

THE STUDENT EVALUATION OF TEACHING: ITS FAILURE AS A RESEARCH PROGRAM, AND AS AN ADMINISTRATIVE GUIDE

Robert Sproule¹

Călin Vâlsan²

^{1) 2)} Bishop's University, Sherbrooke, Québec, J1M 0C8, Canada

e-mail: cvalsan@ubishops.ca



Abstract

This paper points up the methodological inadequacy of the “student evaluation of teaching” as a research program. We do this by reference to three, interrelated arguments. The first is that the student evaluation of teaching cannot claim to capture the wisdom of a crowd because, as a research program, it fails to meet Surowiecki’s conditions for the existence and articulation of the wisdom of a crowd. The second argument extends the first, by stating: (a) the “student evaluation of teaching” research program fails to provide the methodological controls needed to differentiate cause from effect, or put differently (b) the methodological underpinnings of this research program is tantamount to the misapplication of a closed-system paradigm to an open social system. The third argument has two parts. These are that this research program is predicated on: (a) a false analogy between the workings of a business and a university, and therefore (b) on a mischaracterization of the student-professor relationship. These three arguments, these three failures, suggest that the “student evaluation of teaching” research program is methodologically ill-conceived and incoherent, and therefore cannot, with any credulity, serve as a guide to the administration and governance of a university.

Keywords: *Student evaluation of teaching, validity, biases, fallacies*

JEL classification: *A2, H83, I23, and L3*

Introduction

Imre Lakatos (1978a and 1978b) argued that a scientific theory can be seen to be comprised of two parts: (a) a collection of “research programs” with a “hard core”, and (b) a protective belt of auxiliary hypotheses. He also argued that the latter is a shield designed by the proponents or advocates of the theory to defend the former from falsification.

The objective of this paper is to penetrate the belt that protects the research program related to the student evaluation of teaching (SET), in the hope of silencing its proponents or advocates. In particular, this paper mounts a fresh and systematic attack on three methodological weaknesses inherent to the SET Research Program -- weaknesses that in the past its proponents or advocates have dismissed or ignored. The intent of this paper is to deny them the use of these two stratagems henceforth.

In this effort, this paper presents three, interrelated arguments. The first argument uses as its starting point a recent book called *The Wisdom of Crowds* (2004). Its author, James Surowiecki, maintained that wisdom is dispersed throughout society, often very thinly. *But where and when it can be articulated*, this wisdom is *usually* more reliable than the wisdom of elites. In the first argument of this paper, we maintain that the SET Research Program: (a) fails to meet Surowiecki’s conditions for the existence and articulation of the wisdom of a crowd, and therefore (b) the data captured in the SET are apt not to reflect the wisdom of a crowd, but its folly. And so we term this first argument, “The SET Research Program as pedagogical folly.”

In a sense, the second argument extends the first. The second states that: (a) the SET Research Program fails to provide the methodological controls needed to differentiate cause from effect, or stated differently (b) the methodological underpinnings of the SET Research Program is the misapplication of a closed-system paradigm to an open social system. We term this argument, “The SET Research Program as pseudoscience.”

The third argument advances the claim that the SET Research Program is predicated on a false analogy between the nature and substance of the typical business and university, and therefore on a mischaracterization of the true nature of the student-professor relationship. Because of this, we term our third argument, “The SET Research Program as a logical fallacy.”

The remainder of this paper is organized as follows. Section 2 presents some preliminary statements about the origins, nature, and substance of the SET research agenda. Section 3 outlines the first argument, “The SET Research Program as pedagogical folly.” Section 4 outlines the second argument, “The SET Research Program as pseudoscience.” And Section 5 outlines the final argument, “The SET Research Program as a logical fallacy.” Concluding remarks are offered in Section 6.

1. Some preliminary comments

The history of the SET Research Program spans at least the past eight decades. One of the earliest uses of the SET was by the psychologist, E.T. Guthrie, at the University of Washington in the 1920s. By the late 1960s or early 1970s, the SET was widely adopted in most colleges and universities in North America. Its widespread adoption stemmed from the fact that it was supported by a coalition of three groups: (a) students who wanted a say in teaching,¹ (b) administrators who were concerned with accountability and good public relations, and (c) young faculty who wanted their salary, promotion and tenure evaluations to depend on something other than number of publications [Murray (2005, p. 1)].

The utility of the SET data is considered by most to depend on which of two functions the data are put. These two, “formative” and the “summative”, are due to Scriven (1967). The term, “formative”, applies when the SET data are used by an individual instructor to obtain student feedback on his or her “teaching effectiveness” [e.g., Adams (1997), Blunt (1991), and Rifkin (1995)]. The term, “summative”, applies when the SET data are used by an administrative committee (most often comprised of members of the aforementioned collation) in the determination of reappointment, pay, merit pay, tenure, and/or promotion of an individual instructor [e.g., Hills (1974), Rifkin (1995), and Grant (1998)]. In the paragraphs that follow, all references to the SET data and to its research program have the latter, the summative, function in mind.

While the questionnaire used to collect the SET data varies from university to university, certain commonalities exist. One summary of these commonalities is offered in Table 1 for the reader’s benefit.

A characterization of the commonalities in the questionnaires used to collect the SET data due to Sproule (2000)

Table 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.
 - 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.”

¹ Much of this demand arose out of the caldron of student unrest in the late 1960’s and early 1970’s [Frankel (1968), Bay (1988), Platt (1993), and Black (2001)].

- 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).
- In the open-ended section of the SET survey instrument, students are invited to offer short critiques of the course content and of the teaching effectiveness of the instructor.
- The completion of the SET survey instrument comes with a guarantee to students; that is, the anonymity of individual respondents.
- 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.
- 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 cardinal- or ordinal- measured minimal level of “teaching effectiveness.”

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, ..), on course characteristics, etc.

2. The SET Research Program as pedagogical folly

A touchstone in Surowiecki’s book, *The Wisdom of Crowds*, is a paper by the English Victorian polymath, and a half-cousin of Charles Darwin, Sir Francis Galton (1822-1911). This paper, entitled “Vox populi”, was published in *Nature* in 1907. It records Galton’s observations and analysis of a weight-judging contest at the *West of England Fat Stock and Poultry Exhibition*.

In this contest, a competitor viewed a live ox, and then paid a fee to submit a written estimate of the weight of the ox after it has been slaughtered and dressed. In this contest, a cash prize was awarded to the competitor whose estimate came closest to the dressed weight. About the contest, Galton observed and reported that the dressed weight of the ox was 1198 pounds, and (from an analysis of the written

estimates) that the “midmost estimate” of the competitors was 1207 pounds. From these facts, Galton noted that “the vox populi was in this case 9 lbs., or 0.8 per cent. of the whole weight too high”, and he then concluded that, “this result is ... more creditable to the trustworthiness of a democratic judgment than might have been expected.”

In his recent review of Surowiecki’s book, Roger Kerr (2004) observed that, “Galton had stumbled on what Surowiecki calls a ‘simple but powerful truth’ ... that “under the right circumstances, groups are remarkably intelligent, and are often smarter than the smartest people in them” ... that “when our imperfect judgments are aggregated in the right way, our collective intelligence is often excellent” .. and that “the wisdom of a crowd doesn’t depend on any member of the crowd being wise; indeed, every member of the crowd could be hopelessly foolish. *But in a crowd the mistakes are likely to run in all directions and cancel one another out*, leaving a sound judgment in the aggregate” (p. 2).² Kerr (2004) cautions that, “crowd wisdom is not infallible”, and notes that “Surowiecki devotes much of his book to showing how the absence of any of his .. conditions diminish it” (p. 5).

What then are the preconditions to which Kerr refers? What are Surowiecki’s preconditions that separate the wisdom from the folly of crowds? There are four, and these are as follows:

- *Diversity of opinion*: Each person should have private information even if it’s just an eccentric interpretation of the known facts.
- *Independence*: People’s opinions aren’t determined by the opinions of those around them.
- *Decentralization*: People are able to specialize and draw on local knowledge.
- *Aggregation*: Some mechanism exists for turning private judgments into a collective decision.

² A parallel source of inspiration in Surowiecki’s book is the work of the Austrian economist, Frederick von Hayek. According to Hayek (1937, 1945, and 1975), there is a division of knowledge within society. As the noted Hayekian scholar and biographer observed, “Actually-existing knowledge is dispersed; different individuals have access to different bits of it. Their actions are based on their varying subjective beliefs, beliefs that include assumptions about future states of the world and about other people’s beliefs and actions. The ‘central question of all social science’ .. is how such dispersed knowledge might be put to use, how society might coordinate the knowledge that exists in many different minds and places” [Caldwell (1997), p. 1865]. Clearly, in Caldwell’s comment, one can hear echoes of Galton (1907) and Surowiecki (2004).

For purposes of the present discussion, we synthesize these four conditions into two criteria. And these two will serve as our device to test the viability of any claim that the SET data contain what-may-be-termed Galton's "Vox Populi". These two criteria are:

- *Criterion 1 - Verifiability:* Galton's argument about a crowd's wisdom is hinged on the proximity or nearness of the crowd's estimate of the weight of the dressed ox to a verifiable, objective measurement of ox's weight. Put differently, *a verifiable, objective measurement is a necessary, but not a sufficient, condition for Galton's argument about the wisdom of a crowd to persuade.*
- *Criterion 2 - Unbiasedness:* The second criterion requires that (in the aggregate) overestimates of the ox's weight cancel out all underestimates. Stated differently, (a) *the average of all of the participants' estimates must represent an unbiased estimate of true weight of the ox*, or (b) the aggregate estimate of the ox's weight must be unbiased estimate of the ox's actual weight. This unbiasedness criterion is (as before) a necessary, but not a sufficient, condition for Galton's argument about the wisdom of a crowd to persuade.

Next, it is instructive to ask, how does Galton's ox-weighing contest relate to the SET Research Program? The answer to this question has two parts: (a) Common to both Galton's contest and in the SET Research Program is an analyst who solicits the wisdom of the crowd by asking its individual members to estimate a property of an object of analysis. In Galton's case, *the object is an ox*, and *the property of interest is the ox's dressed weight*. In the SET Research Program, *the object is an instructor*, and *the property of interest is the instructor's "quality of teaching"*. (b) But beyond this point, the comparison of the Galton's contest and the SET Research Program breaks down. How? *The SET Research Program fails to satisfy Criteria 1 and 2, as the following review of the research literature on the SET makes this clear.* In particular,

2.1. The failure of the SET Research Program to satisfy Criterion 1

The SET Research Program fails Criterion 1 in that while the dressed weight of an ox can be verified, the "quality of teaching" of an instructor cannot. Three comments about this present claim are warranted.

Remark 1: At the core of this section (indeed, at the core of this paper) is the following logic: (a) "Good teaching" or "teaching quality" cannot be defined [e.g., Sproule (2000 and 2002), and Trout (2000)]. (b) *What cannot be defined cannot be measured* [e.g., Weissberg (1993, p. 8)]. (c) *What cannot be measured cannot be verified for accuracy.* And (d) what cannot be verified for accuracy cannot pass (nor fail) the "wisdom of a crowd" test.

Remark 2: Our claim that "good teaching" or "teaching quality" cannot be defined, and therefore cannot be measured, is echoed elsewhere. For example, in his survey

of the SET, Cashin (1995) wrote, “In educational measurement, the basic question concerning validity is: does the test measure what it is supposed to measure? For student ratings, this translates into: To what extent do student rating items measure some aspect of teaching effectiveness or teaching quality? *Unfortunately there is no agreed upon definition of ‘effective teaching’, nor any single, all-embracing criterion.* The best one can do is to try various approaches, collecting data that either support or contest the conclusion that student ratings reflect effective teaching.” Likewise, Olivares (2003) noted that, “.. teacher performance is a dynamic criterion predicated on an ill-defined notion of teacher effectiveness. Attempts to measure and draw conclusions regarding teacher effectiveness, without a more complete explication of ‘teacher effectiveness’ is bound to result in continued controversy ..” (p. 237). In summary, Harvey and Green (1993) contended that “teaching quality” is *not a unitary concept* and its meaning is “relative to the user of the term and the circumstances in which it is invoked” (p. 10).³

Remark 3: Kerlinger (1973) noted, “The commonest definition of validity is epitomized by the question: Are we measuring what we think we are measuring? The emphasis in this question is on what is being measured” (p. 457). So to say that “good teaching” or “teaching quality” is not measurable and not verifiable is more damaging than answering in the negative to Kerlinger’s question, “Are we measuring what we think we are measuring?”. Likewise, to say that “good teaching” or “teaching quality” is not measurable is to say that the SET Research Program lacks “criterion validity”, to use the

jargon of psychologists.⁴ In summary, Olivares (2003) noted that, advocates for the SET “.. have created a body of evidence that suggests student ratings are a valid measure of teacher effectiveness. Yet, this research rests upon a weak foundation; namely, the inadequate definition and operationalisation of teacher effectiveness. This weakness has compromised the integrity of validity evidence and inferences drawn therefrom” (p. 238).

³And it gets worse: In *Zen and the Art of Motorcycle Maintenance*, Robert Pirsig (1974) observed, “I think there is such a thing as Quality, but that as soon as you try to define it, something goes haywire. You can’t do it.” (p. 184). [For reflections on Pirsig’s assessment of Quality, see Hoyer and Hoyer (2001), Rodino (1980), Shields (1999), and Willis (2000).]

⁴These claims are echoed in a dated paper on criterion-centered research by Astin (1964). He noted that, “(e)ven in higher education, there are no criteria for assessing teaching proficiency; a professor’s competence is usually measured in terms of the number of articles he has published” (pp. 808-09).

If any fallback position from these criticisms exists, then surely it is merely *an opening for the divining of the high priests of the SET Research Program, and for the mongering of pseudoscience*.⁵

2.2. The failure of the SET Research Program to satisfy Criterion 2

The SET Research Program fails Criterion 2 in that the average of students' estimates is not an unbiased estimate of whatever-it-is-that-the-SET-data-actually-measures. To buttress this statement, we present in Table 2 (below) a simplistic overview to Wachtel's (1998) recent literature survey of the SET. In particular, this table presents four overarching types of biases inherent in, or background variables thought to influence, the measurement of "good teaching" in the SET Research Program. These types of contaminating variables are: (a) the characteristics associated with the administration of student evaluations, (b) the characteristics of the instructor, (c) the characteristics of the students, and (b) the reactions of students and faculty to the use of the SET. Within each of the overarching types reported in Table 2, one finds a listing of specific examples of biases reported in the research literature. We shall refer to a few of these in the pages below.

The core framework used in Howard Wachtel's (1998) literature review

Table 2

The background variables thought to influence the SET include:

- *The characteristics associated with the administration of the SET:* The timing of evaluation, the anonymity of student raters, the instructor presence in classroom, the stated purpose of the evaluation, the characteristics of the course (e.g.,

⁵ We choose our words carefully: (a) Hubley and Zumbo (1996) provided a useful, current day overview to "validity" in Psychology, and Olivares (2003, pp. 234-36) outlined the role of three types of validity in the context of the SET [viz., criterion, content, and construct]. (b) While criterion validity is accessible to all including the informed layperson, the other two are not. These are the province of the "experts", or the "high priests" of Psychology. (c) So when a situation denies the use of criterion validity, the informed layperson becomes a captive of the "high priests". (d) Situations that separate the informed layperson from the "high priests" may have helped to prompt Richard Feynman (1986), Nobel laureate in physics, to characterize Psychology as a "Cargo Cult" pseudoscience [cited in Richters (1997, pp. 207-208)], and Paul Meehl (1978) to state that most "theories" in Psychology "... suffer the fate that General MacArthur ascribed to old generals - They never die .. they just slowly fade away" [cited in Richters (1997, p. 195)].

electivity), the class meeting time, the level of course, the class size, the subject area, and the workload of the course.

- *The characteristics of the instructor:* The instructor's rank and experience, the reputation of instructor, the instructor's research productivity, the personality of instructor, the seductiveness of instructor (the 'Dr. Fox' effect), the gender of instructor, the minority status of instructor, and the physical appearance of instructor.
- *The characteristics of the students:* The personality characteristics, the prior subject interest, the gender of the student, the expected grade, the leniency hypothesis, the student's expectations and emotional state, and the student's age
- *The reaction to the dissemination and use and the SET:* The reaction by faculty, the reaction from students, and the publicizing of the student ratings.

3. The SET Research Program as pseudoscience

While the SET data are cursed by contaminating variables [Table 2], the advocates for the SET Research Program rarely make an effort to control for, or to even acknowledge the presence of, them [e.g., Sproule (2000 and (2002)] . And while controlling for these contaminating variables may not be a viable option in that the SET data are extracted from a non-experimental setting, so too is not acknowledging their presence. Academic honesty requires no more, and no less. Stated differently, academic honesty requires a clear admission that the analysis of the SET data is predicated on the application of a closed-system paradigm in the analysis of data from (what is in fact) an open social system.⁶

How does this directly relate to the SET Research Program? The first of two key ontological presumptions that advocates of the SET Research Program hold is that there exists a unique entity, termed "good teaching" or "teaching quality", and that this can be defined and measured operationally. An argument against this assertion has been made already in Section 3. The second presumption is that if "good teaching" or "teaching quality" does exist (which it does not) and if it can be defined and can be measured (which it cannot), then there must also exist a monotonic and a time-space invariant relationship between "teaching quality"

⁶ This specie of discussion appears in Psychology [e.g., Meehl (1990a and 1990b), Lykken (1991), and Richters (1997)], in Educational Psychology [e.g., Elsworth (1994), Kemp and Holmwood (2003), and Nash (2005)], and in Economics [e.g., Lawson (1989, 2001, and 2003), Dow (2002), and Chick and Dow (2005)]. Common to all of these (either implicitly or explicitly) are elements of several overarching frameworks: (a) the part-whole dualism, (b) the mechanistic-organismic dualism [e.g., Wheeler (1935)], (c) methodological individualism-holism dualism vs. systematism [e.g., Bunge (2000a and 2000b), and Denis (2003)], and (d) Critical Realism [e.g., Bhaskar (1989), Collier (1994), and Sayer (1992, 1995, 1997, and 1999)].

itself, and the students' reports of it. In other words, the link between "teaching quality" and students' reports of "teaching quality" is found in, and is defined by, what methodologists term a closed system.

However, the gulf between claim and reality is unbridgeable. Rather than being provided by a setting that assures of time-space invariance and monotonicity, the analyst is confronted by a situation in which the data are contaminated by a "crud factor", to use the phrase of the revered psychologist, Paul Meehl (1990a); that is, the analyst is confronted by a non-experimental setting in which "everything is correlated with everything else". The next two remarks, it is hoped, will drive home the point about the importance and presence of Meehl's "crud factor".

Remark 4: If the classroom setting does indeed provide for time-space invariance and monotonicity, then one ought to be able to expect that those characteristics of the instructor that are irrelevant to the pedagogical process *will not* impact students' reports of "teaching quality" [e.g., Feldman (1977), and Olivares (2003, p. 238)]. But the research literature provides a different story, as Wachtel's (1998) review points up. [See again Table 2 above.] In particular, the literature documents the existence of relationships between students' reports of "teaching quality" and *the instructor's age* [e.g., Goebel and Cashen (1979)], *the instructor's gender* [e.g., Goebel and Cashen (1979), Buck and Tiene (1989), and Feldman (1993)], *the instructor's attractiveness or beauty* [e.g., Goebel and Cashen (1979), Buck and Tiene (1989), and Hamermesh and Parker (2005)], *the instructor's demeanor or personality* [e.g., Damron (1994)], the instructors' charisma [e.g., Ambady and Rosenthal (1993)], and *the instructor's gestural mannerisms, better known as the "Dr. Fox Effect"* [e.g., Ware and Williams (1975), Williams and Ware (1977), and Williams and Ceci (1977)].

Remark 5: Likewise, if the classroom setting does indeed provide for time-space invariance and monotonicity, then one ought to be able to expect that those characteristics of the student that are irrelevant (or even relevant) to the pedagogical process will not impact students' reports of "teaching quality" [e.g., Feldman (1977), and Olivares (2003, p. 238)]. Again, the research literature provides a different story, as Wachtel's (1998) review points up. [See once more Table 2 above.] In particular, *the literature documents cases in which student ratings of "teaching quality" are impacted by a student's personality and emotive state* [e.g., Clayson and Sheffet (2006), Small et al. (1982), and Trout (2006)]. In addition, *the literature documents cases in which student ratings of "teaching quality" are impacted by a student's grade or expected grade* [e.g., Glenn (2007), Krautmann and Sander (1999), and Langbein (2005)]. In summary, Olivares (2003) noted that, "... student ratings are not objective evaluations of teacher effectiveness but rather a reflection of the psychosocial needs, feelings and satisfaction of the raters" (p. 237).

It may prove instructive for some to spotlight the details of a few of the studies cited in Remarks 4 and 5 above:

Remark 6: In the Ambady and Rosenthal (1993) study, non-student raters were exposed to six-second silent video clips of instructors, and their teaching-evaluation scores were correlated with the scores recorded by students in a regular classroom setting. The researchers found that the two sets of scores were correlated positively and highly. They wrote, "On the basis of observations of video clips just half a minute in length, complete strangers were able to predict quite accurately the ratings of teachers by students who had interacted with them over the course of a whole semester! Furthermore, these predictions retained their accuracy after we adjusted for physical appearance of the teachers, indicating that the judges were picking up very subtle nonverbal cues." For us, this study begs two questions: (a) What information about "teaching effectiveness" or "teaching quality" is contained in the numerical evaluations by students at the end of the term that is not contained in the numerical evaluations by non-students after an exposure to a six-second silent video clip? (b) What information about "teaching effectiveness" or "teaching quality" is contained in the numerical evaluations by students at the end of the term? The findings of Ambady and Rosenthal (1993) would suggest that the answer to both questions is, "At best, not much."

Remark 7: As stated above, Glenn (2007), Krautmann and Sander (1999), and Langbein. (2005), report that student ratings of "teaching quality" are impacted by a student's grade or expected grade. In the case of Glenn, he wrote:

"Student evaluations of instructors are deeply imperfect tools that are often misused by college administrators -- but the evaluations should not be scrapped, two scholars said here on Saturday at the annual meeting of the Association for Psychological Science. Each scholar sketched a model for reforming the faculty-assessment system.

"'At best, student ratings provide a weak measure of instructional quality,' said Anthony G. Greenwald, a professor of psychology at the University of Washington. 'They're heavily influenced by grades, and they're also influenced by class size.'

"Mr. Greenwald presented an analysis of the student ratings of more than 14,000 courses given at his university from 1997 to 2001. He was especially interested in exploring how average course ratings varied by department. Courses in Washington's dance department, for example, typically received high student evaluations, while chemistry, physics, and mathematics courses tended to rate poorly.

"Not surprisingly, Mr. Greenwald said, he found that *70 percent of the variance in departments' average course-evaluation scores could be explained by differences in students' grades*. In departments where professors' grading was more rigorous, students' evaluation scores were lower. Like many other analysts of student

evaluations, Mr. Greenwald worries that, if evaluations are misused, they could give instructors perverse incentives to inflate their grades or dumb down their courses.”

For us, one statement that greatly damages the position of those who advocate the use of the SET is Glenn’s statement that “Mr. Greenwald .. found that 70 percent of the variance in departments’ average course-evaluation scores could be explained by differences in students’ grades.”⁷ Clearly, according to Greenwald, the variation in the SET data is swamped by the grade effect to the exclusion of all other effects, including the “teaching quality” effect, if indeed “teaching quality” could be measured. In a related vein, the econometric studies by Krautmann and Sander (1999), and by Langbein (2005), offer results that are consistent with the hypothesis that instructors can “buy” better evaluations through more lenient grading.⁸

Remark 8: In a recent paper, Youmans and Jee (2007) outline experimentally another method by which better evaluations can be bought. They write “that something as trivial as offering chocolates” to students before an evaluation can affect the scores that students assign their teacher.

In summary, in the analysis of an open social system, one cannot do the impossible: one cannot “unscramble an egg” [Sproule (2000)].⁹ And so, if the claims made by the advocates of the SET Research Program are not seen as bad

⁷ Greenwald’s related research includes Gillmore and Greenwald (1999), Greenwald (1997), and Greenwald and Gillmore (1997a, 1997b, and 1998).

⁸ These findings can be combined with other arguments cited in this paper. For example, since “teaching quality” is not-observable, the players can “game the system”. That is, not only can faculty curry more favorable SET scores from students by engaging in grade inflation, but this same stratagem enables students to behave as “customers” of (not as partners in) the educational process, and to participate passively (not actively)]. For more, see Fischer (2006).

⁹ Here, echoes of Critical Realism can be heard. For example, Sayer (1992) explained that positivism is derived from a philosophy of atomism, which has two branches: “The ontological branch – concerning the theory of what exists – holds that the world consists of discrete, distinct atomistic elements existing at discrete, distinct points in time of space. Being atomistic these basic elements have no internal structure or differentiation and no causal powers. The various objects that we know are nothing but different combinations of these atoms. All relations between objects are external and contingent, so that all sequences are accidental. These assumptions are matched by the epistemological branch – concerning the theory of knowledge – which depicts observation as fragmented into simple, unproblematic, indivisible ‘readings’. These two branches are mutually reinforcing: if objects or events are ‘punctiform’ their observation as such is also more plausible and vice versa” (p. 155). Likewise, Richter (1997, pp. 195-97) made clear: (a) that the closed system paradigm presumes that the objects of analysis are like atoms [that is, homogenous and passive], but (b) that human organisms are unlike atoms in that they are heterogeneous and active.

science, then one must conclude: (a) that no research program can be, and (b) that everyone engaged in the analysis of the SET data is entitled to subscribe to the dictum of the scientific anarchist, Paul Feyerabend (1975), which is, “anything goes”.

Next, we ask, can one enlist standards other than those invoked in the above to strengthen the present argument that the SET Research Program is bad science? Here, we say, “yes”, by turning to Stephen Toulmin’s (2003) *Uses of Argument*.¹⁰ And here, we ask, does the SET Research Program conform to Toulmin’s model of good argument and good science? To address this question, we begin by offering two pithy comments by Hansen and Eschelbach Hansen (2005) about Toulmin’s model of good argument and science. They wrote,

“In his influential book the *Uses of Argument* (first published in 1958), the philosopher Stephen Toulmin argues that good arguments, regardless of what field they are made in, have common features. The common features are a *claim*, *evidence*, a *warrant*, and *qualifications*. The claim is the central statement of the argument. A claim is a statement that is disputable. If the claim was obviously true, there would be no need for argument. The evidence provides support for the claim. The warrant explains why the evidence should be persuasive. The qualifications are the limits of the argument: where it should apply and where it may not.” (p. 5)

“Toulmin presents the claim and the evidence as the leading elements in the argument. He explains that ‘the explicit appeal in this argument goes directly from the claim to the data relied on as foundation: the warrant is, in a sense, incidental and explanatory’ (Toulmin 2003, 92). He goes on to state that ‘data are appealed to explicitly, warrants implicitly’ (Toulmin 2003, 92). The researcher, Toulmin argues, should make a claim, present the evidence in support of the claim, and then explain why that evidence is persuasive. The approach is lawyerly: present piece after piece of evidence and then in summation explain why the jury must be persuaded by the evidence.” (p. 5)

A third, useful comment provided by Hansen and Eschelbach Hansen (2005) is that, and here we paraphrase: (a) the warrant comes in the form of a theory and a model, (b) the warrant is the horse, and the evidence the cart, and (c) rather than being implicit and following the data, the warrant is explicit and leads the search for evidence (p. 7). To summarize, *the warrant* is a general, hypothetical (and often implicit) logical statement that serves as *the bridge between the claim and the data*.

Next, we ask, how can Toulmin’s definition of “good argument” and “good science” be used to assess the methodological viability of the SET Research Program? Our response is this: since the SET Research Program has no (formal) model, then (by definition) it can have no warrant, and hence it has no bridge

¹⁰ As evidenced by Footnotes 6 and 9, Critical Realism could be enlisted also.

between its claim(s) and its data. This assessment is borne out by our ongoing, lengthy, and detailed review of the SET literature: we have been unable to identify a single study that defines, or otherwise offers, a coherent model or theory. And so, we conclude that in their failure to meet the standards laid out by Toulmin, all arguments within the SET Research Program are incomplete, and therefore unpersuasive.

Here, it should be noted that this absence of a model or theory, this absence of a warrant, within the SET Research Program is representative of a chronic problem within *soft psychology*. In this very connection, Lakatos (1978a) observed,

“This requirement of continuous growth..., hits patched-up, unimaginative series of pedestrian ‘empirical’ adjustments which are so frequent, for instance, in modern social psychology. Such adjustments may, with the help of so-called ‘statistical techniques,’ make some ‘novel’ predictions and may even conjure up some irrelevant grains of truth in them. *But this theorizing has no unifying idea, no heuristic power, no continuity. They do not add up to a genuine research programme and are, on the whole, worthless.*” (p. 88)

To summarize, in its attempt to apply a closed-system paradigm in its analysis of an open social system, and in its failure to offer a bridge between its claim(s) and the data, the SET Research Program must be viewed as a failure, and as pseudoscience. As the widely-respected astronomer, Carl Sagan (1997), once noted,

“Pseudoscience is embraced, it might be argued, in exact proportion as real science is misunderstood -- except that the language breaks down here. If you’ve never heard of science (to say nothing of how it works) you can hardly be aware that you’re embracing pseudoscience. You’re simply thinking in one of the ways that humans always have.”

“What skeptical thinking boils down to is the means to construct, and to understand, a reasoned argument and - especially important - to recognize a fallacious or fraudulent argument. *The question is not whether we like the conclusion that emerges out of a train of reasoning, but whether the conclusion follows from the premise or starting point and whether that premise is true.*”

“Baloney, bamboozles, careless thinking, flimflam, and wishes disguised as facts are not restricted to parlor magic and ambiguous advice on matters of the heart. Unfortunately, they ripple through mainstream political, social, religious, and economic issues in every nation.”

4. The SET Research Program as a logical fallacy

Here, we shall argue that the SET Research Program is predicated: (a) on a false analogy between the nature and substance of a business enterprise and those of a university, and therefore (b) on a mischaracterization of the true nature of the student-professor relationship. As such, we claim here that the SET Research Program is predicated on a logical fallacy.

This false analogy, this logical fallacy, may be defined as this: (a) a university and a business firm are two organizational entities, (b) a business firm is successful if it is product-quality, customer, and bottom-line oriented, and (c) a university can be “successful” if it too becomes product-quality, customer, and bottom-line oriented.¹¹

Remark 9: The use of this analogy is evident in the literature. For example, Nicklin (1995) stated, “The university is just another business organization, and ‘educating people is a process, just like making a car is a process’” (A34). Likewise, in his monograph on the use of management fads in higher education, Birnbaum (2000) wrote, “Most business leaders think that colleges and universities would become more efficient and more productive by adopting business practices” (p. 215). Finally, Carlin (1999) wrote, “I have been a businessman for over 35 years, and I was a trustee of the University of Massachusetts and chairman of the Massachusetts Board of Higher Education for a total of 12 years. I am, or have been, a director of eight public corporations, and was chief executive officer of a transit system with an annual budget of \$1 billion. I have also founded four businesses, in separate fields, that were recognized by Inc. magazine for their rapid growth and success. I think I’ve learned something about management and controlling costs. Never have I observed anything as unfocused or mismanaged as higher education.”

Remark 10: Claims that this analogy is patently false are also evident in the literature. For example, Olivares (2003) noted that, “the subjectivity in student ratings of teachers is illimitable. To think that students, who have no training in evaluation, are not content experts, and possess myriad idiosyncratic tendencies, would not be susceptible to errors in judgment is specious” (p. 237). Likewise, Birnbaum (2000) observed that businesses and universities are non-comparable. He notes that, “Businesses look the way they do because firms with this form have proven to be more successful than firms with alternative forms. Universities look the way they do for the same reason: their form has proven to be superbly suited to what they do” (p. 217). Finally, Felder and Brent (1999) offered four reasons on

¹¹ This claim represents the core premise underlying the recommendation that institutions of higher education adopt the principles of Total Quality Management (TQM). For a detailed analysis and critique of this, see Birnbaum (2000), especially his Chapter 4. For a similar slant, see Valsan and Sproule (2007).

why the application of corporate model to tertiary education is inappropriate. They wrote:

- “In industry, the true mission is relatively clear, and quality is relatively straightforward to define. In education, the true mission is complex and subject to endless debate, and quality is therefore almost impossible to define in an operationally useful manner.” (page 9)
- “In industry, quality is relatively easy to assess. In education, even if a definition of quality can be formulated and agreed upon, devising a meaningful assessment process is a monumental task.” (page 9)
- “In industry, the customer is relatively easy to identify and is always right, at least in principle. In education, those who might be identified as “customers” have contradictory needs and desires and may very well be completely wrong.” (page 10)
- “In industry, a clear chain of command usually exists, on paper and in fact. In education, a chain of command might exist on paper, but it is in fact relatively amorphous and nothing at all like its industrial counterpart.” (page 10-11)

Remark 11: This false analogy leads to a mischaracterization of the true nature of the professor-student relationship. While the true relationship is one of master-apprentice or (better still) mentor-mentee, the false analogy leads to the misconceptualization of the “student as customer”. In this vein, Haskell (1997) noted that “.. the simple student-as-consumer metaphor is inappropriate. Thus the ‘consumer’ of higher education is in fact a wide constituency of groups distributed in both space and time. The metaphor of student as consumer is more appropriately replaced by the metaphor of student as worker or as apprentice.” Singh (2002) noted, the SET “reduces the student-teacher relationship to a customer-producer relationship, which sanctions each to profit at the expense of the other, whereas quality in education calls for a commitment both from students and teachers. To maintain quality in education, we need an approach that involves communicative action, an approach that brings back students and teachers as a community of scholars, where both sides recognise their commitment towards the process of inquiry” (p. 681). Singh (2002) also noted that, “.. a teacher-student relationship is gift relationship, and it is two-sided. In this relationship, both the students and teachers have their reciprocal rights and responsibilities towards the process of scholarly inquiry. If our classrooms and lecture theatres are meant to be places of such inquiry, then our students are meant to be authentic participants of academic community, whose rights cannot be premised on the customer metaphor” (p. 698). Additional insight into the ludicrous nature of the student-as-customer metaphor can be found in Cheney et al. (1997), and in McMillan and Cheney (1996).

Perhaps the most damning and damaging criticism of the student-as-customer metaphor is provided by Crumbley (1995), when he observed,

“Steven M. Cahn, Provost and Vice-President at City University of New York, debunks this ludicrous consumer argument by pointing out that passengers on a airplane do not certify pilots, and patients do not certify physicians. ‘Those who suffer the consequences of professional malfeasance do not typically possess the requisite knowledge to make informed judgments’ [Cahn, 1986, p. 37]. Imagine the chaos if we certified dentists, nurses, CPAs, lawyers, engineers, architects, air conditioning repair people, etc. with questionnaires from customers.” (p. 5)

5. Concluding Remarks

This paper has pointed up the methodological inadequacy of the SET Research Program, by reference to three, interrelated arguments. The first argument is that the SET Research Program can never capture the wisdom of a crowd because it fails to meet Surowiecki’s preconditions for the existence and articulation of the wisdom of a crowd. The second extends the first, by stating: (a) the SET Research Program fails to provide the methodological controls needed to differentiate cause from effect, or stated differently (b) the methodological underpinnings of the SET Research Program is the misapplication of a closed-system paradigm to an open social system. The final argument is that the SET Research Program is predicated on a false analogy between business and the university, and therefore on a mischaracterization of the student-professor relationship.

These three arguments should and can be viewed now against an even wider backdrop. In particular, there are those who suggest that regardless of the methodological flaws outlined above, there must be a way out. For example, what if one was to adopt Stanley Fish’s (2005) recommendations for the redesign of the SET questionnaire? [For an overview to such questions, to Fish’s questions, see Table 3 below.] That is, what if “fairer” questions were adopted? What if the SET questions focused exclusively on the functionality versus the dysfunctionality of the relationship between students and their instructor? What if one begins with the premise that both students and their instructor have *responsibilities as well as rights*, and then designs a questionnaire that offers a reading on whether or not both parties have honored their respective responsibilities? Our view is that this redirection of effort may be all well and good, but there would still remain a fly in the ointment. And this is the no-small issue of *the reliability of self-reported data offered by anonymous respondents*, especially in a situation that is known not to be devoid of incentives to misreport [e.g., Maxwell and Lopus (1994), Schwarz (1999), Sproule (2000 and 2002), and Wright (2006),]. Enough said.

Some of the more salient questions recommended by Stanley Fish (2005)

Table 3

Questions related to the instructor's pedagogical responsibility:

- Were examinations and other graded materials returned on a timely basis?
- Were students tested on materials covered in the course?
- Was there sufficient feedback on tests and papers?
- Were course materials well prepared?
- Did the course unfold as promised in the catalog and syllabus?
- Was the instructor accessible to students during office hours?
- Did the instructor give lectures that facilitated note taking?

Questions related to the student's pedagogical responsibility:

- Did you attend class regularly?
- Did you read the assignments?
- Did you spend much time doing research for your final paper?

A question which conflates student and teaching performance, and which implies that the latter is always responsible for the former.

- Have you learned and understood the subject materials of this course?
-

To sum up, we offer one last comment. To the hard-core advocates of the SET Research Program, we ask: address the issues raised in the above paragraphs in a manner that persuades and convinces the informed layperson, or admit that you cannot, and abandon this research program once and for all. Now since we hold that such persuasive and convincing arguments cannot be found, we therefore recommend that you swallow a bitter pill; that being, you admit that the architecture of your research program is incoherent, ill-conceived, and fatally flawed, and therefore it cannot, with any credulity, serve as a constructive guide to the administration and governance of a university. And finally, we ask that you admit that the unintended consequences of your efforts not only include the

successful but wrong-headed institutionalization of pseudoscience, but also the successful but wrong-headed transformation of the pedagogical environment into one that is more consumer- and hence less learning-orientated.

References

1. Adams, J.V. (1997), "Student evaluations: The ratings game," *Inquiry* 1 (2), 10-16.
2. Ambady, N., and R. Rosenthal (1993), "Half a minute: Predicting teacher evaluations from thin slices of behavior and physical attractiveness," *Journal of Personality and Social Psychology* 64, 431-441.
3. Astin, A.W. (1964), "Criterion-centered research," *Educational and Psychological Measurement* 24 (4), 807-822.
4. Bay, C. (1988), "University educational reform in the sixties - Ideals, goals, and results," *Interchange* 19 (3-4), 163-76.
5. Bhaskar, R.A. (1989), *Reclaiming Reality: A Critical Introduction to Contemporary Philosophy* (London: Verso).
6. Birnbaum, R. (2000), *Management Fads in Higher Education: Where They Come From, What They Do, Why They Fail* (San Francisco: Jossey-Bass).
7. Black, J.N. (2005), "Restoring the language of truth," A paper presented May 30th at Indiana Wesleyan University, Marion, Indiana.
8. Blunt, A. (1991), "The effects of anonymity and manipulated grades on student ratings of instructors," *Community College Review* 18, Summer, 48-53.
9. Buck, S., and D. Tiene (1989), "The impact of physical attractiveness, gender, and teaching philosophy on teacher evaluations," *Journal of Educational Research* 82, 172-7.
10. Bunge, M. (2000a), "Systemism: the alternative to individualism and holism," *Journal of Socio-Economics* 29, 147-157.
11. Bunge, M. (2000b), "Ten modes of individualism – none of which works – and their alternatives," *Philosophy of the Social Sciences* 30, 384-406.
12. Caldwell, B.J. (1997), "Hayek and socialism," *Journal of Economic Literature* 35, 1856-1890.

13. Carlin, J.F. (1999), "Restoring sanity to an academic world gone mad," *Chronicle of Higher Education*, November 5, A76.
14. Cashin, W. (1995), "Student ratings of teaching: The research revisited," Idea Paper No. 32, Center for Faculty Evaluation and Development, Kansas State University. September.
15. Cheney, G., J.J. McMillan, and R. Schwartzman (1997), "Should we buy the 'student-as-consumer' metaphor?" *The Montana Professor* 7 (3).
16. Chick, V., and S. Dow (2005), "The meaning of open systems," *Journal of Economic Methodology* 12 (3), 363-381.
17. Clayson, D.E., and M.J. Sheffet (2006), "Personality and the student evaluation of teaching," *Journal Of Marketing Education* 28 (2), 149-160.
18. Collier, A. (1994), *Critical Realism: An Introduction to Roy Bhaskar's Philosophy* (London: Verso).
19. Crumbley, D.L. (1995), "Dysfunctional effects of summative student evaluations of teaching: Games professors play," *Accounting Perspectives* 1 (1), Spring, 67-77.
20. Damron, J.C. (1996), "Instructor personality and the politics of the classroom," Working Paper, Accessed online in May 2007.¹²
21. Denis, A. (2003), "Methodology and policy prescription in economic thought: A response to Mario Bunge," *Journal of Socio-Economics* 32, 219-226.
22. Dow, S. (2002), *Economic Methodology: An Inquiry* (Oxford: Oxford University Press).
23. Elsworth, G.R. (1994), "Arguing challenges to validity in field research: A realist perspective," *Science Communication* 15 (3), 321-343.
24. Felder, R.M., and R. Brent (1999), "How to improve teaching quality," *Quality Management Journal* 6 (2), 9-21.
25. Feldman, K.A. (1977), "Consistency and variability among college students in rating their teachers and course: A review and analysis," *Research in Higher Education* 6, 223-274.
26. Feldman, K.A. (1993), "College students' views of male and female college teachers: Part II. Evidence from students' evaluations of their classroom teachers," *Research in Higher Education* 34, 151-211. Feyerabend, P. (1975), *Against Method* (London: Verso).

¹² < <ftp://ftp.csd.uwm.edu/pub/Psychology/BehaviorAnalysis/educational/politics-of-instructor-evaluation-damron> >

27. Feynman, R. (1986), *Surely you're joking, Mr. Feynman!* (New York: Bantam Books).
28. Fischer, J.D. (2006), "Implications of recent research on student evaluations of teaching," *The Montana Professor* 17 (1).
29. Fish, S. (2005), "Who's in charge here? The evaluation of teaching by students amounts to a whole lot of machinery with a small and dubious yield," *The Chronicle of Higher Education*, February 4.
30. Frankel, C. (1968), *Education and the Barricades* (New York: W.W. Norton).
31. Galton, F. (1907), "Vox populi," *Nature* 75, 450-451.
32. Gillmore, G.M., and A.G. Greenwald (1999), "Using statistical adjustment to reduce biases in student ratings," *American Psychologist*, July, 518-519.
33. Glenn, D. (2007), "Method of using student evaluations to assess professors is flawed but fixable, 2 scholars say," *Chronicle of Higher Education*, May 29.
34. Goebel, B., and V. Cashen (1979), "Age, sex and attractiveness as factors in student ratings of teachers: A developmental study," *Journal of Educational Psychology* 71, 646-53.
35. Grant, H. (1998), "Academic contests: Merit pay in Canadian universities," *Relations Industrielles/ Industrial Relations* 53 (4), 647-664.
36. Greenwald, A.G. (1997), "Validity concerns and usefulness of student ratings," *American Psychologist* 52, 1182-1186.
37. Greenwald, A.G., and G.M. Gillmore (1997a), "Grading leniency is a removable contaminant of student ratings," *American Psychologist* 52, 1209-1217.
38. Greenwald, A.G., and G.M. Gillmore (1997b), "No pain, no gain? The importance of measuring course workload in student ratings of instruction," *Journal of Educational Psychology* 89, 743-751.
39. Greenwald, A.G., and G.M. Gillmore (1998), "How useful are student ratings?," *American Psychologist*, Nov., 1228-1229.
40. Hamermesh, D. S. and A. Parker (2005), "Beauty in the classroom: Instructors' pulchritude and putative pedagogical productivity," *Economics of Education Review* 24 (4), 369-376.
41. Hansen, B.A., and M. Eschelbach Hansen (2005), "Don't put the cart before the horse: Teaching the economic approach to empirical research," Working Paper, Department of Economics, American University.

42. Harvey, L., and D. Green (1993), "Defining quality," *Assessment and Evaluation in Higher Education* 18 (1), 9-34.
43. Haskell, R.E. (1997), "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.
44. Hayek, F.A. (1937), "Economics and knowledge," *Economica* 4, 33-54.
45. Hayek, F.A. (1945), "The use of knowledge in society," *American Economic Review* 35, 519-530.
46. Hayek, F.A. (1975), "The pretence of knowledge," *Swedish Journal of Economics* 77, 433-442.
47. Hills, J.R. (1974), "On the use of student ratings of faculty in determination of pay, promotion and tenure," *Journal Research in Higher Education* 2 (4), 317-324.
48. Hoyer, R.W., and B.Y. Hoyer (2001), "What is quality?: Learn how each of eight well-known gurus answers this question," *Quality Progress*, July, 53-62.
49. Hubley, A.M., and B.D. Zumbo (1996), "A dialectic on validity: Where we have we been and where we are going," *The Journal of General Psychology* 123, 207-215.
50. Kemp, S., and J. Holmwood (2003), "Realism, regularity and social explanation," *Journal for the Theory of Social Behaviour* 33(2), 165-187.
51. Kerlinger, F.N. (1973), *Foundations of Behavioral Research* (New York: Holt, Rinehart and Winston).
52. Kerr, R. (2004), "The wisdom of crowds," The New Zealand Business Roundtable, Christchurch, New Zealand.
53. Krautmann, A. and W. Sander (1999), "Grades and student evaluations of teachers," *Economics of Education Review* 18, 49-53.
54. Lakatos, I. (1978a), "Falsification and the methodology of scientific research programs," in J. Worrall and G. Currie, eds., *The Methodology of Scientific Research Programs: Imre Lakatos Philosophical Papers, Vol. 1* (Cambridge, England: Cambridge University Press), 8-101.
55. Lakatos, I. (1978b), *The Methodology of Scientific Research Programmes* (Cambridge: Cambridge University Press).
56. Langbein, L. (2005), "Management by results: Student evaluation of faculty teaching and the mismeasurement of performance," Working Paper, School

- of Public Affairs, American University, presented at Annual Meeting of Public Choice Society, New Orleans, March 10-13.
57. Lawson, T. (1989), "Realism and instrumentalism in the development of econometrics," *Oxford Economic Papers* 41, 236-258.
 58. Lawson, T. (2001), "Two responses to the failings of modern economics: The instrumentalist and the realist," *Review of Population and Social Policy* 10, 155-81.
 59. Lawson, T. (2003), *Reorienting Economics* (New York: Routledge).
 60. Lykken, D.T. (1991), "What's wrong with psychology, anyway?," in D. Chicchetti and W. Grove, eds. *Thinking Clearly About Psychology, Volume 1* (Minneapolis: University of Minnesota Press), 3-39.
 61. Maxwell, N., and J. Lopus (1994), "The Lake Wobegon effect in student self-reported data," *American Economic Review* 84 (2), 201-05.
 62. McMillan, J., and G. Cheney (1996), "The student as consumer: The implications and limitations of a metaphor," *Communication Education* 45 (1), 1-16.
 63. Meehl, P. E. (1978), "Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology," *Journal of Consulting and Clinical Psychology* 46, 1-42.
 64. Meehl, P.E. (1990a), "Why summaries of research on psychological theories are often uninterpretable," *Psychological Reports* 66 (1), 195-244.
 65. Meehl, P.E. (1990b), "Appraising and amending theories: The strategy of Lakatosian defense and two principles that warrant it," *Psychological Inquiry* 1, 108-141.
 66. Murray, H.G. (2005), "Student evaluation of teaching: Has it made a difference?," A paper presented at the Annual Meeting of the Society for Teaching and Learning in Higher Education, Charlottetown, Prince Edward Island, Canada, June.
 67. Nash, R. (2005), "Explanation and quantification in educational research: The arguments of critical and scientific realism," *British Educational Research Journal* 31 (2), 185-204.
 68. Nicklin, J.L. (1995), "The hum of corporate buzzwords," *The Chronicle of Higher Education* 41 (20), 33-34.
 69. Olivares, O.J. (2003), "A conceptual and analytic critique of student ratings of teachers in the USA with implications for teacher effectiveness and student learning," *Teaching in Higher Education* 8 (2), 233-245.

70. Platt, M. (1993), "What student evaluations teach," *Perspectives In Political Science* 22 (1), 29-40.
71. Pirsig, R.M. (1974), *Zen and the Art of Motorcycle Maintenance: An Inquiry into Values* (New York: Bantam Books).
72. Richters, J.E. (1997), "The Hubble hypothesis and the developmentalist's dilemma," *Development and Psychopathology* 9, 193-229.
73. Rifkin, T. (1995), "The status and scope of faculty evaluation," *ERIC Digest*.
74. Rodino, R.H. (1980), "Irony and earnestness in Robert Pirsig's *Zen and the art of motorcycle maintenance*," *Critique: Studies in Modern Fiction* 22, 21-31.
75. Sagan, C. (1997), *The Demon-Haunted World: Science as a Candle in the Dark* (New York: Ballantine Books).
76. Sayer, A. (1992), *Method in Social Science: A Realist Approach*, 2nd ed. (London: Routledge).
77. Sayer, A. (1995), *Radical Political Economy: A Critique* (Oxford: Blackwell).
78. Sayer, A. (1997), "Critical realism and the limits to critical social science," *Journal for the Theory of Social Behaviour* 27 (4), 473-488.
79. Sayer, A. (1999), *Realism and Social Science* (London: Sage).
80. Schwarz, N. (1999), "Self-reports: How the questions shape the answers," *American Psychologist* 54, 93-105.
81. Scriven, M. (1967), "The methodology of evaluation," in R. Tyler, R. Gagne, and M. Scriven, eds., *Perspectives in Curriculum Evaluation* (Skokie, IL: Rand McNally).
82. Shields, P.M. (1999), "Zen and the art of higher education maintenance: Bridging classic and romantic notions of quality," *Journal of Higher Education Policy and Management* 21 (2), 165-172.
83. Singh, G. (2002), "Educational consumers or educational partners: A critical theory analysis," *Critical Perspectives on Accounting* 13 (5-6), 681-700.
84. Small, A.C., A.R. Hollenbeck, and L. Haley (1982), "The effect of emotional state on student ratings of instruction," *Teaching of Psychology* 9, 205-211.
85. Sproule, R. (2000), "The student evaluation of teaching: A methodological critique of conventional practices," *Education Policy Analysis Archives* 8 (50), November 2.

86. Sproule, R. (2002), "The underdetermination of instructor performance by data from the student evaluation of teaching," *Economics of Education Review* 21 (3), 287-294.
87. Surowiecki, J. (2004), *The Wisdom Of Crowds: Why the Many Are Smarter than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations* (New York: Doubleday).
88. Toulmin, S. (2003), *The Uses of Argument* (Cambridge: Cambridge University Press).
89. Trout, P. (2000), "Flunking the test: The dismal record of student evaluations," *Academe*, July–August, 58–61.
90. Trout, P. (2006), "Shame on you," *Chronicle of Higher Education* 53 (11).
91. Valsan, C., and R. Sproule (2007), "The invisible hands behind the student evaluation of teaching: The rise of the new managerial elite in the governance of higher education," Working Paper, Bishop's University, Sherbrooke, Québec.
92. Wachtel, H.K. (1998), "Student evaluation of college teaching effectiveness: A brief review," *Assessment & Evaluation in Higher Education* 23 (2), 191-211.
93. Ware, J. E., and R.G. Williams (1975), "The Dr. Fox effect: A study of lecturer effectiveness and ratings of instruction," *Journal of Medical Education* 50 (2), 149-156.
94. Wheeler, R. (1935), "Organismic versus mechanistic logic," *Psychological Review* 42, 335-353.
95. Williams, R.G., and J.E. Ware (1977), "An extended visit with Dr. Fox: Validity of student satisfaction with instruction ratings after repeated exposures to a lecturer," *American Educational Research Journal* 14 (4), 449-457.
96. Williams, W.M., and S.J. Ceci (1977), "How'm I doing? Problem with student ratings of instructors and courses," *Change* 29 (5), 13-23.
97. Willis, J. (2000), "A personal response to: Zen and the art of motorcycle maintenance by Robert Pirsig," *Journal of Medical Ethics* 26 (2), 110-112.
98. Wright, R.E. (2006), "Student evaluations of faculty: Concerns raised in the literature, and possible solutions," *College Student Journal*, 40 (2), 417-22.
99. Youmans, R.J., and J.D. Jee (2007), "Fudging the numbers: Distributing chocolate influences student evaluations of an undergraduate course," *Teaching of Psychology*, 34 (4), 245 - 247.

Acknowledgement

The authors thank Stuart McKelvie and Leo Standing for their detailed comments on an earlier draft, while assuming the responsibility for all remaining errors and omissions.