Education Tech Research Dev (2012) 60:821–837
DOI 10.1007/s11423-012-9240-y

DEVELOPMENT ARTICLE

Developing a comprehensive teaching evaluation system for foundation courses with enhanced validity and reliability

Yueyu Xu

Published online: 22 March 2012

Abstract This study aims at developing a comprehensive teaching evaluation system, a more useful educational technology, for achieving relatively reliable and valid results that can be well acknowledged by instructors, students, and by administrators. Adopting multi-method approaches, the study integrates student evaluation, expert evaluation and regular examination of teaching into teaching rating. Especially, a novel method of "evaluation on the same platform", as an important part of the comprehensive evaluation system, is designed as a fair approach for teaching evaluation. All the parameters in the proposed system are determined by implementing the Analytic Hierarchy Process to get the weight of each factor. Empirical studies are implemented in the experimental university for more than six semesters. The result demonstrates that the evaluation system may be more reliable if performance of the instructors is assessed by more comprehensive raters. The whole evaluation system is implemented conforming to a series of serious procedures to ensure democracy, fairness and effectiveness.

Keywords Teaching evaluation · Evaluation system · Teaching quality · Fair · Reliability

Introduction

Measures of teaching evaluation are popularly implemented in advanced education. After decades of quantitative growth in higher education, consensus is emerging on the need to establish a valid and reliable evaluation system of teaching for improving teaching quality (Wolfer and Johnson 2003; Ma 2005). Especially, the general decline of curriculum quality measured by the indicators of academic performance, teachers' assessment, employers' assessment and social assessment and defects existing in the current evaluation methods accentuate the importance of appropriate teaching evaluation (Lei 2005; Cai 2005).

Y. Xu (✉)
Department of Fundamental Education, Zhejiang Shuren University, #8 Shuren St., Hangzhou, Zhejiang Province, People's Republic of China
e-mail: xyy.zjsru@gmail.com

The rapid development in higher education sector in China

China's higher education sector is in an era of great change. In the past several years, the Chinese government has implemented many reforms to improve higher education access, resulting in higher education for the masses. This has led to a scarcity of educational resources, reducing educational quality and student satisfaction (People's Daily 2009; Ma 2005). This issue has attracted broad public attention, since after all, teaching quality is positively related to the rating of universities and talent cultivation (Darling-Hammond 2000). Consequently, the Chinese government has prioritized educational quality improvement (Zhou 2006; Hu 2007). These demonstrate that higher education in China has entered the stage of integration and optimization.

The significance of teaching evaluation

The quality of higher education may be enhanced by effectively performing functions of evaluation systems (Yang and Nie 2010). First, teaching evaluation provides participating instructors with the opportunity to consider teaching processes, assess teaching effectiveness, and identify areas of growth as a measure of teaching development. Teaching evaluation can help instructors actually recognize that both teachers' subject matter knowledge and pedagogical skills are pivotal determinants for student achievement (Sanders and Rivers 1996), and teachers' professional education coursework is sometimes more influential than teachers' subject matter knowledge (Begle 1979; Ashtom and Crocker 1987; Monk 1994). Second, "higher education administrators and tenure committees depend on evaluation of teaching for making decisions about instructor hiring, promotion, tenure, salary adjustment, and retention" (Wolfer and Johnson 2003). Also, based on the information obtained from the evaluation system, a university may select specific programs of professional development to aid instructors in building new pedagogical theories and practices.

The limitations of the conventional method of teaching evaluation

The traditional method of teaching evaluation is student evaluation. The popularity of student evaluation may be attributed to at least the following three reasons: (1) it is valid based on the fact that students' ratings correlate positively with teaching effectiveness; (2) it is convenient, feasible, and inexpensive to administer; and (3) it gives an impression of objectivity compared with any other alternative alone (Cahn 1986). Despite these beneficial characteristics, its disadvantages are obvious. First of all, factors outside the control of instructors are not considered in this procedure. These factors include class size (Hanna et al. 1983), course content (Cashin 1990), gender of instructor (Anderson and Miller 1997), and so on. Some researchers also present their worries that instructors' rights to free speech would be compromised by the pressure to secure favorable student evaluations (Williams and Ceci 1997; Schueler 1988; Huemer 2005). These factors causing biases should be considered when making an administrative decision (Jirovec et al. 1998; Wolfer and Johnson 2003); otherwise, a reliable result in teaching evaluation cannot be expected. Secondly, students' motivation towards teaching evaluation may negatively influence the impartiality of evaluation. In empirical research, scholars find that some students' motivation may cause unfairness of the evaluation result, such as snobbish psychology, self-centered psychology, and revenge psychology (Chen 2007). Also, students' academic motivation influences ratings and students tend to give higher ratings in appreciation for lenient grading (Goldman 1985; Greenwald and Gillmore 1997; Greenwald 1997).

As discussed above, student perception of teaching may not accurately reflect the performance of instructors. Several researchers analyze the use of supplemental methods to evaluate teaching. For example, some studies find that both instructors and evaluators have similar perceptions of the peer review system (Hansen et al. 2007). The 360 degree feedback system have been developed to use multiple data sources for documenting performance to provide a more realistic picture of teacher performance Dyers (2001, Ortega et al. 2007). Similarly, scholars propose that the multiple sources of evidence include student ratings, peer ratings, self-evaluation, videos, student interviews, alumni ratings, employer ratings, administrator ratings, teaching scholarship, teaching awards, learning outcome measures, teaching portfolios, and so on (Berk 2005; Berk et al. 2004). It is suggested to have greater weight attached to student and peer rating and less weight attached to self-evaluation, administrator ratings, and others (Arreola 2000). Despite these efforts, few alternatives are favored overwhelmingly against others to date. Furthermore, the multiple source feedback system has not been used uniformly in China considering that many raters cannot provide impartial evaluations of higher education due to Guanxi between different parties and instructors' self-assessments may negatively influence the objective evaluation because the rating results are pertinent to promotion, honor, bonus, and other interests.

Purpose of the study

This study aims at developing a comprehensive teaching evaluation system for achieving a reliable and valid result that can be respected by instructors, students, and administrators. In particular, two research questions drive this project. First, how to establish a multidimensional evaluation system integrating the merits of different rating approaches? Second, what measures should be taken to ensure and enhance the reliability, effectiveness, and fairness of such a teaching evaluation system? In order to achieve these goals, the study integrates student evaluation, expert evaluation and regular examination of teaching into teaching rating, and especially, a novel method of "evaluation on the same platform" (EOTSP) is introduced as a fair approach for teaching evaluation to reduce bias generated by variable evaluation entities. Additionally, a series of procedures are designed to ensure the whole evaluation system is implemented in a democratic, fair and effective environment.

In this article, the second section of describes the methods of this study, including instruments, participants, procedures and parameters; the third section discusses the findings and verifies the reliability of the proposed system; and, the last section reaches the conclusions and the future development of this new evaluation technology.

Methods

Responding to an increase of college students in the age of higher education for the masses, universities have to offer more courses to satisfy the needs of students. Across all higher education institutions in China, the teaching team for foundation courses is usually the biggest group in universities because almost every student is required to take foundation courses to develop a base for further acquiring professional knowledge. The foundation courses including "Calculus", "Probability and Statistics", "Linear Algebra", "Physics", "English Language", "Politics Theory" and "Philosophy" are offered for junior students of the university. Generally, a teaching class has 70–120 students. Each instructor shall teach 1–2 foundation courses and give lectures for 2–4 classes. Involving more students and instructors, the evaluation of foundation courses seems quite challenging.

Instruments

In order to reduce bias, this study proposes a new evaluation system to obtain fair teaching ratings. The proposed evaluation model is composed of three parts: student evaluation, expert evaluation, and regular examination of teaching in order to involve multiple stakeholders: students, teachers, and administrators. It integrates the merits of student evaluation, peer review, and moral evaluation, and the strict procedures of implementation ensure the fairness.

Student evaluation

The proposed student evaluation is composed of two parts, EOTSP and semester evaluation. EOTSP is designed to evaluate instructors who teach the same or similar foundation courses, and refers to comparing similar lectures in the same course. Administrators divide the whole course into several smaller sections and choose 1–2 classes for students to evaluate their instructor. Each instructor provides one or two sections in the selected class. The reason that the researchers made students compare instructors' performances is to achieve an impartial result of evaluation. Although this proposed approach puts teachers in a competitive situation, it may achieve a more fair result compared with the traditional method of teaching evaluation because the entity giving ratings is invariable.

In order to create a collegial environment, instructors may collaborate with colleagues to determine the contents of the rated classes. Meanwhile, the administrator encourages instructors to strengthen the sense of cooperation and competition for ensuring that all the course content is well connected and unnecessary redundancy is avoided. For example, instructors may discuss the contents and teaching techniques with others in the regular meeting or in private. The procedures of implementing evaluation system ensure the rights and privacy of teachers and are helpful for creating a harmonious atmosphere, which is discussed in "Procedure of evaluation" section.

Within the 1–2 opportunities to present lectures, the teaching methods taken by the instructors include but are not limited to teacher-centric approaches as well as approaches that encourage active student participation. EOTSP does not prevent instructors from using contemporary teaching techniques, such as case study, concept mapping, and even problem based learning only if they are suitable for the teaching content and can be well organized. As for the student evaluation of teaching techniques, students merely need to evaluate the appropriateness and effectiveness of the concerned techniques. At the end of the semester, the administrator assigns a non-stakeholder staff member to administer the student evaluation questionnaire. Every student in the class is requested to rate all the instructors giving lectures in the course.

Indeed, the procedure of EOTSP can be optimized. In this study, students' rating of EOTSP was arranged at the end of each semester. Although the photos of instructors and the topics of lectures were shown repeatedly for helping students recall the information and performance of each instructor, it would be better to evaluate each lecture soon after it was delivered so that students have the same memory of all the instructors.

The student evaluation questionnaire is developed based on the popular standard of student evaluation such as effectiveness of instructors, preparation and organization, knowledge of subject matter and ability to stimulate interest in the course, clarity and understandability, and so on (York University 2002; Frick et al. 2009, 2010). It also incorporates some comments from students, instructors, and experts on teaching and policy research. In order to ensure the validity of the instrument, the questionnaire is piloted with 20 junior students for testing clarity in the designing process. It consists of 20 items and the score of each item ranges from 1 to 5.

Despite advantages of EOTSP, it should be noted that because each instructor has only 1–2 opportunities to present lectures by such an approach, the result of EOTSP cannot be regarded as the whole picture. Therefore, the second method of student evaluation, namely "semester evaluation", is undertaken to remedy the defect mentioned above. Because each instructor shall teach 1–2 foundation courses throughout the whole semester, instructors have opportunities to establish a close relationship with students through teaching, discussion, and personal guidance over a period of about 16 weeks. This allows students' perceptions of the instructors to be more comprehensive than in the first method. At the end of semester, the students in classes that instructors teach on their own are required to answer the student evaluation questionnaire used the same as in EOTSP.

Expert evaluation

Expert evaluation is another element of this evaluation system. Experienced teachers with a deep insight into teaching are expected to provide summative and formative evaluations from a peer perspective. In this study, the peer observation team consists of teaching experts from other universities. The observation team learns little about the rated instructors to avoid involving irrelevant factors due to interpersonal relations. All the peer experts are selected according to The Rules on Experts Evaluation of Teaching regulated by the experimental university. They are professors or associate professors whose academic titles are granted by their universities or the provincial educational authority and are recognized across China. These experts have established credibility among their peers because of their outstanding achievements. In order to achieve fairness, experts attend the lectures randomly, and they do not need to inform the instructors in advance. Generally, the instructor's character can be observed in this situation. After auditing the course, experts provide ratings and formative comments referring to the evaluation criteria described in "The Record of Expert Evaluation" (Table 1). This process is important to prevent observers from varying in their definition of teaching quality. The evaluation criteria in Table 1 are derived from the long-term experience of the experimental university and from general standards in peer evaluation such as quality of learning environment, level of student engagement, clarity of presentation, instructor's ability to convey course content, and other factors (York University 2002). As determining the student questionnaire, the content of Table 1 was discussed among instructors and experts, and a test of draft questionnaire was conducted before determining the final version. These efforts aim to ensure the validity of the instrument. In this study, experts consider the elements in Table 1: achievement of teaching objectives, content and techniques of teaching, organization, presentation, and final results of teaching. There are 20 items in Table 1, each of which is scored on a scale ranging from 2 to 5. The final expert evaluation score is calculated by adding the score of each item. Additionally, experts give general comments and advice for reflection and professional development.

Table 1 The record of expert evaluation

The purpose of this record is to evaluate teaching and instructors. There are 20 items in this record. Please consider the questions carefully and choose the suitable rating. Furthermore, please provide overall comments on instructors.

SD- Strongly disagree (score =2)

D – Disagree (score =3)

A – Agree (score =4)

SA – Strongly agree (score =5)

Name of instructor

Subject

Class

Evaluation items

Objective of teaching

1. The objective of teaching is well achieved.

2. Students' self-directive capabilities are developed in teaching.

Content of teaching

3. Follow the syllabus and text book.

4. Contents are organized systematically and logically.

5. Contents are closely related to cutting edge of technologies or social problems.

6. Academic orientation is adopted.

7. Adequate information is delivered in a lecture.

8. The level of difficulty is appropriate.

Techniques of teaching

9. Exploration and innovation of teaching techniques are reflected.

10. The instructor improves students' engagement in learning.

11. Teaching techniques are applied appropriately.

In order to ensure the inter-rater reliability of the peer experts, the experts were given clear and concise instructions on the rating criteria. Before rating, the peer raters of the same instructor were required to experimentally observe the same five video episodes of teaching to test the degree of deviation, and then, the experts discussed the rating standards. After training and experimental experience, the inter-rater reliability was calculated

Table 1 continued

12. The instructor tries to improve students' capabilities of cooperation and self-direction.

Instructor's presentation

13. The information is delivered clearly.

14. Key points and difficult points are interpreted correctly and explicitly.

15. Experimental demonstration is skillful.

16. The instructor has an ability of constructing and organizing teaching.

Results of teaching

17. An enthusiasm is created for topics.

18. Majority of students actively engage in leaning and discussion.

19. Majority of students understand the knowledge taught by instructors.

20. Students' capabilities of thinking, innovation and cooperation are improved.

Total score

Comments of expert

Signature:
Date:

on observations of five video episodes. Each expert was required to fill out the Record of Expert Evaluation. If both the observers agreed on an item, the inter-rater score is 1. If they did not agree, the score 0.5 should be given. When all 20 items were scored, the scores should be totaled. The final inter-rater score that is the average score of all the items turns out to be 0.88.

Regular examination of teaching

The ethical dimensions of instructors' professional practices, one determinant of teaching quality, are discernable in daily work. Evaluation of instructors' professional ethics is of substantial significance not only because professional ethics may potentially influence all participants in classrooms by means of informal and spontaneous interactions (Freire

1998), but it has an inevitable impact on the overall moral climate in universities (Campbell 2003). More attention to professional ethics in Chinese education is necessary because on some occasions, some instructors seriously violate ethical principles regarding teaching and education. Examples are irresponsible teaching, apathetic response to students' questions, and so on. It is worth noting that moral nature should be defined narrowly to focus on duty, diligence and moral responsibility to others (Campbell 2003). Based on these principles, an evaluation of instructors' ethical attitudes to teaching is used to examine their professional performance. This evaluation is composed of five items. The first three items emphasize instructors' duty and diligence, while the last two focus on their moral responsibility to students.

- Does the instructor follow the syllabus?
- Does the instructor have an integrated and systematic schedule for teaching? And, does the progress of teaching exactly adhere to the schedule?
- Does the instructor fully prepare for presentation or demonstration of experiments?
- Does the instructor read students' assignments and give comments on time?
- Does the instructor provide necessary supervision and answer students' questions?

Because these items can be easily documented, instructors are required to keep records and ensure their authenticity. For example, the university requires the instructor to make an integrated and systematic schedule for the whole semester and let students know about it at the beginning of the semester. Moreover, instructors should take the necessary materials related to teaching into the classroom such as teaching documents, lists of students, textbooks, etc.; and should record the teaching progress after each lecture. In addition, instructors should keep grade books and written reports about homework completion. As for the former two items, the teaching committee attends the lectures randomly and exams whether the instructor follows the syllabus and the schedule, but does not evaluate teaching contents or techniques. Because the syllabus merely outlines the essential content of the subject, it does not hinder instructors to adapt to student needs. As for the latter three items, the teaching committee provides assessments for instructors mainly based on their records. The teaching committee also investigates a small amount of students to assess each instructor's response to students' requirements. Each evaluation question is scored on a scale from 0 to 20 and summed for a total score ranging from 0 to 100.

The framework of the proposed evaluation system

The framework of this evaluation system for foundation courses is composed of three parts organized as follows:

$$I = S \times \rho_1 + E \times \rho_2 + R \times \rho_3 (\rho_1 + \rho_2 + \rho_3 = 1; \quad \text{and} \quad \rho_1, \rho_2, \rho_3 > 0). \qquad (1)$$

I = the integrated score, S = the score of student evaluation, E = the average score of expert evaluation, R = the score of regular examination of teaching

The score of student evaluation is calculated by the following formula:

$$S = P \times K_1 + M \times K_2 \quad (K_1 + K_2 = 1; \, K_1, K_2 > 0) \qquad (2)$$

S = the score of student evaluation, P = the average score of EOTSP, M = the average score of "semester evaluation", K_1 and K_2 are the weighting factors of the student evaluation system and are calculated by using the Analytic Hierarchy Process (AHP), which is discussed in "Parameters" section.

In this framework, students' final grades were not included in order to avoid exam oriented classes. In china, exam oriented education in a long term has brought about many problems, such as preventing free speech, hindering academic freedom, and so on. For example, under the background of exam oriented education, instructors would like to focus on the content that will be tested in the examination and give the so-called correct answer for students rather than provide potential solutions to the problems or offer more academic information irrelevant to the examination. Moreover, if students' final grades are incorporated into the evaluation system, there is one possibility that instructors would not give objective scores of students' general performance related to academic study that is one part of student's final grade, because they probably tend to raise their evaluating scores by giving higher ratings to students' performance.

ρ_1, ρ_2 and ρ_3 are the weighting of corresponding parts, and are set separately. Considering that students are the most important participants in teaching activities, ρ_1 is suggested to be larger than the other two parts. Of course, the proportion can be adjusted according to the specific situation.

Substitute Eq. 2 into Eq. 1,

$$\begin{aligned} I &= (P \times K_1 + M \times K_2) \times \rho_1 + E \times \rho_2 + R \times \rho_3 \\ &= P \times K_1 \times \rho_1 + M \times K_2 \times \rho_1 + E \times \rho_2 + R \times \rho_3 \end{aligned} \tag{3}$$

Thus, the integrated score is determined by the result of EOTSP, semester evaluation, expert evaluation and regular examination of teaching, and the weight of each part is respectively $K_1 \times \rho_1$, $K_2 \times \rho_1$, ρ_2, and ρ_3, which is discussed in "Parameters" section.

Participants

This comprehensive system of teaching evaluation was carried out in a famous university of Zhejiang Province of China in the six semesters from 2006 to 2008. The university implemented this method to examine teaching quality in three selected foundation courses: Calculus, Physics, and Politics Theory. These courses are offered to all junior level undergraduate students, and each course had a teaching staff of more than 10 instructors who are professors, lecturers or tutors so the EOTSP was suitable for these courses.

Procedure of evaluation

It is important to establish and follow fair and reasonable procedures to achieve a successful evaluation. An underlying principle in many fields suggests that it is more likely to obtain good outcomes if good practice is employed (Lubawy 2003). In order to maximize benefits of evaluation activity for all participants, some principles of evaluation ethics should be highlighted including due process, privacy, equality, openness of procedures to public, and humaneness (Strike 1990). In recent years, effective procedures have been established and are working properly.

- Establish a special working team for teaching evaluation. The working team takes charge of organizing the student questionnaire survey and expert evaluation, confirming the result of evaluation, and hearing the appeal of instructors.
- Implement withdrawal of interested personnel in order to avoid conflicts of interest.
- Select 1–2 classes that are suitable for implementing the method of EOTSP.
- Interpret the student evaluation in detail for students before they answer the questionnaire.

- Provide feedback to instructors. In order to protect instructors' self-respect and privacy, the related personal information is confidential to other people. The administrator also arranges a meeting at the end of each semester where instructors exchange ideas, discuss course content and teaching techniques.

Parameters

There are five parameters (K_1, K_2, ρ_1, ρ_2, ρ_3) involved in the system of evaluation. According to Eq. 3, four values should be determined first: $K_1 \times \rho_1$, $K_2 \times \rho_1$, ρ_2, and ρ_3. These values are the weight of EOTSP, semester evaluation, expert evaluation and regular examination of teaching respectively.

In order to determine these parameters, a test bench is constructed by implementing AHP to get the weight of each factor. AHP enables a person to make comparisons of importance between decision elements. An n-order reciprocal matrix A is created in which "n" equals the number of factors ($C_1 \ldots C_n$), and the element a_{ij} is the ratio of the importance value of C_i to the importance value of C_j. Ratios concerning comparisons of importance between decision elements are decided by subjective perception based on substantial experience.

$$A = \begin{vmatrix} C_1/C_1 & C_1/C_2 & \cdots & C_1/C_n \\ C_2/C_1 & C_2/C_2 & \cdots & C_2/C_n \\ \vdots & \vdots & & \vdots \\ C_n/C_1 & C_n/C_2 & \cdots & C_n/C_n \end{vmatrix} = \begin{vmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{vmatrix} \tag{4}$$

The eigenvector corresponding to the maximum characteristic root ($\lambda_{\max}$) of the matrix is used as the weighting factor. To verify the rationality of results, a method of examining Ordinal Consistency proposed by Saaty is used (Saaty 1980). Under Saaty's rule, the process of verifying is to calculate the consistency index, $\mathrm{CI} = (\lambda_{\max} - n)/(n - 1)$, and then compare it with the corresponding n-order random consistency index (RI) to obtain the consistency ratio CR. Here RI means random consistency index obtained from a large number of simulation runs and varies depending upon the order of matrix. Table 2 shows the value of RI for matrix of order 1–10 obtained by approximating random indices using a sample size of 500 (Saaty 1980).

If $\mathrm{CR} = \mathrm{CI}/\mathrm{RI} < 0.1$, it can be concluded that the consistence of matrix A is acceptable, meaning that the eigenvectors of such a matrix can be adopted as the weighting factors.

Therefore, referring to AHP, a fourth-order reciprocal matrix A is constructed since there are four parts of the proposed system of teaching evaluation. In this matrix, a_{ij} is set according to the mutual relationship of the four factors (P, M, E, R). In the experimental university, a_{ij} is the ratio of the importance value of C_i ($i = 1, 2, 3, 4$) to the importance value of C_j ($j = 1, 2, 3, 4$). The ratios subjectively perceived by the university are determined based on massive statistical data of evaluation over more than 3 years. Many factors are considered in the statistic work: size of classes, major of students, grade of

Table 2 Average random index (RI) based on matrix order

Order of matrix (n)	1	2	3	4	5	6	7	8	9	10
RI	0	0	0.52	0.89	1.11	1.25	1.35	1.40	1.45	1.49

Source: Saaty (1980)

students, and gender of instructors. Using Linear Regression Analysis, the teaching committee finally determines a_{ij} as follows:

$$A = \left| \begin{array}{ccccc} & P(C_1) & M(C_2) & E(C_3) & R(C_4) \\ P(C_1) & 1 & 5 & 3 & 6 \\ M(C_2) & 1/5 & 1 & 1/2 & 1.2 \\ E(C_3) & 1/3 & 2 & 1 & 2 \\ R(C_4) & 1/6 & 1/1.2 & 1/2 & 1 \end{array} \right| \qquad (5)$$

After calculating, the results are: CI = 0.0014;

In a fourth-order matrix, RI = 0.89 (it is given by the AHP theory) CR = CI/RI = 0.0014/0.89 = 0.0015 < 0.1.

From this respect, the consistence of matrix A is acceptable, and its eigenvectors can be adopted as the weighting factors: 0.59, 0.11, 0.20, 0.10.

Therefore, $K_1 \times \rho_1 = 0.59$; $K_2 \times \rho_1 = 0.11$; $\rho_2 = 0.2$; $\rho_3 = 0.10$.

From Eq. 2, it can be calculated that $\rho_1 = 0.7$. Hence, K_1 and K_2 can be concluded as well. After rounding up within the tolerance of error band, these two parameters are:

$$K_1 = 0.85; \quad \text{and} \quad K_2 = 0.15.$$

Substitute these parameters into Eq. 3,

$$\begin{aligned} I &= (P \times K_1 + M \times K_2) \times \rho_1 + E \times \rho_2 + R \times \rho_3 \\ &= (P \times 0.85 + M \times 0.15) \times 0.7 + E \times 0.2 + R \times 0.1 \end{aligned}$$

In conclusion, EOTSP is given 60% of the final weight, the student semester evaluation is given 10% of the final weight, the expert score receives 20% of the final weight, and the regular examination of teaching is given 10% of the final weight.

Results

Comparison of evaluation made by students and experts

In order to examine the reliability of this system, instructor ratings given by students and experts are compared firstly. The comprehensive system of teaching evaluation is carried out in six successive semesters, and it is found that the rating provided by experts largely parallels the one by students; particularly, they have similar opinions on instructor characteristics and course organization. By contrast, there is an apparent divergence between their perceptions of student's ability development, academic oriented teaching, and teaching technique of instructors. It is should be noted that this comparison, with linguistic interpretation based on the statistical analysis, is suggestive rather than absolute.

For instance, in the course Politics Theory, the number of instructors participating in evaluation in the recent two semesters is respectively 12 and 13. The results of teaching evaluation during these two times are shown in Fig. 1. The horizontal axis is the identification number of each instructor, and the vertical axis is the score given by students, experts and the teaching committee. The figure shows that the curves of the regular teaching examination exhibit a very small variation over the whole range which implies minor differences between instructors. From the figure, it can also be learned that the tendency of the two curves describing student and expert evaluation is roughly equal. For scores ranging from 70 to 90, the scores given by students and experts match well.

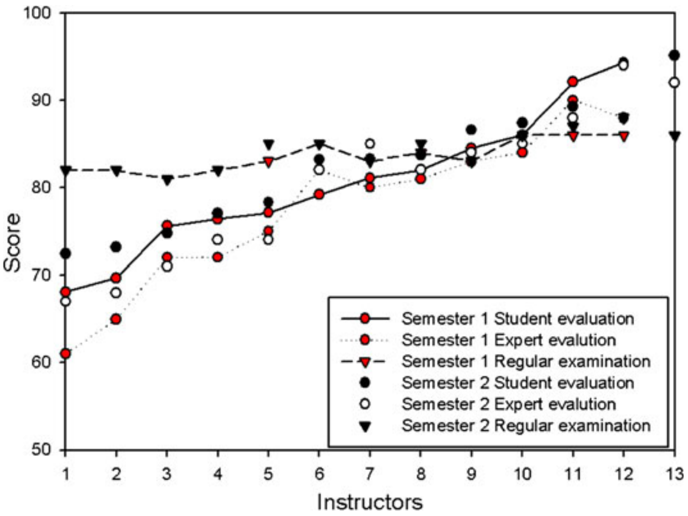

Fig. 1 Evaluation in Semester I and II of the academic year 2007–2008

However, they hold strong divergent views on individual instructor's performance. Especially, the instructor ranked first in the student evaluation does not obtain the highest score from experts. Experts especially appreciate the instructor who has a solid background of academic knowledge, implements academic oriented teaching, and tries to develop students' capability of critical thinking and self-direction to achieve the ultimate objectives of higher education. Experts also argue that educational techniques shall be used for improving the effectiveness of lectures instead of concentrating on adding attraction. For these three items examining the situation of developing students' ability in teaching, the student rating shows little difference between instructors because the difference between highest and lowest score is less than 0.4. However, in the expert evaluation, the difference is more than 2.1, and the rating results show high inter-rater reliability.

Comparison between the evaluation results and student learning outcomes

Although the student grades are not included in the proposed system to avoid the exam oriented teaching, it does not prevent them from being used as a stick yard for measuring the reliability of the evaluation system. A comparison between the evaluation results and student learning outcomes is carried out to verify the effectiveness of the proposed system here. For each instructor, three values are compared: the student evaluation score, the integrated evaluation score, and the average score of students' final grade. For the purpose of this study, student learning is defined as a function of student achievement including his/her understanding of the concepts being taught, attitudes towards science, problem solving skills and abilities to employ critical thinking in science. Therefore, the student's final grade indicating student learning is composed of two parts, the regular score and the score of final examination. The regular score accounts for 30% of the final grade and is determined based on their homework completion and other performance related to academic study such as engagement in class, proposals of new concepts and participation in academic research. The final examination accounts for 70% of the final grade. Examination

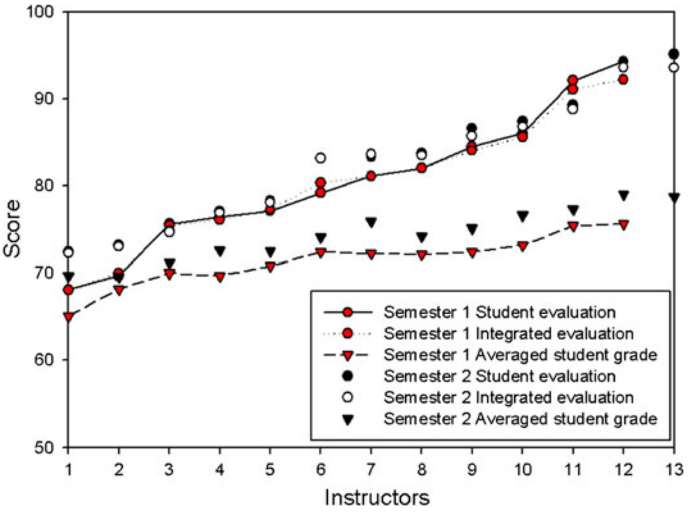

Fig. 2 Comparison of three sources in Semester I and II of the academic year 2007–2008

questions are drafted by an instructor according to curriculum directorate specifications and then reviewed by subject specialists. The final examination is secret to all students and staff, other than the instructors involved in setting the exam questions, to avoid exam orientation in class. Hereby, the average score of the students' final grade is used as an important indicator of students learning. Figure 2 shows the relationship among student evaluation, integrated evaluation, and average students' learning in the two semesters mentioned in Fig. 1.

Logically, the instructor who obtains a higher score in teaching ratings should correspondingly produce superior learning, provided that the approach of teaching evaluation is valid and reliable (Begle 1979; Ashtom and Crocker 1987; Monk 1994). For comparing different methods of teaching evaluation and identifying the better one, one feasible way is to compare the score an instructor receives through two methods to the average score of his/her students' final grade to find which approach creates less deviation from student achievement. Here, it cannot be ignored that the precondition of such a comparison is that the student achievement is measured reasonably. This comparison is feasible because student achievement is well defined and the exam system is successful in avoiding exam oriented classes. In Fig. 2, the horizontal axis is the identification number of each instructor, and the vertical axis shows the score of student evaluation, score of integrated evaluation, and average score of their students' final grade. First, for each instructor, calculate the difference in value between the score of student evaluation and the average score of their students' final grade, and then take the absolute value. Sum all the absolute values corresponding to each instructor together, and calculate the average value which is the average deviation between the score of student evaluation and student achievement. Similarly, calculate the average deviation between the score of integrated evaluation and student achievement. In this study, the former average deviation is larger than the latter in both semesters. That demonstrates the integrated evaluation aligns with student achievement better than the student evaluation so the integrated evaluation system is more valid and reliable than the simple student evaluation.

Effects of implementing the proposed system

Furthermore, this method of teaching evaluation is widely favored and highly appreciated by students and instructors. Students believe that the teaching rating itself increases interaction between them and instructors. In the first semester of the academic year 2007–2008, the experimental university sent out 640 questionnaire forms concerning the student evaluation and 601 were returned. Of the 640 students surveyed, 86% highly agree with such a system of teaching evaluation. Plenty of students take the view that a good approach to teaching rating will benefit teaching and learning in the future. This comprehensive system of teaching evaluation is greatly appreciated by instructors as well. They appraise that such a method is more reasonable than conventional methods since more fair elements and procedures are incorporated, such as the same platform for ratings, expert evaluation, and so on.

The proposed evaluation system contributes to instructor professional development. In this research, after rating, the formal result debriefing session with instructors was carried out to enhance the professional experience and develop confidence in their teaching competence. Through debriefing, sharing of feedback, discussing, encouraging and inquiring, the rating results provided instructors with a supportive environment, access to peer support and a sense of belonging. It is found that the instructors participating in this evaluation perform better and are more active in pursuing professional development than before. They actively seek opportunities of training on content and teaching methods. Moreover, the instructors begin to recognize that professional development is a long-term process and its most effective form is based on daily activities. As a result, they often carry out discussion and research concerning teaching and learning. Furthermore, the evaluation system provides the experimental university with valuable information to promote instructor professional development. Since implementing this research, professional development has been built into the working plans of the experimental university. All rating results are recorded in the instructor's personal file which is not only the reflection of professional experience, but also an impetus to future development. For instructors receiving high ratings, the university grants a series of honors and bonuses to encourage them to make continual improvement. On the other hand, the university adopts assistance policies for instructors receiving low ratings. They are provided with opportunities to observe colleagues to see best practices on a regular basis, receive support from colleagues on group goals, and spend time and funds on professional training. According to the rating results, the university selects specific programs of professional development to aid instructors in building new pedagogical theories and practices. It also endeavors to increase meaningful interactions between instructors, administrators, students and other relevant entities as an additional method of professional development.

To some extent, this approach is time and resource intensive because it involves substantial work and special staff assignments. Especially, in the initial stage when the participants including instructors, students, administrators, and experts, are not familiar with this system, implementing this system seems like a laborious task. Nevertheless, the cost decreases significantly after the system is established since some efforts are not needed once the participants are used to the system and accept it. Several tasks in the system have been embedded into the regular administration of the university. For example, the survey of students' feedback and the regular examination of teaching have been included as regular duties of administration and do not add cost. The main cost of implementing this system is for experts and data analysis, and that it is quite acceptable.

Conclusion and further development

This study is aimed at establishing a useful method for teaching evaluation. It proposes a comprehensive system of teaching evaluation for foundation courses with a systematic application of knowledge to resolve an important and recurring problem—validity and reliability of teaching evaluation. This comprehensive system is composed of three parts: student evaluation, expert evaluation and regular examination of teaching. Also, it introduces a novel method of student evaluation by which instructors teaching the same or similar courses are evaluated by the same group of students.

The proposed system has been proved to be more reliable than the single student evaluation. Overall, this system is simpler to implement and not too time-consuming for staff. Instructors have acquired valuable information in the context of teaching, and the university has earned a reputation for a reasonable system of teaching evaluation.

With further implementation of this teaching evaluation system, it can be gradually popularized. It is generally feasible for most universities since the resources needed in implementing this system, such as experts, teaching committee, and special staff, are available and not very cost-consuming. The formula to calculate the instructor rating score can be adjusted according to the characteristics of subjects, level of students, classes, and so on. Universities may implement this system to achieve a fair and reasonable result of teaching evaluation, which is significant for teaching improvement, professional development, and administrative decision. Moreover, although this method is currently used only for teaching evaluation of foundation courses since a large number of instructors are involved and the same rating platform can be established, it is adaptive and flexible. It is commended that universities can use this method for teaching evaluation of specialized courses in similar subjects to compare performance of instructors who teach different courses belonging to the same discipline. For instance, instructors teaching contract law, tort law, family law and property law perhaps can be classified as a civil law group; consequently, it is reasonable to compare their performance in the same specialized foundation course of civil law. The foundation courses in engineering can be taken as another example. Circuit theory is a foundation course in electrical and electronics engineering and is applicable for the proposed evaluation system.

The proposed system cannot be unquestioningly applied world-wide without seriously understanding the culture, tradition, and educational level of different countries. However, it is still significant to study whether this system is feasible in other countries by appropriately adjusting the weight of each part of the system.

Acknowledgment The author is thankful to Ms. F. Wang for her work on mathematical model.

References

Anderson, K., & Miller, E. D. (1997). Gender and student evaluations of teaching. *Political Science & Politics, 30*, 216–219.

Arreola, R. A. (2000). *Developing a comprehensive faculty evaluation system: A handbook for college faculty and administrators on designing and operating a comprehensive faculty evaluation system* (2nd ed.). Bolton, MA: Anker.

Ashton, P., & Crocker, L. (1987). Systematic study of planned variations: The essential focus of teacher education reform. *Journal of Teacher Education, 38*, 2–8.

Begle, E. G. (1979). *Critical variables in mathematics education.* Washington, D.C.: Mathematical Association of American and National Council of Teachers of Mathematics.

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*(1), 48–62.

Berk, R. A., Naumann, P. L., & Appling, S. E. (2004). Beyond student ratings: Peer observation of classroom and clinical teaching. *International Journal of Nursing Education Scholarship, 1*(1), 1–26.

Cahn, S. M. (1986). *Saints and scamps: Ethics in Academia*. Totowa, NJ: Rowman & Littlefield.

Cai, Y. H. (2005). Experience and exploration of constructing the system of teaching evaluation. *China University Teaching, 3*, 48–49.

Campbell, E. (2003). Moral lessons: The ethics role of teachers. *Educational Research and Evaluation, 9*(1), 25–50.

Cashin, W. E. (1990). Students do rate different academic fields differently. *New Directions for Teaching and Learning, 43*, 113–132.

Chen, B. Y. (2007). Review on teaching evaluation. *Journal of Educator, 16*, 25–27.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8*(1), 1–50.

Dyers, K. M. (2001). The power of 360 degree feedback. *Educational Leadership, 58*(2), 35–39.

Freire, P. (1998). *Pedagogy of freedom: Ethics, democracy and civic courage*. Lanham: Rowman and Littlefield Publishers, INC.

Frick, T. W., Chadha, R., Watson, C., Wang, Y., & Green, P. (2009). College student perceptions of teaching and learning quality. *Educational Technology Research and Development, 57*(5), 705–720.

Frick, T. W., Chadha, R., Watson, C., & Zlatkovska, E. (2010). Improving course evaluations to improve instruction and complex learning in higher education. *Educational Technology Research and Development, 58*(2), 115–136.

Goldman, L. (1985). The betrayal of the gatekeepers: Grade inflation. *Journal of General Education, 37*, 97–121.

Greenwald, A. G. (1997). Validity concerns and usefulness of student ratings of instruction. *American Psychologist, 52*(11), 1182–1186.

Greenwald, A. G., & Gillmore, G. M. (1997). Grading Leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209–1217.

Hanna, G. S., Hoyt, D. P., & Aubrecht, J. D. (1983). Identify and adjusting for biases in student evaluations of instruction: Implication for validity. *Educational and Psychological Measurement, 43*, 1175–1185.

Hansen, L. B., McCollum, M., & Paulsen, S. M. (2007). Evaluation of an evidence-based peer teaching assessment program. *American Journal of Pharmaceutical Education, 71*(3), 45–68.

Hu, J. T. (2007). *Report to the seventeenth national congress of the communist party of China*. Retrieved May 23, 2008 from http://news.xinhuanet.com/newscenter/2007-10/24/content_6938568.htm.

Huemer, M. (2005). *Student evaluations: A critical review*. Retrieved May 28, 2008 from http://home.sprynet.com/~owl1/sef.htm#N_15_.

Jirovec, R. L., Ramanathan, C. S., & Alvarez, A. R. (1998). Course evaluations: What are social work students telling us about teaching effectiveness? *Journal of Social Work Education, 34*, 229–236.

Lei, M. (2005). Method and strategies of improving quality of teaching evaluation by students. *Higher Education Exploration, 1*, 54–57.

Lubawy, W. C. (2003). Evaluating teaching using the best practices model. *American Journal of Pharmaceutical Education, 67*(3), 1–3.

Ma, X. Y. (2005). Establish Internet student-assessing of teaching quality system to make the assessment perfect. *Heilongjiang Researches on Higher Education, 6*, 94–96.

Ma, J. (2010). A mass of students dilute the source of faculty, and the quality of Chinese higher education is generally criticized. *China News Weekly, 474*, 1.

Monk, D. H. (1994). Subject matter preparation of secondary mathematics and science teachers and student achievement. *Economics of Education Review, 13*(2), 125–145.

Ortega, S., Baptiste, L., & Beauchemin, A. (2007). A model for 360° teacher Evaluation in the context of the CSME. In *Reconceptualising the agenda for education in the Caribbean, Proceedings of the 2007 biennial cross-campus conference in education* (pp. 581–586), 23–26 April 2007.

People's Daily (2009). *In the process of massification, who should be responsible for the quality of college students?* Retrieved September 12, 2010 from http://edu.people.com.cn/GB/8216/174479/index.html.

Saaty, T. L. (1980). *The analytic hierarchy process: Planning, priority setting, resource allocation*. New York: McGraw-Hill.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement*. Knoxville: University of Tennessee Value-added Research and Assessment Center.

Schueler, G. F. (1988). The evaluation of teaching in philosophy. *Teaching Philosophy, 11*, 345–348.

Strike, K. A. (1990). The ethics of educational evaluation. In J. Millman & L. Darling-Hammond (Eds.), *The new handbook of teacher evaluation: Assessing elementary and secondary school teachers*. Newbury Park, CA: Sage.

Williams, W. M., & Ceci, S. J. (1997). "How am I doing?" Problems with student ratings of instructors and courses. *Change: The Magazine of Higher Learning, 29*, 12–23.

Wolfer, T. A., & Johnson, M. M. (2003). Re-evaluating student evaluation of teaching: The teaching evaluation form. *Journal of Social Work Education, 39*(1), 111–121.

Yang, J., & Nie, J. (2010). Evaluation of teaching quality in class—a misconception in actual college work. *Development and Assessment of Higher Education, 26*(1), 15–20.

York University (2002). *The teaching assessment and evaluation guide*. Retrieved September 12, 2007 from www.yorku.ca/univsec/senate/committees/scotl/tevguide.pdf.

Zhou, J. (2006). *Evaluation of teaching is the key for improving quality of education*. Retrieved September 14, 2007 from http://www.pgzx.edu.cn/main/webShowDoc?channel=syxw_syxwnr&docID=2006/04/19/1145414467745.xml.

Yueyu Xu is a professor and the vice dean in the Department of Fundamental Education, Zhejiang Shuren University of P.R. China. His interests focus on curriculum theory and educational evaluation.