Student evaluation of teachers in an engineering department

Robin Bradbeer, Department of Electronic Engineering, City University of Hong Kong, Hong Kong,

Aman Shah, Tracy Lo and Philip Wong, Educational Development Office, City University of Hong Kong, Hong Kong,

Background

- Criticism within university had been building for a number of years
- No validity or reliability analysis had ever been carried out on CityU's TFQ – the "instrument" was "uncalibrated"
- Hundreds of studies worldwide has shown Student Rating of Teachers to have many potential biases
- Senior management in many academic departments and the faculties had no understanding of TFQ – using the "uncalibrated instrument" for personnel decisions
- In EE, criticisms along these lines were voiced very strongly at the last Staff Retreat

Background

- 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:
  - 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
  - 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
  - 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.

Background

- A survey was thereby conducted in CityU 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 analysed, and the findings have been published in a QAC Report.
- Parallel to the main study analyses were done on a relatively limited set of data provided by the Department of Electronic Engineering (EE) for the same purpose. (FSE did not take part in the main study)

Possible biases

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.

Possible biases

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.

The EE study

- 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.

- These looked at:
  - Factor structure and reliability
  - The influence of potential/background factors
  - The relationships between the TFQ overall rating item and the TFQ dimensional relationships

Factor structure and reliability

- 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 relationships between the TFQ overall rating item and the TFQ dimensional relationships

- The strong relationships found between the TFQ overall rating item and the TFQ dimensional rating items serve as clear evidence to support the claim that one can use the former to represent the latter.

The influence of potential/background factors

		"Student overall ratings"	
Factors/groups	*N*	*M*	*SD*
1. Year of study			
a. Yr. 1	43	2.77	.58
b. Yr. 2	45	2.72	.52
c. Yr. 3	62	2.42	.45
2. Full time vs. part time students			
a. Full time	110	2.63	.56
b. Part time	65	2.48	.50

The influence of potential/background factors

		"Student overall ratings"	
Factors/groups	*N*	*M*	*SD*
3. Class size			
a. 1 – 20	35	**2.07**	.40
b. 21 – 50	93	**2.41**	.62
c. 51 – 100	116	**2.69**	.51
d. 101 or above	24	**2.66**	.40

The influence of potential/background factors

		"Student overall ratings"	
Factors/groups	*N*	*M*	*SD*
4. Class meeting time			
a. 8:00 – 11:59	28	2.49	.52
b. 12:00 – 15:59	29	2.63	.73
c. 16:00 – 19:59	76	2.44	.54
5. Elective vs. required			
a. Required course	198	**2.85**	.55
b. Elective course	198	**2.44**	.47

Notes

1. Unit of analyses: Class average scores.
2. “Teaching context” groups are not included, as there are too few cases found under “Tutorial” (2), “Studio” (1) and “Laboratory” (4) groups.
3. “Semester” groups are also not included, as this factor has never been regarded as a potential bias factor in previous literature.
4. “Service” courses not included as there are too few cases. (Anecdotal evidence suggests that the variation in class average scores is even greater for non-EE “service” courses.)

Conclusions

- The analyses have offered clear support to the reliability of the TFQ and a unidimensional structure underlying the TFQ items.
- “Class size”, and “Required/elective course”, as background factors, were found to be correlated with student ratings.
  - Nonetheless, at this stage, it is premature to conclude that either of these acts as a bias to student ratings in the EE department because the mere existence of the correlation does not necessarily imply causal relationship.

Conclusions

- Strong relationships were found between the TFQ overall rating item and the TFQ dimensional rating items. These 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.