TEACHING EVALUATION - A COMPLICATED ISSUE

Z Lin, K H Chan and Q Xue
Division of Building Science and Technology,
City University of Hong Kong

Index: Teaching evaluation, factors, statistics

Abstract: To justify/quantify/qualify on teaching evaluation may be difficult and not objective as there are so much variables. Comparing teaching evaluations obtained by individual lecturer may not be meaningful, as each staff facing independent course with different students, course and other variables (or statistically, each is an independent event and a single mean value must not be compared directly with another).

INTRODUCTION

The City University of Hong Kong has a policy statement about evaluation and the Quality Assurance Committee has also published guidelines. The evaluation schemes are formulated and implemented within the Division of Building Science and Technology and College of Higher Vocational Studies. Theoretically, the schemes are the responsibility of the College. The Centre for the Enhancement of Learning & Teaching (CELT) provides services for processing the Teaching Feedback Questionnaires (TFQ) and Teaching Evaluation and Improvement Packages (TEIP). TFQ is a summative evaluation of teaching performance, whereas TEIP is for formative evaluation which aims at providing teachers with information to improve their teaching.

OBJECTIVES

To review existing teaching evaluation schemes with possible modification for achieving better quality and fairness in teaching evaluation. The term of "teaching evaluation" is hereof defined basically as the assessment of performance of an individual staff member. The Teaching Evaluation Scheme of the College (TES) issued by College of Higher Vocational Studies (CHVS 1996) states that:

a. The Scheme serves 2 primary functions:
   Formative Evaluation for the improvement of teaching; and
   Summative Evaluation for the appraisal and personnel-related assessments
b. Major Teaching Evaluation Tools in Practice are TEIP and TFQ.
c. Evaluation data are only worth collecting if it is used to promote improvement. Students are required to invest significant time evaluating teaching. Their response can only remain serious if the effects of their efforts are made apparent. This will also be true for the staff. It is required that students be given sufficient feedback and that follow-up actions are taken with the staff as a result of teaching evaluations.
d. In practical terms, the TEIP is formative with emphasis on self-evaluation and self-improvement, whereas the TFQ is mostly used for summative purposes. The basic rationale/emphasis on improving teaching seems to have been ignored in the TFQ exercise. Therefore, consideration should be given to the possibility of merging the aforementioned two instruments for the following reasons:
   i) Part A of TFQ is actually formative;

ii) Students, especially first year students are often unable to appreciate the differences between the TEIP and the TFQ;

iii) Frequent filling in of questionnaire fatigues students thus results in that students take the process less seriously.

The suggested new questionnaire (SNQ) should incorporate the essences of the current two forms and more descriptive information should be provided, *e.g.*, marks 1 to 7 may denote VERY POOR, POOR, UNSATISFACTORY, SATISFACTORY, GOOD, VERY GOOD and EXCELLENT respectively. The SNQ therefore will be not only formative and developmental, but also summative and judgmental.

A staff member may conduct the SNQ twice in a course, one at an intermediate occasion and the other at the end. Only the second result will be kept in the record. And the staff member may discuss the result of the first one with the students.

POSSIBLE PROBLEMS AFFECTING TFQ RESULTS

A. Statistics Related Matters

For a large sample size (*i.e.* number of response $n \geq 50$), it probably is reasonable to assume that the random variable of error follows the standard normal distribution (Chen *et al.* 1980, Freedman *et al.* 1998).

*Assumptions***:** For small sample sizes, we have to assume that like most practical problems, the matter of student feedback could be treated as a normal distribution, otherwise the technical problem is beyond the expertise of the taskforce. Furthermore, we have to assume that different sample sizes (student groups, often of different group size) are of identical nature, which is very much questionable statistically and practically, in order to compare between teaching staff members or between courses.

The following discussion will use three examples to illustrate the techniques that are considered suitable for tackling the problems if the aforementioned assumptions are valid.

Example 1 Staff Somebody's TFQ feedback for BST0000 is as the follows:

	Mean ($\bar{u}$)	*Standard Deviation (SD)*	*Number of Response (n)*
Overall Rating	*4.80*	*1.05*	*61*

Find the 95% confidence interval for the result.

Solution:

$\bar{u}$ = 4.8, SD = 1.05, n = 61.

$$SD^{+} = SD\sqrt{\frac{n}{n-1}} = 1.05\sqrt{\frac{61}{61-1}} = 1.06$$

The expected value u for the 95% confidence level should fall in the interval of

$$\left(\bar{u} - l\,\frac{SD^{+}}{\sqrt{n}},\quad \bar{u} + l\,\frac{SD^{+}}{\sqrt{n}}\right)$$

For degree of freedom N = n - 1 = 60, $\alpha = 1 - 95\% = 0.05$, from the *t*-distribution (Student curve) table:

$\lambda = 2.000$, therefore $(4.53 < u < 5.07)$ for a 95% level of confidence.

Example 2 Staff Someone's TFQ feedback for BST9999 is as the follows:

	Mean ($\bar{u}$)	*Standard Deviation (SD)*	*Number of Response (n)*
Overall Rating	*6.29*	*0.49*	*7*

Find the 95% confidence interval for the result.

Solution:

$\bar{u} = 6.29$, $SD = 0.49$, $n = 7$.

$$SD^+ = SD\sqrt{\frac{n}{n-1}} = 0.49\sqrt{\frac{7}{7-1}} = 0.53$$

The expected value *u* for the 95% confidence level should fall in the interval of

$$\left(\bar{u} - l\,\frac{SD^+}{\sqrt{n}},\quad \bar{u} + l\,\frac{SD^+}{\sqrt{n}} \right)$$

For degree of freedom N = n -1 = 6, $\alpha = 1 - 95\% = 0.05$, from the *t*-distribution (Student curve) table:

$\lambda = 2.447$, therefore $(5.80 < u < 6.78)$ for a 95% level of confidence.

Example 3 Compare the aforementioned two feedback results with different sample sizes.

Solution:

Step 1:

$$T = \frac{\left|\bar{u}_1 - \bar{u}_2\right|}{\sqrt{SD_1^{\,2} + SD_2^{\,2}}}\sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}}$$

where $\bar{u}_1 = 4.80$, $SD_1 = 1.05$, $\bar{u}_2 = 6.29$, $SD_2 = 0.49$ and $T = 26.178$

Step 2: $N = n_1 + n_2 - 2 = 66$, $\alpha = 0.05$ (for 95% confidence level)

From t-distribution: $\lambda = 1.990$

Step 3: If $T > \lambda$, then $u_1 \neq u_2$

i.e. u_2 represents a better result than u_1 does.

(If $T < \lambda$, then $u_1 = u_2$).

Based on the above discussion, we concluded that:

a. From the viewpoint of statistics, the TFQ result is an interval, not a digital figure (Example 1 & 2). Different results have different intervals;

b. Based on Point a, values related to arithmetic average of various TFQ mean expected values ($\bar{u}$), *e.g.* the divisional mean is less significant statistically;

c. To simply compare any two TFQ mean expected values ($\bar{u}$) has little meaning statistically. A proper method for comparison is illustrated via Example 3.

R`ecommendations:

1. The TFQ result should be presented in the straightforward 'interval' format rather than in the 'sample mean + standard deviation' format;
2. A reference point for Divisional standard somewhere below the Divisional median value could be set. A resultant interval around or beyond the set point should be deemed as acceptable.

B. The impact of oral medium of instruction on evaluation results:

The City University of Hong Kong is committed to excellent teaching, which includes using English as teaching medium. Fluency in English will no doubt add weight for students in the keen job market. As known to all, Hong Kong students feel easier and more comfortable in listening to their home dialect Cantonese in class (Actually this argument has been used for the government to promote teaching with mother tongue in primary and secondary schools). In the situation of Cantonese popularizing in class, those teachers who use English as teaching medium will naturally be in a disadvantaged position.

To encourage more use of English in class, we suggest adding one more item in TFQ: Does the teacher use English in class?
_____ 100%, ______75%,_______50%, ________25%, ________None.
100%, multiply 1.2; 75%, 1.1; 50%, 1; 25%, 0.9; none, 0.8. These coefficients will be applied to the overall evaluation (Part B) as encouragement for sticking to the University's official policy on language of instruction.

C. Reliability of Quantified Teaching Evaluation

Up to date, teaching cannot be fully quantified. Staff members feel threatened by the use of students' evaluation as the major measurement instrument to judge their capacity, performance and contribution as a teacher, yet students see little improvement as a result of evaluation. Quantification of teaching needs more detailed research, as teaching mode, role of the particular teaching staff (for example, Course Examiner), class size etc. These variations should be reflected in the calculation of teaching workload. TFQ is vital in stimulating teaching but sometimes produce pressure for teaching staff. For example, strictness to students and persistency in standards may offend students and result in low TFQ. Therefore, TFQ could only be one of the factors in evaluating a teaching staff. The other performance in teaching may also be quantified besides TFQ.

D. Factors affecting feedback of a lecturer's true performance in teaching:

***Nature of course*:** Most learners/students tend to dislike tough (theoretical, abstract, and/or complex) subjects/courses, especially those not closely related to their own discipline/core profession according to their perception. Their inability to learn/absorb effectively from the lecturer would be the ultimate fault of the lecturer, but not themselves (Cashin 1990, Cranton and Smith 1986, Feldman 1978). They may have not paid the effort in devoting to learn hard; instead they put the blame on the lecturer. This may be more severe for new students when they first join the University from post secondary schools, facing a new transition of education system e.g. credit unit system; and just acclimatized themselves into higher vocational training or para-professional courses where they have not studied previously. Most Level one courses are fundamental courses and integrated with Out of Discipline, Chinese Civilization, Language courses which in turn may make their curriculum tougher, posting a negative effect towards teaching evaluation e.g. TFQ. The situation may turn better when they are more accustomed to this mode of study environment in higher/final years.

Maturity of students: It has been argued that majority of the students are not qualified to give feedback on (or to evaluate) lecturers' teaching, as they do not possess necessary knowledge to judge objectively. However, they could, to some extent, feedback on the way a lecturer delivers his lecture/tutorial/studio, etc., but not on the academic standing yet. A serious lecturer may pose a very serious attitude towards lecturing while not smiling nor being humorous. There has been research stating that students tend to like the course when they like the lecturer/style of lecturing, thus may learn better. Each evaluation instrument should be designed for a specific audience and should only include items for which that audience is capable of giving informed responses. Evaluation should be explained to all students. For the elicitation of reliable and accurate results, students should be properly briefed on the purpose of the evaluation before undertaking the questionnaire completion. The overall average score of the TFQ for academic staff is released to students as a reference (CHVS). According to Knapper (1997), it is believed that students who obtain good grades from their teachers are likely to mark their teachers higher and vice versa. Students may not have certain limitation in making fair judgement. There was a perception by some staff members that results on the TFQ largely reflected students' liking of a lecturer, which in turn was fostered by “mollycoddling” students.

Way of teaching: An active lecturer who employs lively teaching methods together with his friendly, humorous, open-minded, helping character (not sacrifice on academic matters to students) may score more positive feedback from students. Alternatively, a tough/blunt lecturer who is not easy-going or being difficult to approach by his students may score more negative feedback. Students tend to find the easiest way to pass or obtain high marks in a course, without spending too much time/effort. A lecturer will have to (or bound to) make their learning more convenient and effective, in order to obtain positive feedback/better TFQ results. Too much or too less notes could be blamed by students. Involvement of too advanced technology in teaching *e.g.* certain computer software where students are difficult to follow, may receive negative feedback. Joint/shared teaching for a particular course may also affect students' marking (being the evaluator in TFQ) for each and every of the lecturers, where one lecturer may teach the tough part of the syllabus while the other(s) may teach the easy part. A responsible lecturer may not be a popular lecturer. To what extent shall we serve/suit/satisfy the students? Shall we treat them only as customers ultimately?

New Course/Old Course: A lecturer who has taught the same course for several times may be more beneficial to a lecturer who teaches new course, as he understands more the syllabus, developed more useful materials and grasped the needs or taste of students. Then he may obtain positive feedback/evaluation from students versus another lecturer who may not possess all these advantages. The latter may end up with negative feedback/ evaluation, especially when facing a transition period of new syllabus under credit unit system, new types of students, *etc*. But a teacher's evaluation for a particular course should not be a one-time practice. It may continue over a period to see the progress if it is allowed practically. Another matter is whether the courses wherein TFQ is conducted should be the choices of the staff member or the management. This will be affected largely by the factors such as teaching of core or non-core, fundamental or practical courses, class size, teaching load, and *etc*.

Bargain: It may be a fallacy that bargains exist between lecturers and students. Students may not truly reflect a lecturer's teaching where they fear that their feedback/evaluation may affect the lecturer's grading, in turn affect their examination grading, or vice versa. There are established mechanism for academic matters/grading e.g. moderation of examination papers/coursework/design projects where a bargain may not be that easy, unless a lecturer is not professional/genuine. A lecturer who fails many students in a course may still obtain positive feedback/evaluation. Alternatively, will a lecturer who give high marks to many students surely not receive negative feedback or evaluation? There may not be absolute evidence/verdict to prove a bargain exist or not exist.

OVERALL COMMENT

With the forgoing discussion, to justify/quantify/qualify on teaching evaluation may be difficult and not objective with so many variables. Comparing teaching evaluations obtained by individual lecturer may not be meaningful, as each staff is facing independent course with different students, course and other variables (or statistically, each is an independent event).

Those entrusted with using the information from teaching evaluations for decision making related to career progression should be skilled in interpreting and drawing together the difference sources of information.

It has been commented that students receive little feedback from the information which they provide. Students see little improvement as a result of evaluation. Students are overloaded with questionnaires and do not take them seriously. Staff members feel threatened by the use of student evaluation only to judge their capability as teachers. All these factors may have further weakened the degree of trustworthiness of teaching evaluation.

RECOMMENDATIONS

a. TEIP and TFQ should be merged and a new questionnaire developed in the way that it is formative, summative, developmental and judgmental;
b. Resultant data should be presented in the way of an interval format rather than in the 'sample mean + standard deviation' format;
c. A reference point for Divisional standard (perhaps, somewhere below the Divisional median value) could be established. A resultant interval around or beyond the set point should be deemed as acceptable.

REFERENCES

Chen Jiading, Liu Wanru & Wang Rengong (1980). Probability and Statistics (lecture notes), Department of Mathematics, Peking University.

Cashin, W E (1990) students do rate different academic fields differently. New Direction for Teaching and Learning, no. 43, pp 113 – 121.

CHVS (1996): Teaching Evaluation Scheme of the College (2nd Ed.)

Cranton, P A and Smith, R A (1986) a new look at the effect of course characteristics on student ratings of instruction. American Educational Research Journal, Vol 23, pp 117 – 128.

David Freedman, Robert Pisini & Roger Purves (1998). Statistics, 3-ed. W W Norton & Co., New York.

Feldman, K A (1978) course characteristics and college students' ratings of their teachers and courses: what we know and what we don't. Research in Higher Education, 9, pp 199 – 242.

Knapper, C (1997) a report on interview with CityU faculty and school representatives. Queen's University, Canada.