The underdetermination of instructor performance by data from the student evaluation of teaching

Robert Sproule *

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

Received 22 June 2000; accepted 13 October 2000

Abstract

This paper presents two related arguments. The first is that all models of instructor performance are underdetermined by the student evaluation of teaching data. 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. © 2002 Elsevier Science Ltd. All rights reserved.

JEL classification: A0; J44; K31

Keywords: Faculty performance review; Pedagogy; Teaching effectiveness; Teaching evaluation

1. Introduction

The terms, ‘underdetermination of a model by data’ and ‘observational equivalence’, are central to the philosophy of sciences (Curd & Cover, 1998; Laudan & Leplin, 1991; Newton-Smith, 1978). They apply: (i) when a given model does not provide a unique or an unequivocal explanation of a body of data; (ii) when two or more models provide equally plausible explanations of the same data; or (iii) when “the meaning of any set of results is… up for grabs” (Maxwell & Howard, 1987, p. 331). In such instances, other information is needed to identify which explanation is correct (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), and in doing so it offers a detailed methodological rationale for a statement made by William Becker (2000). 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).

2. The analytical framework, and an outline, of this paper

The analytical framework of the present paper is built on the following elements. One, the SET data serve three purposes (Blunt, 1991; Adams, 1997), one of which is the provision of student input into faculty evaluation committees (FEC hereafter) — a committee whose mission is the adjudication of matters related to reappointment, pay, merit pay, tenure, and promotion (Rifkin, 1995; Grant, 1998).
Two, in the data-collection process, a student is asked to rate (on a scale, say from one to five) entities such as the general quality of lectures, as well as the quality of instruction offered by, and the 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 is this: can one attach to this sample point from the population of all instructors of size \( n \), and consider any one instructor \( i \) passes \( S \) based on the evidence. Thus, for example, if \( S-Y_i>0 \), then the FEC test indicates that the instructor-in-question fails the minimal standard. Alternatively, if \( S-Y_i\leq0 \), the FEC test indicates the instructor passes.

Four, at this juncture, the question must be asked: what is the rationale for such a decision rule? An answer to this question requires the following preliminary statements.

**Assumption 1:** Let \( Y_i \) and \( u_i \) be two independent random variables where \( i=1,n \). And let \( Y_i \) and \( u_i \) be related according to \( Z_i=\alpha+\beta Y_i+u_i \).

**Lemma 1 (Sproule & Sproule, 2001).** If Assumption 1 holds, then
\[
\rho(Z,Y) = \frac{\beta \sigma_Y^2}{\sqrt{(\beta^2 \sigma_Y^2 + \sigma_u^2) \sigma_Y^2}}
\]
where \( \rho(Z,Y) \) denotes the correlation coefficient between \( Y_i \) and \( Z_i \), where \( \sigma_Y^2 \) denotes the variance of \( Y_i \), and where \( \sigma_u^2 \) denotes the variance of \( u_i \).

**Lemma 2:** If Assumption 1 holds, if \( u_i=0 \) for all \( i \), and if \( \beta \neq 0 \), then
\[
\rho(Z,Y) = \frac{\beta}{|\beta|}.
\]

**Proof:** If \( u_i=0 \) for all \( i \), then \( \sigma_u^2=0 \), and by Lemma 1,
\[
\rho(Z,Y) = \frac{\beta \sigma_Y^2}{\sqrt{\beta^2 \sigma_Y^2 + \sigma_u^2} \sigma_Y^2} = \frac{\beta}{|\beta|}.
\]

**Lemma 3:** If Assumption 1 holds, if \( u_i=0 \) for all \( i \), and if \( \beta \neq 0 \), then: (a) \( \beta>0 \) implies \( \rho(Z,Y)=1 \); and (b) \( \beta<0 \) implies \( \rho(Z,Y)=-1 \).

**Proof:** Lemma 2.

**Lemma 4:** If Assumption 1 holds, if \( u_i \neq 0 \) for some \( i \), and if \( \beta=0 \), then \( \rho(Z,Y)=0 \).

**Proof:** Lemma 1.

**Lemma 5:** Suppose \( Y_i \) and \( Z_i \) represent population variables drawn from any distribution. The percentage of the variance in \( Y_i \) and \( Z_i \) that is explained by the model, \( Z_i=\alpha+\beta Y_i+u_i \), is given by the square of the correlation coefficient between \( Y_i \) and \( Z_i \).

---

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 & Radner, 1983).

3 The present use of the notion of a university-specific population can be seen to suppress two legitimate concerns. The first is the fact that a university-specific FEC process is not a closed system, and therefore a university-specific population can and should be viewed as a sample point from the population of all universities. The second is that within any given time period the university-specific population may only be sampled. Therefore, in defense, the notion of a university-specific population should be viewed as a fictional convenience, whose use is merely intended to simplify the present discussion.

4 In the case of a sample, this lemma should be rephrased as: “Suppose \( Y_i \) and \( Z_i \) are random variables drawn from a bivariate normal distribution. Then the percentage of the variance in \( Y_i \) and \( Z_i \) that is explained by the model, \( Z_i=\alpha+\beta Y_i+u_i \), is given by the square of the maximum likelihood estimator of \( \rho(Z,Y) \).” For more, see Sproule and Sproule (2001).
Next, return to the original question: What is the rationale for the FEC decision rule? A reading of the literature suggests that it is built on several (and, as will be argued below, indefensible) assumptions. These are:

**Assumption 2**: (a) \( \beta > 0 \); and (b) \( u_i \) is ‘acceptably small’ (viz. \( u_i \equiv 0 \)) for all \( i \).

**Assumption 3**: \( Y_i \) is determined solely by \( X_i \) (viz. \( Y_i = f(X_i) \)).

Two comments about Assumptions 2 and 3 are warranted. First, it would appear (at least in some limited sense) that Assumptions 2(a) and (b) are synonymous with the methodological terms of ‘validity’ and ‘reliability’. Second, two implications follow immediately from Assumptions 2 and 3, and these in tandem are tantamount to the implicit theoretical model underpinning the FEC decision rule. In particular,

**Proposition 1**: If Assumptions 2 and (b) hold, then \( \rho(Z,Y) \equiv 1 \).

**Proof**: Lemma 3. \( \square \)

**Proposition 2**: If Assumptions 2 and 3 hold, then: (i) ‘indisputable proof’ of teaching effectiveness is uncovered if \( S = (\alpha + \beta Y(X_i)) \leq 0 \); and (ii) ‘indisputable proof’ of ineffectiveness is uncovered if \( S = (\alpha + \beta Y(X_i)) > 0 \).

It can be argued that the model articulated by Propositions 1 and 2 is underdetermined in many ways. The present paper discusses just two,\(^5\),\(^6\) which are as follows.

---

\(^5\) The forms of underdetermination not discussed in this paper include reporting errors, inadequate sample size (rather degrees of freedom), the presence of sample-selection bias, the presence of reverse causation, and the 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 vetting for the accuracy of these self-reported data. Such undetermined reporting errors represent one form of underdetermination. Two, for materials on sample-selection bias, reverse causation, and the teaching-to-tests, in the SET process, the interested reader is directed to Aiger and Thum (1986), Becker and Power (2001), Gramlich and Greenlee (1993), Krautmann and Sander (1999) and Nelson and Lynch (1984).

\(^6\) A personal vignette provides some insight into the potential seriousness of the inaccuracy of self-reported data. In the fall of 1997, the author 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 a 5, 27.7% gave a 4, and 22.2% gave a 3. (One anonymous referee reported that he or she encountered the very same phenomenon.) The import of this? If such self-reported measures of this objective metric are inaccurate (as this case indicates), how can one be expected to trust the validity of subjective measures like ‘teaching effectiveness’?

---

In Section 3, it is argued that Proposition 1 is not supported by the facts. This paper therefore infers that Assumptions 2(a) and (b) are fallacious, and their use gives rise to what is termed ‘underdetermination of the first sort’, the focus of Section 3. Section 4 inspect the plausibility of Assumption 3 (viz. the plausibility of the specification, \( Y_i = f(X_i) \)). This paper then argues that available data do not support Assumption 3, but support an alternative formulation (viz. \( Y_i = f(X_i, V_p, W_p) \)), where \( V_p \) denotes a vector of characteristics of the slate of courses taught by instructor \( i \) and \( W_p \) denotes a vector of characteristics of the students found in the slate of courses taught by instructor \( i \).\(^7\) This paper concludes therefore that Assumption 3 is fallacious, and its use gives rise to what is termed ‘underdetermination of the second sort’, the focus of Section 4. Summary remarks are offered in Section 5.

### 3. Underdetermination of the first sort

Recall Assumptions 2(a) and (b): \( Z = \alpha + \beta Y + u \), where \( \beta > 0 \), and \( u_i \equiv 0 \). Proposition 1 states that if Assumptions 2(a) and (b) hold, then \( \rho(Z,Y) \equiv 1 \). To develop the intuition and import of Proposition 1 more fully, reconsider Lemmas 3 and 4 in a new light.

**Proposition 3**: (a) If \( \beta > 0 \) and \( u_i \equiv 0 \) for all \( i \), then the probabilities of a false negative and a false positive are nil. (b) If \( \beta = 0 \) and \( u_i \equiv 0 \) for some \( i \), then the probabilities of a false negative and false positive are non-trivial and uniformly distributed across \( Z \). (c) If \( \beta < 0 \) and \( u_i \equiv 0 \) for
all i, then the probabilities of a false negative and a false positive are at a maximum.

Proof: (a) If $\beta>0$ and $u_i=0$ for all i, then $\rho(Z,Y)=1$ (Lemma 3) and the probabilities of a false negative and a false positive are nil. (b) If $\beta=0$ and $u_i\neq 0$ for some i, then $\rho(Z,Y)=0$ (Lemma 4) and the probabilities of a false negative and false positive are non-trivial and uniformly distributed across Z. (c) If $\beta<0$ and $u_i=0$ for all i, then $\rho(Z,Y)=-1$ (Lemma 3) and the probabilities of a false negative and a false positive are at a maximum. □

The analytical value of this proposition is as follows. Part (a) is analogous to one of two components in the FEC decision rule (Proposition 1). The question to be raised here is: Does the balance of evidence support both Proposition 1 and part (a) of Proposition 3? That is, does the balance of evidence support the intermediate claim that $\rho(Z,Y)=1$ or $\rho(Z,Y)\approx 1$? This paper argues that the evidence does not. Towards this end, the findings of two studies merit discussion. These are Rodin and Rodin (1972) and Abrami et al. (1990).

For the sake of argument, assume for the moment that Assumption 2(b) (viz. $u_i\equiv 0$ for all i) holds, and then consider some of the evidence related to Assumption 2(a). In particular, using data on student’s performance on a calculus test (viz. an objective measure of teaching effectiveness or Z) and students’ evaluation of the professor (viz. a subjective measure of teaching effectiveness or Y), Rodin and Rodin (1972) found: (i) that the correlation between Y and Z (holding constant students’ initial ability in calculus) was $-0.746$; and (ii) that these variates, Y and Z, accounted for more than half of the variance in the data. How did Rodin and Rodin (1972) interpret their findings? 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” (p. 1166).

What are the wider implications of Rodin and Rodin (1972)? Their findings suggest that parts (b) and (c) of Proposition 3 apply. That is, their data suggest that the value of the subjective measure of teaching effectiveness lies in the interval defined by these two limiting statements: (a) a subjective measure of teaching effectiveness offers no information in predicting an objective measure of teaching effectiveness; and (b) the use of a subjective measure of teaching effectiveness as a proxy for an objective measure of teaching effectiveness maximizes the probability of a false negative and false positive.

That said, one may ask: Should one dismiss Rodin and Rodin (1972) as a solitary outlier in the body of existing research?9 Stated differently, with the exception of the Rodin and Rodin (1972) study, does the research-to-date show that the correlation coefficient between Y and Z is close to unity? The answer to the last question is, no. What is the rationale for this? In a recent review of the literature on the validity of the SET in multisection studies, Abrami et al. (1990, p. 222) conclude that the distinguishing feature of SET research is not that the correlation of Y and Z is close to unity, but that there is great variation in the magnitude of this correlation coefficient.9

This assessment of Abrami et al. (1990) suggests not only that Assumption 2(a) is erroneous (that in reality $\beta>0$), but that Assumption 2(b) is also erroneous (that in reality $u_i$ is not ‘acceptably small’ for all i). Taken together, their assessment leads to the conclusion that Assumption 2 (viz. the maintained hypothesis that $Z=\alpha+\beta Y+u_i$ where $\beta>0$, and where $u_i\equiv 0$ for all i) is simply not supported by the facts.

4. Underdetermination of the second sort

In Section 2, it was stated a second assumption of the FEC decision rule is that $Y=Y(X)$. In the present section, it is argued: (a) that $Y=Y(X)$ can be labeled the ‘consumer model of education’ (McMurtry, 1991; Ritzer, 1996; Rowley, 1996); and (b) that the use of $Y=Y(X)$ is fallacious because it leads to multiple interpretations of the SET data, and hence to ‘underdetermination of the second sort’.

To initiate discussion, reconsider the case of the instructor who obtained a 2.8 for overall performance when the average is 3.5. Suppose the datum-in-question (viz. the 2.8) arose in a course for which university Calculus II is a prerequisite. Suppose the instructor held the (reasonable) expectation that all students should be able

---

9 As one might expect, SET advocates dismiss the importance of the Rodin and Rodin (1972) study. For example, in their comments on Gaski (1987), Maxwell and Howard (1987) write: “It becomes tiresome to hear about the same old 15-year-old study (Rodin & Rodin, 1972) once again. If one has to introduce it by acknowledging 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 last statement warrants two comments. One, while Maxwell and Howard (1987) may be correct in citing “its now universally known methodological flaws,” they offer no argument nor documentation in support of their claim. Two, their statement, “the preponderance of evidence… suggests that student ratings are valid”, tells one nothing. As noted in 2 above, the sheer quantity of evidence does not make up for deficiencies in its quality.

9 Their finding jibes with that of an earlier survey (Dowell & Neal, 1982, pp. 56–58), and contradicts that of another (McCallum, 1984).
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 logarithms, inequalities, and exponents). Finally, suppose students viewed the instructor’s expectation as onerous. One might well ask: Was the SET datum caused by teaching ineffectiveness, or by an inappropriate clash of expectations? Next, suppose the datum-in-question arose because many students in this class exercised one or more of their ‘consumer rights’ like not attending class, refusing to do homework, not reading the textbook, or going so far as not purchasing a copy of the textbook.10,11 Suppose too many students did poorly on the mid-term examination(s). Again, one might well ask: Was the SET datum caused by teaching ineffectiveness, or by students inclined to blame their poor performance on the instructor, and not on their own failure to exercise self-initiative and due diligence?

These two examples serve as an intuitive argument for the inclusion of \( V_y \) and \( W_y \) in \( Y(\cdot) \). To develop this argument further, consider the following:

- A key argument for the inclusion of \( W_y \) is that the enterprise of pedagogy is a partnership (of unequals) between students and faculty (Platt, 1993, p. 31). Clearly, a necessary but not a sufficient condition for this partnership to function (optimally) is that both parties have a clear idea of their respective responsibilities. (Those of students, for example, are outlined in Ludewig (1992) and Thien (1997).) Thus, the correct specification of \( Y(\cdot) \) requires that \( W_y \) contain proxy measures of students’ contributions to this partnership, and that \( W_y \) so measured be included. 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).12

- Arguments for the inclusion of \( V_y \) can be found in the following two studies. (i) Cashin (1990) reports that (in the aggregate) students do not provide SET ratings uniformly across academic disciplines. (ii) Using data on teaching evaluations from the Department of Mathematics at Texas A&M University, Rundell (1996) assesses the correlation coefficients for arrays of variables measuring ‘teaching effectiveness’ and ‘course characteristics’, and concludes that: “(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).13

To sum up, the inclusion of \( V_y \) and \( W_y \) in \( Y(\cdot) \) permits the removal of their contaminating effects, and therefore permits the isolation of the effect of \( X_\alpha \) on \( Y \). The general case for this specification is made clearly by Mason et al. (1995). They write: “A… virtually universal problem with previous research is that the overall rating is viewed as an effective representation of comparative professor value despite the fact that it typically includes assessments in areas that are beyond the professor’s control. The professor is responsible to some extent for course content and characteristics specific to his/her teaching style, but is unable to control for student attitude, reason for being in the course, class size, or any of the rest of those factors categorized as student or course characteristics above. Consequently, faculty members should be evaluated on a comparative basis only in those areas they can affect, or more to the point, only by a methodology that corrects for those influences beyond the faculty member’s control.” They continue, “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” (p. 404). Mason et al. (1995) conclude: “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

---

10 This last sentence squares with the observation that “the top priority of most students is to get through college with the highest grades and least amount of time, effort, and inconvenience” (Stone, 1995).
11 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’ could be taken as a forerunner of the ‘consumer model of education’.
12 Likewise, Barrett (1996, p. 206) notes that the ‘consumer model of education’, and the SET decision rule, “grants to the ignorant the right to overrule the knowledgeable” all in the name of educational ‘quality’. One anonymous referee suggests a hallmark of both is ‘the instructor-as-panderer’.
13 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).
pedagogical tools, peer assessments, and administrative visits” (p. 414).

Mason et al. (1995) extend their analysis in two regards. One, using an ordered-probit model, they demonstrate that \( X, V, \) and \( W \), impact significantly a subjective assessment of teaching effectiveness, \( Y \). Two, Mason et al. (1995) ranked the instructors in their study after removing of the contaminating effects of \( V, \) and \( W, \) on \( Y \). In a comparison of the ranking obtained from this ordered-probit model with the ranking generated from the unstandardized SET data, they discovered several discrepancies. Mason et al. (1995) concluded that: “raw student evaluations at the margin can lead to inappropriate conclusions regarding the relative performances of faculty” (pp. 413–414). This last statement can be reformulated more formally as follows.

**Proposition 4:** Consider a pair of instructors, 1 and 2. Suppose that raw SET readings are obtained for \( Y_{1}(X_{1}), Y_{2}(X_{2}) \), and that corrected SET readings are also obtained for \( Y_{1}(X_{1},V_{1},W_{1}), Y_{2}(X_{2},V_{2},W_{2}) \). It does not follow that if \( Y_{1}(X_{1},V_{1},W_{1})=Y_{2}(X_{2},V_{2},W_{2}) \) is observed, then \( Y_{1}(X_{1})=Y_{2}(X_{2}) \) will be observed also.

**Proof:** Informal logic, and the findings of Mason et al. (1995, pp. 413–414). □

5. Conclusions

This paper has presented the argument that the FEC decision rule is underdetermined by the data in at least two ways. The first sort arises from the erroneous claim that \( Z \), and \( Y \), are related as \( Z=aX+BY+u, \) where: (a) \( \beta > 1 \); and (b) \( u, \) is ‘acceptably small’ for all \( i \). The second arises from the use of \( Y=Y_{i}(X_{i}), \) rather than \( Y_{i}=Y_{i}(X_{i},V_{i},W_{i}). \)

The implications of the above argument cast serious doubt about the methodological soundness of the FEC decision rule. Expressions of similar doubts can be found in the aforementioned statement by Becker (2000), and by Barnett (1996) who 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

14 For an elementary discussion of this, see Pindyck and Rubinfeld (1991).

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 present assessment, it would seem appropriate and reasonable for all constituencies of the university community to come clean by acknowledging publicly and unequivocally that the SET data are contaminated with non-trivial, and incalculable, systemic errors, and that the presence of these errors render the FEC decision rule invalid, unreliable, and otherwise hopelessly flawed.15

**Acknowledgements**

An earlier draft of 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, Leo Standing, two anonymous referees, and two members of the editorial board, for many useful materials, comments, and critiques.

**References**


15 This last assessment stems from the notion of ‘academic honesty’, and the virtue of acknowledging ignorance when the situation permits no more or no less. As Thomas Malthus (1836) 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” (p. 14).


