Does the student evaluation of teaching instrument really measure instructors’ teaching effectiveness? An econometric analysis of students’ perceptions in economics courses

Mohammad Alauddin* and Temesgen Kifle

School of Economics, The University of Queensland
Brisbane, Australia 4072

Abstract

While the student evaluation of teaching (SET) has been an intensely researched area in higher education there has been little research using the individual student responses on their perceptions of instructors' effectiveness (TEVAL) score. This research delivers a methodological breakthrough as it fills this gap by employing individual student responses from an elite Australian university and partial proportional odds model to investigate the influence of students' perceptions of instructional attributes included in the SET instrument and other variables on TEVAL score.

The findings indicate clear presentation and explanation, and well-organized classes were key determinants of TEVAL scores. Emphasis on thinking rather than memorizing was less influential. Intermediate level courses and non-English speaking background instructors received lower ratings. Elective courses and instructors below associate professor attracted higher ratings. SET instrument currently used fails to provide a valid measure of teaching quality as it does little to measure the extent of students' actual learning. This paper underscores the need for incorporate variables typifying diversity of student population including academic performance, discipline destination, ethno-linguistic background, age, sex, indicators of students’ effort. It raises broader implications such as subscales, inclusion of items on course contents, intellectual challenge, real world applications, and problem-solving skills.

Keywords: Higher education, Instructional attributes, Partial proportional odds model, Pragmatism, Student diversity, Teaching effectiveness.

JEL Classification: A2, C49, I2

* Corresponding Author (Email: m.alauddin@uq.edu.au; Fax: +61 7 3365 7299)
The practice of student evaluation of teaching in higher education is so widespread that it is impossible to contemplate a university academic’s life without it. As the end of each semester approaches, millions of students globally complete the SET instrument. Several reasons underlie the rationale of such evaluations (Kember et al., 2002, pp. 411-412; Denson et al., 2010, pp. 339-340).

Firstly, the feedback through SET is likely to enhance teaching quality assuming that instructors take note of their strengths and weaknesses to make appropriate changes to their teaching methods.

Secondly, SET ratings are critically important for personnel decisions such as promotion and tenure. These exercises are supposed to enhance teaching quality as they provide an inducement to academics to improve their teaching or else lose their job. This is analogous to providing a counter example to Gresham’s law of bad money driving good money out of circulation\(^1\). In this instance, ‘good’ teachers or practices crowd out ‘bad’ teachers or practices in the battle for survival of the fittest in an intensely competitive environment.

Thirdly, in a globalised environment dominated by a burgeoning rating ‘industry’ the universities are required to demonstrate to the outside world their seriousness about improving and sustaining teaching standards. Furthermore, the provision of a substantial amount of government funding for universities and increasing university student numbers have resulted in demands for greater accountability for their teaching and other activities.

Fourthly, SET provides a source of data for research on teaching and related issues (Marsh, 1987).

---

\(^1\) The New Encyclopædia Britannica (1986, 15th Edition, vol. 5, p.489) traces the origins of this monetary principle as follows: “… if two coins have the same nominal value but are made from metals of unequal value, the cheaper will tend to drive the other out of circulation. Sir Thomas Gresham, financial agent of Queen Elizabeth I, was not the first to recognize this monetary principle, but his elucidation of it in 1558 prompted the economist H.D. Mcleod to suggest the term “Gresham’s law” in the 19th century. ….”
Notwithstanding the above, it is not clear the extent to which instructors make significant use of these evaluations to improve their teaching. Nasser and Fresko (2002, pp. 193-194) reported that only 2.5% to 10.3% of the faculty members reported having made significant changes because of SET. Course assignments, course organisation and instructional strategies represented the most important areas of change while grading procedures and the level of course demands rarely featured in the changes.

While SET has been one of the most intensely researched areas in the field of higher education, detailed quantitative studies of the relationship between instructional attributes used in SET and TEVAL scores have been rare (see e.g., DeCanio, 1986; Boex, 2000; Alauddin & Tisdell, 2010). DeCanio (1986), and Alauddin and Tisdell (2010) used individual responses while Boex (2000) used class average data. DeCanio (1986) a multinomial logit model while Alauddin and Tisdell (2010) and Boex (2000) used an ordered probit model. However, none of these studies used a partial proportional odds model to account for any violation of the parallel lines assumption (see Section 3.2).

The present study uses the frontier literature on the issue from an array of leading journals from the broad disciplines, education and economics. The array of education journals include *American Psychologist, Assessment and Evaluation in Higher Education, International Journal of Educational Research, and Review of Educational Research* while those in economics include *Applied Economics, Economics of Education Review, Journal of Economic Perspectives, Journal of Political Economy, and Review of Economics and Statistics*. The vast literature indicates that there are potential biases in SETs. Sources of bias in SETs are located in information asymmetry; insufficient basis to judge; course quality, easiness and grade inflation; and comprehensiveness and validity.
This research investigates the influence of students’ perceptions of instructional attributes included in the SET instrument on TEVAL scores. Furthermore, it investigates the influence of variables including course level, instructor status, and instructors’ linguistic background. The investigation, however, has been limited by the lack of available information typifying the diversity of the student population on student attributes. More specifically, this paper uses data from 10,223 SET forms completed for various economics courses at an elite Australian university to:

1. identify instructional attributes that result in statistically significant and substantive variations in TEVAL scores and those that do not;

2. investigate whether their effects on TEVAL scores vary across different levels of courses; and

3. examine the implications of major findings.

The paper is organised as follows: Section 2 reviews the literature while Section 3 presents an interpretive overview of the main features of the data and outlines the analytical framework. Section 4 presents the empirical results while Section 5 discusses and presents their implications. Section 6 provides some further observations. Finally, Section 7 concludes the paper.

2 A REVIEW OF LITERATURE

Several strands of literature have dominated the discourse on the validity or otherwise of the SET as a procedure to determine the effectiveness of an instructors’ teaching or pedagogic standards. The staunch supporters of the procedure include Marsh (1987, 2007), Marsh and Roche (1997) and Feldman (1997, 1998). Marsh (1987, p. 255) derived several broad generalisations about the value of SET including multidimensionality; reliability, stability and
validity; and formative (feedback for the improvement of teaching), and summative (administrative decision-making) utilities. Marsh and Roche (1997, p. 1190) claimed that “the voluminous literature on potential biases in SETs is frequently atheoretical, methodologically flawed, and not based on well-articulated operational definitions of bias, thus continuing to fuel (and be fuelled or fooled by) SET myths.”

Studies that question the validity of SET in measuring teaching effectiveness tend to do so on a number of grounds. This strand of research encompasses a spectrum of views that the remainder of this section briefly presents.

**Information asymmetry:** Students are not fully informed customers (Mason, et al., 1995). Students may not immediately realize the significance of a teaching program but could do so later in life, for example, after the completion of their degrees, subsequent higher studies or well into their experience in the workforce (Alauddin & Tisdell, 2000). In this context, it is worth noting the recollection of a current (2013) PhD student at an elite Australian university:

… It was quite a long time ago (2006), when I was still in my second year of undergraduate for the Bachelor of Economics degree. By that time, I was used to the way of studying and doing assessments in undergraduate courses, so I had no trouble following different courses. However, this course was the first one that had journal article review as a form of assessment. All courses before that, as I recall, only had MCQs, problem sets or essays. Back then, I gave no thought to the benefits of doing journal article reviews instead of other types of assessments, in general; or the benefits especially for me when I became a PhD student in Economics five years later, in particular. To me, they were just assessments, and my duty was to complete them as well as possible. Nevertheless, it was new to every other student, not just me, and challenging as well, since you were required not only to understand the articles (to summarize them), but also give comments, even criticize them and put all those in a piece of writing. Now, looking back, I find that these journal-article-review assessments prepared me very well for the research work that I am currently doing. Actually reviewing articles is a skill that, in my opinion, is necessary for many different types of work other than research since it involves expressing your own critical thinking on a given issue. Therefore, I would like to thank the lecturer for introducing a new form of assessment when I myself even did not know how much I should appreciate at the time. …
The economics literature has widely documented the consequences of asymmetry of information since the pioneering work of the 2001 Economics Nobel Laureate George Akerlof (Akerlof, 1970). Imperfect and asymmetric information leads to drastic differences in the nature of market equilibrium (Varian, 2010, p.738) and to market failure.

**Inadequate basis to judge:** SET excludes a number of important background variables that could potentially influence TEVAL scores. These lead to underdetermination of the instructor performance by SET (Laudan & Leplin, 1991). Mason et al. (1995), Sproule (2002) and Wachtel (1998) suggested inclusion of a range of variables including (i) characteristics associated with the administration of the SET, (ii) characteristics of the instructor, (iii) characteristics of the students, (iv) course attributes, and (v) reaction to the dissemination and use of SET. Sproule (2002, p. 289) provided a mathematical proof of the underdetermination of instructor performance by the SET data. Becker (2000) considered heavy reliance on SET to be problematic on a number of grounds. As Becker (2000, p. 114) put it:

> … 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 the class, with paid staff involved only in the processing of 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.…

**Course quality, easiness and grade inflation:** Felton et al. (2008) found a strong positive association among quality, easiness and ‘hotness’ of courses using web-based SET. It should be noted, however, that Felton et al. (2008) made use of data that came from the website ‘ratemyprofessors.com’ which suffered from a major selection bias concerning the respondents participating in these ratings. There is a positive association between grades given and SET ratings. A teacher who is generous in grading is likely to receive a higher SET score (Johnson 2003). Johnson (2003; see also Felton et al. 2008, p. 47) suggested that ‘showmanship’ rather
than content can have a substantial impact on SET scores (cf. Greenwald & Gilmore, 1997). As stated by McKeachie (1997, p. 1219):

… Many students prefer teaching that enables them to listen passively – teaching that organizes the subject matter for them and that prepares them well for tests. … Cognitive and motivational research, however, points to better retention, thinking, and motivational effects when students are more actively involved in talking, writing and doing. …Thus, some teachers get high ratings for teaching in less than ideal ways.

Mason et al. (2003) also supported the views of McKeachie (1997). Employing a utility maximisation model Mason et al. (2003) demonstrated why students preferred easier courses (a grade maximisation strategy) to those providing greater intellectual challenge (a knowledge maximisation strategy). More recently, Valsan and Sproule (2008, p.944) suggested:

… The mere fact that an implicit threat exists is enough to change the attitude of the professor. The professor, whose career depends in part on her teaching scores, is compelled to re-evaluate her strategy. Would she deliver her lectures and assignments according to her best judgment; or, would she adjust her classroom agenda to maximize her teaching scores?

The above discussion seems to suggest a dichotomy between intellectually challenging course content and high student rating of teachers. On the other hand, the ultimate determinant of the quality of teaching (and by implication, the teacher) is content, not delivery. Describing his own experience at the University of Columbia, Colander (2004, p.64) recalled:

… When I think back to those teachers with great delivery and lousy content and those with great content and lousy delivery, it is the ones with content whom I remember – and the ones who convinced me that what they were doing was important. John Rawls, William Vickrey and Edmund Phelps all had horrendous delivery, but had great content, and changed my life²…

Jewell et al. (2013) identified significant grade inflation at a large American university over a period of two decades since 1984-85. Jewell et al. (2013, p. 1199) attributed the grade inflation to two factors. Firstly, universities have to compete for students and public funding to sustain

² William Vickrey (1914–96) and Edmund Phelps (b.1933) won Nobel Prize in economics in 1996 and 2006 respectively.
their revenue base. Secondly, at the instructor level, grade inflation may in part be due to “universal use of student evaluation scores as inputs into tenure, promotion and merit decisions.” Koper et al. (2014) provided similar evidence of the link between SET and grade inflation (see also Galbraith et al., 2012; Worthington, 2002).³ Koper et al. (2014) stated that many students “are learning less as they study less” (see Arum & Roksa, 2011). Koper et al. (2014) also suggested that grade inflation is real and has a high correlation to SET scores. Significant manifestations of this phenomenon exist in disciplines that draw students of lower academic ability.

**Comprehensiveness and validity:** Although the majority of research shows that SET provides useful information to both teachers and administrators (Marsh, 1987), its validity remains open to question (Clayson, 2009; Engdahl et al., 1993; Hassan, 2009; Kember & Leung, 2008). As Woodhouse (2008) argued that when, choosing indicators, ranking should be based on what is relevant and important to student learning – not what one can be easily measured.

SET instruments measure students’ attitudes toward effective teaching and should be seen as a latent construct. This construct is not immediately observable using a single-item approach that sometimes results in highly stable and reliable estimates (Ginns & Barrie, 2004). It also assumes that all aspects or dimensions of teaching quality can be observed unequivocally (Ginns & Barrie, 2004, p.1029). Spooren et al. (2007) challenged such a view, identified inadequacies of SET and suggested the need for the development of a more comprehensive instrument of evaluation of teaching effectiveness. Spooren et al. (2007, p. 667) criticized it for being “… single-item type of evaluation. ... Most of these instruments lack a theoretical foundation and hardly any instrument was tested with modern tests of reliability and validity” (cf. Marsh, 1987). In a similar vein, Stark-Wroblewski et al. (2007, p. 403) argued that

³ Babcock and Marks (2011, p. 477) found a declining trend in student effort over several decades. The average time allocated by full-time students declined by 32.5% from about 40 hours in 1961 to 27 hours per week in 2003.
“without clear guidelines regarding how to best document effective teaching, faculty members may wonder how to convincingly demonstrate teaching effectiveness ….”. Bursdal and Harrison (2008, p.574) went further by stating that SETE (student evaluation of teaching effectiveness) “ … instruments measure students’ attitudes to teaching effectiveness, not necessarily teaching effectiveness per se. … A reliable and valid measure of a particular group’s perceptions regarding effectiveness is not the same thing as having a valid measure of effective teaching” (cf. Ginns & Barrie, 2004). Decades ago, Rodin and Rodin (1972, p. 177) expressed a similar view by stating “… If how much student 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”.

Kember and Leung (2011) demonstrated the existence of common ideas about good teaching across four disciplines (health, humanities, science and business) at a Hong Kong university. This study, however, reported differences among disciplines some aspects of which (e.g., epistemological differences) within the teaching and learning environment brought into play.

In a comprehensive survey of the post-2000 literature, Spooren et al. (2013) suggested that SET results are subject to bias due to both the content and the structure of these scales. Onwuegbuzie et al. (2009) cautioned questionnaire designers about using neutral categories (neither “agree” nor “disagree”) in SET scales. Based on a range of studies, Onwuegbuzie and Weems (2004) and Weems & Onwuegbuzie (2001) argued that the inclusion of a midpoint option diminishes the internal consistency of SET scores. More recently, Hassan (2009, p. 320) expressed a similar view by questioning whether a ‘3’ on a 5-point scale represents an inability to respond, a middle response or lack of interest.

One of the missing elements in the debate on the validity or otherwise of the SET procedure or the positive correlation between student grades and SET scores is changing contexts of higher
education in the last quarter century. University student populations in the developed world embody much more personal, social, cultural, and intellectual diversity than was the case since the 1990s. The number of international students attending university in Western countries has increased dramatically. In several larger and more reputable Australian universities, overseas enrolments account for between 20% and 25% of the total number of students. In some regional Australian universities, this percentage is well over 50. There has also been a commensurate increase in domestic participation. Overall, massification of higher education has led to its expansion from under 15% of the relevant age group to almost 30% (Yang, 2011). These changes have altered the nature of the student populations with varied academic abilities, learning needs, and aspirations as well as the ethno-linguistic mix (Vardi, 2011). Flexibility of enrolment and entry (part-time vs fulltime, school leavers vs matured age student) has also added new dimensions to this diversity (Martin et al., 2013).

Alauddin and Ashman (2014), employing a survey involving 773 students at a leading Australian university identified Deep, Expedient and Responsible approaches to learning that reflected students’ Study Philosophy domain. Deep and Responsible approaches varied with students’ ethnicity and academic performance. Expedient approach differed according to their ethnicity, study discipline and academic performance. Students in business–related disciplines displayed a greater propensity to the Expedient approach than those in other disciplines treating education like any other commodity.

The diversity of the student population and changing academic environment notwithstanding, the SET instrument has remained essentially the same over the years. The SET questionnaire solicits student’s perception about teaching (Denson et al., 2010, p. 340). These include:

1. A combination of closed (mostly on 5, 6 or 7 point Likert scale) and open-ended questions;
2. A single item asking for an instructor’s overall teaching effectiveness;

3. Seeking written comments on the course and the instructor;

4. Students unidentified;

5. Student participation in the evaluation being voluntary;

6. Responses obtained at the end of the term with the requirement that the instructor remain absent while students complete the evaluation; and

7. Evaluation report available to the instructor with item and class specific measures of central tendency and dispersion after the publication of the examination results.

This paper breaks methodological ground in two important ways. Firstly, its use of individual student responses represents a departure from the aggregative type of analysis relying on class averages (see e.g., Feldman, 1998; Marsh, 1997; Mason et al., 1995, 2003). A disaggregated analysis involving individual student responses can capture the underlying heterogeneity within a student cohort while an analysis based on class averages masks it. Secondly, it uses partial proportional odds model and thus, overcomes the problem of violation of the parallel lines assumption that previous studies (e.g., Boex, 2000; DeCanio, 1986) have not taken into consideration.
3. DATA AND METHOD

3.1 The Data

The data for this study came from the SET surveys for economics courses at an elite Australian university. These encompassed the period 2000-2007 with more than 12,000 SET forms involving 25 courses, 102 student cohorts and 20 lecturers. The data contained a wide range of courses encompassing both large and small classes including eighteen undergraduate level courses (five Level 1, six Level 2, and seven Level 3 courses) and seven postgraduate courses. Missing values reduced the total usable sample size to 10,223.

Based on enrolment data, these courses displayed considerable diversity within and beyond their respective student populations typified, amongst other things, by academic background, age, study discipline, sex, ethnic background and their academic performance. The data from the SET surveys did not provide any information on students’ attributes such as their sex, discipline destination\(^4\), academic performance and ethnicity. Nevertheless, information on some factors that could potentially affect students’ perceptions about an instructor’s effectiveness as a university teacher was available to the authors. These include amongst other things, course level (e.g., introductory or advanced), course electivity (elective or compulsory), ethnicity of the instructor (English or non-English), and status of the instructor (e.g., lecturer or professor).

Table 1 provides the codes and definitions of dependent and independent variables used in this study. Fifty-six per cent of the data related to introductory level courses while 27% and 17% of the responses corresponded to intermediate and advanced level courses, respectively. Instructors ranking senior lecturer or below accounted for 65% of the responses. Instructors

\(^4\) For example, whether the student was doing an economics course for an economics degree, or for a science degree.
with non-English speaking background accounted for 28% of the instructors participating in this study.

Table 1: Variable list and descriptive statistics

<table>
<thead>
<tr>
<th>List</th>
<th>Description</th>
<th>Mean¹/Proportion² (standard deviation)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Independent variables – Course and Instructor Attributes</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Introductory</td>
<td>Coded 1 for introductory level course; and 0 otherwise</td>
<td>.56 (.50)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>Coded 1 for intermediate level course; and 0 otherwise</td>
<td>.27 (.44)</td>
</tr>
<tr>
<td>Advanced</td>
<td>Coded 1 for advanced level course; and 0 otherwise</td>
<td>.17 (.38)</td>
</tr>
<tr>
<td>Position</td>
<td>Coded 1 if the instructor is a lecturer or senior lecturer position; and 0 otherwise</td>
<td>.65 (.48)</td>
</tr>
<tr>
<td>Elective</td>
<td>Coded 1 if the course is elective and 0 otherwise</td>
<td>.14 (.34)</td>
</tr>
<tr>
<td>Nesb</td>
<td>Coded 1 if the instructor is from Non-English Speaking Background; and 0 otherwise</td>
<td>.28 (.45)</td>
</tr>
<tr>
<td><strong>Independent variables – Instructional Attributes used in the SET Instrument</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Organise</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of producing well organised classes; and 0 otherwise</td>
<td>.80 (.40)</td>
</tr>
<tr>
<td>Present</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of presenting materials in an interesting way; and 0 otherwise</td>
<td>.67 (.47)</td>
</tr>
<tr>
<td>Feedback</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of providing adequate feedback on the students’ work; and 0 otherwise</td>
<td>.51 (.50)</td>
</tr>
<tr>
<td>Respect</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of treating students with respect; and 0 otherwise</td>
<td>.87 (.34)</td>
</tr>
<tr>
<td>Knowwell</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of knowing the subject well; and 0 otherwise</td>
<td>.91 (.28)</td>
</tr>
<tr>
<td>Thinkmem</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of emphasising thinking rather than just memorising; and 0 otherwise</td>
<td>.77 (.42)</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of communicating his/her enthusiasm for the subject; and 0 otherwise</td>
<td>.81 (.39)</td>
</tr>
<tr>
<td>Explain</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of giving clear explanations; and 0 otherwise</td>
<td>.74 (.44)</td>
</tr>
<tr>
<td>Consult</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of being available for consultation; and 0 otherwise</td>
<td>.66 (.48)</td>
</tr>
<tr>
<td>Lskill</td>
<td>Coded 1 if the instructor achieved a score of 4 or above for students’ perception of helping students to improve their learning skills; and 0 otherwise</td>
<td>.57 (.50)</td>
</tr>
<tr>
<td><strong>Dependent variable</strong></td>
<td>Students’ perception of teaching effectiveness score of the instructor ranging between 1 (very poor) and 5 (outstanding)</td>
<td>4.0 (.94)</td>
</tr>
</tbody>
</table>

Note: ¹ Only for TEVAL. ² For all other variables.
3.2 The Analytical Framework

A large body of literature recognizes that ordinary least squares (OLS) regression is inappropriate when the dependent variable is categorical as is the case with the data in the present study. Consider a customer survey where responses are coded 1 (worst/strongly disagree), 2, 3, 4 or 5 (best/strongly agree). This is because, as Greene (2003) suggested, the OLS model treats the difference between a ‘4’ and a ‘3’ the same as that between a ‘3’ and a ‘2’, even though they are only a ranking.

The present study used a partial proportional odds model. This estimation technique, one of the special cases of a generalized model, was preferred because only some, but not all the independent variables met the parallel lines assumption. The parallel lines assumption posits that the relationship of independent variables to the odds of a dependent variable being in the next higher order category is the same regardless of which category we are comparing. In other words, the assumption says that a one-unit increase in an independent variable has the same effect on the probability of a response being in a higher category regardless of category. This implies that the coefficients do not vary while the thresholds (or cut-offs) differ across the response values. The use of the ordered logit model often leads to violation of this assumption. This paper first tested using Stata whether or not the parallel lines assumption held. Here, we first run an ordered logistic regression (a method that does not free variables from the parallel lines constraint), then we estimated a generalized ordered logistic regression (a method that frees all variables from the parallel lines constraint) and finally a likelihood-ratio test was performed to compare the fit of the two models. The $\chi^2$ value of 155.94 (with 45 degrees of freedom and $p < 0.001$) shows that the null hypothesis (that the two models are the same) should be rejected, implying that the ordered logit model was too restrictive because at least

---

5 For more information on parallel lines assumption refer to Long and Freese (2006).
one variable did not meet the parallel lines assumption. This, however, did not mean that all independent variables failed to meet this assumption. Out of the 15 independent variables included in this study, we found that nine variables satisfied the parallel lines assumption. These were Introductory, Elective, Organise, Present, Feedback, Knowwell, Enthusiasm, Explain and Lskill (Table 1). Given that some, but not all variables, meet the parallel lines assumption, it is essential to apply a partial proportional odds model because it allows the covariates violating this assumption to have category-specific effects. To confirm that the partial odds model is not too restrictive, a likelihood-ratio test was performed by comparing the fit of the generalized ordered logit and the partial proportional odds models. With the $\chi^2$ value of 30.29 (with 27 degrees of freedom and $p < 0.3015$), our test result shows that the partial proportional odds model is a viable alternative that relaxes the parallel lines assumption only for those variables where it is violated.

In this study, we found that six of the independent variables included in the model violated the assumption of parallel lines. These were Intermediate, Position, Nesb, Respect, Thinkmem and Consult (Table 1). When the parallel lines assumption is violated for some explanatory variables, as is the case in this paper, using an ordered response model leads to a wrong conclusion because one set of coefficients can no longer describe the relationship between different groups of TEVAL score categories. A proper way to overcome this limitation is to apply a partial proportional odds model. This can easily be done using a Stata ‘gologit2’ command with ‘autofit’ option.

Equation 1 expresses the partial proportional odds model for an ordered dependent variable with M categories as follows:

$$
P(Y_i > j | X_i) = \frac{\exp(\alpha_j + (X_i \beta_j + T_i \gamma_j))}{1 + \left[\exp(\alpha_j + (X_i \beta_j + T_i \gamma_j))\right]} , j = 1, 2, ..M - 1
$$

(1)
where $Y_i$ is a dependent variable indicating the observable measure of teaching effectiveness of instructor $i$. To rate the overall effectiveness of a lecturer as a university teacher students were asked at the end of each semester to decide on a TEVAL score on a 1-5 Likert scale, where 1 indicated “very poor” and 5 “outstanding”. $X_i$ is a matrix containing course and instructor characteristics and other instructional attributes. $\beta$ is a vector of estimated parameters and $u_i$ is the error term. $T_i$ contains the values of TEVAL $i$ on the subset of the explanatory variables that did not satisfy the parallel lines assumption, and $\gamma_j$ is a vector of regression coefficients associated with only the $j$th cumulative logit.

The partial proportional odds model is equivalent to a series of binary logistic regressions that combine categories of the dependent variable. Given that TEVAL scores range between 1 and 5, category 1 is contrasted with categories 2, 3, 4 and 5 for TEVAL = 1. For TEVAL = 1 or 2 the contrast is between categories 1 and 2 versus 3, 4 and 5. For TEVAL = 1, 2 or 3, the contrast is between categories 1, 2 and 3 versus 4 and 5; and for TEVAL = 1, 2, 3 or 4, it is categories 1, 2, 3 and 4 versus category 5. Thus, positive coefficients indicate that higher values of the independent variable make it more likely of a higher TEVAL score than the current one, while negative coefficients imply that higher values on the independent variable increase the likelihood of being in the current or a lower category.

4. RESULTS

This paper checked the model for its robustness by estimating a standard ordered logit model as well as a stereotype logistic model.\(^6\) A comparison of the estimated coefficients of these models with those of the partial proportional odds model for variations in coefficient signs and

\(^6\) The results of the standard ordered logit model are presented in Appendix A. A stereotype logit model is a model that does not require the parallel lines assumption. It is a compromise between a multinominal logit model and an ordered logit model. Unlike the partial proportional odds model the stereotype logit regression does not assume ordered categories, although it orders the dependent variable along one dimension.
significance levels suggested the robustness of the partial proportional odds model. The data also did not reveal any evidence of multi-collinearity as evidenced by VIF (variance inflation factor).\(^7\) This section reports the results of the partial proportional odds model.

Table 2 reports the coefficient estimates and their robust standard errors. One could interpret the coefficients of variables that met the parallel lines assumption in the same way as an ordered logit model. An elective course has a greater probability of being awarded a higher TEVAL score. A TEVAL score was higher for instructors who achieved a score of 4 or above (out of 5) for students’ perception of: producing well organised class (Organise); presenting materials in an interesting way (Present); providing adequate feedback on students’ work (Feedback); knowing the subject-matter well (Knowwell); communicating his/her enthusiasm for the subject (Enthusiasm); giving clear explanations (Explain); and helping students to improve their learning skills (Lskills).

The effects of the remaining independent variables varied in both sign and magnitude across the range of TEVAL scores. For instance, intermediate-level courses are less likely to be rated a 5 compared to advanced-level courses (Column 4). Those with a lecturer or senior lecturer position (relative to an instructor employed as associate professor or above) tend to achieve a TEVAL score of 1 or 2. This seems contradictory with the coefficient (for “Position”) presented in column (4) which has a significant positive sign. However, a separate effect of being a senior lecturer (not reported in Tables 2 and 3) is positively associated with high TEVAL score. This is perhaps because senior academics (associate professor or above) have to worry less about job security (tenure), and hence may not have to “buy” a higher student rating with a potential grade inflation. This may also be because seniors may teach courses that are more challenging potentially reducing the odds of achieving a higher student rating.

\(^7\) Alauddin and Nghiem (2010) found evidence of significant multi-collinearity for the course average data.
Instructors from non-English speaking background (NESB) are less likely to achieve a TEVAL score of 5. The effect of the variable “Respect” is consistently positive but declines across cut-points. This means that instructors who achieved 4 or above for students’ perception of treating students with respect achieve a higher TEVAL score, with the greatest differences being that they are less likely to achieve the lowest TEVAL score. The same is true for the variable “Thinkmem”. Instructors who achieved 4 or 5 for being available for consultation are likely to achieve a TEVAL score of 4 or 5.

Table 2: Determinants of TEVAL score: Results of partial proportional odds model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficients (robust standard errors)</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Rating 1 vs Rating 2-5</td>
<td>Rating 1-2 vs Rating 3-5</td>
<td>Rating 1-3 vs Rating 4-5</td>
<td>Rating 1-4 vs Rating 5</td>
</tr>
<tr>
<td>Introductory</td>
<td>-0.07 (0.06)</td>
<td>-0.07 (0.06)</td>
<td>-0.07 (0.06)</td>
<td>-0.07 (0.06)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>0.14 (0.19)</td>
<td>0.17 (0.14)</td>
<td>0.04 (0.09)</td>
<td>-0.38* (0.07)</td>
</tr>
<tr>
<td>Position</td>
<td>-0.88* (0.20)</td>
<td>-0.58* (0.13)</td>
<td>0.10 (0.09)</td>
<td>0.38* (0.06)</td>
</tr>
<tr>
<td>Elective</td>
<td>0.22* (0.06)</td>
<td>0.22* (0.06)</td>
<td>0.22* (0.06)</td>
<td>0.22* (0.06)</td>
</tr>
<tr>
<td>Nesb</td>
<td>0.20 (0.18)</td>
<td>-0.02 (0.13)</td>
<td>-0.20** (0.09)</td>
<td>-0.41* (0.07)</td>
</tr>
<tr>
<td>Organise</td>
<td>1.13* (0.07)</td>
<td>1.13* (0.07)</td>
<td>1.13* (0.07)</td>
<td>1.13* (0.07)</td>
</tr>
<tr>
<td>Present</td>
<td>1.30* (0.06)</td>
<td>1.30* (0.06)</td>
<td>1.30* (0.06)</td>
<td>1.30* (0.06)</td>
</tr>
<tr>
<td>Feedback</td>
<td>0.46* (0.05)</td>
<td>0.46* (0.05)</td>
<td>0.46* (0.05)</td>
<td>0.46* (0.05)</td>
</tr>
<tr>
<td>Respect</td>
<td>1.13* (0.19)</td>
<td>0.47* (0.12)</td>
<td>0.66* (0.10)</td>
<td>0.69* (0.15)</td>
</tr>
<tr>
<td>Knowwell</td>
<td>0.65* (0.10)</td>
<td>0.65* (0.10)</td>
<td>0.65* (0.10)</td>
<td>0.65* (0.10)</td>
</tr>
<tr>
<td>Thinkmem</td>
<td>1.00 (0.25)</td>
<td>0.55* (0.13)</td>
<td>0.52* (0.08)</td>
<td>0.86* (0.10)</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>0.82* (0.07)</td>
<td>0.82* (0.07)</td>
<td>0.82* (0.07)</td>
<td>0.82* (0.07)</td>
</tr>
<tr>
<td>Explain</td>
<td>1.32* (0.06)</td>
<td>1.32* (0.06)</td>
<td>1.32* (0.06)</td>
<td>1.32* (0.06)</td>
</tr>
<tr>
<td>Consult</td>
<td>-0.20 (0.19)</td>
<td>-0.12 (0.12)</td>
<td>0.28* (0.07)</td>
<td>0.22* (0.06)</td>
</tr>
<tr>
<td>Lskill</td>
<td>0.82* (0.05)</td>
<td>0.82* (0.05)</td>
<td>0.82* (0.05)</td>
<td>0.82* (0.05)</td>
</tr>
</tbody>
</table>

Log pseudo likelihood: -8710.30
Wald Chi-square: 4609.82
Pseudo $R^2$: 0.3217
Sample size: 10,223

*p < .01; **p < .05. Note: variables that meet the parallel lines assumption have the same coefficients across response categories.
This paper interprets the effects for the partial proportional odds model in terms of changes in odds (Table 3). The odds of achieving a TEVAL score of 5 are 0.68 times smaller for intermediate-level courses relative to advanced-level courses. The odds of achieving a TEVAL score of 2 or above are 0.41 times smaller for an instructor with a lecturer or senior lecturer position relative to an instructor whose position is associate professor or above. On the contrary, the odds of achieving a TEVAL score of 5 are 1.47 times larger for an instructor with a lecturer or senior lecturer position. The fact that the variable ‘position’ is coded as 1 if the instructor is a lecturer or senior lecturer can have varying effects on the TEVAL score.

The odds of achieving a higher TEVAL score are 1.25 times greater for teaching courses that are elective relative to compulsory courses. Being a NESB instructor decreases the odds of achieving a TEVAL score of 5 by 0.66 times relative to an academic with English-speaking background.

With regard to instructional attributes used in the SET instrument, some of them increased the probability of achieving a higher TEVAL score. For instance, the odds of achieving a higher TEVAL score are 3.66 times greater, if the instructor achieved a score of 4 or higher on a 5-point scale for presenting materials in an interesting way. The odds of achieving a higher TEVAL score are around 3-4 times greater if the instructor achieved a score of 4 or above for students’ perception of producing well organized classes, and providing clear explanations. The odds of achieving a higher TEVAL score are around 2 times greater if the instructor achieved a score of 4 or above for students’ perception of communicating his/her enthusiasm for the subject, and helping students to improve their learning skills. The odds of achieving a higher TEVAL score are less than 2 times greater if the instructor achieved a score of 4 or above for students’ perceptions of providing adequate feedback on the students’ work, and knowing the subject-matter well.
The odds of achieving a TEVAL score of 2 or above are 3.10 times greater if an instructor achieved a score of 4 or above for students’ perception of treating students with respect. For the same attribute, the odds of achieving the highest TEVAL score are 2 times greater. This indicates that there is a direct link between treating students with respect and achieving a higher TEVAL score, with the greatest differences being that instructors who achieved 4 or above with regard to respect are less likely to achieve the lowest TEVAL score.

The odds of achieving a TEVAL score of 2 or above are 2.73 times greater if the instructor achieved a score of 4 or higher for students’ perception of emphasizing thinking rather than just memorizing. For this attribute, the odds of achieving a TEVAL score of 5 are 2.36 times greater. Similar to the instructional attribute measuring respect the greatest differences being that instructors who achieved 4 or above for emphasizing thinking are less likely to achieve the lowest TEVAL score. The odds of achieving a TEVAL score of 5 are 1.24 times greater if the instructor achieved a score of 4 or above for students’ perception of being available for consultation.

Overall, students’ perceptions of instructors giving clear explanations (Explain), presenting materials in an interesting way (Present), and producing well-organized classes (Organise) are the three key instructional attributes that significantly and substantively influenced students’ rating of their instructors. The next two important variables that positively affect the TEVAL score are students’ perceptions of instructors helping them to improve their learning skills (Lskills) and communicating enthusiasm for the subject (Enthusiasm).
Table 3: Odds ratios of TEVAL score: Results of partial proportional odds model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Odds ratios</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(1) Rating 1 vs Rating 2-5</td>
</tr>
<tr>
<td>Introductory</td>
<td>0.94</td>
</tr>
<tr>
<td>Intermediate</td>
<td>1.15</td>
</tr>
<tr>
<td>Position</td>
<td>0.41*</td>
</tr>
<tr>
<td>Elective</td>
<td>1.25*</td>
</tr>
<tr>
<td>Nesb</td>
<td>1.22</td>
</tr>
<tr>
<td>Organise</td>
<td>3.08*</td>
</tr>
<tr>
<td>Present</td>
<td>3.66*</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.58*</td>
</tr>
<tr>
<td>Respect</td>
<td>3.10*</td>
</tr>
<tr>
<td>Knowwell</td>
<td>1.92*</td>
</tr>
<tr>
<td>Thinkmem</td>
<td>2.73*</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>2.28*</td>
</tr>
<tr>
<td>Explain</td>
<td>3.76*</td>
</tr>
<tr>
<td>Consult</td>
<td>0.82</td>
</tr>
<tr>
<td>Lskill</td>
<td>2.26*</td>
</tr>
</tbody>
</table>

*p < .01; **p < .05.

Note: variables that meet the parallel lines assumption have the same odds ratios across response categories.

5. DISCUSSION AND POLICY IMPLICATIONS

5.1 DISCUSSION

The findings of this paper are consistent with and differ from those of some previous studies. Attributes such as Organise, Explain and Present in the present study closely resemble those of the Boex (2000) study that identified organization and clarity as most important. The finding of the present study regarding the importance of organisation and clarity of explanation is also consistent with previous studies (DeCanio, 1986; Mason et al., 1995).

The third characteristic of effective economics instructors identified by Boex (2000) was the ability to motivate students. There was no such comparable attribute in the present study even
though perceived improvement in students’ learning skills (Lskill) was somewhat similar come close. In the Boex (2000) study, a higher student rating on instructor’s intellectual/scholarly dimension attracted a lower instructor rating by students and differed from the present study. A higher student rating of the instructor knowing the subject matter well (Knowell) had a significant positive impact on the TEVAL score.

Denson et al. (2010), in determining student satisfaction in course evaluation using more than 60,000 individual student responses from a leading Australian university found that evaluation questions accounted for a majority of the variation even though student characteristics and reasons for enrolment were statistically significant. However, instructor-selected optional questions were stronger predictors than those that were mandatory. The findings of the present study, despite differences in focus of the investigation, have some similarities with those of Denson et al. (2010). First, the mandatory questions in both studies were significant predictors of student satisfaction as were the cases with course electivity. However, these varied across disciplines.

The key findings of this paper also support the findings of Alauddin and Tisdell (2010) in that TEVAL scores varied significantly, substantively and directly with whether students perceived the instructors as providing well-organised lectures, clear explanation and well-presented lectures. Alauddin and Tisdell (2010) had a sample size of 2,467, on nine economics courses at the same university.

One disturbing finding is that a key variable measuring instructor’s pedagogical responsibility, namely emphasis on thinking rather than memorising (Thinkmem), ranks lower down the order in terms of having a substantive impact on TEVAL score. Thus, an instructor can achieve a high TEVAL score at the cost of some critically important factors in teaching and learning. In this respect, the results are consistent with Everett’s finding that “the literature consistently
shows that high level cognitive skills material plays little role in raising SET scores, and by reducing clarity may actually lower such scores. …” (Everett 1977, pp. 101-102). More recently, Carrell and West (2010) provided comprehensive evidence in this regard. As they stated:

… our results show that student evaluations reward professors who increase achievement in the contemporaneous course being taught, not those who increase deep learning. Using our various measures of teacher quality to rank-order teachers leads to profoundly different results. Since many U.S. colleges and universities use student evaluations as a measurement of teaching quality for academic promotion and tenure decisions, this finding draws into question the value and accuracy of this practice.

5.2 POLICY IMPLICATIONS

The findings of this paper have several implications for the instructor, the school and the university as a whole.

One key question for the lecturer to ask is, how s/he can raise students’ perceived rating for the key TEVAL-enhancing instructional attributes such as Explain, Present and Organise. A way might be to making ‘objective’ improvements in key pedagogic variables such as Thinkmem and Lskills expecting that students will reflect them in higher stated scores (Alauddin & Tisdell, 2010). A problem, however, is that these relationships are not well documented. Secondly, the lecturer might keep the nature of the instruction constant but spend more time trying to convince students how well s/he organized, presented and explained the instruction process. This amounts to little more than a marketing and promotional effort in relation to the three key variables identified in this research. This, however, as Alauddin and Tisdell (2010) stated may not result in ‘quality-enhancing’ teaching practices.
The lecturer should bring it to the attention of the Head of School that the existing SET instrument is inadequate to measure teaching effectiveness in that a lecturer is able to increase her/his student rating by engaging in less than scholarly teaching practices. One of the important criticisms in this paper is the failure to use TEVAL data to ‘instruct’ lecturers on effective means (action) to increase their TEVAL scores. A further major problem is that these scores may indirectly undermine the quality of course procedures and may encourage ‘spoon-feeding’ and reduce ‘independent work’ by students or fail to maintain, let alone enhance, academic standards the very goal that a university system should espouse. This may take the form of some percentage of faculty members being tempted to lower assessment requirements and academic standards to inflate expected grades and to increase their instructor excellence scores on some evaluations (Stapleton & Murkison, 2001).

Our results support broadening the scope of the SET instrument so that information on student attributes, instructor attributes, and course attributes are utilized. Student attributes should include information on student quality, effort level and class attendance, age, and sex. Instructor attributes should include age, ethnicity, gender, and experience. Course attributes ought to include quantitative/non-quantitative, compulsory/elective nature of the course, difficulty level and so on. The instrument should also gather information on teaching philosophy including emphasis on critical thinking, developing students’ analytical ability, their independence, contents, problem-solving skills, real world applications and intellectual challenges. Likewise, information on learning attitude and philosophy including emphasis on deep learning/rote learning, greater dependence on instructor rather than exercising greater independence are of vital importance.

---

8 Lumpkin and Moulton (2013) found a closer link between instructors’ teaching and their research and the synergy between the two facets of a university academic.
6. SOME FURTHER OBSERVATIONS

The proceeding discussion suggests that there are a number of critically important issues that the data gathered through the SET instrument are unable to address. First, the very purpose of a SET survey is unclear about whether it wants to measure the ‘quality of teaching’ or merely reflects students’ perceptions about teaching. If it is the latter, as most likely is the case, then it may be more a subjective measure than an objective measure. In that case, as Judge et al., (1988, p.582) put it:

In some cases in empirical analysis, the variables we measure are not really what we want to measure. … The proxy variables may be subject to large measurement errors. Even for the observable variables, the data may be subject to a variety of errors. Errors may be introduced by the wording of the survey questionnaires. Words such as weak and strong may imply different things to different respondents.

The specification of TEVAL is vague. The student is asked to rate the “instructor’s overall effectiveness as a university teacher”. However, the SET instrument says nothing about effectiveness in what regard. Different students may use different criteria. This is likely to generate errors in data measurement. Griliches (1974, pp.973-74), notes that errors in data measurement arise because of the: (1) separation of the data collection and the analysis processes; (2) fuzziness about what one would like to measure; and (3) complexity of the phenomena that one is trying to measure. On all of these grounds, the SET procedure produces data with significant errors in measurement.

Assuming that the TEVAL score “effectively” identifies all the relevant explanatory variables that determine the TEVAL score, the preceding analysis helps to reveal how students construct their evaluation, that is, how much weight they give to each of the explanatory variables. This raises several questions as to whether these weights are appropriate. For instance, should the

---

9 The collection is largely the responsibility of organizations such as survey research centers within a university system and may be divorced from the researchers who engage in in-depth analysis of the collected data.
weights vary across levels or courses at the same levels or across disciplines? It is probable that instructors, scholars or even the administrators believe that different weights would be more appropriate. For example, the latter may believe that Thinkmem (emphasis on thinking rather than memorizing), should be given considerable weight as some universities regard the development of analytical abilities and critical judgment of students as a central graduate attribute. However, the relatively low influence of Thinkmem on the dependent variable in the present study does not support the attainment of this key graduate attribute. Thus, the SET procedures may favour non-academic styles of teaching that entail less critical analysis than desirable at a university level e.g., a deep learning approach reminiscent of Level 3 teaching (Biggs & Tang, 2011). Thus, an unintended consequence could be that this method of teaching evaluation reduces the emphasis on reading and consideration of competing intellectual ideas and points of view and could reduce the intrinsic quality of university courses.

Some may judge effectiveness on superficial grounds – whether they were entertained, for example. A more appropriate measure of effectiveness would be a measure of what the students learned or how much their understanding of the focal subject that the lecturer was able to advance through an emphasis on intellectually challenging advanced by the lecturer. This (in some cases) may not correlate highly with presenting a subject in an interesting way. Thus, SET procedures could lead to superficial presentation of lectures that could have a significant impact on TEVAL scores. Nevertheless, one does not learn what it is that makes the teaching interesting and how well that variable relates to academic achievement which is somewhat wider than learning and difficult to measure accurately.

The scores provided by SET data are averages. The distribution of those scores and what influences them, would be worthy of consideration. For example, a teacher may be highly rated by one group but not by another. Is it the quality or the nature of the course content rather more
than the quality of teaching that affects the score? No constructive use of SET data of this type appears to be made in this respect. Consequently, judgements based on the results may be superficial, and even erroneous if one uses only the mean TEVAL measure of central tendency for assessment. In its present form, the SET procedure that treats the distribution of scores as unimodal may be unrealistic. Furthermore, the process of averaging implies that each student in the sample receives an equal weight. This is despite that fact that some students are much better informed, intellectually superior, and less inclined to be lured by superficial treatment of the subject matter and more interested in the substance than appearance than those from the other end of the spectrum.

7. CONCLUDING COMMENTS

While TEVALs are simple to apply, there are dangers of using such an indicator to judge the quality of teaching. In the absence of appropriate weighting of independent variables or weighting of student quality by some attributes such as their study habits, their effort level and intellectual capabilities, the use of TEVAL essentially loses much of its real significance. The present paper concurs with the view expressed by Sproule (2002, p.288) that in the absence of

---

10 In one of the courses included in this study, using an identical teaching method two years in a row, the same lecturer received significantly different TEVAL ratings which dropped from 4.19 in the first year to 3.04 in the second year. The distribution of TEVAL score displayed bimodality with just over a third of the sample rating the lecturer in the 1-2 range (very poor to poor) with almost the same proportion rating him in the 4-5 (excellent to outstanding) end of the spectrum. In another case, a lecturer in a third level economics course received a very poor TEVAL score in one year while at the same time being commended for making the most significant impact on the students who were doing the same course but were enrolled in degree programs of a non-economics discipline. The distribution of the TEVAL score for this lecturer displayed bimodality in the preceding years.

11 The educational literature and the administrators alike routinely use the mean rather than median or mode even though it is patently wrong to do so from a statistical point of view in cases of ordinal data. What seems intriguing is that some of the administrators are highly competent mathematicians, statisticians or econometricians who would advise their students to adhere to methodological correctness when they teach. However, when wearing the administrators’ hat such as head of school, or serving on promotion and tenure committees, they stridently defend the use of mean TEVAL score as the indicator of instructors’ teaching “quality”.

27
data on these or suitable proxies, the hypothesis that a particular lecturer has failed or succeeded in her/his pedagogical responsibilities remains underdetermined. Accepting or rejecting the hypothesis based on a certain mean value of TEVAL score is little more than promoting pseudoscience (Radner & Radner, 1983; Sproule 2002).

The extent to which the perceived quality of teaching in prerequisite courses have any bearing on the perception of teaching quality in subsequent courses is quite important. Perusal of narrative comments often portrays the students’ feelings about a sub-discipline. ‘I hate (love) microeconomics or macroeconomics and so on.’ In addition, lecturers in a course often have to build on the prerequisites that may have inadequately prepared the students for higher-level courses. As a result, the average student may find subsequent higher-level course(s) too difficult and this may result in the lecturer receiving a poor rating. Therefore, the teaching consequences could be unsatisfactory (see, e.g., Carrell and West, 2010).

In conclusion, as Becker (2000, p.115) notes that “… in the 21st century, sole reliance on traditional end-of-term student evaluation of teaching should not be tolerated. For starters, student evaluations should focus on what students know, that is, what they have learnt”. The present study views Becker’s tests from a pragmatic viewpoint in that scientific methods are those that result in successful rules for action (Dewey, 1963; see also James, 1946; 1975). Given that the findings of the paper suggest that SET data do not result in successful and meaningful rules for action, one should reject them on pragmatic grounds. There is a danger that they will encourage academic institutions to evolve in undesirable academic directions e.g. to prefer rote learning to critical thinking, scepticism and, exposure to a diversity of views that constitute the fundamental building blocks of teaching and learning in higher education. As stated in Section 5.2, the existing instrument needs a significant revamp if it is to result in successful and meaning rules for action. The revamped instrument needs to be able to minimise
the potential sources of biases, as the extensive survey of the literature and the findings of the present study suggested. This is in order to ensure measuring teaching effectiveness in terms of the strength of the linkage between teaching and learning outcomes. It is also necessary to use a forced response on the Likert scale instead of a “neither agree nor disagree” type of response\textsuperscript{12}. One could also consider asking the students to put an actual number or a percentage mark to various instructional attributes and then put a grade just as we mark students’ assignment. The use of actual numbers have the advantage of deriving more meaningful use of measures of central tendency and dispersion than ranked data can permit.

**ACKNOWLEDGEMENTS**

The authors gratefully acknowledge the anonymous colleagues for access to their *TEVAL* data, two anonymous reviewers, Jackie Robinson for constructive comments and suggestions on an earlier draft, and Farhanaz Alauddin, Fidah Mosaddek and Ranjini Shome for data entry for this research. The usual *caveats* apply.

\textsuperscript{12} This type of scale is in use by some institutions in Australia.
REFERENCES


Appendix A

Odds ratios of TEVAL score: Results of the standard ordered logit model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Odds ratios (robust standard errors)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Introductory</td>
<td>0.97 (0.06)</td>
</tr>
<tr>
<td>Intermediate</td>
<td>0.83 (0.05)*</td>
</tr>
<tr>
<td>Position</td>
<td>1.21 (0.06)*</td>
</tr>
<tr>
<td>Elective</td>
<td>1.22 (0.08)*</td>
</tr>
<tr>
<td>Nesb</td>
<td>0.75 (0.04)*</td>
</tr>
<tr>
<td>Organise</td>
<td>3.07 (0.21)*</td>
</tr>
<tr>
<td>Present</td>
<td>3.71 (0.21)*</td>
</tr>
<tr>
<td>Feedback</td>
<td>1.61 (0.08)*</td>
</tr>
<tr>
<td>Respect</td>
<td>1.85 (0.14)*</td>
</tr>
<tr>
<td>Knowwell</td>
<td>1.85 (0.18)*</td>
</tr>
<tr>
<td>Thinkmem</td>
<td>1.88 (0.11)*</td>
</tr>
<tr>
<td>Enthusiasm</td>
<td>2.22 (0.16)*</td>
</tr>
<tr>
<td>Explain</td>
<td>3.71 (0.23)*</td>
</tr>
<tr>
<td>Consult</td>
<td>1.21 (0.06)*</td>
</tr>
<tr>
<td>Lskill</td>
<td>2.29 (0.12)*</td>
</tr>
</tbody>
</table>

Log pseudo likelihood: -8773.12
Wald Chi-square: 4560.53
Pseudo $R^2$: 0.3168
Sample size: 10,223

*p < .01