Student evaluation of teachers in an engineering department

Authors:

Robin Bradbeer, Department of Electronic Engineering, City University of Hong Kong, Hong Kong, eersbrad@cityu.edu.hk
Aman Shah, Tracy Lo and Philip Wong, Educational Development Office, City University of Hong Kong, Hong Kong, tracy.lo@cityu.edu.hk

Abstract *- Student evaluation of teachers is still a controversial subject even after 70 years of study and evaluation. This paper seeks to put the subject into a historical context. It the n goes on to consider the potential and actual biases inherent in the commonly available testing instruments, attempting to locate these within a cross - cultural context, and then analyses the results from a study into the biases in student ratings over a 5 year period in the Department of Electronic Engineering at City University of Hong Kong.*

Index Terms *- biases in rating scores, cross -cultural contexts, engineering, student evaluation of teachers.*

INTRODUCTION

The formal evaluation of teachers and their teaching effectiveness has been discussed since the beginning of formal education itself. Universities were relatively late in adopting formal evaluation methods for teaching faculty, both as a method of aiding those faculty in improving their own standards of teaching, but also, more controversially, as a means of administrative decisions of promotions and tenure.

The modern era of student evaluations of teaching (SET) can be broken roughly into four periods: the thirty-year period preceding 1960, the 1960s, the 1970s, and the period from the 1980s to the present. Before 1960, most of the research on student evaluations was conducted by Herman Remmers and his colleagues at Purdue University. Remmers correlated student rating and achievement data over 60 years ago [1], and he published the first multisection validity study 37 years ago [2]. The Purdue rating form, published in 1927, was probably the first student evaluation form. Remmers conducted a series of studies with it, investigating such issues as the relationship of students' grades to their ratings of teachers (1930), the reliability of student ratings (1934), and the comparison between alumni and student [3].

ANALYTICAL METHODS

One of the main problems associated with the evaluation of the effectiveness and validity of SET is deciding what is to be measured! Many researchers have attempted to identify the factors that influence effective teaching and learning. Cohen [4], in one of the first major papers on the subject, and the first study to carry out a meta-analysis of previous findings, used what he called the six dimensions of teaching. Kulik and McKeachie [5] identified four of these - skill, rapport, structure and difficulty - and Isaacson [6], two more - interaction and feedback.

Feldman [7]), quoted in Centra and Bonesteel [8], synthesised the results from 31 studies, both of staff and students, and found that both groups gave high rankings to the following factors:

- The teacher's sensitivity to and concern with class level and progress
- The teacher's preparation and organisation of the course
- The teacher's knowledge of the subject
- The teacher's enthusiasm (for the subject or for teaching)
- The teacher's clarity and understandableness
- The teacher's availability and helpfulness
- The teacher's fairness
- Impartiality of the teacher's evaluation of students
- The quality of examinations.

Other reviewers have focused on fewer characteristics. Sherman and others [9], again quoted in Centra and Bonesteel [8], identified five characteristics: enthusiasm, clarity, preparation and organisation, stimulation, and

Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: .
Deleted: .
Deleted: ¶
Deleted: ¶

knowledge. From this it is clear that most education researchers are of the opinion that effective teaching is multi-dimensional.

However, there is a school of thought that, although agreeing the multi-dimensionality of learning, believes that evaluation of effectiveness can be categorised by a single global variable. Abrami [10] has argued that the multidimensional approach is untenable and that summative evaluation should probably be based on global ratings of overall teaching effectiveness.

As Koon and Murray [11] point out, one of the main arguments against the multidimensional approach is that unless it can be shown that certain teaching dimensions have uniform effects on student outcomes throughout the group of teachers in which they are to be used (for example, all faculty members in the social sciences), ratings of specific dimensions of teaching cannot be fairly weighted for use in promotion and tenure decisions.

This basic argument between the two schools determines the methods that they use in evaluating the effectiveness of student ratings. The multi-dimensionalists use multi-trait, multimeasure analysis (MTMM), the globalists multi-section validity studies.

A series of articles in the American Psychologist dealt with some of the major validity issues pertinent to teaching evaluation instruments. The authors disagree on whether teaching and ratings are multidimensional [12] or best described by either a global factor [13] or a global factor with several highly correlated lower order factors (d'Apollonia and Abrami, [14]; McKeachie, [15]). It seems to be a debate about the most appropriate analyses to use - multi-trait, multimeasure or multisection validation designs.

Multidimensionality is important not only because of its obvious diagnostic utility as instructor feedback but also because it provides a more sophisticated and realistic assessment of the various aspects of teaching. Thus, as Marsh states (ibid) various contextual variables possible biasing influences, and validity all can be investigated more systematically and productively, rather than lumping all the different dimensions into a puree and then trying to separate out the causal ingredients! (The analysis of results in this paper uses multidimensionality). Whatever the methodology applied to the evaluation of the effectiveness and validity of SETs, most researchers have shown that there are positive correlations between student ratings and teaching effectiveness.

Cohen's seminal papers in 1980 [16] (on the effectiveness of student-rating feedback for improving college instruction) and 1981 [4] (on the validity of student ratings) applied meta-analysis to a majority of the published literature at the time. The first showed that there was a positive correlations between the increase in final grades after mid-term feedback. The second found that the average correlation between overall instructor rating and student achievement was 0.43; the average correlation between overall course rating and student achievement was 0.47.

Since then numerous other studies have been carried out. These are summarised in the five papers quoted above, (d'Apollonia and Abrami, [14]; Greenwald, [17]; Greenwald and Gillmore, [13]; Marsh and Roche, [12]; McKeachie, [15]). These authors suggest at least a moderate relationship between ratings of teaching effectiveness and measures of student achievement.

Unfortunately, even though all agree that there is something significant in this relationship, there is disagreement on what are the causal factors. Meta-analysis has just confused the issue. Rather than bringing research findings into sharper focus, these meta-analyses of student rating validity have come to strikingly different conclusions. "What we now have are conflicting results in a body of meta-analyses" (Cohen, [18]). Cohen goes on to question the applicability of meta-analysis in this area.

Marsh and Roche [19] suggested that for purposes of feedback to instructors (and perhaps for purposes of teachers' input into personnel decisions), it might be useful to weight SET factors according to their importance in a specific teaching context. Marsh [12] points out that unresolved issues concerning the validity and the utility of importance-weighted averages (e.g., Marsh, [20], however, dictate caution in pursuing this suggestion. Abrami et al [14], quoted in Marsh [12] however raised many concerns about factor analysis that were largely addressed by Marsh[19] [20]. They cited Cashin and Downey [21] as showing that specific ratings add little to global ratings. Marsh's [20] reanalysis of this study showed that the optimal subset of SETs (in relation to their outcome variable of students' progress ratings) did not even include global items. It also implied that specific items were less valid than global ratings in multisection validity, studies, even though Feldman [22] reported nine SET dimensions (.57 for organisation, .56 for clarity, .46 for impact. .38 for interest stimulation, .36 for discussion. .36 for availability, .35 for elocution, .35 for objectives, and .34 for knowledge) that were more highly correlated with achievement than the .32 correlation between achievement and global ratings reported by Abrami et al.

However, one factor that multidimensional studies has identified is that of bias in the evaluations. This is of particular importance in the Hong Kong setting where instruments designed in one cultural milieu (N America) are being used in an entirely different cultural context.

Deleted: ¶

Deleted: ¶

Deleted: ¶

Deleted: ¶

Deleted: ¶

Deleted: .

Deleted: ¶

Deleted: ¶

Deleted: them

Deleted: our

Deleted: (Greenwald have been quoted twice here. Shoud it four or five papers?)¶ ¶

Deleted: ¶

Deleted: but

Deleted: ;

Deleted: and

Deleted: .

Deleted: (This paragraph is difficult to read. A lot of information is packed here.)¶ ¶

BIAS

There is mounting evidence that differences in courses, structure, discipline, age, sex, race, year of study and class size, to identify a few factors, have an effect on the student rating.

For example, Centra [22] reports that one study he did at Syracuse University suggested that according to student ratings, mathematics and statistics and natural science courses (and the teachers who teach them) tend to be less student oriented, less effective in presentation (they are often lectures), and more difficult and quicker paced.

As far as course difficulty, work load, and effort are concerned, a five-college study he conducted revealed differences in teachers' and students' views (Centra, [23]). Teachers in the natural sciences thought the level of difficulty and pace of their courses were appropriate; students found the courses difficult and fast paced. Teachers in the natural sciences also thought that students did not put enough effort into their courses; students disagreed. Students tend to give slightly higher ratings in their major field or that they elect to take than they do to required courses (Centra and Creech, [24]).

Feldman's [25] review of the research concluded that a small positive relationship (correlations in the .10s and .20s) existed between class ratings and the students' average intrinsic interest in the subject area. Intrinsic interest - or prior subject interest as Marsh [26] refers to it - correlated by about .40 with students' own evaluations of their learning in the course. Although the students' prior subject interest probably affects course ratings more than it does teacher ratings, most rating systems do not take it into account.

As for sex and race, Centra [22] reports that on the basis of the classroom studies and most of the laboratory studies, students generally do not rate male and female teachers much differently. If a teacher has a class in which most students are of the same gender as the teacher, however, the ratings could be somewhat higher than with a more mixed group. He also states that:

> "Students who are racially similar to a teacher may rate that teacher more highly than those who are not. Although no studies have been reported that investigate systematic racial bias in student ratings, based on the gender studies, however, the expectation would be that a class of same-race teacher and students would result in a somewhat higher rating than one where race differs". (p :76)

Leeds et al [27] report that from a study they carried out at Temple University in the USA, the results suggest that, all else being equal, students preferred male, native born instructors. Instructors' SETs fell with age until instructors reached 54, at which point the SETs began to turn upward. Part-time instructors had lower SETS. Of these results, however, only the coefficient on the sex of the instructor was significant at the 10% level.

Other researchers have asked faculty to indicate a list of 17 "potential biases" they believed would actually have a substantial impact on student ratings. The most commonly mentioned were course difficulty (72%), grading leniency (68%), instructor popularity (63%), student interest in subject before course (62%), course workload (55%), class size/enrollment (55%), and required versus elective (55%). (Marsh, [28])

However, studies in this area are not as common as those looking at other variables. As Centra [22] notes Gage[29] and Dunkin and Barnes [30] describe four classes of variables that have been used in research on teaching: presage variables (age, sex, social class, background, training, experience); context variables (grade level, subject matter, class size); process variables (the ways in which teachers and students behave and interact); and product or outcome variables (the extent of learning and achievement of educational objectives). In their review of research on teaching in higher education, Dunkin and Barnes conclude:

> "that the vast majority of research at the college level has been conducted with process and product variables, and much of the process part, unfortunately, has been obtained on the basis of prescriptive definitions or ratings from untrained observers, rather than on the basis of careful observation. We not only need alternative ways to document process variables, we also need to do more work with presage, context, and product variables".(quoted in Centra and Bonesteel [8])

In his 1994 paper addressing dimensionality, reliability, validity, potential biases and utility of student evaluations, Marsh attempted to summarise the findings on bias from evaluations carried out at that time. Although many studies have been published since this attempt which tend to show stronger biases in some factors (Centra [22]), and he leaves out bias due to race and language of instruction for lack of evidence, they still make interesting reading - Table 1.

TABLE I

OVERVIEW OF RELATIONS FOUND BETWEEN STUDENTS' EVALUATIONS OF TEACHING EFFECTIVENESS AND SPECIFIC BACKGROUND CHARACTERISTICS

Background Characteristic	Summary of "Typical" Findings
Prior subject interest	Classes with higher prior subject interest are rated more favourably, though it is not always clear if interest existed before the start of course or was generated by the instructor.
Expected/actual grades	Classes expecting (or actually receiving) higher grades give somewhat higher ratings, though this can be interpreted to mean either that higher grades represent grading leniency or that superior learning occurs.
Reason for taking a course	Elective courses and those with a higher percentage taking a course for general interest tend to be rated slightly higher.
Workload/difficulty	Harder, more difficult courses that require more effort and time are rated somewhat more favourably.
Class size	Mixed findings but most find that smaller classes are rated more favourably, though some report curvilinear relations and a few find the effect limited primarily to items related to class discussion and individual rapport.
Level of course/year in school	Graduate level courses rated somewhat more favourably; weak, inconsistent findings suggesting that upper-division courses are rated higher than lower-division courses.
Instructor rank	Mixed findings, but little or no effect.
Sex of instructor &/or student	Mixed findings, but little or no effect.
Academic discipline	Weak tendency for higher ratings in humanities and lower ratings in sciences, but too few studies to be clear.
Purpose of ratings	Somewhat higher ratings if known to be used for tenure/promotion decisions.
Administration	Somewhat higher ratings if surveys not anonymous and/or instructor present when the survey is completed.
Student personality	Mixed findings, but apparently little effect, particularly for class-average responses, since different "personality types" may appear in somewhat similar numbers in different classes.

Note. For most of these characteristics, particularly the ones that have been more frequently studied, some studies have found results opposite to those reported here, whereas others have found no relation at all. The size, and in some cases even the direction, of the relation varies considerably depending on the particular component of students' evaluations being considered. Few studies have found any of these characteristics to be corelated more than .30 with class-average student ratings, and most reported relations are much smaller. (Marsh [31])

Notwithstanding the findings concerning bias of researchers such as Marsh and Centra, they have carried out most of their initial published studies in N America only. Recently, as detailed below, Marsh has published some articles detailing the validity of using N American instruments in non-American universities. Others have even looked at the validity of the instruments taking into account the different cultural settings of the teaching and learning processes. Hence, the validity of blindly applying commonly accepted SETs, like SEEQ and Endeavor, or their derivations, as in most universities in Hong Kong, is currently being questioned. This will be discussed further in the conclusions.

Marsh [28] developed an applicability paradigm and used it to investigate the validity of a US-developed model of teaching effectiveness, using a questionnaire at campuses in six different countries representing distinct and different cultural, economic, and philosophical traditions. The data supported the reliability, appropriateness, and to some extent the convergent and discriminant validity of the instruments.

Watkins[32] of Hong Kong University summarised most of the then-current findings twelve years later. He states that researchers from third world countries have long questioned the assumption that Western educational and psychological theories and measuring instruments are appropriate for non-Western subjects (Enriquez, [33]). All too often in the past a researcher has taken a test developed in one culture and administered and scored it for subjects

Formatted: None

Deleted: ¶

Deleted: Endevour

Deleted: ¶

Deleted: ¶

from another culture without demonstrating the relevance of the construct or the validity of the instrument for the new culture. Triandis [34], in warning against this practice, calls it "pseudo etic" research.

A related problem is the assumption that test scores derived from samples from different cultures are directly comparable in a numerical sense. This assumes what Hui and Triandis [35] refer to as scalar equivalence - that is, the construct of interest is measured on the same metric in the different cultures.

Marsh and Roche [19] report results from six applications of Marsh's applicability paradigm in studies with students from technical and further education (TAFE) colleges in Australia (Hayton, [36]); from a Spanish university (Marsh, Touron, and Wheeler, [37]); from a Papua New Guinea (PNG) university (Clarkson, [38]); from a New Zealand university (Watkins, Marsh, and Young, [39]); from a traditional Australian university (Marsh, [40]); and from a new Australian university (Marsh and Roche, [19]. In each of these studies all but the workload/difficulty items discriminated well between the "good" and "poor" teachers. This latter result should not be surprising as surely a course is "poor" if it is too hard or too easy or has too light or too heavy a workload. Marsh points out that any halo effect caused by the "good"-"poor" selection procedure both tends to exaggerate differentiation between the two groups of teachers and to make it difficult to discriminate between the multiple components of teaching by the MTMM analyses. In each of these six studies most of the items were considered appropriate and the MTMM analyses did support the multidimensionality of the components of teaching effectiveness.

As Watkins [32] points out, Marsh and Roche [19] went further than a simple sum of the other studies by comparing the pattern across studies when their student respondents indicated the importance of individual items. The results indicate that the scales of the SEEQ and Endeavor scales have internal consistency adequate for both research and applied purposes in five of the campuses representing very different university settings and cultures. Only for the Nepalese sample are the alphas rather lower than acceptable for decision-making purposes.

Further, four of the six samples showed clear evidence of convergent and discriminant validity. However, the Filipino and the Nepalese samples suggested a lack of discrimination perhaps caused by the influence of a halo effect due to the nature of the applicability paradigm task. The results also indicated that all but the items relating to assessment are appropriate in the six different settings. The pattern of importance ratings of the questionnaire items from subjects of the 11 studies, when the data reported by Marsh were also considered, suggested some overall similarity in perception of teaching effectiveness. Watkins [32] postulates that the Western studies were more similar to each other because they may reflect the greater campus/cultural similarities in these studies but may also be a function of the higher reliabilities of the scales for these samples (also not surprising since all of the Western students were responding to questions in their first language whereas all the non-Westem students were responding to questions in English, while their language of instruction is at best their second or third language).

CITYU'S TEACHING FEEDBACK QUESTIONNAIRE

At City University of Hong Kong, teaching evaluation is made separately from course evaluation. The University requires all Faculties and the College to assess the teaching performance of their staff on an individual basis. The Quality Assurance Committee (QAC), set up in 1993 to look into all teaching and learning matters of the University, decided that teaching evaluation scheme should be discipline specific. Therefore individual Faculties and the College (which administers non-degree courses) are responsible for devising, implementing and maintaining their own teaching evaluation schemes. The framework for this implementation is outlined in a policy document in which six policies and thirteen principles regarding the design of a valid and reliable teaching evaluation scheme are listed. In general, the University recommends that each teaching staff has to undergo at least one summative student evaluation each year. The evaluation time should be as near as possible to the end of the teaching term/semester. The University also strongly emphasises that summative teaching evaluation should include information from various sources such as graduates, peers, self etc. Students should be used as one source of input.

Since the teaching evaluation scheme is designed and owned by the Faculties and the College, it is their own responsibility to design its own implementation policy and procedures. They have to decide specifically on issues like the ownership of the results, the administration procedures, the follow up activities, the number of evaluations etc.

The resulting Teaching Feedback Questionnaire (TFQ), as stipulated above, should be short with around ten items in total. It should consist of items of general nature about the core teaching responsibilities and also an item seeking students' overall impression of staff 's teaching performance. The first five items in all TFQs should address issues of clear communication of class materials, good preparation for classes, effective organization of class time, stimulation of student interest in the subject and responsiveness to student problems. The rest of the questionnaires

Deleted: ¶

Deleted: ¶

Deleted: ¶

Deleted: ¶

Formatted: Indent: First line: 0"

Deleted: ¶

Deleted: ¶

should be designed by individual departments/divisions to "reflect the particular qualities of teaching" in their own discipline and context.

The Faculty of Science and Engineering introduced a formal summative teaching evaluation scheme in 1994-95. A common TFQ is adopted across the entire Faculty. Each teaching staff who is involved in the evaluation exercise is given his/her evaluation report, the departmental data summary and the faculty data summary. The head of department is provided with a staff report, a departmental data summary and a faculty data summary.

A critical view of the CityU TFQ shows that at first glance many of the "essentials" of good design are in place. Any bias due to discipline have been accommodated by making it discipline specific. Most of the main 'dimensions" of teaching have been addressed, although it is interesting to note that most universities in Hong Kong use only a few questions - around 10- to determine the ratings compared to up to 40 in SETs in N America. There is even an attempt to take into account the different teaching settings - lecture or tutorial. And recently a question on the main language of teaching has been added.

Unfortunately none of these are taken into account when the TFQ results are passed to the Head of Department. The only result that matters is the global one asking whether the teacher is on a range of "excellent" to "poor". And no-one has actually made any attempt to determine what these terms mean in the context of teaching in Hong Kong.

Worse than that, since its inception in 1994, until 2003, no attempt had previously been made to evaluate the validity or reliability of the instrument. As Lo et al [41] have pointed out, this is true of all the instruments used at all Hong Kong's universities:

> "All the standardised instruments are designed by working parties or committees composing of faculty members, administrators and evaluation experts. These instruments are to be reviewed on a regular basis. However in some cases, how a particular instrument was designed and developed is a myth. Many are "just there". None of the collected instruments has gone through reliability or validity tests. They are used as "they are there" with few challenges from both users and students". =(p 60). The research reported in this paper attempts to remedy this deficiency.

THE ANALYSIS

In 2001, the Quality Assurance Committee (QAC) approved a project to study the "Influence of Bias Factors on Student Ratings of Teaching", in an attempt to foster a better understanding of:

1. the psychometric properties of the six common items of the TFQ in order to determine whether they are valid and reliable measures of teaching effectiveness;
2. the relationships between student ratings and a list of potential bias variables as well as student learning variables in order to determine whether such ratings are biased; and
3. the appropriateness of using the ratings of the single global item – the TFQ overall rating item, to represent the five TFQ common dimensional rating items for making summative evaluations.

A survey was thereby conducted in CityU (referred to in this report as "the main study") to collect data from seven departments, namely, Commerce; Chinese, Translation and Linguistics; Computer Studies; English and Communication; Language Studies; Creative Media; and Law. The data were thoroughly analyzed, and the findings have been published in the QAC Report [42]. Parallel to the main study were the analyses done on a relatively limited set of data provided by the Department of Electronic Engineering (EE) for the same purpose. The results in this report are not directly comparable with those in the main study, as the number and types of variables involved are different.

Factor structure and reliability of six TFQ common items

Data used in this study were collected during the period between 1997–1998 and 2001–2002. (Note: Data for semester A, 1998-99, was incomplete, and was therefore excluded from our analyses). All analyses were done on class average scores. Exploratory factor analyses (principal-component analyses) were conducted using six TFQ common items (presentation, preparation, organisation, stimulation, responsiveness and overall rating) in order to determine the number and nature of components or dimensions underlying such items.

A one-factor solution emerged i.e the factor concerned with students' evaluations of a teacher's performance, referred to as "teaching performance". The inter-correlations (Pearson product-moment correlations) among the variables/items under this factor were between .77 and .96. The single factor accounted for 89 percent of the item variance. Each of the six common TFQ items loaded highly on the single factor: all of these loadings exceeded .92 (.92-.99, median = .94).

Deleted: ¶
Deleted: ¶
Deleted: all
Deleted: the
Deleted: ¶
Deleted: ¶
Formatted: Indent: First line: 0"
Deleted: ¶
Formatted: Font: Bold
Deleted: The factor structure of six TFQ common items
Deleted: TFQ data were collected during the school year 1997–1998 up to the school year 2001–2002. All analyses were done on class average scores.
Deleted: only the
Deleted: . (Note: The database for semester A, 1998-99, was incomplete, and was therefore excluded from our analyses).¶
Deleted: ;
Deleted: and is

The same one-factor solution was found across semesters, which accounted for 85-92 percents of the item variance (median = 89 percent). Each of the six common TFQ items loaded highly on the single factor across semesters: all of these loadings exceeded .87 (medians = 97-98A: .94; 97-98B: .91; 98-99B: .95; 99-00A: .93; 99-00B: .95; 00-01A: .95; 00-01B: .95; 01-02A: .93; 01-02B: .96).

Though the sample size for each semester was relatively small, according to Guadagnoli and Velicer [43], factors/components with four or more loadings above .60 are reliable, regardless of sample size. Stevens (1996) also stated that factors/components with at least three loadings above .80 will be reliable.

The internal consistency estimates of reliability of the six TFQ common items for the total sample gave a Cronbach α [45] of .97 - the internal consistency estimate of reliability test indicated that the responses among items within the single factor were consistent for the total sample. The cross-validation across nine semesters gave a Cronbach α of .96-.98 (median = .97), thus the internal consistency estimate of reliability tests indicated that the responses among items within the single factor were consistent across semesters.

An exploratory factor analysis (on the total sample) was conducted using both the six TFQ common items and the five TFQ optional items (communication, understanding of subject matter and three coursework related items) which are asked in the TFQ of the Department of Electronic and Engineering. This analysis was done in order to cross-validate previous findings by adding new variables (i.e. the five TFQ optional items) which are supposed to measure the same construct, namely, "teaching performance", to the analysis. The same one-factor solution was found, which accounted for 90 percent of the item variance. The inter-correlations among the variables/items under this factor were between .77 and .97.

Each of the 11 TFQ items loaded highly on the single factor: all of these loadings exceeded .90 (.90-.99, median = .95). The Cronbach α was .99, showing an internal consistency estimate of reliability test indicated that the responses among the 11 items within the single factor were consistent.

The fact that the six TFQ common items consistently tap into one and only one factor during the series of exploratory factor analyses offers clear evidence for a one-factor, unidimensional structure underlying the said items. The evidence clearly supports the internal consistency of the six TFQ common items.

The influence of potential bias/background factors on student ratings

According to Marsh [31], "The mere existence of a significant correlation between students' evaluations and some background characteristic should not be interpreted as support for a bias hypothesis…An external influence, in order to constitute a bias to student ratings, must be substantially and causally related to the ratings, and relatively unrelated to other indicators of effective teaching."

Unlike the main study, the data from EE provided a very limited set of information. Hence, only a limited set of four background factors was involved in our analyses, namely, class size, class meeting time, year of study, and students' status (i.e. full time vs. part time students). These were supplemented by another - required or elective course after the initial analysis was completed. It was hoped that the language of instruction could be included as a variable, but there were technical difficulties involved in defining the major language of instruction of a class based on students' self-reported responses to the TFQ, as few, if any, classes could reach unanimous agreement in this respect. Hence, this factor was also not included in the analyses.

All analyses were done on class average scores. A multiple regression analysis using the aforementioned four background factors as predictors was conducted to 1) examine the combined effect of all the predictors on "teaching performance"; 2) determine which of the individual predictor(s) made the largest contribution.

Based on the findings of the aforementioned factor analyses, the average of the six TFQ common items under the component "teaching performance" could be employed as the component criterion.

When all the four background variables were entered into a multiple regression to predict "teaching performance", – "class size" was selected into the equation by the stepwise procedure, which accounted for 13.2 percent of the total variance of "teaching performance".

The Stein estimate was then computed to check if our regression equation could generalize well to the population. The Stein estimate of .072 meant that if we applied our prediction equation to many other samples from the same population, on the average we would account for 7.2 percent of the variance on the criterion. In other words, compared to the 13.2 percent found in the present study, the shrinkage in predictive power would be as large as 45 percent, indicating that our equation might not be able to generalise well to the population. Hence, at this stage, a cautious approach to such regression results is indicated, in the form of additional references only.

Among the four background factors, "class size" is the only predictor selected into the equation, and accounts for 13 percent of the variance of the criterion, indicating that the two variables are correlated (r = .36). In general, the smaller the class size, the higher the student ratings. The effect is neither large nor negligible, but is statistically

Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: students,
Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: ¶
Deleted: it was found that one of them
Deleted: ,
Deleted: ¶
Deleted: Using the Stein [46] formula, it was found that
Deleted: was
Deleted: . That
Deleted: were to apply
Deleted: then
Deleted:
Formatted: Tabs: 1.13", List tab
Deleted: ¶
Deleted: is the one that requires attention. It

significant. Such results are consistent with the literature [25], [47], [48], [49]. However, the mere existence of such a correlation does not necessarily support a bias interpretation. Firstly, the existence of the correlation does not necessarily imply causal relationship. Secondly, whether the same results could hold when more predictors (including teaching effectiveness indicators and other background factors) are included in the analysis remains to be seen. Moreover, as we have mentioned earlier, the reliability of the results is questionable. Therefore, at this stage, it is premature to conclude that "class size" acts as a bias to student ratings in the EE department. Obviously, future studies with a larger set of predictors and a larger sample size should be conducted in order to draw firm conclusions on the relationships between the variables.

Deleted: And secondly

Supplementary findings

After adding "required/elective course" to the list of predictors under the same condition as that applied to predictors used in the previous study of EE TFQ data and reported above, we re-ran the multiple regression analysis using a total number of five (instead of four) predictors.

When all the five background variables were entered into a multiple regression to predict "teaching performance", it was found that "required/elective course" had replaced "class size" as the dominant and sole predictor selected into the equation (inter-correlation between the two background variables = -.34), which accounted for 13.8 percent of the total variance of "teaching performance.

The Stein estimate was then computed to check if the regression equation can be used to generalise well to the population. Using the Stein [46] formula, it was found that the Stein estimate was .081. This means that if we apply our prediction equation to other samples from the same population, then, on the average, it is possible to account for 8.1 percent of the variance on the criterion. In other words, when compared to 13.8 percent found in the present study, the shrinkage in predictive power would still be as large as 41 percent, indicating that this equation may be inappropriate as a tool to generalise well to the population. Hence, at this stage, a cautious approach to such regression results is indicated, in the form of additional reference only, for further study of data of a larger population.

Deleted: ¶

The relationships between the TFQ overall rating item and the TFQ dimensional rating items

The close relationships between the TFQ overall rating item and the TFQ dimensional rating items were confirmed by our findings.

The simple bivariate correlations between the TFQ overall rating item and the five TFQ common dimensional rating items were very high (.88 -.96, median = .92), meaning that the former was strongly related to the latter. The simple correlations between the TFQ overall rating item and the five TFQ optional items were found to be even higher (.91 -.97, median = .95).

Deleted: ¶

The close relationship between the TFQ overall rating item and the five TFQ common dimensional rating items was further supported by the fact that the six TFQ items loaded together under a one-factor solution during our previous factor analysis on the data from the total sample. The single factor explained 89 percent of the variance in student ratings. The same one-factor solution could be identified when factor analyses were conducted separately on the data from individual semesters (which explained 85-92 percent [median = 89 percent] of the variance of ratings). Note that the TFQ overall rating item consistently had the highest loadings on the single factor for the total sample (.99) as well as for all the individual semesters (.98-.99, [median = .99]). Even when the five TFQ optional items were involved in the factor analysis, the 11 TFQ items still loaded together under one and the same component which explained 90 percent of the variance of the items, while the TFQ overall rating item still had the highest loading (.99).

Deleted: ¶

Linear regression analysis using the TFQ overall rating item as the predictor was also conducted to determine how well the mean of the five TFQ common dimensional rating items (i.e. the criterion) could be predicted by the former. It was found that the former could significantly predict the latter and that the effect size was substantial, R^2 = .96, F (1, 394) = 8559.34, p = .000, β = .978. Similar results were found when we used the mean of all the ten TFQ dimensional rating items as the criterion instead: R^2 = .96, F (1, 394) = 10697.22, p = .000, β = .982.

Deleted: ¶

The strong relationships found between the TFQ overall rating item and the TFQ dimensional rating items clearly support the claim that the former can be used to represent the latter (including the five TFQ common dimensional rating items) in the context of personnel decisions, provided that the overall rating item is preceded by the dimensional rating items, and that there is corroborative evidence of teaching effectiveness from other sources as required by CityU policy on teaching evaluation.

Deleted: ¶

Deleted: serve as clear evidence to

Deleted: one can use

Deleted: is

DISCUSSION

The analyses have offered clear support to the reliability of the TFQ and a unidimensional structure underlying the TFQ items."

"Class size", a background factor, was found to be correlated with student ratings. Nonetheless, at this stage, it is premature to conclude that "class size" acts as a bias to student ratings in the EE department because the mere existence of the correlation does not necessarily imply causal relationship, and if the same set of predictors used in the main study (which includes a teaching effectiveness indicator in the form of "student competence/learning"; and other background factors like "student motivation") is involved in the analysis, and that a larger pool of data is readily accessible, the results might be different.

"Required/elective course" is found to be correlated with student ratings (r = -.37), since higher ratings are found from classes taking elective courses. The effect is neither large nor negligible, but is statistically significant. Notwithstanding such findings, at this stage, it is premature to conclude that "required/elective course" acts as a bias to student ratings in the EE department.

Strong relationships were found between the TFQ overall rating item and the TFQ dimensional rating items. That serve as clear evidence to support the claim that one can use the former to represent the latter (including the five TFQ common dimensional rating items) in the context of personnel decisions, provided the overall rating item is preceded by the dimensional rating items, and that there is corroborative evidence of teaching effectiveness from other sources as is required by CityU policy on teaching evaluation.

REFERENCES

[1] Remmers, H H (1928), "The relationship between students' marks and students' attitudes toward instructors", *School and S ociety* , v 28, pp 759-760

[2] Remmers, H H, Martin, F D and Elliot, D N (1949), "Are student ratings of instructors related to their grades?", *Purdue Studies in Higher Education* , v 66, pp 17-26

[3] Drucker, A J, and Remmers, H H (1951), "Do alumni and students differ in their attitudes towards instructors?", *Journal of Educational Psychology* , v 42, n 3, pp 129-143

[4] Cohen, P A (1981), "Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies", *Review of Educational Research* , v 51, n 3, pp 281-309

[5] Kulik, J A and McKeachie, W J (1975), "The evaluation of teachers in higher education", In Kerlinger (Ed), *Review of research in higher education* , (Vol. 3), F E Peacock, Itaska, Illinois, USA.

[6] Isaacson, R L et al (1964), "Dimensions of student evaluations of teaching", *Journal of Educational Psychology* , v 55, pp 344-351

[7] Feldman, K A (1988), Effective college teaching from students' and faculty's view: Matched or mismatched priorities?", *Research in Highe r Education* , v 28, pp 291-344

[8] Centra, J A, and Bonesteel, P (1990), "College teaching; an art or a science", *New Directions for Teaching and Learning* , n 43

[9] Sherman, T M and others (1987), "The quest for excellence in university teaching", *Journal o f Higher Education* , v 58, 66-84

[10] Abrami, P C (1989), "How should we use student ratings to evaluate teaching?, *Research in Higher Education* , v 30, pp 221-227

[11] Koon J and Murray, H G (1995), "Using multiple outcomes to validate student ratings of overall teacher effectiveness", *The Journal of Higher Education* , v 66, n 1, pp 61-70

[12] Marsh, H W and Roche, L A (1997), "Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility", *American Psyc hologist* , v 52, n 11, pp 1187-1197

[13] Greenwald, A G and Gillmore, G M (1997), "Grading leniency is a removable contaminant of student ratings", *American Psychologist"* , v 52, pp 1209-1217

[14] Abrami, P C, d'Apollonia, S and Rosenfield, S (1997), "The dimensionality of student ratings of instruction: What we know and what we do not", In Perry and Smart (Eds), *Effective teaching in higher education: Research and practice* , pp 321-367, Agathon Press, New York, USA

[15] McKeachie, W J (1997), "Student ratings: The validity of use", *American Psychologist* , v 52, pp 1218-1225

[16] Cohen, P A (1980), "Using student rating feedback for improving college instruction; A meta-analysis of findings", *Research in Higher Education* , v 13, n 4, pp 321-341

[17] Greenwald, A G (1997), "Validity concerns and usefulness of student ratings of instruction", *American Psychologist* , v 52, pp 1182-86

[18] Cohen, P A (1987), "A critical analysis and re-analysis of the multisection validity meta-analysis", *Paper presented at the Annua l Meeting of the American Educational Research Association* , Washington DC, Apirl 1987

[19] Marsh, H W and Roche, L A (1991), "The use of student evaluations of university teaching in different settings: The applicability paradigm", *University of Western Sy dney,* Sydney, Australia

[20] Marsh, H W and Roche, L A (1994), "The use of students' evaluations of university teaching to improve teaching effectiveness", *Department of Employment, Education and Training* , Canberra, ANT, Australia

[21] Cashin, W E and Downey, R G (1992), "Using global student rating items for summative evaluation", *Journal of Educational Psychology* , v 84, pp 563-572

[22] Centra, J A (1993), "Reflective faculty evaluation", *Josey-Bass Publishers* , San Francisco, California, USA

[22] Feldman, K A (1997), "Identifying exemplary teachers and teaching: Evidence from student ratings", In Perry and Smart (Eds), *Effective teaching in higher eduaction: Research and practice* , pp 368-395, Agathon Press, New York, USA

[23] Centra, J A (1973), " Self-ratings of college teachers: A comparison with student ratings", *Journal of Educational Measurement,* v 10 n 4, pp 287-295

Deleted: **Supplementary findings**¶
¶
After adding "required/elective course" to the list of predictors under the same condition as that applied to predictors used in the previous study of EE TFQ data and reported above, we re-ran the multiple regression analysis using a total number of five (instead of four) predictors.¶

Deleted:

Deleted: ¶
When all the five background variables were entered into a multiple regression to predict "teaching performance", it was found that "required/elective course" had replaced "class size" as the dominant and sole predictor selected into the equation (inter-correlation between the two background variables = -.34), which accounted for 13.8 percent of the total variance of "teaching performance. ¶
¶
The Stein estimate was then computed to check if the regression equation can be used to generalise well to the population. Using the Stein [46] formula, it was found that the Stein estimate was .081. This means that if we apply our prediction equation to other samples from the same population, then, on the average, it is possible to account for 8.1 percent of the variance on the criterion. In other words, when compared to 13.8 percent found in the present study, the shrinkage in predictive power would still be as large as 41 percent, indicating that this equation may be inappropriate as a tool to generalise well to the population. Hence, at this stage, a cautious approach to such regression results is indicated, in the form of additional reference only, for further study of data of a larger population. ¶
(This section should go after "The influence of potential bias/background factors on student ratings¶
¶

Deleted: ¶

Deleted: ¶

Deleted: better

Deleted: ¶

Formatted: Justified, Indent: Left: 0", First line: 0", Space After: 0 pt

[24] Centra, J A and Creech, F R (1976), "The relationship between student teachers and course characteristics and student ratings of teacher effectiveness", *Project Report 76 -1*, Educational Testing Service, Princeton University, Lawrenceville, NJ, USA
[25] Feldman, K A (1978), "Course characteristics and college students' ratings of their teachers abd courses: What we know and what we don't", *Research in Higher Education* , V 9, pp 199-242
[26] Marsh, H W (1987), "Student evaluations of university teaching: Research findings, methodological issues, and directions for future research", *International Journal of Educational Research* , v 11, pp 253-388
[27] Leeds, M, Stull, W A, and Westbrook, J (1998), "Do changes in classroom techniques matter? Teaching strategies and their effects on teaching evaluations", *Journal of Education for Business* , v 74, n 2, pp 75-78
[28] Marsh, H W(1982), "Validity of students' evaluations of college teaching: A multitrait-mutimethod analysis", *Journal of Educational Psychology*, v 74, n 2, pp 264-279
[29] Gage, N L (1978), "The scientific basis of the art of teaching", *Teachers College Press*, New York, USA
[30] Dunkin, M J and Barnes, J (1986), "Research of teaching in higher education", M C Wittrock (Ed), *Handbook of Research on Teaching* , 3rd Ed, Macmillan, New York, USA
[31] Marsh, H W (1984), "Students' evaluations of university teaching: dimensionality, validity, potential biases and utility", *Journal of Educational Psychology* , v 76, n 5, pp 707-754
[32] Watkins, D (1994), "Student evaluations of teaching effectiveness: A cross-cultural perspective", *Research in Higher Education* , v 35, n 2, pp 251-266
[33] Enriquez, V G (1977), "Filipino psychology in the third world", *Philippine Journal of Psychology* , v 10, pp 3-17
[34] Triandis, H C (1972), *The analysis of subjective culture* , Wiley, New York, USA
[35] Hui, C H and Triandis, H (1985), "Measurement in cross-cultural psychology: A review and comparison of strategies", *Journal of Cross - Cultural Psychology* , v 16, pp 131-152
[36] Hayton, G E (1983), "An investigation of the applicability in technical and further education of a student evluation of teaching instrument", *Faculty of Education* , University of Sydney, Australia
[37] Marsh H W, Touron, J and Wheeler, B (1985), "Students' evaluation of university instructors: The applicability of American instruments in a Spanish setting", *Teaching and Teacher Education* , v 1, n 2, pp 123-138
[38] Clarkson, P C (1984), "Papua Mew Guinea students' perception of mathematics teachers", *Journal of Educational Psychology* , V 76, pp 1386-1395
[39] Watkins, D, Marsh, H W and Young, D (1987), "Evaluating tertiary teaching: A New Zealand perspective", *Teaching and Teacher Education* , v 3 pp 41-53
[40] Marsh, H W (1991), "A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia (1991)", *Journal of Educational Psychology*, v 83, pp 416-421
[41] Lo, T, Wong, W and Barrett, J (1999), *An Evaluation Sourcebook for Higher Education in Hong Kong*, CELT, City University of Hong Kong
[42] Shah, A., Lo, T., Rudowicz, E., Smith, R., & Wong, P. (2002). QAC Project: The influence of bias factors on student ratings of teaching. A report to the Quality Assurance Committee. *Quality Assurance Committee Paper no. QAC/44/A6* .
[43] Guadagnoli, E., & Velicer, W. (1988). Relation of sample size to the stability of component patterns. *Psychological Bulletin, 103*, 265-275.
[44] Stevens, J. (1996). *Applied multivariate statistics for the social sciences.* Mahwah: Lawrence Erlbaum.
[45] Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 443-507). Washington, D.C.: American Council on Education.
[46] Stein, C. (1960). Multiple regression. In I. Olkin (Ed.), *Contributions to probability and statistics, essays in honor of Harol Hotelling* . Stanford, CA: Stanford University Press.
[47] Franklin, J., Thell, M., & Ludlow, L. (1991). Grade inflation and student ratings: A closer look. Paper presented at the Annual Meeting of the American Educational Research Association, Chicago.
[48] McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Edu cational Psychology, 82* , 189-200.
[49] Sixbury, G. R., & Cashin, W. E. (1995). *IDEA technical report no. 9: Description of database for the idea diagnostic Form.* Manhattan: Kansas State University, Center for Faculty Evaluation and Development.

Formatted: Justified, Indent: Left: 0", First line: 0", Space After: 0 pt