THE UNDERDETERMINATION OF INSTRUCTOR PERFORMANCE BY DATA FROM THE STUDENT EVALUATION OF TEACHING*

[This paper is to appear in Economics of Education Review.]

*Robert Sproule***

Department of Economics
Williams School of Business and Economics
Bishop's University, Lennoxville, Québec, J1M 1Z7, Canada

October 2000

Abstract: This paper presents two arguments. The first is that all models of instructor performance are underdetermined by the student evaluation of teaching data in at least three ways. The second is the obverse of the first -- that the exclusive use of the student evaluation of teaching data in the determination of instructor performance is tantamount to the promotion and practice of pseudoscience, two activities anathema to the academic mission.

Key Words: Faculty performance review, pedagogy, teaching effectiveness, and teaching evaluation.

JEL Classification Code: A0, J44, and K31

---

* *This paper 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 John Damron, Stuart Mckelvie, and two anonymous referees for many useful materials, comments, and critiques.*

** *The author can be reached via e-mail at rsproule@ubishops.ca.*

I. Introduction

The 'underdetermination of a model by data' and 'observational equivalence' are two related notions pivotal to the philosophy of sciences [Curd and Cover (1998), Laudan and Leplin (1991), and Newton-Smith (1978)]. The first connotes: (a) that a given model does not provide a unique or unequivocal explanation or interpretation of a body of evidence, (b) that two or more models can provide equally plausible explanations of the same evidence, and (c) that "the meaning of any set of results is … up for grabs" [Maxwell and Howard (1987, p. 331)]. The second 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)].

The present paper makes the case that any model of instructor performance is underdetermined by the data on the student evaluation of teaching (SET hereafter). Stated differently, any two competing models of instructor performance are observationally equivalent if the evidence is limited solely to the SET data. By making this case, the present paper offers a detailed methodological rationale for a statement made by William Becker (2000), professor of economics at Indiana University and Editor of The Journal of Economic Education. He writes:

> *"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" (p. 114).*

II. The Analytical Framework, And An Outline, Of This Paper

The origins and analytical framework of the present paper is built on the following elements.[1] One, the SET data serve three purposes [Blunt (1991), and Adams (1997)], one of which is to provide student input into faculty evaluation committees (FEC hereafter) whose mission is the adjudication of matters related reappointment, pay, merit pay, tenure, and promotion [Rifkin (1995), and Grant (1998)].

Two, in the SET process, a student is asked to respond on a scale (say from one to five) with his or her rating of entities such as the general quality of lectures and the quality of instruction offered by, and the

[1] For overviews to the literature (in no particular order), see d'Apollonia and Abrami (1997), Greenwald and Gilmore (1997), Marsh (1987), Marsh and Roche (1997), McKeachie (1997), Damron (1995), and Haskell (1997a, 1997b, 1997c, and 1997d).

overall value of, an instructor. So for example, students in a particular course may give their instructor an average score of 2.8 when the university-wide average may be 3.5. The principal question of interest here is this: In the absence of additional information, is there a unique or unequivocal interpretation that one can attach to these data? More to the point, can one conclude that this particular instructor has failed to meet a minimal standard of pedagogical performance? The present paper offers the argument that this latter question cannot be answered by the SET data alone; that is, the hypothesis that 'this particular instructor has failed his or her pedagogical responsibilities' is underdetermined by the SET data. The present paper also asserts (by implication) that any attempt to appraise an hypothesis like 'this particular instructor has failed his or her pedagogical responsibilities' with the SET data in the absence of additional data is tantamount to the promotion and practice of pseudoscience, two activities anathema to the academic mission.[2]

Three, the SET model and the associated FEC decision rule can be fairly represented in the abstract as follows: Let **X** denote a vector of an instructor's characteristics, let Y denote a subjective assessment of his or her teaching effectiveness (as imputed from the SET data on instructor performance), let Z denote an objective assessment of his or her teaching effectiveness, and let S denote some pre-defined minimal standard of performance set by the FEC. The function of the FEC is the determination of whether or not the instructor-in-question passes the minimal standard, S, based on the evidence. Thus, if $S - Y > 0$, the FEC would conclude the instructor fails the minimal standard. Alternatively, if $S - Y \leq 0$, the FEC would conclude the instructor passes.

Four, at this juncture, the question must be asked: What is the rationale for such a decision rule? A reading of the literature suggests that the rationale is built on several presumptions. In the abstract, these are that $Z = Z(Y)$ and $Y = Y(\mathbf{X})$ [or $Z = Z(Y) = Z(Y(\mathbf{X}))$] such that: (a) $Z(Y)$ is (in a local sense) the affine function, $Z(Y;\alpha,\beta) = \alpha + \beta Y$, where $\beta = 1$ [Hands (1991, p. 112)]; and (b) the function $Y(\mathbf{X})$ has only one argument, the vector **X**. [License is taken hereafter by referring to $Y(\mathbf{X})$ as a 'univariate' function.] In words, the rationale is: (a) that the subjective assessment of effectiveness as imputed from the SET data (Y) is a perfect, or near-perfect, proxy for an objective assessment of teaching effectiveness (Z), and (b) the subjective assessment of effectiveness as imputed from the SET data (Y) is determined solely by instructor characteristics (**X**). Thus, under the SET model, the FEC would argue that: (a) 'proof' of teaching effectiveness is uncovered if $S - Y(\mathbf{X}) \leq 0$, since $S - Y(\mathbf{X}) \leq 0$ implies $S - Z \leq 0$, and (b) 'proof' of teaching ineffectiveness is uncovered if $S - Y(\mathbf{X}) > 0$, since $S - Y(\mathbf{X}) > 0$ implies $S - Z > 0$.

[2] One mark of a pseudoscience is the 'grab-bag approach' to evidence – that the sheer quantity of the evidence makes up for any deficiency in its quality [Radner and Radner (1983)].

Finally, it can be argued that the above SET model and the associated FEC decision rule are underdetermined in many ways. The present paper discusses just three, [3,4] and these are as follows. In Section III, it is argued that the formulation, $Y = Y(\mathbf{X})$, is tantamount to the 'consumer' model of education. Because of this, it is argued that the SET model and the associated FEC decision rule do not capture the pedagogical process in its entirety, and they are therefore underdetermined. This is termed 'underdetermination of the first sort.' In Section IV, it is argued that (even if it could be assumed the SET model and the associated FEC decision rule do capture the pedagogical process in its entirety) the claim that that $Z(Y)$ is a locally affine function with unit slope is not supported by the facts. Thus to presume $Z(Y)$ is a locally affine function with a unit slope when it may not be is termed 'underdetermination of the second sort.' In Section V, it is argued that (even if it could be assumed the SET model and the associated FEC decision rule do capture the pedagogical process in its entirety, and $Z(Y)$ is a locally affine function with unit slope) the claim that $Y(\mathbf{X})$ is a 'univariate' function is also not supported by the facts. Thus to use $Y(\mathbf{X})$ is to omit relevant variables, and therefore to use an underdetermined model. This situation is termed 'underdetermination of the third sort.' Summary remarks are offered in Section VI.

III. Underdetermination of The First Sort

The enterprise of pedagogy is a partnership between student and faculty. While it is a partnership of unequals [Platt (1993, p. 31)], it is a partnership nonetheless. A necessary but not a sufficient condition for this collaborative enterprise to function optimally is that both parties have a clear idea of their distinct responsibilities. Those of students are outlined in Ludewig (1992) and Thien (1997). A complete operational model of pedagogy should be able: (a) to articulate the rights and responsibilities of both students and faculty, and (b) to discern observationally (for both groups) the magnitude of the gap between the responsibilities and the actual delivery on those responsibilities.

[3] The forms of underdetermination not discussed in this paper include reporting errors, sample size (rather degrees of freedom), sample-selection bias, reverse causation, and teaching to tests. Two comments are warranted here. One, reporting errors arise from the fact that students complete the SET questionnaire anonymously, and that there is no, and cannot be any, vetting of the data for accuracy. In the jargon of statistics, this is the issue of the 'accuracy of self-reported data.' Undetermined reporting errors represent one form of underdetermination. Two, for materials on sample-selection bias, reverse causation, and teaching to tests in the SET process, the interested reader is directed to Aiger and Thum (1986), Becker and Power (2000), Gramlich and Greenlee (1993), and Nelson and Lynch (1984).

[4] 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. [One anonymous referee reported to me that he or she encountered the very same phenomenon.] 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?'

However, the underlying premise of the SET model and the associated FEC decision rule is that students have only rights, and faculty have only responsibilities – a situation that has been dubbed the 'consumer' model of education [McMurtry (1991), Ritzer (1996), and Rowley (1996)]. And so the FEC decision rule and the SET model of pedagogy would have one compare the performance of the instructor against his or her responsibilities. And because the model presumes that the student has no responsibilities, no such comparison for students is warranted or even possible. This underlying premise represents wrong-headed thinking of the first order.[5] As Bauer (1997) notes: "Teachers can help a bit and they can hinder a bit, but the chief responsibility rests with the learner. The degree and utility of learning is determined almost exclusively by the learner" (p. 26).

These considerations lead one to conclude that the complete pedagogical model is underdetermined by the SET data. As Dowell and Neal (1982, pp. 59-60) note, the rating of an instructor can arise from several motivations, which the SET process does not and cannot control for. They write:

> *"The process of obtaining ratings requires students to perform a complex judgment task. This task involves a self-assessment of the amount of learning that has occurred (much or little) and an attributional analysis of the reasons for the learning (student effort, teacher ability, situational factors). Attribution theory leads us to expect that a variety of situational and student characteristics would bias this process. For example, a low assessment would provoke cognitive dissonance for poorer students who remain in the course until the end of term. Because of their large investment in the course, they would be reluctant to estimate that they have learned little. Hence, one dissonance-reducing strategy poor students may adopt is overestimating the amount learned. On the other hand, poor students might adopt an alternative strategy: attributing their poor performance to external factors such as the instructor. The poor ratings that would result would likely be an underestimate of the teacher's ability. Which of these two dissonance-reducing strategies is adopted by a poor student probably depends upon that student's perception of other students assessment of the teacher and amount learned. Because this analysis of the perspective of poor students identifies influences not relevant to better students, these factors may not only bias ratings but also moderate the validity of student ratings.*
>
> *"There are numerous other factors that might distort the self-assessment and attributional analysis process. One of the more familiar biases in the literature is the 'enthusiasm' bias. Instructor enthusiasm is often the factor most highly correlated with summary ratings suggesting the potency of the bias. What is not clear from the literature is whether enthusiasm merely raises mean ratings...or whether it interacts with student characteristics to influence self-assessments or attributional analyses of good students differently from poor students. If interactions occur, enthusiasm would not only be a bias but a moderator of the validity of student ratings."*

At this juncture, an example may prove helpful. Consider again the case of an instructor who gets a 2.8 for overall performance, and the average is 3.5. The presumption of the SET model and the associated FEC decision rule is that this instructor has failed (in some ill-defined or undetermined way) to deliver on his or her pedagogical responsibilities. The truth is that the SET model and the associated FEC decision rule overlook an equally plausible explanation for these data – that the students-in-question failed to honor

[5] Kolitch and Dean (1999, pp. 37-38) note that while the preconditions for a functional relationship between student and instructor include 'reciprocity' and 'mutuality,' and the SET document predicates this same relationship on adjectives like "minimal, procedural, one-sided, and bureaucratically convenient." As Barrett (1996) notes, the 'consumer' model "grants to the ignorant the right to overrule the knowledgeable" all in the name of educational 'quality' (p. 206). One anonymous referee suggests a hallmark of the 'consumer' model is 'the instructor-as-panderer.'

their responsibilities in the pedagogical partnership. In the eyes of Stone (1995), this situation may arise because "the top priority of most students is to get through college with the highest grades and least amount of time, effort, and inconvenience." Thus,

- *The data-in-question may have arisen in a course for which university Calculus II is a prerequisite. Because the instructor held the legitimate expectation that all students should be able to recall on demand, and apply, the elementary rules of differentiation (covered in Calculus I), and elements of high school algebra (such as the laws of inequalities and exponents), and because students may have held that this expectation was onerous and burdensome, they may have marked the instructor down accordingly.*

- *The data-in-question may have arisen because many students exercised one or more manifestations of their 'consumer rights' like not attending class, refusing to do homework, not reading the textbook, and going so far as not purchasing a copy of the textbook. So after doing poorly on the mid-term examination(s), some students may have blamed the instructor, and not their own failure to exercise self-initiative and due diligence.*[6]

The presumption that students have only rights and faculty have only responsibilities (the presumption that the 'consumer' model of education applies) we term 'underdetermination of the first sort.' This is claimed because a complete operational model of pedagogy is underdetermined by the SET data in the sense that they do not provide a unique or unambiguous insight into the pedagogical performance of the instructor against the responsibilities of the instructor and his or her students.

IV. Underdetermination of The Second Sort

Contrary to Section III, suppose that the SET model and the associated FEC decision rule do capture the pedagogical process in its entirety. A second sort of underdetermination can be defined thusly. As outlined in Section II, the SET model and the associated FEC decision rule assume the relationship between the subjective assessment of effectiveness (Y) and an objective assessment of teaching effectiveness (Z) is perfect or near perfect [that is, $Y \doteq Z$ except for an additive scaling factor]. The evidence suggests the correct specification is otherwise, which indicates again that the SET model and the associated FEC decision rule are underdetermined.

To make the point, consider the statistical analogue of the claim $Y \doteq Z$. Suppose we have data on Y and Z, and these data are partitioned into two classes: 'high value' and 'low value.' Next, suppose that the data are cross-tabulated according to Table 1. And, suppose the correlation coefficient between Y and Z

[6] Such behaviors speak of an attitude predicated on myopia, narcissism, and unreality – an attitude captured by Allan Ginsberg (1955) in his poem, 'America.' He writes: "When can I go into the supermarket and buy what I need with my good looks? America after all it is you and I who are perfect in this world." Reflection suggests that Ginsberg's 'consumer' model should be viewed a forerunner of the 'consumer' model of education.

under these partitions is determined. An analysis of Table 1 leads to three limiting cases defined over the interval, $-1 \leq \beta \leq 1$:

Case 1: In the case in which Correlation (Y,Z) = 1, the subjective measure of teaching effectiveness is a perfect proxy for the objective measure of teaching effectiveness. In this situation, the probabilities of a false negative and a false positive are nil (that is, Cells B and C are zero filled). Thus, under this scenario, an instructor will be effective (in an objective sense) if the subjective measure of teaching effectiveness is high. In summary, this case is the stochastic analogue of: (a) the claim that $Y \doteq Z$ except for an additive scaling factor, and (b) the model $Z(Y;\alpha,\beta) = \alpha + \beta Y$ where $\beta = 1$.

Table 1:
A Cross-Tabulation of
Objective and Subjective Assessments of Teaching Effectiveness

Objective Assessment Of Teaching Effectiveness, Z	Subjective Assessment of Teaching Effectiveness, Y	
	High Value	Low Value
High Value	Cell A: Correct Assessment	Cell B: Error – False Negative
Low Value	Cell C: Error – False Positive	Cell D: Correct Assessment

Case 2: In the case in which Correlation (Y,Z) = 0, the subjective measure of teaching effectiveness is orthogonal to, or independent of, the objective measure of teaching effectiveness. In this situation, the probability of a false negative and false positive is non-trivial (that is, one would expect the data to be uniformly distributed between Cells A and C, and uniformly distributed between Cells B and D. In this situation, the datum that the subjective measure of teaching effectiveness is high provides no informational value in predicting the objective measure of teaching effectiveness, and vice versa. In summary, this case is the stochastic analogue of: (a) the claim that Y and Z are orthogonal, and (b) the model $Z(Y;\alpha,\beta) = \alpha + \beta Y$ where $\beta = 0$.

Case 3: In the case in which Correlation (Y,Z) = -1, the subjective measure of teaching effectiveness is an 'inverted' perfect proxy for the objective measure of teaching effectiveness. In this situation, the probabilities of a false negative and false positive are at a maximum (that is, Cells A and D are zero filled).

Thus, under this scenario, an instructor is likely to be ineffective (in an objective sense) if the subjective measure of teaching effectiveness is high. In summary, this case is the stochastic analogue of: (a) the claim that $Y \doteq -Z$ expect for an additive scaling factor, and (b) the model $Z(Y;\alpha,\beta) = \alpha + \beta Y$ where $\beta = -1$.

The upshot of this analysis is this: In the model outlined in Section II, we asserted that the SET model and the associated FEC decision rule assume that Z(Y) is a locally affine function with unit slope. In a stochastic sense, this is tantamount to stating that Correlation (Y,Z) is close to one in repeated samples. The question posed here is this: Does the balance of evidence support this assumption? This paper argues that the evidence does not. Towards this end, the findings of two studies merit citation and discussion. These are Rodin and Rodin (1972), and Abrami et al. (1990).

In particular, writing in Science, Rodin and Rodin (1972) present a study in which students' performance on a calculus test (viz., an objective measure of teaching effectiveness or Z) was correlated with students' evaluation of the professor (viz., a subjective measure of teaching effectiveness or Y) holding constant students' initial ability in calculus. What they found is that Correlation (Y,Z) = – 0.746, and these variates, Y and Z, accounted for more about half of the variance in the data. How did Rodin and Rodin (1972) interpret their findings? In the last sentence, they state: "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."

There is more to be said about Rodin and Rodin (1972) – this in the context of Table 1. The findings Rodin and Rodin (1972) suggest that their sample lies somewhere between Cases 2 and 3 above; that is, somewhere between the situation in which the subjective measure of teaching effectiveness provides no informational value in predicting the objective measure of teaching effectiveness, and the situation in which the probabilities of a false negative and false positive are at a maximum – which ought to be a sobering thought for the SET advocates.

While there is no disputing that the findings of Rodin and Rodin (1972) cast grave doubt upon the validity of the SET model and hence the FEC decision rule, the next question to be addressed is this: Can the Rodin and Rodin (1972) study be dismissed as a solitary outlier in the broad sweep of research on the validity of the SET process?

There is good evidence that their study cannot be so dismissed.[7] In a recent review of the literature on the validity of the SET in multisection studies, Abrami et al. (1990, p. 222) conclude there is a great

[7] It seems that SET advocates like Maxwell and Howard (1987) would like to dismiss the import of the Rodin and Rodin (1972) study. Commenting on Gaski (1987), Maxwell and Howard (1987) write: "It becomes tiresome to hear about the same old 15-year-old study (Rodin and Rodin, 1972) once again. If one has to introduce it by acknowledging

variation in the magnitude of the correlation coefficient between objective and subjective measurements of teaching effectiveness. They cite as extremes two other surveys of the literature: (a) the McCallum (1984) study which concludes that there is little variation in the correlation coefficient across studies, and (b) the Dowell and Neal (1982, pp. 56-58) study which concludes that variation in the correlation coefficient is the most striking feature of the literature.

V. Underdetermination of The Third Sort

Contrary to Sections III and IV, suppose that the SET model and the associated FEC decision rule do capture the pedagogical process in its entirety, and suppose Z(Y) is a locally affine function with unit slope. A third sort of underdetermination can be defined thusly. As outlined in Section II, the SET model and the associated FEC decision rule are centered on the sign of $S - Y(\mathbf{X})$, and hence on the assumption that $Y = Y(\mathbf{X})$ is a 'univariate' function; that is, the subjective measure of teaching effectiveness (Y) is solely determined by a vector of the instructor's characteristics (**X**).

There is much evidence that this assumption is fallacious -- that in fact, the correct specification should be $Y = Y(\mathbf{X}; \mathbf{V},\mathbf{W})$ where **V** denotes a vector of 'course characteristics' and **W** denotes a vector 'student characteristics.' In $Y = Y(\mathbf{X};\mathbf{V},\mathbf{W})$, it should be noted that: (a) **V**, **W**, and **X** are antecedent vectors, (b) **X** is the sole treatment vector, and (c) both **V** and **W** are control vectors [that is, vectors whose effects on Y are to be excluded when estimating the impact of **X** on Y or when estimating 'teaching effectiveness']. Thus, the pivot of the correct FEC decision rule and the correct SET model is the sign of $S - Y(\mathbf{X};\mathbf{V},\mathbf{W})$. The omission of the control variables in Y [that is, the use of $Y(\mathbf{X})$ instead of $Y(\mathbf{X};\mathbf{V},\mathbf{W})$] is what we term 'underdetermination of the third sort.'

At this juncture, two questions are warranted: One, how to measure course characteristics, **V**, and student characteristics, **W**? Two, what evidence is there that student characteristics and course characteristics affect the subjective measure of teaching effectiveness, Y? A response to both of these follows next.

In a recent study, Mason et al. (1995) defined student characteristics multi-dimensionally as: (i) the reason for taking the course, (ii) class level of the student, (iii) student effort in the course, (iv) expected

---

its now universally known methodological flaws and when one realizes that its findings are at variance with the preponderance of evidence that suggests that student ratings are valid, one wonders why Gaski needs to bring it up in the first place" (p. 332). This statement warrants two comments here. One, while Maxwell and Howard (1987) may be correct in saying "its now universally known methodological flaws," they offer no supporting documentation for this claim. Two, the statement, "the preponderance of evidence .. suggests that student ratings are valid," tells one nothing. As noted in an earlier footnote, the sheer quantity of evidence does not make up for deficiencies in its quality.

grade in the course, and (v) student gender. Mason et al. (1995) also defined course characteristics multi-dimensionally as: (i) course difficulty, (ii) class size, (iii) whether the course is required or not, and (iv) when the course was offered. [Parenthetically, it may be noted that Mason et al. (1995) defined instructor characteristics multi-dimensionally as : (i) the instructor's use of class time, (ii) the instructor's availability outside of class, (iii) how well the instructor evaluates student understanding, (iv) the instructor's concern for student performance, (v) the instructor's emphasis on analytical skills, (vi) the instructor's preparedness for class, and (vii) the instructor's tolerance of opposing viewpoints and questions.]

The body of evidence that the subjective measure of teaching effectiveness is affected by student characteristics and course characteristics, as well as instructor characteristics, includes many studies. We shall cite and discuss just three.[8] These are the studies of Cashin (1990), Rundell (1996), and Mason et al. (1995).

Cashin (1990): In a review of the literature, Cashin (1990) reports that (in the aggregate) students do not provide SET ratings uniformly across academic disciplines. For example, instructors of fine arts and music receive higher subjective assessments of teaching effectiveness than do instructors of chemistry and economics, all things being equal.

Rundell (1996): In a study of teaching evaluations for the Department of Mathematics at Texas A&M University [a study 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).

Mason et al. (1995): 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 are: (a) instructor characteristics, (b) student characteristics, and (c) course characteristics. 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.' That is, in the determination of teaching effectiveness, the influence of student and course characteristics should be removed. Commenting on prior research, Mason et al. (1995, p. 404) note:

> *"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*

[8] For additional discussions of the magnitude of the effect of student and course characteristics, see Cashin (1988), Barrett (1996), Hoyt (1997), Timpson and Andrew (1997), and Hoyt and Pallet (1999).

> *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."*

Using an ordered-probit model,[9] Mason et al. (1995) then confirm that student characteristics, instructor characteristics, and course characteristics impact the subjective assessment of teaching effectiveness. They write:

> *"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) continue:

> *"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)*

VI. Conclusion

This paper has presented the argument that the SET model and the FEC decision rule are underdetermined in (at least) three ways. The first arises from the adoption of the 'consumer' model of education, a model that does not capture the pedagogical process in its entirety. The second arises from the assumption that Z(Y) is a locally affine function with a unit slope when there is evidence that this may not

[9] For an elementary discussion of the ordered-probit model, see Pindyck and Rubinfeld (1991, pp. 273-274).

be so. The third arises from the unsupportable claim that Y(**X**) is a 'univariate' function. The implication of the doubts raised by these arguments is summarized by Barnett (1996) when he writes:

> *"I do not believe we can yet be confident that we know what is being measured by student-completed teaching evaluation questionnaires. .. Any other conclusion seems to me to give insufficient weight to the serious limitations that characterize existing research on the questionnaires. As in the case of a drug that a pharmaceutical company seeks government approval to market, we should insist on a body of credible evidence that the questionnaires are safe and effective before we use them in personnel matters. Based on my review, such evidence does not presently appear to exist." (p. 342)*

In view of the above arguments and the many assessments like Barnett's (1996), it seems that a university presently using the SET model and the FEC decision rule has a choice borne of moral responsibility. The first is to place an embargo on the use of SET data by its FEC. Since this is unlikely, the second choice is for the university to acknowledge publicly and unequivocally that the SET data are contaminated with sizable and incalculable systemic errors, which by implication render the explanatory power of the SET model and the FEC decision rule indeterminable and hence worthless.

The motivation for the above choice is tied to the notion of 'academic honesty,' and the virtue of acknowledging ignorance when the situation permits no more or no less. As Thomas Malthus (1836, p. 14) asserted over a century and half ago:

> *"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."*

References

Abrami, P., S. d'Apollonia, and P. Cohen (1990), "Validity of student ratings on instruction: What we know and what we do not," Journal of Educational Psychology 82 (2), 219-231.

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.

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

Barrett, R. (1996), "'Quality' and the abolition of standards: Arguments against some American prescriptions for improvements in higher education," Quality in Higher Education 2 (3), 201-210.

Bauer, H. (1997), "A response to 'What the numbers mean: Providing a context for numerical student evaluations of courses'," Change, October/November, 26.

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.

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

Cashin, W. (1988), "Student ratings of teaching: A summary of research," IDEA Center Paper No. 20, Center for Faculty Development and Evaluation, Kansas State University.

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).

Curd, M., and J. Cover, eds. (1998), Philosophy of Science: The Central Issues (New York: W.W. Norton).

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.

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

Gaski, J. (1987), "On the construct validity of measures of college teaching effectiveness," Journal of Educational Psychology 79, 326-330.

Ginsberg, A. (1955), "America," Howl and Other Poems (San Francisco: City Light Books).

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, W. (1991), Introductory Mathematical Economics (Lexington, MA: Heath).

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.

Hoyt, D. (1997), "Studies of the impact of extraneous variables," IDEA Center Paper, Center for Faculty Development and Evaluation, Kansas State University.

Hoyt, D., and W. Pallet (1999), "Appraising teaching effectiveness: Beyond student ratings," IDEA Center Paper No. 36, Center for Faculty Development and Evaluation, Kansas State University.

Kolitch, E., and A. Dean (1999), "Student ratings of instruction in the USA: Hidden assumptions and missing concepts about 'good' teaching," Studies In Higher Education 24 (1), 27-42.

Laudan, L., and J. Leplin (1991), "Empirical equivalence and underdetermination" Journal of Philosophy 88, 449-472.

Ludewig, L. (1992), "Ten commandments for effective study skills," Teaching Professor, December.

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.

Maxwell, S., and G. Howard (1987), "On the underdetermination of theory by evidence," Journal of Educational Psychology 79, 331-332.

McCallum, L. (1984), "A meta-analysis of course evaluation data and its use in the tenure decision," Research In Higher Education 21, 150-158.

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

McMurtry, J. (1991), "Education and the market model," Journal of Philosophy of Education 25, 209-217.

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

Newton-Smith, W. (1978), "The underdetermination of theory by data," Aristotelian Society (Supplement) 52, 71-91

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.

Radner, D., and M. Radner (1983), Science and Unreason (Belmont, CA: Wadsworth Publishing).

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

Ritzer, G. (1996), "McUniversity in the postmodern consumer society," Quality in Higher Education 2 (3), 185-199.

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

Rowley, J. (1996), "Measuring quality in higher education," Quality in Higher Education 2 (3), 237-255.

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

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

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), 1-33.

Thien, S. (1997), "Successful students: Guidelines and thoughts for academic success," unpublished manuscript, Department of Agronomy, Kansas State University.

Timpson, W., and D. Andrew (1997), "Rethinking student evaluations and the improvement of teaching: Instruments for change at the University of Queensland," Studies In Higher Education 22 (1), 55-65.