Evaluating teaching effectiveness using quantitative student feedback

Mandar Chitre
Department of Electrical & Computer Engineering
National University of Singapore
Singapore
mandar@nus.edu.sg

Dipti Srinivasan
Department of Electrical & Computer Engineering
National University of Singapore
Singapore
dipti@nus.edu.sg

Abstract— In spite of many shortcomings of the student feedback system, it is still one of the most valuable tool to measure student learning and the effectiveness of teaching. The student feedback scores, however, provide a biased estimate of teaching evaluation. Commonly cited biases include effect of class size, teaching activity type and level. In this paper, we outline our findings and recommendations on how student feedback scores should be used to produce an unbiased minimum variance estimate of teaching effectiveness. Data from a large engineering department involving 100 academic staff members and about 2500 students was used to validate the model proposed in this paper.

Keywords—student feedback, teaching effectiveness, performance evaluation

I. INTRODUCTION

Evaluation of teaching can take a number of forms, suited to a range of purposes [1-6]. Since assessment and evaluation of teaching is a process that relies on the subjective judgment of evaluation panel, peers, students, and staff member's self-evaluation, a set of truly objective criteria is an impossible goal [2]. However, these multiple sources of information can be used in appropriate ways to reduce and control subjectivity, and develop a fair and equitable system. Many universities around the world use quantitative and qualitative feedback obtained from students for evaluating a staff member's performance, and to encourage and motivate teaching scholarship [7-10]. Student feedback plays an important role in this process, enabling all stake holders to obtain information about the student experience at a number of levels, linking it to reflective practice, action and quality enhancement [11-15].

Student feedback is one of the most common sources of information for assessing teaching, for which the data can be systematically acquired. This information is quantifiable and can be obtained from student feedback data. Student feedback scores and additional sources of information can then be used to provide valid, summative information regarding the quality of teaching and formative direction for faculty improvement efforts [16-18].

Students play a very important role in the evaluation of teaching, since they are, after all, the group most directly affected by the quality of teaching. Student feedback forms given at the end of the semester provide quantitative and qualitative data that can be used for summative as well as formative purposes. Research indicates that although students are in the best position to evaluate specific and critical aspects of classroom teaching, these ratings often contain biases, and should not be directly used to evaluate teaching [16, 18].

Faculty members are often concerned about the use of student feedback in teaching evaluations citing variables such as class size, teaching activity (lecture or tutorial), level (undergraduate or postgraduate), or whether the module was compulsory or elected.

This paper addresses this issue, and proposes a mathematical model for obtaining unbiased minimum variance estimate of teaching effectiveness using feedback scores submitted by students. Statistical analysis was performed using three years of data from student feedback scores from a large university department with over 100 academic staff members who were involved in about 700 teaching activities. This data was analyzed to suggest ways to minimize the effect of commonly cited biases such as the effect of class size, teaching activity type and level. The corrected feedback scores using this approach have been used in evaluating teaching performance for the last three years, and results have been very satisfactory. The findings and recommendations on how student feedback scores should be used to produce an unbiased minimum variance estimate of teaching effectiveness are explained in the paper.

II. STUDENT FEEDBACK

The student feedback system in our university asks the students eight questions on the teaching effectiveness of a teacher. The first seven questions (Q1–Q7) shown as follows ask specific questions, whereas Q8 asks about the overall effectiveness of the teacher.

Q1. The teacher has enhanced my thinking ability.
Q2. The teacher provides timely and useful feedback.
Q3. The teacher is approachable for consultation.
Q4. The teacher has helped me develop relevant research skills.
Q5. The teacher has increased my interest in the subject.
Q6. The teacher has helped me understand how to apply knowledge
Q7. The teacher has enhanced my ability to learn independently

Fig. 1 shows a plot of the average of Q1–Q7 scores against the Q8 scores. The scores are highly correlated and suggest that Q8 adequately captures the information from Q1–Q7. Hence in this paper we focus our analysis only on Q8 feedback scores. Preliminary statistical analysis of Q8 scores from student feedback data for the department over three academic years shows that the commonly believed biases due to the effect of class size, teaching activity type and level are indeed real. The "before correction" plots in Figs. 2 and 3 show these biases clearly. Although the activity type-level has a stronger effect than class size, both effects are considered significant and therefore have to be corrected for in evaluating teaching effectiveness in an unbiased way.

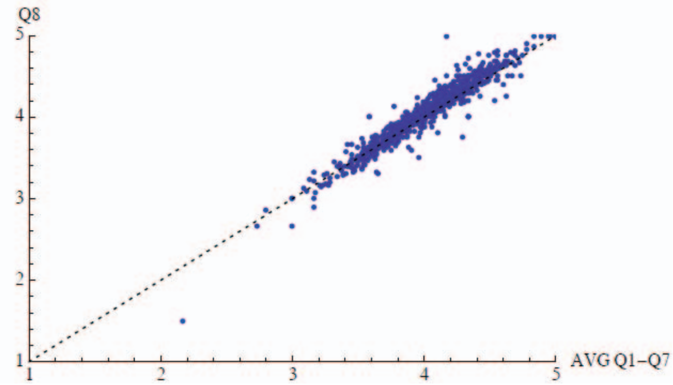

Fig. 1. Q8 feedback scores are highly correlated with the average of Q1-7 feedback scores.

In section III we show that this correction simply involves subtraction of two correction factors determined based on the class size and level/type of each module.

Once the class size and level/type correction is applied, the scores from various modules taught by the same teacher in an academic year can be combined into a single teaching effectiveness score. The way to combine the scores into a minimum variance estimate in light of varying response sizes is the subject of section IV.

For a required precision in the final score, we compute the minimum response size requirements. We find that the response size of an individual module is immaterial to the precision of the final score. What matters is the total number of students who provide feedback on a teacher's effectiveness across all modules taught by the teacher. We show in section IV that only about 25 student responses are required for an accurate measurement of a teacher's effectiveness.

A. *Student feedback – quantitative score*

The student feedback process is modeled as a noisy linear measurement process and an expression for an unbiased minimum variance estimate of teaching effectiveness is obtained as shown in the next section. The combining of scores is via a mean weighted by feedback response size. A total response size of 25 across all modules taught by a teacher in an academic year is sufficient for a good estimate of teaching effectiveness.

B. *Student feedback – qualitative comments*

The quantitative scores obtained above can be combined with another score based on qualitative comments provided by the students. The rubrics used for using the qualitative comments from the students on a 5-point scale is as follows:

5	Student comments indicate that the teacher is excellent. Examples include comments that suggest that the teacher is highly effective at developing student understanding, explaining concepts; provides differentiated ways of engaging with content specific to individual student needs; frequently develops higher-level understanding through various means.
4	Student comments indicate that teacher is effective at developing student understanding and mastery of lesson objectives, demonstrates content knowledge and delivers content that is clear, concise and well-organized; uses multiple ways to increase understanding; adjusts lesson accordingly to accommodate for student prerequisite skills and knowledge so that all students are engaged.
3	Teacher sometimes checks for understanding of content, but does not make any special effort; is often successful in pitching the module at the right level and encourage understanding; teacher may attempt to make adjustments to instruction based on feedback from students but these attempts may be misguided and may not increase understanding for all students
2	Teacher needs improvement at developing student understanding and mastery of lesson objectives although the content is factually correct; the lessons lack clarity and are not well organized; teacher does not adequately emphasize main ideas, explanations are not clear; teacher may make attempts to engaging students, but perhaps not aligned to lesson objective or mastery of content
1	Teacher is ineffective at developing student understanding; ineffective at demonstrating and clearly communicating content knowledge to students; the content is factually incorrect; explanations may be unclear or incoherent and fail to build student understanding of key concepts; teacher is ineffective at modifying instruction as needed

III. Teaching Activity and Class Size Correction

A. *Mathematical Formulation*

We assume that each teacher T has a zero-mean teaching effectiveness score x_T associated with him/her. A positive score indicates an effective teacher, while a negative score indicates an ineffective teacher. We can only make noisy measurements of this score through student feedback score y_j (1–5) for module j taught by teacher T_j. The measurements are biased depending on the level and type L_j of teaching activity and class size N_j . We denote the level–type combination, from year 1 (first year undergraduate course) to year 6 (senior year postgraduate course), and lecture/tutorial activity of a module by $L_j \in \{1,1',2,2',3,4,5,6\}$, where the prime denotes a tutorial activity. We assume that the measurement

$$y_j = x_{T_j} + \alpha(L_j) + \beta(N_j) + \omega_j \tag{1}$$

where ω_j denotes additive zero-mean measurement noise as a result of student variability and response quantization, and

$\alpha(\cdot)$ and $\beta(\cdot)$ are some functions of level/type and class size respectively. Since $\alpha(\cdot)$ is defined on a finite set, we simply define it by tabulating its value α_L for each possible level/type L. We define $\beta(\cdot)$ by quantizing the class size into four different groups (small, medium[1], large and very large):

$$\beta(N) = \begin{cases} \beta_S & \text{if } N \le 20 \\ \beta_M & \text{if } 20 < N \le 60 \\ \beta_L & \text{if } 60 < N \le 100 \\ \beta_X & \text{if } N > 100 \end{cases} \quad (2)$$

Setting

$\phi = \left(\alpha_1, \alpha_{1'}, \alpha_2, \alpha_{2'}, \alpha_3, \alpha_4, \alpha_5, \alpha_6, \beta_S, \beta_M, \beta_L, \beta_X\right)^T$, we can write Eq. (1) as:

$$y_j = x_{T_j} + \alpha_j \phi + \omega_j \quad (3)$$

where α_j is a row vector with all zeros and two ones corresponding to the level/type and class size. A noisy estimate $\hat{x}_{T_j}$ of the teaching score is

$$\hat{x}_{T_j} = x_{T_j} + \omega_j = y_j - \alpha_j \hat{\phi} \quad (4)$$

where $\hat{\phi}$ is an estimate of the correction vector ϕ.

It should be noted that the maximum size for the "medium" class was set to 60 to coincide with threshold to move from a seminar room to a lecture theater.

IV. Estimate of Correction Vector

Writing data for all modules under consideration in vector form, we rewrite (3) as:

$$y = x + A\phi + \omega \quad (5)$$

where A is a matrix with rows α_j. Since x and ω are unknown zero-mean random variables, we obtain an estimate

$$\hat{\phi} = A^+ y \quad (6)$$

where A^+ is the Moore-Penrose pseudo-inverse of A.

A. Historical Data Analysis

We obtained an estimate $\hat{\phi}$ from all student feedback data for the department modules taught during three consecutive academic years:

$$\hat{\phi} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_{1'} \\ \hat{\alpha}_2 \\ \hat{\alpha}_{2'} \\ \hat{\alpha}_3 \\ \hat{\alpha}_4 \\ \hat{\alpha}_5 \\ \hat{\alpha}_6 \\ \hat{\beta}_S \\ \hat{\beta}_M \\ \hat{\beta}_L \\ \hat{\beta}_X \end{pmatrix} = \begin{pmatrix} 1.31147 \\ 1.40173 \\ 1.32857 \\ 1.25245 \\ 1.30304 \\ 1.32538 \\ 1.42938 \\ 1.48337 \\ 2.83918 \\ 2.73185 \\ 2.66025 \\ 2.60411 \end{pmatrix} \quad (7)$$

The correction vector agrees with the general belief that smaller class sizes and postgraduate courses have better feedback scores. By applying (4) we correct for this bias to obtain corrected feedback scores $\hat{x}_{T_j}$.

Fig. 2 shows the feedback scores before and after correction as a function of level/type of teaching activity. Fig. 3 shows the feedback scores before and after correction as a function of class size. It is clearly seen that the bias seen in the feedback scores is completely removed in the corrected scores. Fig. 4 shows a histogram of corrected feedback scores. As expected, the scores are zero-mean but have a slight skew with a longer negative tail. This suggests that it is difficult to be much better than average, but relatively easy to be much worse.

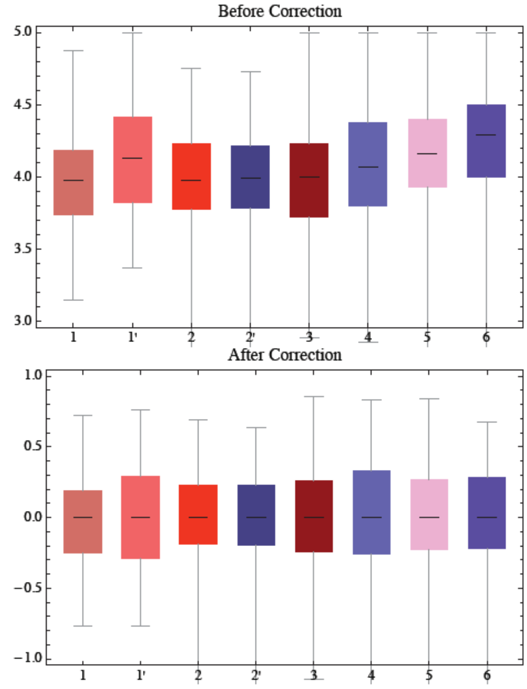

Fig. 2: Average feedback scores before and after correction as a function of level/type of teaching activity. Labels 1–6 denote lecture activities while 1' and 2' denote tutorial activities. The average feedback score after correction is zero irrespective of the level/type of teaching activity.

V. COMBINING MULTIPLE MODULE SCORES

A teacher typically teaches more than one activity during an academic year. Our goal is to combine the corrected feedback scores $\hat{x}_{T_j}$ from all modules $\mathcal{J}_T = \{j \forall T_j = T\}$ taught by teacher T into a single estimate $\hat{x}_{T_j}$ of the teaching effectiveness.

A. Mathematical Formulation

Measurement y_j is obtained by averaging the student feedback responses from R_j students who provide feedback. Hence the standard deviation σ_{ω_j} of the measurement noise ω_j is given by the standard error of the mean, i.e.,

$$\sigma_{\omega_j} = \frac{\sigma}{\sqrt{R_j}} \tag{8}$$

where σ is the standard deviation of the feedback error from a single student. As one would expect, $\sigma_{\omega_j} \to 0$ as $R_j \to 0$. Let us obtain estimate $\hat{x}_{T_j}$ as a weighted sum of estimates from each module:

$$\hat{x}_T = \sum_{j \in \mathcal{J}_T} \lambda_j \hat{x}_{T_j} \tag{9}$$

where $\sum_{j \in \mathcal{J}_T} \lambda_j = 1$.

We wish to find the optimum weights λ_j such that $\hat{x}_{T_j}$ is a minimum variance estimate of x_T. From (4), we have:

$$\hat{x}_T = \sum_{j \in \mathcal{J}_T} \lambda_j \left(x_{T_j} + \omega_j \right) = x_T + \sum_{j \in \mathcal{J}_T} \lambda_j \omega_j \tag{10}$$

Assuming that the errors ω_j are uncorrelated,

$$Var[\hat{x}_T] = \sum_{j \in \mathcal{J}_T} \lambda_j{}^2 \frac{\sigma^2}{R_j} = \sigma^2 \sum_{j \in \mathcal{J}_T} \frac{\lambda_j{}^2}{R_j} \tag{11}$$

We therefore have a constrained minimization problem of minimizing $Var[\hat{x}_T]$ subject to $\sum_{j \in \mathcal{J}_T} \lambda_j = 1$.

This is easily solved using the method of Lagrange multipliers to yield the solution:

$$\lambda_j = \frac{R_j}{\sum_{j \in \mathcal{J}_T} R_j} \tag{12}$$

Hence

$$\hat{x}_T = \frac{\sum_{j \in \mathcal{J}_T} R_j \hat{x}_{T_j}}{\sum_{j \in \mathcal{J}_T} R_j} \tag{13}$$

is the desired estimate.

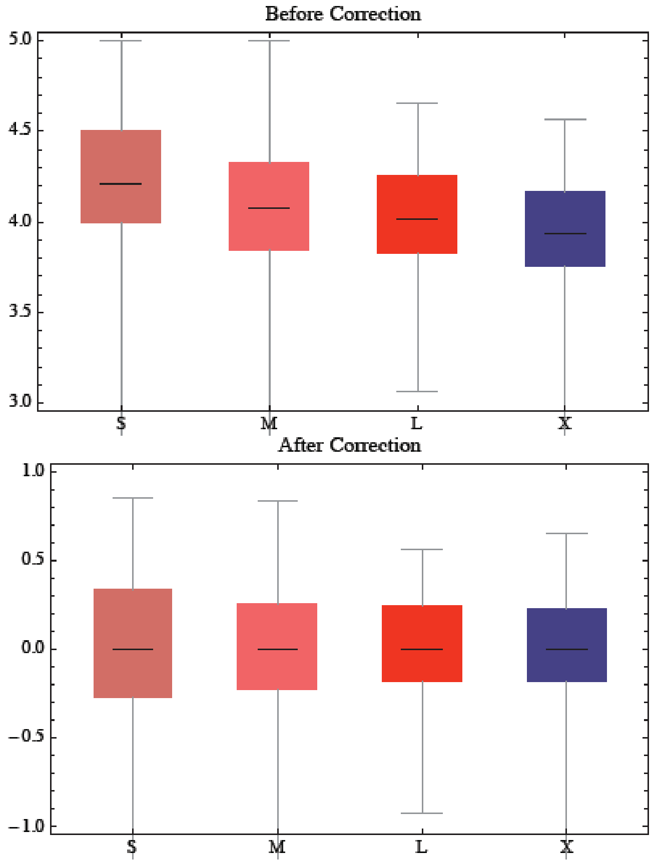

Fig. 3. Average feedback scores before and after correction as a function of class size. The average feedback score after correction is zero irrespective of the class size.

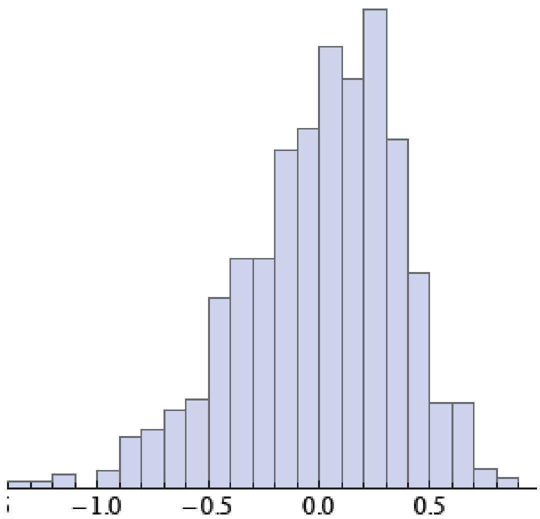

Fig. 4. Histogram of corrected scores are zero-mean but slightly skewed with a longer negative tail.

B. Historical Data Analysis

Data from a large engineering department involving 100, academic staff members, 700 teaching activities, and about 2500 students studying undergraduate (levels 1-4) and postgraduate (levels 5-6) courses was used to validate the model proposed in this paper.

Based on the actual data set for an academic year, we compute the teaching evaluation scores for each teacher using (13) and (4). A histogram of the scores is shown in Fig. 5. The mean score is zero. The median score is 0.04, with 50% of the teaching staff having scores between -0.25 and 0.23.

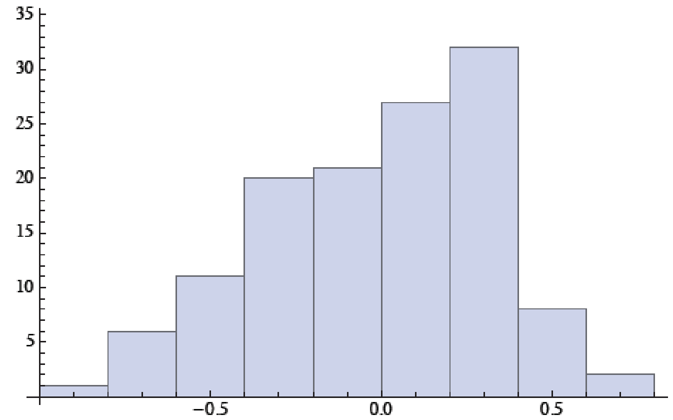

Fig. 5. Histogram of final teaching evaluation scores for teachers.

VI. MINIMUM RESPONSE SIZE

Substituting (12) in (11), we get the variance of our teaching evaluation score estimate:

$$Var\left[\hat{x}_T\right] = \frac{\sigma^2}{\sum_{j\in\mathcal{J}_T} R_j} \tag{14}$$

The variance only depends on the total number of students that provided feedback across all modules taught by the teacher in an academic year.

An estimate of σ^2 can be obtained from historical feedback scores. By sampling a few modules, we believe $\sigma^2 \approx 0.5^2$. If we desire a resolution of 0.1 on the teaching evaluation score,

$$Var\left[\hat{x}_T\right] \le 0.1^2$$

Therefore,

$$\sum_{j\in\mathcal{J}_T} R_j \ge 25 \tag{15}$$

An interesting finding from this analysis is that we need a minimum of 25 students across all modules taught by a teacher in an academic year for a good estimate. Although this is not a very onerous requirement, if the number of student responses is small, some teachers may not have this minimum number requirement met. This number should therefore be managed through careful teaching assignment.

VII. MAPPING TO A SCORE OF 1–5

For the recommended scoring for the purposes of obtaining an overall student feedback scores that combines quantitative and qualitative data, the student feedback score can be mapped to any scale. IN our analysis, this it mapped to range 1–5. From Fig. 5, we see that most of the aggregate teaching evaluation scores are between -0.6 and 0.6. We therefore recommend the following expression to map the teaching evaluation score into a range of 1–5:

$$\hat{x}_T^{(1-5)} = \begin{cases} 1 & \hat{x}_T < -0.6 \\ 3 + 3.33\hat{x}_T & -0.6 \le \hat{x}_T \le 0.6 \\ 5 & \hat{x}_T > 0.6 \end{cases} \tag{16}$$

This mapping should be reviewed based on the histogram of the teaching evaluation scores of all staff during each annual review.

VIII. CONCLUSIONS

This paper presents a mathematical model for obtaining unbiased student feedback scores for assessing teaching in university department offering a variety of courses with varying number of students, at different levels, and involving various teaching activities. The multiple sources of information can be used in appropriate ways to reduce and control subjectivity, and develop a fair and equitable system, as shown in this paper. This paper shows how the student feedback process can be modelled as a noisy linear measurement process and presents an expression for an unbiased minimum variance estimate of teaching effectiveness. The estimate corrects for bias due to teaching activity level, type and class size, and combines the feedback data from all modules taught by a teacher in an academic year. The correction involves two lookups and additions and is therefore easy to implement. The combining of scores is via a mean weighted by feedback response size. A total response size of 25 across all modules taught by a teacher in an academic year is sufficient for a good estimate of teaching effectiveness.

Data from a large engineering department involving about 100, academic staff members, 700 teaching activities, and 2500 students studying undergraduate (levels 1-4) and postgraduate (levels 5-6) courses was used to validate the model proposed in this paper. The corrected feedback scores using this approach have been used in evaluating teaching performance for the last three years, and results have been very satisfactory.

ACKNOWLEDGMENT

We would like to thank members of the teaching evaluation task force, department management committee and academic committee for constructive feedback that helped refine the method outlined in this paper.

REFERENCES

[1]. Remedios, R., Lieberman, D.A., I liked your course because you taught me well: The Influence of grades, workload, expectations and goals on students' evaluations of teaching (2008) British Educational Research Journal, 34 (1), pp. 91-115.

[2]. Ronald A. Berk, Survey of 12 strategies to measure teaching effectiveness (2005) International journal of TeachIng and Learning In Higher Education, 17 (1), pp 48-62.

[3]. John A. Centra, Student Ratings of Instruction and Their Relationship to Student Learning, JA Centra - American educational research journal, 1977

[4]. Ghedin, E; Aquario, D, Moving towards multidimensional evaluation of teaching In higher education: A study across four faculties (2008) Higher Education, 56 (5), pp 583-597.

[5]. Alberto F. Cabrera, Carol L. Colbeck and Patrick T. Terenzini, Developing Performance Indicators for Assessing Classroom Teaching Practices and Student Learning Research In Higher Education (2001), Volume 42, Number 3, 327-352

[6]. Aleamoni, L.M. The usefulness of students' evaluations In improving college teaching. 1978. *Instructional Science*, 7, 95-105.

[7]. Francis F. Balahadia et. Al., Teacher's performance evaluation tool using opinion mining with sentiment analysis, IEEE Region 10 Symposium (TENSYMP), 2016, pp. 95 - 98

[8]. Huma Fawad and Irfan Anjum Manarvi, Student Feedback & Systematic Evaluation of Teaching and its correlation to Learning Theories, Pedagogy & Teaching Skills, IEEE International Conference on Teaching, Assessment and Learning for Engineering (TALE), 2014 , pp. 398 – 404.

[9]. Axel Böttcher, Andreas Kämper and Veronika Thurner, On feedback techniques for the evaluation of teaching effectiveness, IEEE Global Engineering Education Conference (EDUCON), 2015, pp. 668 - 675

[10]. Hardy, N., Online Ratings: Fact and Fiction, In Online Student Ratings of Instruction, D. L. Sorenson and T.D. Johnson (eds.), *New Directions for Teaching and Learning*. Jossey-Bass: San Francisco, 2003, 96, 31- 38.

[11]. Rosni Abu Kassim ; Norlida Buniyamin, Evaluating teaching quality using data from student online feedback system, IEEE 7th International Conference on Engineering Education (ICEED), 2015, pp 64 - 68

[12]. Zenon Chaczko et. Al, Assessment of the Quality of Teaching and Learning Based on Data Driven Evaluation Methods, 7th International Conference on Information Technology Based Higher Education and Training, 2006, pp. 21-36.

[13]. Marsh, H.W. & Roche, L. (1993). The use of students' evaluations and an Individually structured Intervention to enhance university teaching effectiveness. *American Educational Research Journal*, 30(1), 217-251.

[14]. Marsh, H.W., Students' Evaluations of University Teaching: Dimensionality, Reliability, Validity, Potential Biases and Usefulness, *The Scholarship of Teaching and Learning In Higher Education: An Evidence-Based Perspective (2007)*, 319–383.

[15]. James H. Stronge, What Makes Good Teachers Good? A Cross-Case Analysis of the Connection Between Teacher Effectiveness and Student Achievement Journal of Teacher Education (2011), 62: 339-355

[16]. Swapna Gottipati, Venky Shankararaman and Sandy Gan, A conceptual framework for analyzing students' feedback, IEEE Frontiers In Education Conference (FIE), 2017, pp. 1-8.

[17]. Tigelaar, DEH, Dolmans, DHJM, Wolfhagen, IHAP, The development and validation of a framework for teaching competencies In higher education (2004), Higher Education, 48 (2) pp. 253-268

[18]. Casey, R, Gentile, P, Bigger, SW, Teaching appraisal In higher education: an Australian perspective (1997), Higher Education 34 (2), pp 459-482.