Education Policy Analysis Archives

Volume 8 Number 50 **November 2, 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*.

Student Evaluation of Teaching: A Methodological Critique of Conventional Practices

Robert Sproule
Bishop's University (Canada)

Abstract
The purpose of the present work is twofold. The first is to outline two arguments that challenge those who would advocate a continuation of the exclusive use of raw SET data in the determination of "teaching effectiveness" in the "summative" function. The second purpose is to answer this question: "In the face of such challenges, why do university administrators continue to use these data exclusively in the determination of 'teaching effectiveness'?"

I. Introduction

The original purpose of collecting data on the student evaluation of teaching (hereafter SET) was to provide student feedback to an instructor on her "teaching effectiveness" [(Adams (1997), Blunt (1991), and Rifkin (1995)]. This function is

dubbed the "formative" function by some, and is viewed as non-controversial by most. In time, raw SET data have been put to another use—this is to provide student input into faculty committees charged with the responsibility deciding on the reappointment, pay, merit pay, tenure, and promotion of an individual instructor [Rifkin (1995), and Grant (1998)]. This second function, dubbed the "summative" function by some, is viewed as controversial by many. (Notes 1, 2)

The purpose of the present work is twofold. The first is to outline two arguments that challenge those who would advocate a continuation of the exclusive use of raw SET data in the determination of "teaching effectiveness" in the "summative" function. The first argument identifies two conceptual, and the second identifies two statistical, fallacies inherent in their methodology. Along the way, I shall also argue that while both conceptual fallacies cannot be remedied, one of the statistical fallacies can—this by means of the collection of additional data and the use of an appropriate statistical technique of the sort outlined in Mason et al. (1995). The second purpose of the present paper is to answer this question: In the face of such challenges, why do university administrators continue to use these data exclusively in the determination of "teaching effectiveness"?

The general motivation for the present work is located in three classes of statements. The first class is the many reports of the confusion and general disarray caused to the academic mission of many disciplines by the SET process. For example, Mary Beth Ruskai (1996), an associate editor of *Notices of The American Mathematical Society*, wrote:

> Administrators, faced with a glut of data, often find creative ways to reduce it (the SET process) to meaningless numbers. I encountered one who insisted that it sufficed to consider only the question on overall effectiveness, because he had once seen a report that, on average, the average on this question equaled the average of all other questions. He persisted in this policy even in cases for which it was patently false ... Advocates often cite a few superficial studies in support of the reliability of student evaluations. However, other studies give a more complex picture ... Many experienced faculty question the reliability of student evaluations as a measure of teaching effectiveness and worry that they may have counter-productive effects, such as contributing to grade inflation, discouraging innovation, and deterring instructors from challenging students.

The second concerns what constitutes admissible and inadmissible evidence in legal and quasi-legal proceedings related to the "summative" function. For example, over fifteen years ago, Gillmore (1984) wrote: "If student ratings are to qualify as evidence in support of faculty employment decisions, questions concerning their reliability and validity must be addressed" (p. 561). In recent times, it seems that the issue of admissibility has been clarified in the U.S. courts. For example, Adams (1997) wrote:

> Concerning questions about the legal basis of student evaluations of faculty, Lechtreck (1990) points out that, "In the past few decades, courts have struck down numerous tests used for hiring, and/or promotions on the grounds that the tests were discriminatory or allowed the evaluator to discriminate. The question, How would you rate the teaching ability of this instructor, is wide open to abuse" (p. 298). In his column, "Courtside,"

> Zirkel (1996) states, "Courts will not uphold evaluations that are based on subjective criteria or data" (p. 579). Administrative assumptions to the contrary, student evaluations of faculty are not objective, but rather, by their very nature, must be considered subjective. (p. 2) (Note 3)

That said, the present work should be seen as an attempt to further reinforce two views: that SET data are not methodologically sound, and that they ought not be treated as admissible evidence in any legal or quasi-legal hearing related to the "summative" function.

And the third motivation stems from *the notion of academic honesty, or from the virtue of acknowledging ignorance when the situation permits no more or no less* – a notion and a virtue the academic community claims as its own. This motivation is captured succinctly by Thomas Malthus (1836) in a statement made over a century and half ago. He wrote:

> To know what can be done, and how to do it, is beyond a doubt, the most important species of information. The next to it is, to know what cannot be done, and why we cannot do it. The first enables us to attain a positive good, to increase our powers, and augment our happiness: the second saves us from the evil of fruitless attempts, and the loss and misery occasioned by perpetual failure. (p. 14)

This article is organized as follows. In the second section, I offer a characterization of the conventional process used in the collection, and processing, of the SET data. This is done for the benefit of those unacquainted with the same. This is then followed by an outline of fallacies inherent in the conventional SET process of the conceptual sort. Similarly, in the fourth section, I outline fallacies inherent in the same of the statistical sort. The next to last section addresses this question: In the face of such challenges, why do university administrators continue to use these data exclusively in the determination of "teaching effectiveness"? Final remarks are offered in a concluding section.

II. The Conventional SET Process

The conventional process by which the SET data (on a particular instructor of a particular class) are collected and analyzed may be characterized as follows (Note 4)

1. The SET survey instrument is comprised of a series of questions about course content and teaching effectiveness. Some questions are open-ended, while others are closed-ended.
2. Those, which are closed-ended, often employ a scale to record a response. The range of possible values, or example, may run from a low of 1 for "poor," to a high of 5 for "outstanding."
3. In the closed-ended section of the SET survey instrument, one question is of central import to the "summative" function. It asks the student: "Overall, how would you rate this instructor as a teacher in this course?" In the main, this question plays a pivotal role on the evaluation process. For ease of reference, I term this question the "single-most-important question" (hereafter, the SMIQ).
4. In the open-ended section of the SET survey instrument, students are invited of

offer short critiques of the course content and of the teaching effectiveness of the instructor.

5. The completion of the SET survey instrument comes with a guarantee to students; that is, the anonymity of individual respondents.
6. The SET survey instrument is administered: (i) by a representative of the university administration to those students of a given class who are present on the data- collection day, (ii) in the latter part of the semester, and (iii) in the absence of the instructor.
7. Upon completion of the survey, the analyst then takes the response to each question on each student's questionnaire, and then constructs question-specific and class-specific measures of central tendency, and of dispersion – this in an attempt to determine if the performance of a given instructor in a particular class meets a cardinally- or ordinally- measured minimal level of "teaching effectiveness." (Note 5)
8. It seems that, in such analyses, raw SET data on the SMIQ are used in the main. More likely than not, this situation arises from the fact that the SET survey instrument does not provide for the collection of background data on the student respondent (such as major, GPA, program year, required course?, age, gender, …), and on course characteristics. (Note 6)

An example of the two-last features may prove useful. Suppose there are three professors, A, B, and C, who teach classes, X, Y, and Z, respectively. And suppose that the raw mean of the SMIQ for A in X is 4.5, the raw mean value of the SMIQ for B in Y is 3.0; and the raw mean value of the SMIQ for C in Z is 2.5. Suppose too that the reference-group raw mean score for the SMIQ is 3.5 where the reference group could be either: (i) all faculty in a given department, or (ii) all faculty in the entire university. In the evaluation process, C's mean score for the SMIQ may be compared of with that of another [say A's], and will be compared with that of her reference group. The object of this comparison is the determination of the teaching effectiveness, or ineffectiveness, of C. The questions addressed below are: (a) are the data captured by the SMIQ a valid proxy of "teaching effectiveness," and (b) can the raw mean values of the SMIQ be used in such comparisons?

III. Fallacies Of A Conceptual Sort Inherent In The SET Process

In this section, I outline two fallacies of a conceptual sort inherent in the SET process. These are: (a) that students are a, or alternatively are the only, source of reliable information on teaching effectiveness, and (b) there exists a unique and immutable metric termed "teaching effectiveness."

III.1. Students As A, Or The Only, Source Of Reliable Information on Teaching Effectiveness

Let us return to the example of the three professors, A, B, and C, who teach classes, X, Y, and Z, respectively. There are two questions to be addressed here: (a) Would one be justified in believing that students provide reliable information on teaching effectiveness? (b) If yes, would one be justified in believing that students provide the only source of reliable information on teaching effectiveness? In my view, one would not be justified in holding either belief. There are four reasons:

- The Public-Good Argument: The advocates of the SET process would argue: The university is a business, and the student its customer. And since the customer is always right, customer opinion must drive the business plan. Mainstream economists would argue that this is a false analogy. Their reason is that these same advocates are assuming that the provision of tertiary education is a "private good." This (economists would argue) is not so: It is a "public good."(Note 7) As such, students are not solely qualified to evaluate course content, and the pedagogical style of a faculty member.

- The Student-Instructor Relationship Is Not One of Customer-Purveyor, And Hence Not A Relationship Between Equals: As Stone (1995) noted,

> Higher education makes a very great mistake if it permits its primary mission to become one of serving student "customers." Treating students as customers means shaping services to their taste. It also implies that students are entitled to use or waste the services as they see fit. Thus judging by enrollment patterns, students find trivial courses of study, inflated grades, and mediocre standards quite acceptable. If this were not the case, surely there would have long ago been a tidal wave of student protest. Of course, reality is that student protest about such matters is utterly unknown. Tomorrow, when they are alumni and taxpayers, today's students will be vitally interested in academic standards and efficient use of educational opportunities. Today, however, the top priority of most students is to get through college with the highest grades and least amount of time, effort, and inconvenience.

As Michael Platt (1993) noted:

> The questions typical of student evaluations teach the student to value mediocrity in teaching and even perhaps to resent good teachers who, to keep to high purposes, will use unusual words, give difficult questions, and digress from the syllabus, or seem to. Above all, such questions also conceive the relation of student and teacher as a contract between equals instead of a covenant between unequals. Thus, they incline the student, when he learns little, to blame the teacher rather than himself. No one can learn for another person; all learning is one's own .... (p. 31)

While the student-instructor relationship is not one of customer-purveyor, and hence not a relationship between equals, the SET process itself offers the illusion that it is. As Platt (1993) noted:

> Merely by allowing the forms, the teacher loses half or more of the authority to teach. (p. 32)

- Students Are Not Sufficiently Well-Informed To Pronounce On The Success Or Failure of the Academic Mission: Because of age and therefore relative ignorance,

students are not sufficiently well-informed about societal needs for educated persons, and employers' needs for skill sets. Therefore, students are not in a position to speak for all vested interests (including their own long- term interests). For example, Michael Platt (1993) noted:

> Pascal says: while a lame man knows he limps, a lame mind does not know it limps, indeed says it is we who limp. Yet these forms invite the limpers to judge the runners; non-readers, the readers; the inarticulate, the articulate; and non-writers, the writers. Naturally, this does not encourage the former to become the latter. In truth, the very asking of such questions teaches students things that do not make them better students. It suggests that mediocre questions are the important questions, that the student already knows what teaching and learning are, and that any student is qualified to judge them. This is flattery. Sincere or insincere, it is not true, and will not improve the student, who needs to know exactly where he or she stands in order to take a single step forward. (p. 32)

In the same vein, Adams (1997) noted,

> Teaching, as with art, remains largely a matter of individual judgment. Concerning teaching quality, whose judgment counts? In the case of student judgments, the critical question, of course, is whether students are equipped to judge teaching quality. Are students in their first or second semester of college competent to grade their instructors, especially when college teaching is so different from high school? Are students who are doing poorly in their courses able to objectively judge their instructors? And are students, who are almost universally considered as lacking in critical thinking skills, often by the administrators who rely on student evaluations of faculty, able to critically evaluate their instructors? There is substantial evidence that they are not. (p. 31)

- The Anonymity of The Respondent: As noted above, the SET process provides that the identity of the respondent to the SET questionnaire would or could never be disclosed publicly. This fact contains a latent message to students. This is, in the SET process,

> there are not personal consequences for a negligent, false, or even malicious representation. There is no "student responsibility" in student evaluations. It is as if the student was being assured: "We trust you. We do not ask for evidence, or reasons, or authority. We do not ask about your experience or your character. We do not ask your name. We just trust you. Your opinions are your opinions. You are who you are. In you we trust." Most human beings trust very few other human beings that much. The wise do not trust themselves that much. [Platt (1993, p. 34)]

III.2. Opinion Misrepresented As Fact Or Knowledge

A major conceptual problem with the SET process is that opinion is misrepresented as fact or knowledge, not to mention the unintended harm that this causes to all parties. As Michael Platt (1993) noted:

> I cannot think that the habit of evaluating one's teacher can encourage a young person to long for the truth, to aspire to achievement, to emulate heroes, to become just, or to do good. To have one's opinions trusted utterly, to deliver them anonymously, to have no check on their truth, and no responsibility for their effect on the lives of others are not good for a young person's moral character. To have one's opinions taken as knowledge, accepted without question, inquiry, or conversation is not an experience that encourages self-knowledge. (pp. 33-34)

He continued:

> What they teach is that "Opinion is knowledge." Fortunately, the student may be taught elsewhere in college that opinion is not knowledge. The student of chemistry will be taught that the periodic table is a simple, intelligible account of largely invisible elements that wonderfully explains an enormous variety of visible but heterogeneous features of nature. (p. 32)

This misrepresentation of opinion as fact or knowledge raises problems in statistical analysis of the SET data in that any operational measure of "teaching effectiveness" will not be, by definition, a unique and immutable metric. [This is one of the concerns raised in the next section.] In fact, I claim that the metric itself does not exist, or the presumption that it does is pure and unsubstantiated fiction. The assessment of these claims is the next concern.

To initiate discussion, return to the example of the three professors, A, B, and C, who teach classes, X, Y, and Z, respectively. From data extracted from the SMIQ, recall that A in X scored 4.5, B in Y scored 3.0; and C in Z scored 2.5. Two premises of the conventional SET process are: (i) there exists a unique and an immutable metric, "teaching effectiveness," and (ii) the operational measure of this metric can be gleaned from data captured by the SMIQ, or by a latent-variable analysis (most commonly, factor analysis) of a number of related questions. The question to be addressed here is: Would one be justified in believing that these two premises are true?

In my view, neither premise is credible. The first premise is not true because to assume otherwise is to contradict both the research literature, and casual inspection. There are three inter-related aspects to this claim:

> 1. The first premise contains the uninspected supposition that through introspection, any student can "know" an unobservable metric called "teaching effectiveness," and can then be relied upon to accurately report her measurement of it in the SET document. (Note 8)
>
> 2. The literature makes quite clear that within any group of students one can find multiple perceptions of what constitutes "teaching effectiveness" (e.g., Fox (1983)). (Note 9)
>
> 3. If a measure is unobservable, its metric cannot be claimed to be also unambiguously unique and immutable. (Note 10) To argue otherwise is to be confronted by a bind: A measure cannot be subjective, and its metric

objective.

That said, what could account for the subjective nature of the term, "teaching effectiveness"? One explanation arises from the existence of two distinct motivations for attending university, or alternatively for enrolling in a given program. The details are these:

> 1. One motivation is the "education-as-an-investment-good" view. This is tantamount to the view that "going to university" will enhance one's prospects of obtaining a high-paying and/or an intellectually-satisfying job upon graduation. Latent in this view is the fact or belief that many employers take education as a signal of the productive capability of a university graduate as a job applicant [Spence (1974), and Molho (1997, part 2)]. (Note 11)
>
> 2. The other motivation is the "education-as-a-consumption-good" view. This view is tantamount to some mix of these five views: (a) that education is to be pursued for education's sake, (b) that "going to university" must be above all else enjoyable, (c) higher education is a democracy, and (d) in this democracy, learning must be fun, and (e) to be educated, students must like their professor. (Notes 12, 13)

Thus, any student can be seen holding some linear combination of these two views. What differentiates one student from the next (at any point in time) is the weighting of this combination.

Next, consider the second premise. It states that the operational measure of the metric, "teaching effectiveness," can be gleaned from data captured by the SET data in general, and by the SMIQ in particular. In my view, one is not justified in assuming the second premise is true because the metric, "teaching effectiveness," is unobservable and subjective. (Note 14) As such, the data captured by the conventional SET process in general, and the SMIQ in particular, can at best measure "instructor popularity" or "student satisfaction" [Damron (1995)]. An example of this subjectiveness can be found in the following passage from Cornell University's (1997) *Science News*,

> Attention teachers far and wide: It may not be so much what or how you teach that will reap high student evaluations, but something as simple as an enthusiastic tone of voice and beware, administrators, if you use student ratings to judge teachers: Although student evaluations may be systematic and reliable, a Cornell university study has found that they can be totally invalid. Yet many schools use them to determine tenure, promotion, pay hikes and awards.
>
> These warnings stem from a new study in which a Cornell professor taught the identical course twice with one exception—he used a more enthusiastic tone of voice the second semester—and student ratings soared on every measure that second semester.
>
> Those second-semester students gave much higher ratings not only on how knowledgeable and tolerant the professor was and on how much they say they learned, but even on factors such as the fairness of grading policies, text quality, professor organization, course goals and professor accessibility. And although the 249 students in the second-semester course said they

> learned more than the 229 students the previous semester believed they had learned, the two groups performed no differently on exams and other assessment measures.
> "This study suggests that factors totally unrelated to actual teaching effectiveness, such as the variation in a professor's voice, can exert a sizable influence on student ratings of that same professor's knowledge, organization, grading fairness, etc.," said Wendy M. Williams, associate professor of human development at Cornell. Her colleague and co-author, Stephen J. Ceci, professor of human development at Cornell, was the teacher evaluated by the students in a course on developmental psychology that he has taught for almost 20 years.

The assertion that the data captured by the conventional SET process in general, and the SMIQ in particular, measure at best "instructor popularity" or "student satisfaction" is echoed by Altschuler (1999). He wrote:

> At times, evaluations appear to be the academic analogue to "Rate the Record" on Dick Clark's old "American Bandstand," in which teen-agers said of every new release, "Good beat, great to dance to, I'd give it a 9." Students are becoming more adjectival than analytical, more inclined to take faculty members' wardrobes and hairstyles into account when sizing them up as educators.

IV. Fallacies Of A Statistical Sort Inherent In The SET Process

In this section, I outline potential fallacies of a statistical sort inherent in the SET process. There are two: (a) under all circumstances, the SMIQ provides a cardinal measure of "teaching effectiveness" of an instructor, and (b) in the absence of statistical controls, the SMIQ provides an ordinal measure of "teaching effectiveness" of an instructor. (Notes 15,16)

IV.1. Ascribing A Cardinal Measure of Teaching Effectiveness To An Instructor Based on The SMIQ

Return to the example of the three professors, A, B, and C, who teach classes, X, Y, and Z, respectively. Recall that A in X scored 4.5, B in Y scored 3.0; C in Z scored 2.5, and the reference group scored 3.5. A premise of the SET process is that these averages are *cardinal measures* of "teaching effectiveness." The question to be addressed here is: Would one be justified in believing that this premise is true? That is, would one be justified in believing that A is 50% "more effective" than B, that B is 20% "more effective" than C, or that A is 28% "more effective" than the average? (Note 17)

In my view, one would not be justified in believing any such claim simply because of the argument outlined in the previous section; that is, a unique and an immutable metric, "teaching effectiveness," does not exist.

IV.2. The Rank Ordering Of Instructors By Teaching Effectiveness Based on The SMIQ

Return again to the example of three professors, A, B, and C, who teach classes, X, Y, and Z, respectively. An alternative premise of the conventional SET process is that

the averages of the data captured by the SMIQ serve as a basis for an *ordinal measure* of "teaching effectiveness." The question to be addressed here is: Would one be justified in believing that this premise is true? That is, would one be justified in believing that A is "more effective" than B, or that B is "more effective" than C? In my view, this belief could be seen as justifiable: (a) if the SMIQ captures an unequivocal reading of "teaching effectiveness" (see above), and (b) if the subsequent analysis controls for the many variables which confound the data captured by the SMIQ.(Note 18)

What are these confounding variables that require control? To answer this question, two studies are worthy of mention. One, in a review of the literature, Cashin (1990) reports that (in the aggregate) students do not provide SET ratings of teaching performance uniformly across academic disciplines. (Note 19)

Two, in their review of the literature, Mason et al. (1995, p. 404) note that there are three clusters of variables, which affect student perceptions of the teaching effectiveness of faculty members. These clusters are: (a) student characteristics, (b) instructor characteristics, and (c) course characteristics. (Note 20) They also note that only one of these clusters ought to be included in any reading of "teaching effectiveness." This is the cluster, "instructor characteristics." Commenting on prior research, Mason et al. (1995, p. 404) noted:

> A …virtually universal problem with previous research is that the overall rating is viewed as an effective representation of comparative professor value despite the fact that it typically includes assessments in areas that are beyond the professor's control. The professor is responsible to some extent for course content and characteristics specific to his/her teaching style, but is unable to control for student attitude, reason for being in the course, class size, or any of the rest of those factors categorized as student or course characteristics above. Consequently, faculty members should be evaluated on a comparative basis only in those areas they can affect, or more to the point, only by a methodology that corrects for those influences beyond the faculty member's control.
>
> By comparing raw student evaluations across faculty members, administrators implicitly assume that none of these potentially mitigating factors has any impact on student evaluation differentials, or that such differentials cancel out in all cases. The literature implies that the former postulate is untrue.

The true import of the above is found again in Mason et al. (1995). Using an ordered-probit model, (Note 21) they demonstrate that student characteristics, instructor characteristics, and course characteristics do impact the response to the SMIQ in the SET dataset. They wrote:

> Professor characteristics dominated the determinants of the summary measures of performance, and did so more for those summary variables that were more professor-specific. However, certain course- and student-specific characteristics were very important, skewing the rankings based on the raw results. Students consistently rewarded teachers for using class time wisely, encouraging analytical decision making, knowing when students did not understand, and being well prepared for class. However, those professors who gave at least the impression of lower grades, taught more difficult courses, proceeded at a pace students did not like, or did not

stimulate interest in the material, fared worse. (p. 414)

Mason et al. (1995) then wrote:

> Based on the probit analysis, an alternative ranking scheme was developed for faculty that excluded influences beyond the professor's control. These rankings differed to some extent from the raw rankings for each of the aggregate questions. As a result, the validity of the raw rankings of faculty members for the purposes of promotion, tenure, and raises should be questioned seriously. … Administrators should adjust aggregate measures of teaching performance to reflect only those items within the professors' control, so that aggregates are more likely to be properly comparable and should do so by controlling for types of courses, levels of courses, disciplines, meeting times, etc. … Administrators failing to do this are encouraged to reconsider the appropriateness of aggregate measures from student evaluations in promotion, tenure, and salary decisions, concentrating instead on more personal evaluations such as analysis of pedagogical tools, peer assessments, and administrative visits. (p. 414)

It may be useful to ask: To what extent are the findings of Mason et al. (1995) unique? Surprisingly, they are not; they echo those of other studies, some recent, and some more than a quarter-century old. For example, Miriam Rodin and Burton Rodin (1972) writing in Science present a study in which they correlated an objective measure of "good teaching" (viz., a student's performance on a calculus test) with a subjective measure of "good teaching" (viz., a student's evaluation of her professor) holding constant the student's initial ability in calculus. What they found is that these two measures were not orthogonal or uncorrelated as some might expect, but something more troublesome. These two variables had a correlation coefficient less than –0.70, and these two accounted for more about half of the variance in the data. How did they interpret their findings? The last sentence in their paper states: "If how much students learn is considered to be a major component of good teaching, it must be concluded that good teaching is not validly measured by student evaluations in their current form." How might others interpret their findings? They suggest the individual instructor is in a classic double-bind: If she attempts to maximize her score on the SMIQ, then she lowers student performance. Alternatively if she attempts to maximize student performance, then her score on the SMIQ suffers. This begs the question: In such a dynamic, how can one possibly use SET data to extract a meaningful measure of "teaching effectiveness?"

In a different study (one concerned with the teaching evaluations for the Department of Mathematics at Texas A&M University, and one which entails the analysis of the correlation coefficients for arrays of variables measuring "teaching effectiveness" and "course characteristics"), Rundell (1996) writes: "(T)he analysis we have performed on the data suggests that the distillation of evaluations to a single number without taking into account the many other factors can be seriously misleading" (p. 8).

V. Why Has The Conventional SET Process Not Been Discarded?

Given that the likelihood of deriving meaningful and valid inferences from raw SET data is nil, the question remains: Why is the conventional SET process (with its conceptual and statistical shortcomings) employed even to this day, and by those for

who highly revere the power of critical thinking?

To my mind, there are three answers to this question. The first answer concerns political expediency; that is, while fatally flawed, raw SET data can be used as a tautological device; that is, to justify most any personnel decision. As a professor of economics at Indiana University and the Editor of *The Journal of Economic Education* noted:

> End of term student evaluations of teaching may be widely used simply because they are inexpensive to administer, especially when done by a student in class, with paid staff involved only in the processing of the results…Less-than-scrupulous administrators and faculty committees may also use them … because they can be dismissed or finessed as needed to achieve desired personnel ends while still mollifying students and giving them a sense of involvement in personnel matters. [Becker (2000, p. 114)]

The second is offered by Donald Katzner (1991). He asserted that in their quest to describe, analyze, understand, know, and make decisions, western societies have accepted (for well over five hundred years) the "myth of synonymity between objective science and measurement" (p. 24). (Note 22) He wrote:

> [W]e moderns, it seems, attempt to measure everything…. We evaluate performance by measurement…. What is not measurable we strive to render measurable, and what we cannot, we dismiss it from our thoughts and justify our neglect by assigning it the status of the "less important." … A moment's reflection, however, is all that is needed to realize that measurement cannot possibly do everything we expect it to do. … by omitting from our considerations what cannot be measured, or what we do not know how to measure, often leads to irrelevance and even error. (p. 18)

The third reason is offered by Imre Lakatos (1978) in his explanation as to why prevailing scientific paradigms are rarely replaced or overthrown. This contains these elements:

1. What ought to be appraised in the philosophy of the sciences is not an isolated individual theory, but a cluster of interconnected theories, or what he terms "scientific research programs" (hereafter SRP).
2. An SRP protects a "hard core" set of unquestioned and untestable statements. These statements are accepted as "fact."
3. Stated differently, the hard core of a SRP is surrounded by a "protective belt" of "auxiliary hypotheses."
4. One or more of the hard core statements cannot be refuted without dismantling the entire cognitive edifice, which happens in practice only very rarely. That said, it follows that any departure from the hard core of a SRP is tantamount to the creation of a new and different SRP.

Thus, in my view, the conventional SET process is the artifact of an SRP. Judging from the substance of its protective belt, and from the disciplinary affiliations of its proponents or advocates, this is an SRP defined and protected by a cadre of psychologists and educational administrators. (Notes 23,24)

VI. Conclusion

In the present work, I have advanced two arguments, both of which question the appropriateness of using raw SET data (as the only source of data) in the determination of "teaching effectiveness." The first argument identified two types of fallacies in this methodology. One is conceptual, and the other statistical. Along the way, I argued by implication that the conceptual fallacies cannot be remedied, but that one of the statistical fallacies can – this by means of the collection of additional data and the use of an appropriate statistical technique of the sort outlined in the study of Mason et al. (1995), which I also discussed.

The second argument is centered on the question, why do the current practices used in the determination of the "teaching effectiveness" ignore these two fallacies? I offered three answers to this question. These are: (a) that the conventional SET process offers to any university administration a politically-expedient performance measure, and (b) that the conventional SET process may be seen as an example of: (i) Katzner's (1991) "myth of synonymity between objective science and measurement," and (ii) Lakatos' (1978) general explanation of the longevity of SRPs.

Two implications flow from these arguments, and the related discussion. These are as follows: One, the present discussion should not be seen as tantamount to an idle academic debate. On the contrary, since the SET data have been entered as evidence in courts of law and quasi-legal settings [Adams (1997), Gillmore (1984), and Haskell (1997d)], and since the quality and the interpretation of these data can impact the welfare of individuals, it is clear that the present paper has import and bearing to the extent that: (i) it explicates the inadequacies, and unintended implications, of using raw SET data in the "summative" function, and (ii) it explains the present resistance of the conventional SET process to radical reform.

Two, given the present assessment of the conventional SET process, and given the legal repercussions of its continued use, the question becomes: What to do? Here, the news is both good and bad. The bad news is that nothing can be done to obviate the conceptual fallacies outlined in the above pages. The inescapable truth is that the SMIQ in particular, and the SET dataset in general, do not measure "teaching effectiveness." They measure something akin to the "popularity of the instructor," which (it must be emphasized) is quite distinct from "teaching effectiveness." [Recall the discussion of Rodin and Rodin (1972) in the above.] The good news is that one of the statistical fallacies inherent in the conventional SET process can be overcome – this by capturing and then using background data on student, instructor, and course characteristics, in the mold of Mason et al. (1995). That said, I leave the last word to what (in my opinion) amounts to a classic in its own time. Mason et al. (1995) state, and I repeat:

> Administrators should adjust aggregate measures of teaching performance to reflect only those items within the professors' control, so that aggregates are more likely to be properly comparable and should do so by controlling for types of courses, levels of courses, disciplines, meeting times, etc. … Administrators failing to do this are encouraged to reconsider the appropriateness of aggregate measures from student evaluations in promotion, tenure, and salary decisions, concentrating instead on more personal evaluations such as analysis of pedagogical tools, peer assessments, and administrative visits. (p. 414)

Notes

This article was prepared during the winter semester of 2000 while the author was on a

half-year sabbatical at the University of Manitoba (Winnipeg, Canada). Without implicating them for any remaining errors and oversights, the author thanks Donald Katzner, Paul Mason, Stuart Mckelvie, and three anonymous referees, for many useful comments and critiques.

1. For reviews of the literature that are essentially supportive of the SET process, see d'Apollonia and Abrami (1997), Greenwald and Gilmore (1997), Marsh (1987), Marsh and Roche (1997), and McKeachie (1997). And for reviews of the literature that are highly critical of some mix of the conceptual, statistical, and legal foundations of the SET process, see Damron (1995), and Haskell (1997a, 1997b, 1997c, and 1997d).
2. The terms "formative" and "summative" are due to Scriven (1967).
3. On such matters, the position of the Canadian Association of University Teachers on the admissibility of SET data appears unambiguous in light of statements like these: "Appropriate professional care should be exercised in the development of questionnaires and survey methodologies. Expert advice should be sought, and reviews of the appropriate research and scientific evidence should be carried out. Comments from faculty and students and their associations or unions should be obtained at all stages in the development of the questionnaire. Appropriate trials or pilot studies should be conducted and acceptable levels of reliability and validity should be demonstrated before a particular instrument is used in making personnel decisions" [Canadian Association of University Teachers (1998, p. 3)]. In a footnote to this passage, this document continues, "Most universities require at least this standard of care before investigators are permitted to conduct research on human subjects. It is unacceptable that university administrations would condone a lesser standard in the treatment of faculty, particularly when the consequences of inadequate procedures and methods can be devastating to teachers' careers."
4. The present characterization represents an amalgam of three sources: (a) first-hand knowledge of the SET documents used at three Canadian universities; (b) a small, non-random sample of SET documents for four universities taken from the internet [viz., University of Minnesota, University of British Columbia, York University (Toronto), and University of Western Ontario]; and (c) non-institutional-specific comments made in the voluminous literature on the SET process.
5. The phrase "a cardinally- or ordinally- measured minimal level of "teaching effectiveness"" requires four comments. One, examples of cardinal measures are: The heights of persons A, B, and C are 6'1", 5'10", and 5'7" respectively. And using the same data, examples of ordinal measures are: A is taller than B, B is taller than C, and A is taller than C. Two, the present measurement terminology is used in economics [Pearce (1992)], and (it can be said) is distinct from that used in other disciplines [e.g., Stevens (1946), Siegel (1956, p. 30), and Hands (1996)]. Three, it is the existence of a unique and an immutable metric (in the above examples, distance or length) that makes both cardinal and ordinal measures meaningful. Four, as the above examples make clear, an ordinal measure can be inferred from a cardinal measure, but not the reverse.
6. An example of this statement is the instrument used by York University (Toronto). An exception to this statement is that used by the University of Minnesota.
7. The distinction between a "private good" and a "public good" can be rephrased in several, roughly equivalent ways. These are: (i) tertiary education has

externalities; (ii) that the net social benefits of tertiary education differ from the net private benefits, (iii) that the benefits of tertiary education do not accrue to, nor are its costs borne by, students solely, and (iv) that students do not pay full freight. Because of this, one could argue that (in the evaluation of "teaching effectiveness") the appropriate populations of opinion to be sampled are all groups who share in the social benefits and social costs. These would include not only students, but also members of the Academy, potential employers, and other members of society (such as taxpayers). In sum, because tertiary education is not a private, but a public good, students are not solely qualified to evaluate course content, and the pedagogical style of a faculty member.

8. A personal vignette provides some insight into the potential seriousness of the inaccuracy of self-reported data. In the fall of 1997, I taught an intermediate microeconomics course. The mark for this course was based solely on two mid-term examinations, and a final examination. Each mid-term examination was marked, and then returned to students and discussed in the class following the examination. Now, the course evaluation form has the question, "Work returned reasonably promptly." The response scale ranges from 0 for "seldom," to 5 for "always." Based on the facts, one would expect (in this situation) an average response of 5. This expectation was dashed in that 50% of the sample gave me a 5, 27.7% gave me a 4, and 22.2% gave me a 3. The import of this? If self-reported measures of objective metrics are inaccurate (as this case indicates), how can one be expected to trust the validity of subjective measurements like "teaching effectiveness?"
9. Indeed, it appears that students and professors can hold different perceptions as to what constitutes "appropriate learning," and hence "appropriate teaching," in tertiary education. For example, Steven Zucker (1996), professor of Mathematics at Johns Hopkins University, laments the gulf between the expectations of students and instructors. He writes: "The fundamental problem is that most of our current high school graduates don't know how to learn or even what it means to learn (a fortiori to understand) something. In effect, they graduate high school feeling that learning must come down to them from their teachers. That may be suitable for the goals of high school, but it is unacceptable at the university level. That the students must also learn on their own, outside the classroom, is the main feature that distinguishes college from high school." (p. 863).
10. Alternatively, Weissberg (1993, p. 8) noted that one cannot measure what one cannot define.
11. These assertions have been borne out empirically under the rubric, "sheepskin effect." The interested reader is directed to Belman and Heywood (1991 and 1997), Heywood (1994), Hungerford and Solon (1987), and Jaeger and Page (1996).
12. Some of these views contradict the *raison d'être* and the *modus operandi* of tertiary education. For example, Frankel (1968) wrote: "Teaching is a professional relationship, not a popularity contest. To invite students to participate in the selection or promotion of their teachers exposes the teacher to intimidation." (pp. 30-31) In fact, the Canadian Association of University Teachers (1986) speaks of the irrelevance of "popularity" as a gauge of professional performance by stating: "The university is not a club; it is dedicated to excellence. The history of universities suggests that its most brilliant members can sometimes be difficult, different from their colleagues, and unlikely to win a popularity contest. 'The university is a community of scholars and it is to be expected that the scholars will

hold firm views and wish to follow their convictions. Tension, personality conflicts and arguments may be inevitable by-products.'"

13. As Crumbley (1995) noted: "There is another universal assumption that students must like an instructor to learn. Not true. Even if they dislike you and you force them to learn by hard work and low grades, you may be a good educator (but not according to SET scores). SET measures whether or not students like you, and not necessarily whether you are teaching them anything. Instructors should be in the business of educating and teaching students--not SET enhancement. Until administrators learn this simple truth, there is little chance of improving higher education."
14. It seems that some psychologists would argue that latent measures of "teaching effectiveness" can be uncovered by a factor analysis of the SET data [e.g., d'Apollonia and Abrami (1997)]. Also, it seems that the motivation for such a claim is the intellectual appeal and success of studies of a completely different ilk. A case in point is Linden (1977) who uses factor analysis to uncover dimensions, which account for event-specific performances of athletes in the Olympic decathlon. However, the expectation that the success found in studies such as Linden (1977) can be replicated in the factor analysis of SET data is unwarranted in that this expectation ignores the fact that the SET data (unlike Linden's data) are opinion based or subjective, have measurement error, and are in need of statistical controls. In brief, it is my view that the use of factor analysis on SET data to uncover latent measures such as "teaching effectiveness" is analogous to trying to "unscramble an egg" in that it just cannot be done. Besides, as the authors of a popular text on multivariate statistics observe, "When all is said and done, factor analysis remains very subjective" [Johnson and Wichern (1988, p. 422)].
15. The terms, ordinal and cardinal measures, are defined in a footnote above. In conjunction with that, it should be noted that the type of a variable governs the statistical manipulations permissible [Hands (1996, pp. 460-62)], and "(T)he use of ordinally calibrated variables as if they were fully quantified .. results in constructions that are without meaning, significance, and explanatory power. Treating ordinal variables as cardinal … can mislead an investigator into thinking the analysis has shed light on the real world" [Katzner (1991, p. 3)]. This latter point captures an important dimension of the present state of research on SET data, and of the present paper.
16. For reasons of brevity, I have concentrated on only two of several statistical problems. These are "measurement error" and "omitted variables." By doing so, I have overlooked other statistical problems inherent in the SET data like the unreliability of self- and anonymous-reporting, inadequate sample size, sample-selection bias, reverse causation, and teaching to tests. The reader interested in a more complete treatment of some of these issues may wish to consult readings such as Aiger and Thum (1986), Becker and Power (2000), Gramlich and Greenlee (1993), and Nelson and Lynch (1984).
17. As Rundell (1996) noted, in actual practice, this would mean: "…'Jones had a 3.94 mean on her student evaluations, and since this is 0.2 above the average for the Department, we conclude she is an above average instructor as judged by these questionnaires' is a statement that appears increasingly common" (p. 1).
18. Statistical controls are needed to the extent that they eliminate "observational equivalence." In this connection, two comments are warranted here. One, observational equivalence is said to exist when "alternative interpretations, with different theoretical or policy implications, are equally consistent with the same

data.. No analysis of the data would allow one to decide between the explanations, they are observationally equivalent. Other information is needed to identify which is the correct explanation of the data" [Smith (1999, p. 248)]. Two, Sproule (2000) has identified three distinct forms of observational equivalence in the interpretation of raw data from the SMIQ.

19. Cashin (1990) reports, for example, professors of fine arts and music receive high scores on the SMIQ, and professors of chemistry and economics receive lower scores, all things being equal.
20. Mason et al. (1995) contend that those variables which fall under the "student-characteristics" rubric include: (i) reason for taking the course, (ii) class level of the respondent, (iii) student effort in the course, (iv) expected grade in the course, and (v) student gender. Those variables which fall under the "instructor-characteristics" rubric include: (i) the professor's use of class time, (ii) the professor's availability outside of class, (iii) how well the professor evaluates student understanding, (iv) the professor's concern for student performance, (v) the professor's emphasis on analytical skills, (vi) the professor's preparedness for class, and (vii) the professor's tolerance of opposing viewpoints and questions. Those variables which fall under the "course- characteristics" rubric include: (i) course difficulty, (ii) class size, (iii) whether the course is required or not, and (iv) when the course was offered.
21. For an elementary discussion of the ordered-probit model, see Pindyck and Rubinfeld (1991, pp. 273-274.).
22. Katzner (1991) also states that this "blind pursuit of numbers" can lead to unintended, and unjust, outcomes. For example, "(W)hen the state secretly sterilizes individuals only because their 'measured intelligence' on flawed intelligence tests is too low, then bitterly dashed hopes and human suffering becomes the issue." (p. 18). That said, it would not be too difficult to claim that the "blind pursuit of numbers" by those responsible for the "summative" function has also led to unintended, and unjust, outcomes. [In fact, see Haskell (1997d) for details.]
23. Three comments seem warranted here. One, the enterprise of science can been seen as a "market process" [Walstad (1999)]. Two, the SRP of this cadre of psychologists and educational administrators could be viewed as barrier to entry (of the epistemological sort) into the marketplace of ideas. Three, that said, perhaps the recommendation of Paul Feyerabend (1975) applies in this instance; that competition between epistemologies, rather than the monopoly of a dominant epistemology, ought to be encouraged.
24. While it is clear from the above that the protective belt of the SRP associated with the SET has survived many types of logical appraisals (or epistemological attacks), the question remains: Can this protective belt, and this SRP itself, continue to withstand such repeated attacks? I would hazard the opinion that, no, it cannot.

References

Adams, J.V. (1997), Student evaluations: The ratings game, *Inquiry 1 (2),* 10-16.

Aiger, D., and F. Thum (1986), On student evaluation of teaching ability, *Journal of Economic Education*, Fall, 243-265.

Altschuler, G. (1999), Let me edutain you, *The New York Times*, Education Life Supplement, April 4.

Becker, W. (2000), Teaching economics in the 21st century, *Journal of Economic Perspectives 14 (1),* 109-120.

Becker, W., and J. Power (2000), Student performance, attrition, and class size, given missing student data, *Economics of Education Review,* forthcoming.

Belman, D., and J.S. Heywood (1991), Sheepskin effects in the returns to education: An examination on women and minorities, *Review of Economics and Statistics 73 (4*), 720-24.

Belman, D., and J.S. Heywood (1997), Sheepskin effects by cohort: Implications of job matching in a signaling model, *Oxford Economic Papers 49 (4)*, 623-37.

Blunt, A. (1991), The effects of anonymity and manipulated grades on student ratings of instructors, *Community College Review 18*, Summer, 48-53.

Canadian Association of University Teachers (1986), What is fair? A guide for peer review committees: Tenure, renewal, promotion, Information Paper, November.

Canadian Association of University Teachers (1998), Policy on the use of anonymous student questionnaires in the evaluation of teaching, CAUT Information Service Policy Paper 4-43.

Cashin, W. (1990), Students do rate different academic fields differently, in M. Theall and J. Franklin, eds., *Student Ratings of Instruction: Issues for Improving Practice,* New Directions for Teaching and Learning, No. 43 (San Francisco, CA: Jossey-Bass).

Cornell University (1997), Cornell study finds student ratings soar on all measures when professor uses more enthusiasm: Study raises concerns about the validity of student evaluations, *Science News*, September 19th.

Crumbley, D.L. (1995), Dysfunctional effects of summative student evaluations of teaching: Games professors play, *Accounting Perspectives 1 (1)*, Spring, 67-77.

Damron, J.C. (1995). The three faces of teaching evaluation, unpublished manuscript, Douglas College, New Westminster, British Columbia.

d'Apollonia, S., and P. Abrami (1997), Navigating student ratings of instruction, *American Psychologist 52 (11)*, 1198-1208.

Feyerabend, P. (1975), *Against Method* (London: Verso).

Fox, D. (1983), Personal theories of teaching, *Studies in Higher Education 8 (2),* 151-64.

Frankel, C. (1968), *Education and the Barricades* (New York: W.W. Norton).

Gillmore, G. (1984), Student ratings as a factor in faculty employment decisions and

periodic review, *Journal of College and University Law 10*, 557- 576.

Gramlich, E., and G. Greenlee (1993), Measuring teaching performance, *Journal of Economic Education*, Winter, 3-13.

Grant, H. (1998), Academic contests: Merit pay in Canadian universities, *Relations Industrielles / Industrial Relations 53 (4)*, 647-664.

Greenwald, A., and G. Gilmore (1997), Grading leniency is a removable contaminant of student ratings, *American Psychologist 52 (11)*, 1209-17.

Hands. D.J. (1996), Statistics and the theory of measurement, *Journal of the Royal Statistical Society – Series A 159 (3)*, 445-473.

Haskell, R.E. (1997a), Academic freedom, tenure, and student evaluations of faculty: Galloping polls in the 21st century, *Education Policy Analysis Archives 5 (6),* February 12.

Haskell, R.E. (1997b), Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty (SEF): (Part II) Views from court, *Education Policy Analysis Archives 5 (6)*, August 25.

Haskell, R.E. (1997c), Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty (SEF): (Part III) Analysis and implications of views from the court in relation to accuracy and psychometric validity, *Education Policy Analysis Archives 5 (6),* August 25.

Haskell, R.E. (1997d), Academic freedom, promotion, reappointment, tenure, and the administrative use of student evaluation of faculty (SEF): (Part IV) Analysis and implications of views from the court in relation to academic freedom, standards, and quality of instruction, *Education Policy Analysis Archives 5 (6),* November 25.

Heywood, J.S. (1994), How widespread are sheepskin returns to education in the U.S.?, *Economics of Education Review 13 (3)*, 227-34.

Hungerford, T., and G. Solon (1987), Sheepskin effects in the returns to education, *Review of Economics and Statistics 69 (1)*, 175-77.

Jaeger, D., and M. Page (1996), Degrees matter: New evidence on sheepskin effects in the returns to education, *Review of Economics and Statistics 78 (4),* 733-40.

Johnson, R., and D. Wichern (1988), *Applied Multivariate Statistical Analysis*, Second Edition (Englewood Cliffs: Prentice-Hall).

Katzner, D. (1991), Our mad rush to measure: How did we get there?, *Methodus 3 (2),* 18-26.

Lakatos, I. (1978), *The Methodology of Scientific Research Programmes* (Cambridge: Cambridge University Press).

Linden, M. (1977), A factor analytic study of Olympic decathlon data, *Research Quarterly 48 (3*), 562- 568.

Malthus, T. (1836), *Principles of Political Economy*, 2nd Edition.

Marsh, H. (1987), Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research, *International Journal of Educational Research 11*, 253-388.

Marsh, H., and L. Roche (1997), Making students' evaluations of teaching effectiveness effective: The central issues of validity, bias, and utility, *American Psychologist 52 (11)*, 1187-97.

Mason, P., J. Steagall, and M. Fabritius (1995), Student evaluations of faculty: A new procedure for using aggregate measures of performance, *Economics of Education Review 12 (4),* 403-416.

McKeachie, W. (1997), Student ratings: The validity of use, *American Psychologist 52 (11)*, 1218- 1225.

Molho, I. (1997), *The Economics of Information: Lying and Cheating in Markets and Organizations* (Oxford: Blackwell).

Nelson, J., and K. Lynch (1984), Grade inflation, real income, simultaneity, and teaching evaluations, *Journal of Economic Education*, Winter, 21-37.

Pearce, D.W., ed. (1992), *The MIT Dictionary of Modern Ecoonomics*, 4th Edition (Cambridge, MA: MIT Press).

Pindyck, R. and D. Rubinfeld (1991), *Econometric Models & Economic Forecasts* (New York: McGraw-Hill).

Platt, M. (1993), What student evaluations teach, *Perspectives In Political Science 22 (1),* 29-40.

Rifkin, T. (1995), The status and scope of faculty evaluation, *ERIC Digest.*

Rodin, M., and B. Rodin (1972), Student evaluations of teaching, *Science 177*, September, 1164- 1166.

Rundell, W. (1996), On the use of numerically scored student evaluations of faculty, unpublished working paper, Department of Mathematics, Texas A&M University.

Ruskai, M.B. (1996), Evaluating student evaluations, *Notices of The American Mathematical Society 44 (3)*, March 1997, 308.

Scriven, M. (1967), The methodology of evaluation, in R. Tyler, R. Gagne, and M. Scriven, eds., *Perspectives in Curriculum Evaluation* (Skokie, IL: Rand McNally).

Siegel, S. (1956), *Nonparametric Statistics For The Behavioral Sciences* (New York:

McGraw-Hill).

Smith, R. (1999), Unit roots and all that: The impact of time-series methods on macroeconomics, *Journal of Economic Methodology 6 (2),* 239-258.

Spence, M. (1974), *Market Signaling* (Cambridge, MA: Harvard University Press).

Sproule, R. (2000). The underdetermination of instructor performance by data from the student evaluation of teaching,*Economics of Education Review* (in press).

Stevens, S.S. (1946), On the theory of scales of measurement, *Science 103*, 677-680.

Stone, J.E. (1995), Inflated grades, inflated enrollment, and inflated budgets: An analysis and call for review at the state level, *Education Policy Analysis Archives 3 (11).*

Walstad, A. (1999), Science as a market process, unpublished paper, Department of Physics, University of Pittsburgh—Johnstown.

Weissberg, R. (1993), Standardized teaching evaluations, *Perspectives In Political Science 22 (1),* 5-7.

Zucker, S. (1996), Teaching at the university level, *Notices of The American Mathematical Society 43 (8),* August, 863-865.

About the Author

Robert Sproule
Department of Economics
Williams School of Business and Economics
Bishop's University
Lennoxville, Québec, J1M 1Z7, Canada
The author can be reached via e-mail at rsproule@ubishops.ca

Robert Sproule is a Professor of Economics at Bishop's University. He received his Ph.D. in Economics at the University of Manitoba (Winnipeg, Canada). His research interests include statistics, econometrics, and decision making under uncertainty. His research appears in journals such as the *Bulletin of Economic Research, Communications in Statistics: Theory and Methods, Economics Letters,* the *Economics of Education Review*, the *European Journal of Operations Research, Metroeconomica, Public Finance,* and *The Statistician*.

---

casey.cobb@unh.edu .

EPAA Editorial Board

EPAA Spanish Language Editorial Board