of the 'gate keeping' function encompasses scientific, technical,

99. The courts use the term "reliability" as a synonym for validity.

100. *See*, *e.g.*, Major Victor Hansen, *Rule of Evidence 702: The Supreme Court Provides a Framework for Reliability Determinations*, 162 MIL. L. REV. 1 (1999); Mangrum, *supra* note 98; Leslie Morsek, *Get on Board for the Ride of Your Life! The Ups, the Downs, the Twists and the Turns of the Applicability of the "Gatekeeper" Function to Scientific and Non-Scientific Expert Evidence:* Kumho's *Expansion of* Daubert, 34 AKRON L. REV. 689 (2001).

101. Morsek, *supra* note 100, at 693 (citations omitted).

102. Daubert v. Merrell Dow Pharm., Inc. 509 U.S. 579 (1993).

103. Morsek, s*upra* note 100, at 707–09 (citations omitted).

104. *See*, *e.g.*, Mark McCormick, *Scientific Evidence: Defining a New Approach to Admissibility*, 67 IOWA L. REV. 879, 911–12 (1982) (identifying eleven factors that could be considered); *see also*, UNIFORM R. OF EVID., R. 702 (1974) (Tentative Draft No. 1, 1997), *available at:* http://www.law.upenn.edu/bll/ulc/ure/ev702.pdf. *See also* The National Conference of Commissioners on Uniform State Laws recommendations as to the factors and other elements of Rule 702 *available at* www.law.upenn.edu/bbl/ulc/ure/ev702htm.

105. Mangrum, *supra* note 100, at 39 (citations omitted).

106. Kumho Tire Co., Ltd. v. Carmichael, 526 U.S. 137 (1999).

107. *Id.* at 147.

and other specialized knowledge."[108] The determination as to whether a professor is good at helping students learn by using proper pedagogy is certainly "specialized knowledge," at a minimum.

These cases create a "gate keeping" function in regard to what an expert witness could testify about. Four foundation levels are relevant to the admissibility of testimony by an expert witness: competency, theory, technique, and application.[109] As already discussed, the first level, competency, establishes the expertise of the witness and the "competency" of that person's testimony based on Rule 702.[110]

The second level of inquiry in the "gate keeping" function is to inquire whether the theory is reliable. If the theory is new this may be shown by the expertise of the witness. The theory in student evaluations is that the student evaluations measure the ability to teach well, which results in increased learning by the students. That theory, as discussed above, has not been shown to be valid and reliable. Evidence of its reliability might include whether it has been subject to recent peer review and/or publication, whether it has an established rate of error, and whether the relevant professional community still generally accepts this theory. None of these can be shown to exist, since there is no valid evidence to show the student evaluations measure the teaching ability of faculty, nor that teaching ability actually results in higher levels of learning.

The third level of inquiry that the "gatekeeper" must determine is whether the technique or procedure was properly used. To show this the "expert" (each student) may testify that he/she is qualified to use the technique or procedure properly based on knowledge, skill, experience, training, or education. No student could testify that he/she is qualified to properly use the theory that teaching ability will increase the amount of learning. In addition, there must be evidence that the technique or procedure used is reliable because the technique or procedure has been reliably tested, has been subject to peer review and/or publication, has an established rate of error, and the technique is generally accepted in that profession, whether there are safeguards in the characteristics of the technique, whether there are existing standards governing its use, whether there is some continuing maintenance/update of the standards governing the theory, to what extent the basic data that is being used by the fact finder is verifiable, whether there are other experts available to test and evaluate the theory, and questions to establish the degree of care taken by the expert to prepare the information. Clearly, none of these standards can be met.

In addition to testifying that the formula—the use of the medians to establish effective teaching—was properly used, the expert must also explain the technique

108. Morsek, *supra* note 100, at 723 (citation omitted).

109. Mangrum, *supra* note 100, at 34. In addition, some of the questions come from factors for reliability that come from an article by Mark McCormick. *See* McCormick, *supra* note 104.

110. *See* FED. R. EVID. 702 ("If scientific, technical, or other specialized knowledge will assist the trier of fact to understand the evidence or to determine a fact in issue, a witness qualified as an expert may testify thereto in the form of an opinion or otherwise, if (1) the testimony is based upon sufficient facts or data, (2) the testimony is the product of reliable principles and methods, and (3) the witness has applied the principles and methods reliably to the facts of the case.").

or procedure itself to the fact finder and explain how the information developed by the use of the formula (the medians) relates to his or her testimony regarding the theory that teaching ability increases learning. This last step is the application function—the fourth level of inquiry by the gatekeeper. In regard to the specific application the expert must be qualified based on knowledge, skill, experience, training, or education to be able to apply the principle and to be able to interpret the results. The expert must testify as to the proper use of the statistical methods employed to arrive at the results that he or she is testifying to. The expert must also testify as to why he or she is capable of interpreting and/or explaining the application of the method to the case, and is able to explain the application of the result to the opinion that arises therefrom.[111] There are several problems with this requirement. First, the evaluations are anonymous, so no student could be called to "testify" to any of this. Second, even if they could be called, they have virtually no expertise to testify to any of this. They know nothing about whether the use of the medians to establish effective teaching is a proper application of this information—nor does anyone else for that matter. Even though the cases and the Federal Rules of Evidence have expanded the concept of who qualifies as an expert, this is not broad enough by any stretch of the imagination to include anonymous students as experts in pedagogy.[112]

D. Discussion

As the opening quote by Justice Marshall said,

> [B]efore the decision is made to terminate an employee's wages [and in our scenario, an employee's position and his/her future reputation], the employee is entitled to an opportunity to test the strength of the evidence "by confronting and cross-examining adverse witnesses and by presenting witnesses on his own behalf, whenever there are substantial disputes in testimonial evidence."[113]

The above discussion has shown that currently in the faculty review process, the situation is that in many faculty positions there is a "legislatively" created right to an expectation of continued employment, which based on U.S. Supreme Court decisions, becomes a fundamental right subject to due process.

Even though procedural due process exists, by using student evaluations as the primary source of information as to the faculty member's teaching ability, the right to substantive due process is taken away in an arbitrary manner by the use of the raw numbers that the student evaluations provide. If the faculty member is denied renewal of the contract and/or tenure, the faculty member is, ultimately, entitled to a hearing. At these retention hearings there is generally no shortage of procedural due process. Therefore, the focus is not on procedural due process; instead the focus is on substantive due process. As we have seen, substantive due process and

111. Based on the principles in the Mangrum law review article. *See* Mangrum, *supra* note 100, at 34–36.

112. *See generally* Morsek, *supra* note 100.

113. Cleveland Bd. of Educ. v. Loudermill, 470 U.S. 532, 548 (1985) (Marshall, J., concurring) (quoting Arnett v. Kennedy, 416 U.S. 134, 214 (1974) (Marshall, J., dissenting)).

the use of student evaluations has yet to be adequately addressed relative to the content, drafting, completion, and use of student evaluations.

We have shown that student evaluations do not measure what they are intended to measure—effective teaching. They may measure any number of other things, to some degree, including whether the students liked the process. We have also shown that there are numerous factors which may bias the input from the students as they rate the professor. None of those factors are considered in the final use of the student ratings. The final use of these ratings is to reduce them to some statistical medium and use that median score to determine whether the professor is an effective teacher (by ranking above the median) or an ineffective teacher (by ranking below the median.) This number is used by administrators to make life-changing decisions such as pay raises, retention, promotion, and tenure. These decisions are critical in the professional life of faculty members, yet the main piece of information used to make those decisions is seriously flawed and cannot be challenged by the professor in any substantive way.

One of the most critical flaws is that the student evaluations may not be valid. For example, Stapleton and Murkison demonstrated the limits of the term "valid" as applied to student ratings.[114] The data, from this study, revealed that some instructors confounded the general trend: of the twenty-nine instructors studied, four who produced learning in the top half received ratings in the bottom half, while four who produced learning in the bottom half received ratings in the top half.[115]

> Had personnel decisions been made on the basis of these data, with a cutoff at the median, four of the more effective professors would have been punished or dismissed, while four of the less effective ones would have been rewarded. This study highlights an important point about statistical data: an overall correlation between two variables does not mean that one variable is always correlated with the other in particular instances.[116]

Another study showed that, at best, there is a 50/50 chance that how high the professor was rated was correlated to how much the students learned.[117]

If the outcome of the classroom experience is supposed to be increased learning by the students, as claimed by the way the student evaluations are used, and student evaluations supposedly measure learning by the students, then using such invalid data certainly creates a "fundamental fairness" issue in these situations. Most of the scientific literature considers any correlation below .70 as unreliable,

114. Richard John Stapleton & Gene Murkison, *Optimizing the Fairness of Student Evaluations: A Study of Correlations between Instructor Excellence, Study Production, Learning Production, and Expected Grades*, 25 J. MGMT. EDUC. 269 (2001).

115. *Id*. at 279.

116. Fischer, *supra* note 85, at 125 (citations omitted).

117. In a study that used the students' grades on an external exam on the subject matter (one the professor did not prepare) as a basis for how much was learned, and correlating that with the various student rating items on the evaluation, the best result was a .50 correlation. *See* Cashin, *supra* note 33. *See also* P.A. Cohen, *Student Ratings of Instruction and Student Achievement: A Meta-Analysis of Multisection Validity Studies*, 51 REV. EDUC. RES. 281 (1981).

and often even higher correlations are required or expected for legitimate conclusions. Instead of teaching effectiveness, did the student evaluations instead measure student happiness with the process, the affect of the professor,[118] or something else? What was really measured? There is no reliable research that shows that the evaluations actually measure how much the students have learned from a particular professor. Yet, the assumption is that they measure the professor's teaching effectiveness even though none of the questions on the evaluation document actually determine how much learning took place. These studies and the related issues with validity show that these life-changing decisions are made in an arbitrary and fundamentally unfair manner, in violation of substantive due process.

In addition, the discussion relevant to the use of experts and their expert testimony clearly shows that under the Supreme Court standards for both of these, the students, and what they are "testifying" to, could not qualify as experts or as expert testimony. Therefore, the use of students as experts and the use of the medians as expert testimony as to effective teaching is also arbitrary, fundamentally unfair, and a violation of substantive due process.

The Due Process Clause requires that when the government takes away a fundamental right, it is done in a fair manner. What is currently done, by using these medians, is unfair both from a validity viewpoint and from the viewpoint of the expert testimony not meeting any of the requisite standards for such testimony. Due process and other legal issues arise when student questionnaires ask students to anonymously reflect upon "ill-informed expectations and comparisons with some hidden benchmark which differs from one student to the next . . . . [Proper use of evaluations must reflect the] individuality of our students [and] we need to acknowledge diversity and lack of homogeneity within a student group in terms of teaching."[119] The current use of student evaluations must be changed to make their use constitutional and provide appropriate protection for faculty members with a constitutionally protected interest.

E. Recommendations

Ultimately, what is necessary is a well-conceived assessment program, which will require considerable time, energy, and resources. It is essential that this drive for reform come from within the academy itself.[120] Higher education needs to take the lead in overall assessment reform, which includes defining, evaluating, and rewarding valid teaching behaviors linked to teaching effectiveness.

Teaching effectiveness can be adequately assessed only when multiple indicators of effectiveness are utilized. A direct-observation peer evaluation component performed by an expert evaluator skilled in pedagogical assessment, (which could include videotaping) is critical, as are additional multiple direct and indirect assessment measures. Additional measures of teaching effectiveness

118. Cashin, *supra* note 33, at 3–4.
119. Clouder, *supra* note 22, at 192.
120. Hersch, *supra* note 14.

should include the development of a teaching portfolio by the faculty member that permits an examination of class materials such as syllabi, assignments and examinations, handouts, and assorted deliverables produced by students in the class, as well as a statement of teaching philosophy.[121] Pedagogical and technological innovations utilized in the course that are proposed to enhance learning should be examined. Finally, and perhaps most importantly, the student teaching evaluations, if utilized at all, must be redesigned to reliably and validly assess specific teaching behaviors considered desirable by the institution and peer experts as much as possible, with an awareness that teaching evaluations are to be used only as one of a number of measures of an assessment triangulation process because they are subject to substantial biases and most likely measure the faculty members' ability to generate positive affect in the classroom. Future use of student evaluations ought to be constrained by the institution's ability to develop truly valid and reliable instruments.

As stated above, one component of a better and more constitutionally valid evaluation would be the proper use of peer evaluations. Peer evaluators would be known, rather than anonymous, would be expected to be experts in pedagogy, and could be asked the reasons for their scoring and calculation decisions. Peer evaluations may also be professionally valid if those completing the evaluations are teaching professionals with proper credentials and maturity, instead of eighteen to twenty-two year olds without such credentials or maturity to make constitutionally valid decisions as to the quality of the teaching received or the qualifications of their teachers. Of course, utilizing peer evaluations also requires meeting the same rigorous standards student evaluations are currently not meeting, and assumes that faculty, administrators, and peers must truly want fair, valid, reliable assessment of both teaching and learning.

CONCLUSION

Since many faculty members have a constitutionally protected interest in their teaching positions, in order to protect that interest there has to be both a proper process and a fair process, including procedural and substantive due process, in regard to a review of whether or not their contract will be renewed. Therefore, given that the current uses of the anonymous summative evaluations are invalid because they do not reflect the complexity of the teaching/learning experience and the "evidence" that they provide is not challengeable, the evaluations themselves and their use violate the substantive due process rights of those faculty who are constitutionally protected. Substantive due process rights are violated precisely because such evaluations cannot and do not measure what they purport to measure (quality teaching and teacher qualifications), are without meaningful statistically valid standards, and because the scoring and numerical comparisons of such evaluations cannot be challenged by accurately discovering which factors each anonymous student considered important when scoring each particular evaluation question.

121. B.W. Kemp & G.S. Kumar, *Student Evaluations: Are We Using Them Correctly?* 66 J. EDUC. BUS. 106 (1990).

Ambiguous and anonymous information thus collected is not considered factual evidence in other legal proceedings affecting fundamental constitutional rights. Likewise, it should not be allowed for use in the decision-making process when a professor's and/or other constitutionally protected university teacher's fundamental rights of life, liberty, and reputation are at stake. In addition, students are put into the role of being "experts" as to proper pedagogy, without actually being experts on pedagogy.

The entire use of student evaluations needs to be re-assessed in light of the substantive due process issues raised by their current use. They may have some appropriate use in making decisions about faculty performance, especially in terms of a faculty members' ability to generate positive affect, but they are not appropriate for their current use of assessing faculty performance, especially when such use results in life-changing decisions for a faculty member. Such re-assessment of their use is now even more important as we face an era of increased accountability, where heightened demands on faculty teaching performance are advocated, including raising the bar for measurable student performance and learning. Faculty and teaching professional often run amok of student evaluations by creating more challenging courses and insisting students increase their level of learning far beyond rote memorization. Student evaluations that do little to measure desirable teaching and learning outcomes are likely only to become even more problematic for colleges and universities in the future. With reliance on them unfounded, continued usage will result in more unjustifiable attacks on faculty members' teaching performance. Even though no specific lawsuits challenging their usage exist as of yet, as more faculty become affected by their unfair use, universities and administrators will increasingly find themselves in court, unless they make essential changes to the teaching evaluation process.