Education Policy Analysis Archives

Volume 8 Number 48 **October 1, 2000** **ISSN 1068-2341**

A peer-reviewed scholarly electronic journal
Editor: Gene V Glass, College of Education
Arizona State University

Articles appearing in **EPAA** are abstracted in the *Current Index to Journals in Education* by the ERIC Clearinghouse on Assessment and Evaluation and are permanently archived in *Resources in Education*.

"Put Teaching on the Same Footing as Research?" Teaching and Learning Policy Review in Hong Kong and the U.S.

Orlan Lee
Hong Kong University of Science & Technology

Abstract
The Research Assessment Exercises (RAEs) in hugely expanded universities in Britain and Hong Kong attempt mammoth scale ratings of "quality of research." If peer review on that scale is feasible for "quality of research," is it less so for "quality of teaching"? The lessons of the Hong Kong Teaching and Learning Quality Process Reviews (TLQPRs), of recent studies on the influence of grade expectation and workload on student ratings, of attempts to employ agency theory both to improve teaching quality and raise student ratings, and of institutional attempts to refine the peer review process, all suggest that we can "put teaching on the same footing as research" and include professional regard for teaching content and objectives, as well as student ratings of effectiveness and personality appeal, in the process.

> …in the winter term of 1992, the Simon School faculty passed a

> resolution, that determined: "[T]o establish a faculty committee to evaluate teaching content and quality on an on-going basis. *The intent of the proposal is to put the evaluation of teaching on the same footing as the evaluation of research.* The committee will have the responsibility to evaluate both the content and presentation of each faculty member on a regular basis to be determined by the committee…. The output of this process should be reports designed to provide constructive feedback to faculty and evaluations to be considered in promotion, tenure, and compensation decisions." (Faculty Meeting Minutes, University of Rochester, William E. Simon Graduate School of Business Administration, February 26, 1992, cited: Brickley and Zimmerman, 1997, p. 5, emphasis added).

Introduction

"Put teaching on the same footing as research?" I can hear my scholarly colleagues ask, "You mean another attempt to credit those who do 'teaching' to the detriment of their 'research'?" No, my friends, what I understand from the quote in the box is that administration would measure the quality of "teaching" on the same basis they demand from "research."

In 1997, in response to growing concern about maintaining quality of teaching and learning in expanding institutions of higher education—not only in Hong Kong, but worldwide—the University Grants Committee (UGC) (Hong Kong), undertook a study of the process by which teaching and learning quality was to be evaluated in Hong Kong institutions of higher education. This became known as the Teaching and Learning Quality Process Review (TLQPR) of 1997.

A series of institutional studies addressed critical problems bound to arise in an atmosphere of democratic interest in promoting expansion of economic opportunity and social mobility by means of wider access to higher education. It also revealed concerns within the institutions and the academic profession at large regarding free exercise of the functions of research and teaching, and their survival in light of calls for greater public accountability.

The UGC panel assigned to conduct the Teaching and Learning Quality Process Review (TLQPR) of the author's own University expressed its concern about the institution's reliance, almost exclusively, on mean quantified scores of student responses to course surveys to assess the quality of teaching and learning. This has also been a significant problem in teaching quality assessment in U.S. institutions since adoption of formalized "student evaluation" mechanisms as the result of student protest movements in the late 1960s and the 1970s.

No doubt, every teacher likes to be appreciated by his or her students. Similarly every student has an interest in minimizing risk in evaluation of his or her own course performance. But surely this situation describes a source of conflict of interest—likely on both sides—as much as a demonstration of the "validity" of "student evaluation" of teaching and learning on the theory that "the customer is always right."

A considerable volume of published research in this area attributes a "validity" to figures that are allegedly replicable because of their apparent "consistency and stability." Yet, we are also told that: "The literature on validity, though extensive, remains very

fluid and not perfectly conclusive." Still other researchers find that teaching ratings and learning are only "weakly related."

Some authorities on the literature tell us that in part this predicament arises from research concentrating on "construction of instruments to yield items and subscales which [are] intended to measure student learning outcomes." Yet they also report that others have found "content validity," i.e., "positive relationships between student ratings and achievement."

Chief factors that would establish "validity," these experts tell us, are that evidence suggests that students and instructors seem to agree on what constitutes "effective teaching" and on the qualities of "an ideal professor." This conclusion must be flawed if, as the present author suspects, the literature of education theory, and practical experience of student responses indicate that these two do not always share agreement on what "achievement" is, what "good teaching" is, and perhaps even on what "education" itself should aspire to.

This article compares the presumption of "validity" of "student evaluation" of teaching quality with the results of recent studies at the University of Washington on the influence of grade expectation and workload on student ratings, on the results of attempts, at the University of Rochester, to employ agency theory both to improve teaching quality and raise student ratings, and with the peer review model employed at the City University of Hong Kong.

I. Concerns about Quality of Teaching and Research in Expanding Institutions in Times of Contracting Budgets

In the Plenary Address of an International Conference on the Application of Psychology to the Quality of Learning and Teaching held in Hong Kong, Professor Robert J. Sternberg of Yale University (Sternberg, 1998) warned that universities that have used IQ tests, and other standardized measures of practical intelligence or practical experience as sole standards of university admissions, have created self confirming systems. "Only those with high IQs succeed, because only those with high IQs are admitted." The "tragedy" of this self selection as a "social goal," he said, is that "in our emphasis on skills that benefit the individual, we have created societies in which. . .the optimization of our individual outcomes at the expense of common well-being is becoming ever more pervasive."

The point of this paper is similar: if by "Quality of Teaching and Learning" we mean what style of Teaching and Learning is most popular with our students, or most satisfies the expectations they bring with them from their schools, or what they believe most readily facilitates their immediate needs in getting jobs or obtaining professional certification, that is what they will confirm to us in student ratings.

If, on the other hand, our goal is to contribute to modifying the tendency to the rote learning and recitation method, and to promoting critical thinking and general education—as the Vice Chancellors of both sponsoring institutions of the Hong Kong Conference, the University of Hong Kong, and the Hong Kong University of Science & Technology, urged in their opening addresses—then we better attempt to balance student input, with reasonable professional efforts to meet those expectations.

In response to numerous and growing concerns about maintaining quality of teaching and research in expanding institutions of higher education, not only in Hong Kong, but worldwide (see, e.g., Clark, 1995)(Note 1), the University Grants Committee

(UGC) (Hong Kong),(Note 2) has undertaken studies that will affect the funding of both the research and teaching sides of university functions. Three Research Assessment Exercises (RAEs), studies of the research being done in Hong Kong universities, were carried out in 1994, 1996, and 1999. A study, not of teaching and learning quality as such, but of the process for reviewing the quality of teaching and learning in Hong Kong institutions of higher education—the Teaching and Learning Quality Process Review (TLQPR) (see: Massy ; French)—followed in 1997, and a second is proposed for 2000-2001.

Both sets of studies addressed critical problems bound to arise in an atmosphere of democratic interest in promoting expansion of economic opportunity and social mobility by wider access to higher education. Both also reveal concerns within the institutions and the academic profession at large regarding free exercise of the functions of research and teaching, and their survival in light of calls for greater public accountability. The author has already described some of the professional concerns arising in the Research Assessment Exercises, the RAEs (see: Lee, 1998). The following discussion will address similar concerns with respect to the TLQPR. Whereas the author has expressed some reservation with respect to the former (the RAEs), he is generally in agreement with the latter (the TLQPR)—and especially as it affects his home university.

II. Measuring Teaching and Learning Quality

The announcement of an International Conference on the "Application of Psychology to the Quality of Learning and Teaching" (Hong Kong, June, 1998), indicated that it "strongly emphasize[d] cutting-edge research on the application of psychological principles to improving learning and teaching quality, with the aim of developing a global perspective on learning and achieving motivation" (HKU; HKUST, 1997).

With research on psychology of teaching and learning so highly specialized that a paper submitted to the Hong Kong conference required at least one of 27 keyword codes to classify it before it could be considered, it would appear that there are at least that many psychological perspectives alone from which to evaluate quality of teaching and learning. No wonder the TLQPR was troubled to find institutions with only student ratings in place.

II. A. Standardized Student Ratings Surveys

II. A. 1. Sole Use of "Student Evaluations"

It is understandable, in light of the multiplicity of just the psychological perspectives on teaching and learning, that the UGC (Hong Kong) panel assigned to conduct the 1997 Teaching and Learning Quality Process Review (TLQPR) of the author's own University expressed its concern about our University's reliance, almost solely, on mean quantified scores of students responding to semester surveys to assess the quality of teaching and learning in our various courses: "There appears to be little systematic monitoring of teaching and learning quality [at HKUST] other than through the [student] teaching evaluation questionnaires..." (TLQPR, 1998, para. 16). This phenomenon is doubtless far more pervasive than only at HKUST, or only in Hong Kong. The problem surely reflects not only that universities do not know better ways to

evaluate teaching, but probably also that they have no clear idea of what they want to accomplish in their courses either.

Despite the University response to the TLQPR, this imbalance was still reflected in the subsequent HKUST, *Faculty Handbook, 1997*, where, after indication that review of faculty performance for retention or promotion would involve consideration of "research, teaching, and service," it is made clear that unlike the case with "research" and "service": Reviews of teaching performance rely to a greater or lesser extent on *student evaluations* . . . (HKUST, 1997, p. 169, emphasis added).

The appearance of being responsive to student concerns is such a pre-occupation with university administrations that follow the American model, that finding a professionally acceptable method of evaluating what reasonable people recognize to be the essential characteristics of good teaching continues to elude them. One of the leading American authorities on "student evaluation"—who has great hopes of reforming the prevailing system—concedes privately:

> Most universities in the USA give lip service to using information other than student ratings for teaching evaluation. However, at most places the information obtained by other means (teaching portfolios, peer evaluation) is rarely put into a form that permits ready use for evaluation. Consequently most places end up relying primarily on student ratings.

That was precisely the HKUST administration's response to the TLQPR. Despite elaborate verbal acknowledgment of the existence of all other means of evaluating teaching in theory, the official "Progress Report to the University Grants Committee" (2 March, 1998), comes full circle to student ratings, and essentially concedes that at HKUST there *is* nothing else—students evaluate teaching. The university administration then lists "repeat offenders" and "monitors" faculty "accountability":

> A more formal use of the student evaluation results to monitor Department accountability for teaching performance was introduced in the past year. It involves the identification, by the Academic Affairs office, of a group of instructors with particularly poor records of performance in the previous year. Department Heads were provided with a list of any faculty members in their own Departments who have been so identified, and asked to take appropriate corrective actions to help these instructors improve. In subsequent years, Department Heads will have to provide, for any instructor who turns up on the list as a "repeat offender," details on what actions, if any, were taken, and a statement of planned future actions to address the problems. (TLQPR Progress Report, 1998, p. 2).

Surely, every teacher likes to be appreciated by students. But, is that why our University relies almost exclusively on that one measure—what our students say about us—to assess our teaching competence? I doubt it seriously.

In Hong Kong, as elsewhere, institutional growth accompanied growth of student population. A subsequent dramatic change in the rate of student population growth, together with declining economic growth, means that there is, now, a heightened awareness of inter-institutional competition for student applicants (see: e.g., JUPAS, 1997), which leads inevitably to greater sensitivity to student tastes and student demands—doubtless one of the chief sources of the "student evaluation of teaching" movement in the first place (cf. Imrie, 1996).

Institutional growth, especially in Hong Kong, had been phenomenal in recent

years (see: UGC, 1996). We are told, that full time equivalent enrollments (FTEs) in higher education increased from 42,000 in 1990-91 to 62,000 in 1995-96, or an increase of roughly 47% in only five years, giving rise to concerns about how institutions would be able to maintain the quality of teaching and learning (HKU, 1997, para. 3), but also about how new institutions would fare in regard to competition for student enrollments.

II. A. 2. Why Is There No Other Established Measure?

Over the years, there has been a great deal written about the overemphasis on, and inherent conflict of interest in, "student evaluation" of professional performance—for which there is no parallel in any other profession (see: Appendix: "Conflict of Interest," 1974-82, and "Formative"and "Summative" uses, 1970s). But how did it happen that there was no existing institutional system of measurement of teaching and learning effectiveness in the first place, that would have addressed quality of teaching and learning concerns suitably, prior to the massive expansion of the use of "student evaluation"? Ask any college or university teacher and you are bound to get a sense of why: "Academic freedom" (Note 3) (cf. Flexner, 1967)—i.e., from the perspective of what the Germans call, "*Lehrfreiheit*," the "freedom to teach without interference." None of us is particularly fond of having other colleagues, or administrators, poking their noses into how or what we teach.

As a consequence of our profession's concern with generations of political and ideological attempts to control what we can do or say in the classroom, we have been brought up with an academic legacy of resistance to thought control and, therefore, have developed no mechanism or standard, universally accepted, for assessing what we do, professionally, or how well we do what we do in the classroom. Consequently, the teaching profession was an easy target for institutions seeking to satisfy reformist demands in this area in the late 1960s and early 70s. For this reason, and because of our even greater subservience to those in the education schools, in teaching technology, and in educational testing, we have allowed new professions to arise which specialize in telling those of us who teach "how to do it better." (cf. UGC, 1996, p.8) (Note 4)

All of us in the academic world know that our students will observe and react to our flaws and weaknesses as much as to our strengths. Yet, when it comes to assessment of our professional performance and abilities, most of us expect the same courtesy in evaluation as is accorded to other professionals (cf. Appendix, "Consumerism," 1976-91) (Note 5)—and to our students:

- evaluation by those who understand what we are attempting to do;
- evaluation by those who have a professional understanding of what we should do;
- evaluation without conflict of interest; —as well as, of course,
- evaluation for effectiveness.

II. A. 3. Need for Student Feedback

There is no need to convince the present author—at one time or another a candidate for five university degrees—that students often have valid opinions and cogent arguments. Which one of us, as a student or a faculty member, has not sat through

lectures, and even whole courses, that we would be ashamed to have given ourselves. Simply being boring is a malady that even the best of us suffers from at times. These are concerns, which certainly should not be silenced, and perhaps also deserve some greater outlet for discussion on all campuses.

The Harvard Crimson Confi-Guide once served a function like this. At one time the independent Harvard University student newspaper gathered and published student comments on their Harvard courses—a short web search revealed that they still do. But that is all it purports to be. It makes no pretense of being a "survey," of being "scientific," or even of being "quantitative" in its results. It refers to itself as embodying: "Irreverent and honest appraisals of your favorite (and not so favorite) Harvard courses":

> Be very careful what you do with this guide. Read. Enjoy. Laugh out loud. The goal of the *Confidential Guide to Courses* is . . . to help students by giving them the lowdown on classes. Is it good? Is it a gut? Does the professor give interesting lectures? Are the exams difficult?
>
> This guide generally succeeds in providing that information, but that doesn't mean the articles have all the answers. They are meant to be helpful, but they can't necessarily be taken at face value.
>
> Each article is an opinion piece written by a student who took the class recently. The author can say whatever he or she wants, no matter how big the chip on his or her shoulder. It's important to remember that different people can come away from the same class with different impressions. . . .(*Confi-Guide*, 1998).

Instructors know, or ought to know, that they can get feedback from their students on how effective their teaching style is. Some do this by survey; some by private chat; some by instinct. But this does not mean that every student comment is good as gold or ought to be taken to heart. A professional person has to know for himself or herself what to make of such comments. That is not what standardized testing or survey research does, however. As we all know, you cannot argue with the question where you already know that the tested population is so large that the examiners—or the survey experts—are only looking for a positive or negative response pre-defined to carry specific conclusory meaning. That may sound like poor survey or test writing. Nevertheless, practically speaking, any teaching rating questionnaire will call for these same up or down responses. Professor Wilbert McKeachie, probably the most authoritative figure in the student ratings genre writes critically of this technique:

> . . . effective teachers come in many shapes and sizes. Scriven (1981) has long argued that no ratings of teaching style (e.g., enthusiasm, organization, warmth) should be used, because teaching effectiveness can be achieved in many ways. Using characteristics that generally have positive correlations with effectiveness penalizes the teacher who is effective despite less than top scores on one or more of the dimensions usually associated with effectiveness. Judging an individual on the basis of characteristics, Scriven says, is just as unethical as judging an individual on the basis of race or gender (McKeachie, 1997, p. 1218).

With all respect, there is something disingenuous about this admission. Those who have done most to promote the concept of "validity" of measures here admit they may be accurate *only* for what they measure literally. Then they argue that they do not measure what administrators are known to want to apply their quantifiable results for. They give

teaching assessment committees a howitzer and tell them to use it like a smart bomb:

> Almost as bad as dismissal of student ratings, . . . is the opposite problem—attempting to compare teachers with one another by using numerical means or medians. Comparisons of ratings in different classes are dubious not only because of between-classes differences in the students but also because of differences in goal, teaching methods, content, and a myriad of other variables . (McKeachie, 1997, p. 1222).

In other words, (1) ratings are considered "valid," yet, (2) the quantified results relate only to individual performance. That is, they may presumably be used for "formative" and "summative" purposes—i.e., to advise that particular instructor how to *improve teaching*, and, ultimately, to advise the personnel committee how to *judge effectiveness* of that instructor. However, whereas results are expressed in quantified form, the scores for identical qualities are to be considered "not comparative."

It may be that schools with great sophistication in the use of student survey scores express such a qualification as to how student numerical ratings are to be applied—publicly. In practice, however, I do not see any hesitation in considering an 80% rating of one instructor equivalent to an 80% rating of another. At the author's University, for example, both get congratulatory letters from the Dean. Similarly, with a 40% rating for two years in a row, any instructor is bound to be considered a "repeat offender."

Accordingly, with regard to survey sophistication at HKUST, we are forewarned: "Note that the descriptions of the ratings should not be taken literally." (HKUST, 1998) Read further, however, and one is told that: "The average scores for all courses is in the range 60-70, so that the 'average' course has an 'above average' rating (HKUST, 1998)."

Does this mean that our administrators are so sophisticated about statistical and survey measures that they count these scores for no more than a simple exercise in measuring student opinion? Not on your life. We already know from Section II.A.1. above, that "Reviews of teaching performance rely to a greater or lesser extent on *student evaluations*," and "repeat offenders" will be dealt with.

Let me say first of all that the Hong Kong University of Science & Technology would rate itself as among the top universities in Asia—if not in the world. But "the average scores for all courses," judged by our students, we are told here, are rated between D+/C- and C+/B-. Heaven help the instructors whose average grades for their own students actually looked like that! But perhaps you may say that our students are more honest about us than we are about them.

What is the source of this disparity in ratings between faculty of students and students of faculty? Grade inflation can also have varying sources—since, according to this report, at least, it is not simply producing higher faculty ratings. Presumably the faculty believe that they are achieving better results with students than students give them credit for. Does it go too far to suggest that the two may have different concepts or goals of teaching and learning in mind, and that that is what their respective grades and ratings scores are measuring?

This disparity in concepts and goals of education will be dealt with further below (at Section II.A.6). In this connection, however, let us take a closer look at something else Wilbert McKeachie alludes to in passing in his paper in the "Current Issues" section of the *American Psychologist* (November, 1997) devoted to controversy over findings in the students' ratings research. McKeachie is willing to admit exactly the inherent contradiction of goals and objectives in student evaluation of teaching:

> There are . . . two problems that detract from the usefulness of ratings for improvement. . . . Many students prefer teaching that enables them to listen passively—teaching that organizes the subject matter for them and that prepares them well for tests. . . .
> Cognitive and motivational research, however, points to better retention, thinking, and motivational effects when students are more actively involved in talking, writing, and doing.

This inherent conflict of interest, notwithstanding, McKeachie justifies the continued reliance on the ratings survey system on the basis of what it is conceptually intended to achieve, i.e., "feedback":

> The second problem is the negative effect of low ratings on teacher motivation. . . .A solution for both of these problems is better feedback (McKeachie, 1997, p. 1219:1).

Only one set of convictions can conceivably attempt to justify knowingly relying on a system of assessment that you concede is based on conflict of interest: (1) the persuasion that an institutional system of measurement of teaching effectiveness is mandatory for personnel decisions; and (2) that no professional measurement compares in "validity" (as we shall see shortly, he says as much) with student ratings.

Here, I suspect we do have the root of the dichotomy in the grading and ratings problem: "Many students prefer teaching that enables them to listen passively. . .and that prepares them well for tests," and judge faculty on that basis. On the other hand, many faculty members are persuaded that "retention, thinking, and motivational effects" are greater "when students are more actively involved in talking, writing, and doing." I suspect that they also tend to grade on the belief that they are achieving results of this kind. While each scoring system may be perfectly honest as far as what it purports to measure is concerned, as McKeachie says, ". . .the two problems detract from the usefulness of ratings for improvement," i.e., for the much vaunted "formative" effect. McKeachie, further on, gingerly admits, the two systems simply do not relate to each other: "However, student ratings are not perfectly correlated with student learning. . . . " (McKeachie, 1997, p. 1219: 2)

The "solution for both of *these* problems [may be] better feedback." However, while educational technologists may believe that they are promoting feedback, there is in reality little communication about these matters in large public institutions, either between faculty and students, or between each among themselves. Student ratings are an *educational technology product* that, regardless of the mildly qualified claims of those who argue "validity," provide academic administrators with what purports to be quantitative measurements of teaching effectiveness—and that is precisely how the survey technologists expect them to be used:

> But what about the use of student ratings for personnel decisions? Here again the authors of the articles in this *Current Issues* section [of *American Psychologist*, November, 1997] provide reassurance. All of the authors (and I join them) agree that student ratings are the single most valid source of data on teaching effectiveness. In fact, as Marsh and Roche (1997) point out, there is little evidence of the validity of any other sources of data. (McKeachie, 1997, p. 1219:2).

II. A. 4. Attractiveness of "Student Evaluation" Surveys

The beauty of student ranking surveys for a college or university administration is that they are cheap, and that they purport to offer exact *quantitative*, and, like it or not, *comparative* figures between faculty members. On their face, they appear to be the unqualified ranking by a representative sampling of students taking a course—without need for discursive explanations—moral, legal, or professional. The president of the author's university also reports that instructors have been fired because of low ranking in student evaluation surveys: ". . . In terms of system, all courses are evaluated by students and the results are disclosed on the World Wide Web; unsatisfactory teaching performance has resulted in many cases of contract non- renewal or salary bar. . . ."(Woo, 1997)

In a note in reaction to the foregoing observations, the President seems to take a more balanced view: "We certainly cannot just rely on student evaluation *scores*. Good teachers often get remembered only long after the students have graduated." This was despite subsequent publication of the "Report to the University Grants Committee" (2 March, 1998) cited above. Obviously the President has sensibilities as a teacher as well as an administrator.

II. A. 5. Crucial Variables and Consistency and Stability of Results

With the exception of some actually sometimes crucial variables(Note 6)—prior subject interest, class size, time of day a course is taught, rank of the instructor, grades expected, and course load which educational measurement investigators acknowledge affect student ratings of faculty in some way (cf. Appendix)—there have been a number of student ratings researchers who have argued that the student survey system is "consistent and stable." That is, they argue, similar ratings are seen to be attributable to the same faculty members, regardless of the subject matter they teach, and from year to year. Moreover, some investigators attribute close correlations to more professional appearing reviews by peers, administrators, and alumni (cf. Appendix).

Yet, while such correlations between results of different groups of survey subjects may exist at times, other researchers tell us that, teaching ratings and learning are only "weakly related" (Gramlich; Greenlee, 1993). To the extent that this is true, it would tend to link the rating with the faculty member's teaching style or personality, and would tend to obviate one supposed major purpose of ratings, i.e., that they are "formative," that they can be used to assist the instructor to achieve improvement either in the teaching itself, or in its reception by students.

Nevertheless, some researchers in this area attribute a "validity" to figures that are supposedly replicable because of their apparent "consistency and stability." Yet, the same authority tells us: "The literature on validity, though extensive, remains very fluid and not perfectly conclusive" (Arubayi, 1987, p. 270).

In what A.G. Greenwald has called "the best of the largest group of construct-validity studies" (Greenwald, 1997, p. 1184) there seemed to be evidence to support correlational validity between student ratings in multisection courses. Here the results of student ratings were compared for different instructors giving different sections of the same course, where similar or identical examinations were given to different sections with students with similar ability (Abrami; Cohen; d'Apollonia, 1988).

The present author, who has, heretofore, limited himself to reviewing the literature on this subject, must interject at this point that he has observed completely unforeseen but sharply conflicting statistical results on this particular kind of experiment. The

II. A. 6. Is There Validity If There Is No Agreement on Outcomes?

The same authority on the literature who argued "validity" because of apparent "consistency and stability" tells us that part of the predicament of "fluidity in research results" lies in the research concentrating on "construction of instruments to yield items and subscales which were intended to measure student learning outcomes" (Arubayi, 1987). He reports that others have found "content validity," i.e., "positive relationships between student ratings and achievement" (Arubayi, 1987).

Other factors that would establish "validity," this expert tells us, are that:

> Evidence suggests that *students and instructors seem to agree as to what leads to good teaching.* Similarly, . . . very close *similarity between the perceptions of students . . . on what constitutes a[n] "ideal professor." If students can agree with their instructors* as to *what constitutes effective teaching* and the qualities of an ideal professor then one might be sage to conclude that students are mature enough to rate or evaluate instructors and instruction (Arubayi, 1987, p. 270f. emphasis added).

Reliance on near-exclusive use of "student evaluation" of teaching is bound to arouse concern for those of us in Hong Kong—where there are also faculty members to be found, who, while deeply attached to the region, their students, and the subject matter of their fields**,** do not share agreement with their students on what "achievement" is, what "good teaching" is, and perhaps even on what "education" itself represents.

In no way does it dispose of the issue to say that those faculty members are themselves out of joint, and that the situation will be cured by localizing expatriates out and putting local people in their place. The definitions of "education" and "achievement" are not simply heritage and culture-bound. An institution like the Hong Kong University of Science & Technology is overwhelmingly staffed by PhDs from the world's leading universities. Are we to believe that they are prepared to abandon the educational values they hold for themselves—and upon which they want their own research and career accomplishments to be judged—when they instruct their students?

"We ought to teach every course the same way we would teach majors in the United States," our University President Woo Chia Wei is reported to have opined—somewhat at odds with what as an administrator he seems to be telling us. Are we to believe that there is one set of values for the world, and another for our own students?

How would I teach in the U.S.? Like an Ivy League graduate would be expected to:

- Evaluating how we GATHER FACTS;
- Establishing how we DEFINE A PROBLEM;
- IDENTIFYING ISSUES and METHODS leading to various SOLUTIONS of a problem;
- STRESSING REASONING over factual information;
- STRESSING HOW WE REACH CONCLUSIONS—NOT OPINIONS (Lee, 1997).

Does this form of teaching offer an advantage to Hong Kong and to China? Many of us believe it does—not least of all the Vice Chancellors who keynoted the international conference in Hong Kong on Teaching and Learning Quality.

By no means do all Western educated scholars in Hong Kong pursue this method. But, those who do, know that this style of teaching is not the mainstream tradition of the region. The instructor dedicated to this approach is, therefore, faced with the deliberate choice—of attempting to bring his or her students out of their protection of silence and anonymity to develop discursive verbal abilities (Lee, 2000) or—of abandoning what he or she believes is both sound practice—and attainable with persistence—in order to pursue the more accepted purely didactic approach that will gain him better ratings.

Many of our students are afraid that departure from their accepted learning habits—and how such a change in them will be received by their peers—will create a disadvantage to them in competing: first with their own classmates for grades, then with their fellow graduates, for jobs. They are, therefore, more at home with the standardized testing and curved grading results aspect of the American heritage, believing that they must receive and repeat exact information to be "testable," and that it is, therefore, "unfair" to them to introduce new standards of teaching and learning that suddenly give away their "place on the curve."

These conclusions are not based upon a formal scientific survey, but do derive from years of listening to student comments, both personal and anonymous. However, more formal case studies in Hong Kong have produced similar results. In a case study on law student learning in English at the University of Hong Kong, for example, three language use researchers conclude: "…by the time students reach the end of their secondary education and probably well before that point, they have internalised a set of unstated survival strategies for choosing which language to use [Cantonese or English] or, indeed, whether to communicate at all in a given situation." (Corcos; Churchill; Lam, 1998).

They refer to a set of implicit socio-cultural rules derived by an earlier researcher in this area:

- If you want to talk to another student in a friendly way and without seeming superior, you must not use English;
- Do not show off your language proficiency in front of your peers;
- You should deny such proficiency if anyone praises you;
- You must hesitate and show difficulty in arriving at an answer when called upon by the teacher;
- You must not answer the teacher voluntarily or enthusiastically in English;
- You must not speak in fluent English (Wong, 1984, as cited).

Similar defenses to class response techniques apply in other parts of the world (even in some parts of the U.S. where "class participation" is established doctrine), however, in Hong Kong, university instruction in English, a foreign language, though still the basis for official and business communication, serves as cover for non-participation. Actually response in Cantonese is no better—if students are not accustomed to verbal reasoning.

II. B. Measurement and Enhancement of Teaching by Peer Review

Of course you listen to your students—and you adjust to whoever comes. But is that all there is? If better teaching and enhanced learning are desired, experience tells us that they can be encouraged or cultivated—the elements are all well-known. (Note 7)

We may agree that there is a difference between encouraging enhanced quality of teaching and learning, and merely conducting a survey to see whether teaching conforms to students' established expectations. However, encouraging better teaching by whatever

method may involve changing incentives and investing greater resources, and may, therefore, discourage administrators from pursuing such a course too vigorously in times of contracting budgets. But testing is cheap, and appears to satisfy the student constituency.

II. B. 1. Changing Incentives from Research to Teaching

The process by which incentive structure can be changed in a university environment has been described in the literature in the same terms as changes in incentive structure in business. This process was employed in efforts to reinforce the teaching and learning environment at the William E. Simon School of Business Administration at the University of Rochester, and apparently in other leading American business schools, when the administrations determined that environmental factors affecting them, leading to competition for public funding and for student applicants, were similar to those described at the outset of this paper as leading to the Research Assessment Exercises (RAEs) and Teaching and Learning Quality Process Review (TLQPR) in Hong Kong (see: Brickley; Zimmerman, 1997—the following relies on that report).

The birth rate has long been declining in the United States, leading, over the years, to declining numbers of children in schools, and, as a result, declining numbers of students in colleges and universities. In the late 1980s this reduction in numbers of applicants was also felt in the graduate schools of business—combined with a lower demand for MBAs as a result of economic conditions.

Competition for applicants among American business schools first led to enhanced spending on public relations, then on scholarships, and, finally, on enhanced spending on incentives to improve the teaching environment. At about that time, *Business Week* began publishing a biannual list of top-20 business schools, and asked graduating students and recruiters to rate the schools according to opportunities 1) either in class or in extracurricular activities, and 2) to nurture and improve your skills in leading others (Byrne; Leonhardt, 1996).

Focus on Research emphasis, so important in the competitive standing of former years, received no special mention, and seemed to have fallen by the wayside in a competition fired expressly by students' interests.

Concern with media rankings seems to have been quite intense. The Simon School at Rochester, was for example, listed in the *Business Week* top-20 business schools in 1988, and 1990, but not in 1992. As a result, a number of business schools, including Rochester, were led to serious reconsideration of their academic programs—emphasizing enhanced incentives to improve teaching. A faculty report at Rochester called for efforts to:

> . . . increase teaching incentives, and make the change clearly visible to applicants, students, administrators and faculty
> ("MBA Program Status Report," University of Rochester, William E. Simon Graduate School of Business Administration [June 14, 1991] cited: Brickley and Zimmerman, 1997. Cf. also: "The Report of the Task Force on Improvement," M.I.T., Sloan School of Management [May 7, 1991]).

To meet the demands of that situation, the School of Business Administration at the University of Rochester determined to become more competitive in the market for business school applicants. In the process, they determined to enhance their standing as a

top-20 business school by seeking to attract student applicants by an enhanced teaching and learning environment—a significant change from the emphasis on advanced Research in the 1980s, when the applicant level was strong and rising.

II. B. 2. Changing to a Peer Review Measurement System

It is interesting to observe that at about the same time as The Simon School at Rochester was engaged in the process of re-assessing its system for teaching evaluation, a similar process was underway at the City University of Hong Kong—for different reasons.

In 1993, the year before full university status was conferred on the then City Polytechnic, the Academic Board (now the Senate) established a Quality Assurance Committee which laid down guidelines for, among other things, teaching evaluation (QAC, 1993). While emphasizing that teaching evaluation "*must* include student feedback as a substantial primary element in the process," the Guide makes clear that teaching evaluation must also be an institutional determination: conforming with stated "*policy*" and "*principles*," based on all available "*evidence*," fully "*documented*," and "*accessible*":

> Teaching evaluation must conform to the *Principles* stated. . . . Teaching evaluation schemes must be documented. . . .
>
> The *primary* purpose of any teaching evaluation scheme should be to improve teaching. Teaching evaluation schemes *must* include student feedback as a substantial primary element. . . . Where a scheme is designed to evaluate teaching for assessment purposes, evidence *must* be included from other appropriate sources such as peer review, individual reflection, expert observation, etc., *in addition* to student feedback. . . . Those entrusted with using the information from teaching evaluations for decision-making related to career progression should be skilled in interpreting and drawing together the different sources of information. . . . In all cases the staff member being evaluated must be fully consulted. . . . Provisions should exist for regular review of the . . . evaluation schemes and of the institution's evaluation procedures (QAC, 1993, p. 1f.)
>
> (The first paragraph is taken from "policy," the remainder from "principles." The Guide is undated, but acknowledges Hall; Cedric; Fizgerald, 1994, as the source from which its principles were developed.)

This policy has been applauded in the TLQPR at City University. Yet, both from the TLQPR, and from faculty comments, one gets the impression that this system has not been fully implemented at City University either.

In both cases cited above, recourse to a peer review measurement system was motivated by new roles of the institution—calling for greater attention to the teaching and learning mission. On the other hand, both institutions (or their faculties?) were remarkably sensitive to the implication that either matters of professional competency or career decisions might be driven purely by reaction to data arising solely from student inputs. Clearly, both institutions were acutely attentive to the importance of maintaining ultimate *institutional responsibility* for professional decision-making, and correspondingly, professional information gathering.

As a result of the situation described in the foregoing section, the Simon School made a significant decision to change from dependence solely on the student quantitative

rating system for course and instructor, to a highly organized qualitative peer review system.

Based on the evidence of the cited study that teaching ratings and Learning was only "weakly related" (Gramlich; Greenlee, 1993), and on the concern that "some instructors game student ratings by reducing course work loads and cutting analytic content," or "...hand out cookies, bagels, and wine and cheese the last day of class when student ratings are administered" (Brickley; Zimmerman, p. 5), in the winter term of 1992, the Simon School faculty passed a resolution, that determined:

> [T]o establish a faculty committee to evaluate teaching content and quality on an on-going basis. *The intent of the proposal is to put the evaluation of teaching on the same footing as the evaluation of research*. The committee will have the responsibility to evaluate both the content and presentation of each faculty member on a regular basis to be determined by the committee. . . . The output of this process should be reports designed to provide constructive feedback to faculty and evaluations to be considered in promotion, tenure, and compensation decisions.
> ("Faculty Meeting Minutes," University of Rochester, William E. Simon Graduate School of Business Administration [February 26, 1992], Brickley; Zimmerman, p. 5 emphasis added).

In the case of City University of Hong Kong, the faculty Quality Assurance Committee (QAC) took a more systematic approach, in a manner befitting its role in determining future guidelines for policy of a major university, it devoted its early efforts to outlining statements of principles on quality and quality assurance. While these principles clearly were to acknowledge the role of students and other "stakeholders," e.g., employers and professional bodies, they were not to be construed in such a way as would utterly disenfranchise the teaching faculty: "The systems of quality assurance must be capable of operating independently of the participation of particular individuals and have an integrity which enables judgements to be formed that are unaffected by other managerial imperatives." (QAC, 1993, p. 4)

What is recognizable from the City University statements and principles is that these derive from faculty deliberations and are not simply imposed from above. In this respect, they are unique in circumscribing the activities of the whole institution:″Quality assurance policies should embrace all activities of the institution (QAC, 1993, p. 4). These principles not only recognize the institution's public roles and obligations to student's and other "stakeholders," they declare that they will apply "in all aspects of the staff's role including teaching, research, and administration" (QAC, 1993, p. 4).

II. B. 3. Implementation of the Peer Review System

As long as an informal quantitative student rating of course and faculty member was the only goal, it could be accomplished with comparative ease by passing out and collecting questionnaires at the end of the semester. If the evaluation of teaching were now to be put "on the same footing as evaluation of Research," then an objective means of qualitative measurement of the work of the course and the faculty member had to be found. For this purpose, the Rochester Business School faculty established a "Committee on Teaching Excellence" (CTE). The Committee developed a set of procedures, following the example of psychoanalysis, by first setting about evaluating six of the courses taught by members of the Committee itself:

> By the end of the 1993 academic year the CTE established a process, that except for minor changes, remains in effect through 1997. This process includes *benchmarking* the class with other top business schools: using a two-person evaluation team to *observe lectures*, *review material*, and *conduct student focus groups*; *video taping* several classes; *full committee discussion of the course*; and a *final written report* which goes to the instructor and the Dean's office and which is included in the faculty member's personnel file.
>
> . . . In addition to evaluating nine individual courses each year, the CTE held several seminars to discuss teaching. These forums allowed faculty to share their experience on various topics including: teaching cases, using computer- based presentation packages, and managing class discussion ("cold" calling). These seminars in the 1995 academic year were the first faculty seminars devoted to teaching (Brickley; Zimmerman, 1997, p. 5, emphasis added).

Evaluating the teaching process—involving analysis of quality of inputs or preparation and materials, form of classroom delivery, and measurement of effect upon students and their achievement—is necessarily a time intensive effort for all Committee members. The opportunity cost to evaluate one course was estimated at (US)$15,000.

In the case of the City University of Hong Kong, as well, the section of the *CityU Policy and Guide for Developing Teaching Evaluation Schemes* dealing with peer review specifically refers to evidence drawing on the following topics, and calls for citation of evidence in each case:

1. **subject expertise**: (up-to-dateness of content material);
2. **module design**: (relationship between content and objective, sequence, etc.);
3. **enhancing student learning**: (activities included, assessment requirements, etc.);
4. **module organisation**: (variety of experiences, reading lists, availability of materials, etc.);
5. **supporting departmental goals**: (from departmental objectives);
6. **research supervision** (QAC, 1993, sec. 2.2.2).

The guidelines conclude with the admonition that any peer review scheme must emphasize "expertise," "integrity," and "training" (QAC. 1993, sec. 2.2.2), both in the collection of data and its interpretation. No doubt this system, as well, must require a considerable "opportunity cost" that the institution considers is justified.

III. An Assessment System that Dwells on the Past? Or Education Policy with Increased Incentives for Teaching?

It should not be necessary here to enumerate the extent of the literature on opinion survey research. Neglect of comparative validation of an investigator's particular empirical method, or neglect of the potential impact of pre- existing biases—both among the research subjects, and among the investigators—would ordinarily arouse sufficient consternation among scholars of the field that such results would receive little credibility.

As the foregoing has suggested, however, there has been little attempt to obtain general agreement on the standards of psychometric validity of student ratings of

teaching despite the fact that investigators are well aware that their findings are being put to practical use in so-called "formative" and "summative" evaluation of members of their own profession.

Very simply, there appear to be two camps: 1) Those who treat student ratings as a reasonable "input" to "formative" and/or "summative" teaching assessment—along with all other professionally accepted indices; and 2) Those who consider that student ratings are the "valid" and sufficient basis for "formative" and "summative" evaluation of teaching by themselves. Institutions that employ student ratings alone tend to be interested primarily in quantitative and comparative results—i.e., numerical values that can be employed across the board to gauge and reward faculty performance.

Within the context of the empirical research reports, however, little interest is shown in qualitative criticism of the formulation of survey questions in student opinion surveys—and little attention is given to the impact of value systems in interpretation of survey questions. The foregoing has shown that leading authorities in the area: e.g., Scriven and McKeachie recognize the danger of confusing "characteristics that generally have positive correlations with effectiveness" with either "effectiveness" per se, or as all there is to be said for good teaching, or, more important, what teaching policy should aspire to.

Recognizing the needs of students in acquiring the skills to comprehend and master the subject matter of their field, and response of the instructor to the needs of a particular body of students is certainly one aspect of good teaching. But formation of forward looking education policy, cannot endlessly avoid the necessity of considering the obligation of the instructor—and of the institution—to the public and to the profession of teaching, to pursue clear educational goals which reflect the ambitions of our civilization and not simply those of any one generation of students whose priority is solely admission to professional qualification.

III. A. Haskell's Survey of the Literature of Pschometric Validity of Student Ratings and of Whether There is a Cause of Action for Violation of Academic Freedom for Reliance on Student Ratings in Personnel Decisions to the Exclusion of Everything Else

The serious omission of a qualitative discussion of psychometric validity of student ratings has been addressed in a comprehensive, at times rambling, series of four articles, a study of the literature of student ratings theory by Robert E. Haskell, Professor of Psychology at the University of New England in the United States (Haskell, 1997a,b,c,d).

Haskell is clear about his own personal position, "SEF [student evaluation of faculty] is deceptive regarding its negative implications for higher education" (1997b, p.3), and that the present system ". . .sets up a conflict of interest between the instructor and quality of education. . .[the] opposite of the original intent of SEF which was the improvement of instruction" (1997a, p. 16). It is inescapable that these considerations must return to the forefront of academic discussion at the turn of the century as democratization of access to higher education, now combined with increasing budgetary constraint, forces institutions to concentrate on issues of "*quality*" and "*accountability*."

Haskell's contribution lies in providing a kind of qualitative comparative survey of the ratings literature. He also recognizes that improper use of student ratings can result, and has resulted, in litigation over abuse of process in renewal, salary, and tenure

decisions. He has attempted to study the possible remedy of use of the issue of violation of "academic freedom" in such litigation where litigants have attempted to identify academic freedom with freedom of speech, which enjoys unqualified protection under the American Constitution.

Haskell points out the conspicuous disregard of faculty rights throughout the period in which reliance on student ratings of faculty has been associated with student and minority rights causes: "A recent booklet on '*The Law of Teacher Evaluation*' (Zirkel, 1996) contains no mention of SEF cases. Nor does a recent comprehensive legal guide for educational administrators (Kaplin and Lee, 1995), nor do other reports (Poch, 1993) on the legalities of academic freedom, tenure and promotion" (Haskell, 1997b, p. 2).

Haskell's insight into the value of considering how the courts have reacted to cases based on student ratings could have led to a more significant contribution if his results had been more systematic and analytical. The second article, particularly, would have benefited from closer collaboration with a person trained in handling this kind of material. The colossal labor represented by this vast qualitative review of the literature of the field, notwithstanding, the value of the author's discussion of judicial opinion, is practically limited to the enumeration of 78 cases where the issue of over reliance on, or neglect of, student ratings has been raised. Some of the cases are properly cited, others are not. High level court reports are listed side by side with low level. There is no attempt to distinguish between where reference to ratings would support the faculty member's case but are ignored, and cases where negative results are relied on to make decisions that should have been supported by professional opinion. There is little analysis of whether arguments for use of ratings on either side were well-taken.

There is, furthermore, no distinction made between *decisions* based upon use of ratings, and mere *obiter dicta*, or comments in passing mentioning ratings. Nevertheless, from Haskell's investigation of this problem we can begin to recognize that the concept of "academic freedom" does not seem to have been developed very far by the American courts themselves as a First Amendment (i.e., freedom of speech) category in connection with student ratings. (Note 8) On the other hand, there appear to be a number of efforts to combine complaints supported by reliance on student ratings with a theory of discrimination on the basis of sex or race—which is statutorily based and has a more consistent jurisprudence. Courts have developed measures such as "disparate impact" of policies on protected groups to support claims of illegal discrimination.

Haskell makes the valid point that whereas some lower courts have, in the past, distinguished between "freedom of speech," that was protected, and "action" in connection with expression of opinion, that was not protected (notably in *Lovelace v. S.E. Mass. Univ.*, 793F.2d 419 [1st Cir.1986]), the U.S. Supreme Court has overtaken them (Haskell, 1997d, p. 5). In 1989, the U.S. Supreme Court ruled that flag burning could be seen as political expression, and would, in that sense, be protected under the First Amendment (*Texas v. Johnson*, 491 U.S. 397 [1989]; see also: *United States v. Eichman*, 496 U.S. 310 [1990]).

On the other hand, there appears to be no American case law expressly protecting what the Germans call "*Lehrfreiheit*," i.e., freedom to teach with respect to methodology, coverage or organization of material, and grading. Indeed the cases cited suggest that some courts would allow interference in this area on the basis of institutional or public policy.

A teacher's right to say, or teach, what he or she believed to be professionally defensible would be protected. Of course, the requirement that a faculty member's expression of opinion be professionally defensible is clearly a limitation that would not

apply to others—students, for example, or student ratings. Students, and other interested members of the public, can say whatever comes into their heads—providing that it is not outright defamation.

Perhaps because of lack of a sufficient number of appeals one does not learn whether any of these cases has led to a rule adopted either in the American state or federal courts. However, we do learn that numerous judicial reservations can be cited against relying on student ratings alone—to the exclusion of professional opinion—in faculty personnel decisions (Haskell, 1997b, *passim*). Impressively, the Canadian examples cited seem to stress the need for balance between student ratings and professional assessment more than the American cases.

At the same time, we see the courts' hesitation to interject themselves into institutional decision-making. Haskell quite accurately characterizes the courts' unwillingness (unlike juries) to inquire into substantive criteria an institution applies for personnel evaluation as long as the procedural safeguards appear adequate—i.e., that the standard is applied generally to all faculty members (Haskell, 1997c, p. 4)—even though such criteria may appear to be incompetent when applied for the purpose. That was the case for a schoolteacher previously renewed over a 10 year period but terminated because her pupils ranked too low on the Iowa Test of Basic Skills (ITBS) and Iowa Test of Educational Development (ITBD). If measuring teaching effectiveness of the teacher on the basis of the performance of her pupils in standardized testing could be shown to be totally absurd or incompetent, the teacher might have been successful in thwarting dismissal. On the other hand, if a political decision, or public policy, calls for such a measure of teaching effectiveness, courts tend to leave judgment to the political arm, public policy, or simply institutional practice.

Yet, we must take care in characterizing judicial perspective. For, whereas course content and grading standards *may* be treated as a matter of institutional policy (Haskell, 1997d, p. 7), we also hear: "assignment of a letter grade is protected speech" (Haskell, 1997d., p. 6):

> [B]ecause the assignment of a letter grade is symbolic communication intended to send a specific message to the student, the individual professor's communicative act is entitled to some measure of First Amendment protection. (*Parate v. Isibor*, 868 F.2d 821, at 828 [6th Cir. 1989])(Note 9)

More disturbing is an allegation of professional incompetence in use of ratings by institutions which should know better, such as:

> According to Thompson (1988, p. 217), "Bayes Theorem shows that anything close to an accurate interpretation of the results of imperfect predictors is very elusive at the intuitive level. Indeed, empirical studies have shown that persons unfamilliar with conditional probability are quite poor at doing so (that is interpreting ratings results) unless the situation is quite simple." It seems likely that the combination of less than perfect data with less than perfect users could quickly yield completely unacceptable practices, unless safeguards were in place to insure that users knew how to recognize problems of validity and reliability, understood the inherent limitations of ratings data and knew valid procedures for using ratings data in the context of summative and formative evaluation (Franklin & Theall, 1990, pp. 79f.) (Haskell, 1997c, p. 6).

It asks a great deal of a court to assess an argument of this kind. Yet, there appears

to be accumulating evidence that educational institutions, which are capable of evaluating psychometric standards, choose to ignore such weaknesses in favor of the efficiency of the continued unquestioned reliance on student polling results. All-in-all, we see diversity of judicial opinion may be comparable to the diversity of opinion in the psychometric survey discipline. Yet, what does appear from these citations is that while courts have not equated freedom of speech with academic freedom in all its manifestations, nor created a protected zone around assessment of teaching effectiveness, they have, from time to time, expressed clear reservations about reliance on student ratings in personnel decisions to the exclusion of everything else.

III. B. Should Forward Looking Education Policy Concentrate on Goals and Incentives to Improve Teaching?

The two authors of the study of the report on the shift to peer review of teaching at the Simon School of Business at Rochester tell us that there was a very rapid adjustment to changes in incentives—that was reflected by a corresponding rapid rise in student teaching evaluations:

> During the 1990s, there was a substantial environmental shift that increased the importance of teaching relative to academic research at top business schools. The Simon School, like other business schools, changed its performance evaluation and reward systems to increase the emphasis on teaching. One might have expected the effects of these changes to be gradual, given the human capital constraints implied by the composition of existing faculty.
>
> Our results, however, suggest a very rapid adjustment to the changes in incentives. Average teaching ratings increased from about 3.8 to over 4.0 (scale of 5) almost immediately. Teaching ratings continue to rise after the changes in incentives, suggesting additional learning and turnover effects (Brickley; Zimmerman, 1997, p. 21).

They believe this dramatic effect was owed to incentives rather than peer review. Whereas they had found that: "Some evidence suggests that research output fell" (Brickley; Zimmerman, 1992, abstr.) they continue that, thereafter: ". . .we find some evidence that faculty substituted research for teaching following the incentive changes" (Brickely; Zimmerman, 1997, abstr.).

On the other hand, these authors find that, in the long run, peer review may support "quality"— the declared objective of efforts in Hong Kong associated with the TLQPR, and with the City University QAC. But they are forced to recognize an inherent conflict of interest when it comes to recognition of these efforts in student ratings:

> . . . Intense peer review of classes had no obvious effect on either teaching ratings for the evaluated classes or subsequent classes.
>
> One possible *reason peer review is not associated with higher student evaluations in the reviewed or subsequent courses might be due to the complementary nature of performance evaluation and compensation* [citing: Milgrom; Roberts, 1995]. The Deans' office did not formally announce that CTE reviews would explicitly enter the compensation policy of the School. An alternative explanation of the lack of statistical association is that "good" teaching as perceived by faculty evaluators and by students are

orthogonal. For example, *faculty evaluations value courses with more intellectual rigor and greater work loads, whereas students value courses with more current business content, more entertaining lectures, and lower work loads*. (Brickley; Zimmerman, 1997, p. 22, emphasis added).

The turnaround process is described for us in terms of agency theory by the two faculty members of the Simon School:

> Agency theory suggests that the principal is interested in both the amount of effort exerted by the agent, as well as the agent's allocation of effort across tasks. As environments change, firms are expected to adjust incentive contracts on both dimensions. For example, the 1990s witnessed significant developments in information technology, which lowered the costs of measuring performance. These changes potentially help to explain why many firms increased their use of incentive compensation over this period. Similarly, changes in competition and technology motivated numerous firms to *increase their focus on quality over quantity***,** for example, through the adoption of TQM programs (Brickley; Zimmerman, 1997, p. 22, emphasis added).(Note 10)

Changing incentives and "focus on quality over quantity" to concentrate more on teaching and learning—particularly in an environment which esteems research and/or technological development higher—is, perhaps, just as likely to involve more than merely issuing letters of congratulation to those who score high on student ratings polls.

IV. Open Decisions Openly Arrived At

Teachers may be stung by what students say if they ask for their students' opinions and find that they are significantly out of keeping with their own expectations. Of course, students have a right to their own opinions. But teachers would be foolish to let themselves become ruled by *everything* students have to say—especially on those occasions when what they have to say derives from wholly different concepts of educational goals and/or is based on teaching practices contrary to wise learning patterns. They are students, and students test what they are thinking by saying it aloud.

If there are legitimate differences about teaching and learning, they must be addressed by the institution as well as the individual instructor. On the other hand, if low "student evaluation" figures reflect that an instructor comes into a class drunk, or is on drugs, perhaps does not come at all, or does not prepare, or preys upon those in his or her charge, then that instructor ought to be fired—you do not put his or her name up on the world wide web!.

But it is not students who post their opinions on the web. It is a university administration, which does this in place of deeper thought or due diligence. If a student calls me a fool, it may be an inept way to open a conversation—about what fools are. If a university administrator calls me a fool—he robs me of my right to teach.

Is there an inherent problem in recognizing a qualitative measurement for rating of teaching? For putting teaching evaluation "on the same footing as evaluation of research"? Isn't that what Universities do? In the 1996 Research Assessment Exercise (RAE) in Hong Kong, we are told, the research "output" of all research academics in the territory's then seven traditional "tertiary" institutions—covering 14,000 publications of

3,300 academic personnel—was assessed by 110 experts, many chosen worldwide, and all in less than nine months. If there is a way of obtaining assent of universities to standards for a monumental task of that kind, there must surely be an acceptable means of, at least, setting the standards for a professional teaching and learning quality review.

There is a reason, however, why the *CityU Policy and Guide for Developing Teaching Evaluation Schemes* takes such a judicious stand on the collecting of concrete evidence for teaching evaluation—this is a step that cannot be undone. And there is a reason why it calls for "expertise," "integrity," and "training," and applies the "quality" standards to the administration as well as the faculty. Too often these decisions are made behind closed doors not simply to protect confidentiality, but because ill-defined standards applied in secret leave no trace.

There may be a right of appeal. But no appeal ever corrected injustice that should not have been done in the first place. If we know the standards of "quality," and they are as clear as, for example, those in the *CityU Policy and Guide*, or those pursued by the Committee on Teaching Excellence at the Simon School, then let the sun shine in.

Notes

1. These concerns are well illustrated and documented by Clark. He considers the difficulties facing universities around the world from loss of funding for research and emphasis on mass education. He describes the situation in universities in the United States, Britain, France, Germany and Japan, also as they form a model for their areas of cultural influence.
2. The UGC is an advisory committee appointed by the Chief Executive of the Hong Kong Special Administrative Region (SAR). Although the UGC has neither statutory nor executive powers, it administers public funds to the eight leading institutions of higher education in Hong Kong through its Secretariat, which is "staffed by civil servants."
3. The ideals of "academic freedom" derive from many sources: They were formalized as a pre-requisite of the research and teaching functions of the modern university by Wilhelm von Humboldt in the establishment of the University of Berlin in 1810. These ideals of "Lernfreiheit," the "freedom of inquiry, or advanced study," and "Lehrfreiheit," "the freedom to teach what one perceives to be the principles of one's special field," became institutional ideals not only of the German universities (until 1933, and again in the Federal Republic), but also, in a way, of the American graduate schools created on the German model. Intellectually, they derive from the same background of the European philosophers of the Renaissance and the Enlightenment that led to the creation of political institutions in the United States of America. (Cf. Flexner, 1967).
4. Importance of Educational Technology: All technology has to recommend itself to users to be adopted. There have been enormous changes in business and the professions, including education, as the result of improvements in technology in the last generation. Angela Castro of the Social Sciences Research Centre, of the University of Hong Kong writes on adoption of new technology: I do not believe professional development can be externally imposed on an individual, it must come from a personal prioritising of needs and values. If that passionate conviction is there, then the individual will seek ways to improve him/herself. (Castro, 1996) Even the authors of the "TLQPR Review" cannot resist referring to the fear of "Educational development units" being "cast in the role of ‘teach police’ " (TLQPR Review, 1996, p. 8).

5. Of course there are some who believe that, even in education, "the customer is always right." See: "Consumerism" in Appendix.
6. Other variables: sex of the student, sex of the instructor, personality of the student, and mood of the student, have also been studied in this context. More will be said about "personality" and "mood" of the student as they appear in Hong Kong student culture below.
7. Elements of Better Teaching Defined: e.g., breadth and depth of subject matter covered, development of understanding by students, amount and quality of such understanding retained, development of case material and textbooks, etc., and cooperation and collegiality between teachers and teachers and students.
8. The authors of the Basic Law (i.e., the mini-Constitution) of Hong Kong, had the foresight to include reference to the concept of "academic freedom," which "institutions" may retain and enjoy:

   > Art. 137: Educational institutions of all kinds may retain their autonomy and enjoy academic freedom. They may continue to recruit staff and use teaching materials from outside the Hong Kong Special Administrative Region. Schools run by religious organizations may continue to provide religious education, including courses in religion. Students shall enjoy freedom of choice of educational institutions and freedom to pursue their education outside the Hong Kong Special Administrative Region.

   As is apparent, however, even with statutory protection of a specific right, it can not be foreseen how a court might interpret that right—or indeed whether a court might limit that right to what is immediately ascertainable within the four corners of Art. 137 itself.
9. With respect, this decision should not be written in stone either. On the one hand, what a faculty member ought to be able to bring to an institution is professional perspective on course design and grading standards. Yet, whereas a professional person should certainly enjoy a right to expression of professional opinion with respect to a grade, he or she cannot be said to have a right to create or destroy a career with that opinion. Even judicial decisions are subject to appeal.
10. On the application of agency theory, they refer to: Holmstron. B., and Milgrom, P. (1991); and Feltham, G., and Xie, J. (1994). For focus on quality over quantity, see also: Wruck, K., and Jensen, M. (1994); and Brickley, J., Smith, C.; Zimmerman, J.(1997).

References

Abrami, P.C.; Cohen, P.A.; d'Apollonia, S**.** (1988). Implementation Problems in Meta-Analysis, *Review of Educational Research* , 58: 151-79.

Aleamoni, L.M. and Graham, M.H. (1978). The Relationship between CEQ Ratings and Instructor's Rank, Class Size, and Course Level, *Journal of Educational Measurement* 34: 189-202.

Archibold, R.C. (1998). Payback Time: Give Me an 'A' or Else, *The New York Times*, Week in Review, May 24, 1998, and http:// www. nytimes. com/ library/ review/ 052498 students-evaluate-review.html, a review of the situation on American university

campuses

Arubayi, Eric A. (1987). Improvement of Instruction and Teacher Effectiveness: Are Student Ratings Reliable and Valid? *Higher Education* 16: 267-78, at 270.

Arubayi, Eric (1985). Subject Disciplines and Student Evaluation of a Degree Programme in Education, *Higher Education* , 114: 683-91.

Avi-Itzhak, T. (1982). Teaching Effectiveness as Measured by Student Ratings and Instructor Self-Evaluation, *Higher Education* 11: 629-37.

Barnoski, R.P. and Sockloff, A.L. (1976). A Validation Study of the Faculty and Course Evaluation (FACE) Instrument, *Educational and Psychological Measurement* 36: 391-400.

Brickley, James A. and Zimmerman, Jerold L. (1997). Changing Incentives in a Multitask Environment: The Case of Teaching at a Top-25 Business School. Working Paper, William E. Simon Graduate School of Business, University of Rochester.

Brickley, J.; Smith, C.; Zimmerman, J. (1997). *Managerial Economics and Organizational Architecture* , Irwin.

Byrne, J. and Leonhardt, D. (1996). The Best B-Schools, *Business Week* (October 21, 1996).

Castro, Angela (1996). Professional Development in IT for Tertiary Academics. Available: http://www.ugc.edu.hk/UGCweb/inte.

Clark, Burton R. (1995). *Places of Inquiry: Research and Advanced Education in Modern Universities,* Berkeley: University of California Press.

Corcos, R. Churchill, D. and Lam, A., (1998). Enhancing the Participation of Law Students in Academic Tutorials, in Kember, D.; Lam, D.-H.; Yan, L.; Yum, J.C.-K.; Liu, S.B., *Case Studies of Improving Teaching and Learning from the Action Learning Project,* Hong Kong: Hong Kong Polytechnic University, p. 358.

Crittenden, K.S.; Norr, J.L.; Lebailly, R.K. (1975). Size of University Classes and Student Evaluations of Teaching, *Journal of Higher Education* 46: 461-70.

Danielson, A.L. and White, R.A. (1976). Some Evidence on the Variables Associated with Student Evaluations of Teaching, *Journal of Higher Education* 7: 117-19.

Downie, N.W. (1952). Student Evaluation of Faculty, *Journal of Higher Education* 23: 495-96, 503.

Doyle, K and Whitely, S. (1974). Student Ratings as Criteria of Effective Teaching, *American Educational Research Journal* 11: 259-74.

Feltham. G., and Xie, J. (1994). Performance, Measure, Congruity, and Diversity in Multitask Principal/Agent Relations, *Accounting Review* , 69: 429-53.

Franklin, J, and Theall, M. (1990). Communicating Student Ratings to Decision Makers: Design for Good Practice, in Theall, M. and Franklin, J., eds., *Student Ratings of Instruction: Issues for Improving Practice*, San Francisco: Jossey-Bass.

Flexner, Abraham (1967). *Universities: American, English, German*, with an introduction by Robert Ulich (New York: Teachers College Press, 1967).

Gage, N. L. (1974). Students' Ratings of College Teaching: Their Justification and Proper Use, in N.S. Glasman and B.R. Killait, eds., *Second UCSB Conference on Effective Teaching*, University of California at Santa Barbara, pp. 72-86.

Gage, N.L. (1961). The Appraisal of College Teaching: An Analysis of Ends and Means, *Journal of Higher Education* 32: 17-22.

Gayles, A.R. (1980). Student Evaluation in a Teacher Education Course, *Improving College and University Teaching* , 28: 128-31.

Gillmore, G.M. (1973). Estimates of Reliability Coefficients for Items and Subscales of the Illinois Course Evaluation Questionnaire, *Research Report, No. 341,* (Urbana, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois).

Gramlich, E. and Greenlee, G. (1993). Measuring Teaching Performance, *Journal of Economic Education ,* Winter: 3-13 (based on a study of over 15,000 economics students who study the same subject matter, are graded in a common examination, and have instructors who each receive a similar form of student evaluation).

Greenwald, A.G. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction, *American Psychologist,* 52:11: 1182-86, at p. 1184.

Greenwald, A.G. and Gillmore, G.M. (1997a). Grading Leniency is a Removable Contaminant of Student Ratings, *American Psychologist* 52:11: 1209-17.

Greenwald, A.G. and Gillmore, G.M.. (1997b). No Pain, No Gain? The Importance of Measuring Course Workload in Student Ratings of Instruction, *Journal of Educational Psychology* 89:4: 743-51.

Greenwald, A.G. and Gillmore, G.M.. (1997c). Supplement to UW's December 4 Press Release, http:// weber. u. washington. edu/ ~agg/ paingain/ supplement. Html.

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

Hall, Cedric and Fizgerald, C. (1994). *Student Summative Evaluation of Teaching: Code of Practice* , Wellington, N.Z.: Association of University Staff of New Zealand.

Harris, E.L. (1982). Student Ratings of Faculty Performance: Should Departmental Committees Construct the Instruments, *Journal of Educational Research* 76: 100-106.

Harvard Crimson (1998). *Confi-Guide* 1/25/98, at http://www.thecrimson.harvard.edu/cgi-bin/ . . . .

Haskell, R.E. (1997a,b,c,d). Academic Freedom, Tenure and Student Evaluation of Faculty, *Education Policy Analysis Archives* 5: 6,17,18, 21; a: Galloping Polls in the 21st Century; b: Views from the Court; c: Accuracy and Psychometric Validity; d: Analysis and Implications of Views from the Courts in Relation to Academic Freedom, Standards, and Quality Instruction. Available: http://epaa.asu.edu/epaa/v5n6.html and v5n17.html, v5n18.html, v5n21.html.

Hillery, J.M. and Yuk, G.A. (1974). Convergent and Discriminant Validation of Student Ratings of College Instructors, *JSAS Catalog of Selected Documents in Psychology* 4: 26.

Hogan, T.P. (1973). Similarity of Student Ratings across Instructors, Course, and Time, *Research in Higher Education* 1: 149-54.

Holmstron, B., and Milgrom, P. (1991). Multitask Principal- Agent Analysis: Incentive Contracts, Asset Ownership and Job Design, *Journal of Law, Economics & Organization* (special issue) 7: 24-52.

Hong Kong University of Science & Technology (HKUST) (1998). Progress Report to the University Grants Committee (2 March, 1998). Available http://www.ust.hk/%7Ewebaa /TLQPR/account.htm, p.2.

Hong Kong University of Science & Technology (HKUST) (1998). Course Evaluations, http://www.ust.hk/~webaa/courseeval/#SOURCE.

Hong Kong University of Science & Technology (HKUST) (1997). *Faculty Handbook, 1997*.

Imrie, Bradford W. (1993). Professional Development as Quality Assurance, a summary of official thinking of the last 30 years leading up to the TLQPR. Available: http://www.ugc.edu.hk/documents/tlqpr/INQAAHE.html.

Joint University Programmes Admissions System (JUPAS) (1997). 1997 Admissions Grades Achieved by the 'Median' & 'Lower Quartile' Applicants in Programmes Offered by the 7 Institutions.

Kaplin, W.A. and Lee, B. (1995). *The Law of Higher Education: A Comprehensive Guide to Legal Implications of Administrative Decision Making*, 3rd ed., San Francisco: Jossey Bass.

Kennedy, R.W. (1975). Grades Expected and Grades Received—their Relationship to Students' Evaluations of Faculty Performance, *Journal of Educational Psychology* 57: 109-15.

Kohlan, R.G. (1973). A Comparison of Faculty Evaluations Early and Late in the Course, *Journal of Higher Education* 44: 587-95.

Lee, O. (1998). 'Accountability' and 'Quality' in the Research Assessment Policy of the University Grants Committee (Hong Kong), which appeared as: Scrutineers perpetuate reign of error, in *The Times Higher Education Supplement*, August 14, 1998, Commonwealth Section, p. xiii.

Lee, O. (1997). *Hong Kong Business Law in a Nutshell*, New York: Juris Publishing, pp. xiii ff.

Lee, O. with the multimedia assistance of She, James (1999). 'I Want To See—Not To Be Seen!' Teaching 'Moot Court' Debating Skills through Interactive Multimedia, in James, Jeff, ed., *Quality in Teaching & Learning in Higher Education,* Hong Kong: University Grants Committee, 2000.

Marsh, H.W. (1983). Students as Evaluators of Teaching, *International Encyclopaedia of Education: Research and Studies* (New York: Pergamon Press).

Marsh, H.W. and Overall, J.U. (1981). The Relative Influence of Course Level, Course Type, and Instructor, on Students' Evaluations of College Teaching, *American Educational Research Journal* 18: 103-11.

Marsh, H.W. and Roche, L.A. (1997). Making Students' Evaluations of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility, *American Psychologist*, 52: 1187-97.

Massy, William F. and French, Nigel J. (1993). Teaching and Learning Quality Process Review: A Review of the Hong Kong Programme (hereafter, TLQPR Review), cited from the UGC internet home page, http://www.ugc.edu.hk/documents/papers/wfm_njf5.html.

McKeachie, Wilbert J. (1997a). Student Ratings: The Validity of Use, *American Psychologist* 52: 1218-25.

McKeachie, W.J. (1997b). Student Ratings of Faculty: A Reprise, *Academe*, 65: 384-97.

Milgrom, P. and Roberts, J., (1995). Complementarities and Fit: Strategy, Structure, and Organizational Change in Manufacturing, *Journal of Accounting and Economics* , 19:179-208.

Murray, H.G. (1980). *Evaluating University Teaching: A Review of Research ,* (Toronto: Confederation of University Faculty Associations).

Nichols, A, and Soper, J.C. (1972). Economic Man in the Classroom, *Journal of Political Economy,* 80: 1069- 73.

Payne, D.A. and Hobbs, A.M. (1979). The Effect of College Course Evaluation Feedback on Instructor and Student Perceptions of Instructional Climate and Effectiveness, *Higher Education* 8: 525-33.

Perry, R.R. and Baumann, R.R. (1973). Criteria for Evaluation of College Teaching: Their Reliability and Validity at the University of Toledo, in A.I Sockloff, ed.,

Proceedings: Faculty Effectiveness as Evaluated by Students (Philadelphia: Temple University).

Poch, R.K. (1993). Academic Freedom in American Higher Education: Rights, Responsibilities, and Limitations, *ASHE-ERIC Higher Education Report No. 4.*

Pohlmann, J.T. (1975). A Description of Teaching Effectiveness as Measured by Student Ratings, *Journal of Educational Measurement* 12: 49-54.

Quality Assurance Committee (QAC), City University of Hong Kong, (1996). CityU Policy and Guide for Developing Teaching Evaluation Schemes, http:// www. cityu. edu.hk/ QAC/ scheme. htm.

Quality Assurance Committee (QAC), City University of Hong Kong (1996). Statement of Principles on Quality and Quality Assurance. Available http://www.edu.hk/QAC/aboutQAC.htm

Riggs, R.O. (1975). The Prevalence and Purposes of Student and Subordinate Evaluations among AACTE Member Institutions, *Journal of Teacher Education* , 26: 218-21.

Rosenshine, B.; Cohen, A.; Furst, N. (1974). Correlates of Student Preference Ratings, *Journal of Economic Education* 4: 90-99.

Seldin, F. (1976). New Ratings Names for Professors, *The Peabody Journal of Education,* 53: 254-59.

Scott, C.A. (1977). Student Ratings and Instructor-Defined Extenuating Circumstances, *Journal of Educational Psychology* 69: 744-47.

Schwab, D.P. (1975). Course and Student Characteristic Correlates of the Course Evaluation Instrument, *Journal of Applied Psychology* 60: 742-47.

Scriven, M. (1981). Summative Teacher Evaluation, in J. Millman, ed*., Handbook of Teacher Evaluation*, Beverley Hills, CA: SAGE, pp. 244-71.

Sternberg, Robert J. (1998). Plenary Address, Practical Intelligence: Wisdom, Schooling, and Society, International Conference on the Application of Psychology to the Quality of Learning & Teaching, Hong Kong, June 13-18, 1998.

Sullivan, A. and Skanes, G. (1974). Validity of Student Evaluation of Teaching and the Characteristics of Successful Instructors, *Journal of Educational Psychology* 66: 584-90.

TLQPR of Hong Kong University of Science & Technology (1996).

University Grants Committee (UGC) (1996). Higher Education in Hong Kong—A Report by the University Grants Committee, Hong Kong, November, 1996.

University of Hong Kong (HKU), Department of Psychology, and Hong Kong University of Science & Technology (HKUST), Division of Social Science (1997). Call

for Submissions, *International Conference on the Application of Psychology to the Quality of Learning and Teaching*, Hong Kong, June 13-18, 1998.

University of Hong Kong (HKU) (1997). Response to the TLQPR Report, http://www.hku.hk/acad/hku-tlqpr/response.htm , 11/23/97,

University of Washington (1997). Press Release: Student Evaluations Don't Get a Passing Grade: Easy-Grading Professors Get Too-High Marks, new UW Study Shows, http://www. washington. edu/ newsroom/news/k120497.html .

Walker, B.D. (1968). An Investigation of Selected Variables Relative to the Manner in which a Population of Junior College Students Evaluate their Teachers, Diss., University of Houston, *Dissertation Abstracts,* 29: (1969) 3474B.

Weyrauch, W.O. (1971). The 'Basic Law' or 'Constitution' of a Small Group, *Journal of Social Issues* 27: 49, examining the emergence of community standards of behavior in a group isolated for a nutritional experiment, although many members saw themselves beyond culturally imposed rules.

Weyrauch, W.O. (1969). Governance Within Institutions, *Stanford Law Review* 22:141 (1969), reviewing Rubenstein and Lasswell, *The Sharing of Power in a Psychiatric Hospital* (1966).

Wong, C., (1984). Sociocultural Factors Counteract the Instructional Efforts of Teaching through English in Hong Kong*,* Seattle: University of Washington.

Woo Chia Wei (1997). The President's Progress Report, *HKUST Newsletter*, Fall, 1997, and email 9/10/97.

Wruck, K., and Jensen, M. (1994). "Science, Specific Knowledge, and Total Quality Management, *Journal of Accounting and Economics*, 18: 247-87.

Zirkel, P.A. (1996). *The Law of Teacher Evaluation*, Bloomington, IN: Phi Delta Kappa Educational foundation.

About the Author

Orlan Lee
Hong Kong University of Science & Technology
School of Business & Management; and
Visiting Fellow, Clare Hall, University of Cambridge

A.B. (hons.), Harvard; M.A., Yale; Ph.D., Freiburg (Germany); JurisDr., Pennsylvania; LL.M., Virginia. Dr. Lee teaches business law and cyberlaw in the School of Business and Management of the Hong Kong University of Science & Technology, and is also Visiting Fellow at Clare Hall, the college of advanced studies at the University of Cambridge. He is trained in both the civil law and common law systems, and as a social scientist, and has had extensive practical field experience. He has published widely on emergence of law and issues of public policy.

Acknowledgments

This article was prepared in part with support of a Direct Allocation Grant of the University Grants Committee of Hong Kong, and with the rare opportunity for research and reflection provided by Clare Hall, of the University of Cambridge. The author would like to thank Mr. Chan Tai Yat, and Mr. Ng Yiu Fai for their assistance in research and production of the manuscript.

Appendix
Divergent Findings

- **Those Duscussing the Conflict of Interest in Student Evaluation:**
  Gage, N. L. (1974);
  Harris, E.L. (1982).

- **Those Studying the Widespread use of Student Evaluation for Formative and Summative Purposes:**
  In the 1970s, the American Council on Education surveyed 669 American colleges and universities and found 65% using such student ratings; 35% used these for so-called "summative" purposes, i.e., for faculty hiring, tenure, termination or promotion. See: Payne, D.A. and Hobbs, A.M. (1979).
  Obviously this form of questionnaire was even more at home in schools of teacher education, where 86% of the American Association of Colleges for Teacher Education (AACTE) reported using these measures. See: Riggs, R.O. (1975).

- **Those Advocating "Consumerism" in Education:**
  Seldin, F. (1976);
  Gayles, A.R. (1980);
  Arubayi, Eric (1985).

- **Those Attributing High Rating to Impact of Prior Interest in Subject:**
  Marsh, H.W. (1980);
  Greenwald, A.G. (1997).

- **Those Believing that Ratings are Consistent for the Same Faculty Members from Year-to-Year:**
  Guthrie, E.R. (1954).

- **Those Finding that Smaller Class Size Produced Higher Ratings:**
  Danielson, A.L. and White, R.A. (1976);
  Crittenden, K.S.; Norr, J.L.; Lebailly, R.K. (1975);
  Scott, C.A. (1977);
  Perry, R.R. and Baumann, R.R. (1973);
  Avi-Itzhak, T. (1982).

- **Those Still Arguing that Class Size Has NO Effect:**
  Aleamoni, L.M. and Graham, M.H. (1978).

- **Those Finding Student Ratings Correlate with Professional and Alumni**

Evaluation:
Marsh, H.W. (1983);
Murray, H.G. (1980).

- **Those Finding that Time of Day Affects the Survey (Afternoon Ratings Lower than Morning):**
  Nichols, A, and Soper, J.C. (1972).

- **Those Finding that Lecturers are Rated Lower than Professors:**
  Downie, N.W. (1952);
  Gage, N.L. (1961);
  Walker, B.D. (1968).

- **Those Finding that Students at Lower Levels Tend to Rank Lecturers Less Favorably than Professors:**
  Downie, N.W. (1952);
  Gage, N.L. (1961);
  Pohlmann, J.T. (1975);
  Kohlan, R.G. (1973).

- **Those Finding that Students at Lower Levels Do NOT Tend to Rank Lecturers Less Favorably than Professors:**
  Hillery, J.M. and Yuk, G.A. (1974).

- **Those Finding that "Grades Expected" Affect Ratings:**
  Barnoski, R.P. and Sockloff, A.L. (1976);
  Kennedy, R.W. (1975);
  Schwab, D.P. (1975);
  Sullivan, A. and Skanes, G. (1974);
  Hillery, J.M. and Yuk, G.A. (1974);
  Perry, R.R. and Baumann.R.R. (1973);
  Rosenshine, B.; Cohen, A.; Furst, N. (1974).

- **Those Finding that "Grades Expected" Do NOT Affect Ratings:**
  Doyle, K and Whitely, S. (1974).

- **Those Finding that Ratings Are Consistent for the Same Faculty Members Regardless of Subject Matter Taught:**
  Marsh, H.W. and Overall, J.U. (1981);
  Gillmore, G.M. (1973);
  Hogan, T.P. (1973).

- **Those Finding that Teaching Ratings and Learning are Only "Weakly Related":**
  Gramlich, E. and Greenlee, G. (1993).

- **Those Who Surveyed the Literature on Validity:**
  Arubayi, Eric A. (1987);
  McKeachie, W.J. (1997b).
  Haskell, R.E. (1997a, b, c, d).

- **Current Research Returning to the Conclusion that Grades Expected and Course Workload are Dominant Factors:**
  Greenwald, A.G. (1997);
  Greenwald, A.G. and Gillmore, G.M. (1997a);
  Greenwald, A.G. and Gillmore, G.M.. (1997b);
  University of Washington (1997);
  Greenwald, A.G. and Gillmore, G.M.. (1997c);
  Archibold, R.C. (1998).

- **Those Discussing the Disparity in the Concepts of Teaching and Learning:**
  Lee, O. with She, James, (2000);
  Haskell, R.E. (1997a,b,c,d).

The World Wide Web address for the *Education Policy Analysis Archives* is **epaa.asu.edu**

General questions about appropriateness of topics or particular articles may be addressed to the Editor, Gene V Glass, glass@asu.edu or reach him at College of Education, Arizona State University, Tempe, AZ 85287-0211. (602-965-9644). The Commentary Editor is Casey D. Cobb: casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board

Daniel Schugurensky
(Argentina-Canadá)
OISE/UT, Canada
dschugurensky@oise.utoronto.ca

Jurjo Torres Santomé (Spain)
Universidad de A Coruña
jurjo@udc.es

Simon Schwartzman (Brazil)
Fundação Instituto Brasileiro e Geografia
e Estatística
simon@openlink.com.br

Carlos Alberto Torres (U.S.A.)
University of California, Los Angeles
torres@gseisucla.edu