Pakistan Journal of Psychological Research, 2013, Vol. 28, No. 1, 155-178

Development and Validation of University Teacher's Evaluation Scale

Iram Altaf, Anila Kamal, and Bushra Hassan

Quaid-i-Azam University

The present research aimed at the development and validation of an indigenous scale for evaluation of effectiveness of University teaching. The research has been carried out in two studies. First study dealt with the development of University Teacher's Evaluation Scale (UTES). The items of the scale were empirically determined for content validation, and factor analysis on university students (N = 300) including male (n = 150) and female students (n = 150). The results indicated that UTES as an internally consistent single factor scale. Study II of the present research was conducted on independent sample of university students (N = 30) to establish the psychometric properties of UTES. The convergent validity was established with the help of Peshawar University Teacher's Rating Scale (PUTRS; Riaz, 2000) and both scales showed UTES as valid and reliable instrument for measuring teaching effectiveness. There exist nonsignificant difference between gender of students in evaluation of male teacher and female teachers respectively.

Keywords: development, validation, teaching effectiveness, factor analysis, convergent validity

Higher education plays an important role in the development of a country. Evidence shows that in both developed and developing countries higher education has contributed substantially to their socio-economic, political, and cultural development (Narula, 2000; Regel, 1992). Evaluation of instruction at the university level has become a common phenomenon; whereas evaluation of teaching involves collecting evidence from various stakeholders for the purpose of improving the effectiveness of the teaching-learning process. A

Irum Altaf, Anila Kamal and Bushra Hassan, National Institute of Psychology, Quaid-i-Azam University, Islamabad, Pakistan.

Bushra Hassan is now at School of Psychology, University of Sussex Brighton, England.

Correspondence concerning this article should be addressed to Bushra Hassan, Department of Psychology, International Islamic University, Islamabad, Pakistan. E-mail: bushimalik@gmail.com

successful evaluation generates outcomes that are valid, reliable and indicate directions and action for improvement (Chen & Hoshower, 2003). Assessing teacher effectiveness is a complex issue and has social and historical dimensions. Effective teaching is a multidimensional construct (e.g., a teacher may be organized but lack enthusiasm (Marsh & Roche, 1993). What constitutes effective teaching in the context of higher education has proven rather elusive to describe. Teaching is a multidimensional and complex activity while, traditionally assessment of teacher effectiveness was never considered as an important concept (Khandelwal, 2009; White, 2011). Defining and measuring teaching effectiveness plays an important role in many of the decisions made in higher education (Chen & Hoshower, 2003). During the 1970s, however many universities began requiring student's evaluations, standardizing evaluation instruments, and scoring the evaluation results for performance appraisal purposes (Buskist, 2001; Centra, 1993).

Validity of research productivity as a measure of teaching effectiveness remains unclear. While some (Gavlick, 2006; Hong, Xuezhu, & Zhao, 2007; Stack, 2003) have found research productivity and teaching effectiveness to be positively correlated, while others (e.g., Feldman, 1993) have found measures of research to share little or no variance with measures of teaching. This leaves us with student evaluations, a rather complicated measure of teaching effectiveness. Proponents of student evaluations (Cashin, 1988, 1992; Cohen, 1981; d'Apollonia & Abrami, 1997; Dunkin & Barnes, 1986; Greenwald & Gilmore 1997) have argued that student ratings are generally both reliable and valid. Student's feedback and importance of assessment of teaching is now a vital component in the formal faculty performance appraisal systems of most universities (Clift, 1981; Lersch, & Greek, 2001; Trout, 2000).

A significant body of research related to this issue has been accumulated over the last many years. These studies focus on students' evaluation as accounted for by the interaction between students' and teachers' attitudes toward higher education (Hofman & Kremer, 1980), the nature of feedback to teachers, effects of feedback upon instruction, the efficiency of students' evaluation in improving instruction. Students' evaluations are used for two main purposes (Stack, 2003); summative (those used to evaluate teachers for rank, salary, and tenure purposes) and formative (those that diagnose in ways that allow teachers to improve their teaching (Khandelwal, 2009). The present research uses students' evaluations for formative purposes. Inarguably, students represent the most important stakeholders in any given classroom, and their satisfaction is not a

trivial matter. The present study is an additional attempt to deal with this issue in the indigenous context of Pakistan.

Keeping in view the importance of teacher and teaching evaluation in the development of nation, present study aimed at the development of scale for the evaluation of their performance. The scale intends to identify components of effective teaching that would help teachers to play a more meaningful role in the sacred profession of teaching. The assessment of teachers has been done by gathering students' opinion, studying the educational product, analyzing teaching practices, and by taking opinions of colleagues as proposed by previous researches (Abrami & d' Apollonia, 1997; Bhantanger & Jain, 1994; Murray, Rushton, & Paunonen, 1990; Riaz, 2000). The student views of teaching effectiveness are important for evaluating teacher's effectiveness (Braskamp, & Ory, 1994; Wheeler, Haertel, & Scriven, 1992) With the surge in public demand for accountability in higher education and the great concern for quality of university teaching, the practice of collecting student ratings of teaching has been widely adopted by universities all over the world as part of the quality assurance system (Kwan, 1999).

The key individual in the learning process is a teacher. Since his/her qualities and characteristics can highly affect the student's learning, there is an urgent need for the assessment of the qualities of teaching and teachers. In some universities of Pakistan, like many other under developed countries, the American model of academic credit system was adopted. This system gives considerable independence to teachers in determining what and how they can teach; and this is cost-effective system. But unfortunately the component of evaluation is missing from this system in most of these universities (Riaz, 2000). However, after nearly seven decades of research on the use of student evaluations of teaching effectiveness, it can safely be stated that the majority of researchers believe that student ratings are a valid, reliable, and worthwhile means of evaluating teaching (Koon & Murray, 1995; Marsh, 1990; Marsh & Dunkin, 1992, 1997; McKeachie, 1997; Murray et al., 1990; Seldin, 1999). Teaching evaluation seems to be an important topic. However, there are two important issues related to teaching effectiveness. The first issue is related to the accuracy of evaluations. Therefore, the major concern is related to the reliability and validity of the measuring instrument itself.

Results of several studies (i.e., Braskamp & Ory, 1994; Feldman, 1993; Marsh & Dunkin, 1992; Marsh, 2007; Murray, 1983; Perry, 1997, Seldin, 1999; Sproule, 2002; Wachtel, 1998) provide a general consensus about some apparent dimensions of teaching effectiveness.

In Pakistan research on teacher's effectiveness has not been undertaken systematically (Riaz, 2000). The new developments in the field of education have created very different kind of needs, both on the part of students and teachers. Students' evaluations of teaching effectiveness (SETs) are variously collected to provide (a) diagnostic feedback to faculty that will be useful for the improvement of teaching; (b) a measure of teaching effectiveness to be used in personnel and administrative decision making; (c) information for students to use in the selection of courses and teachers; and (d) an outcome or a process description for research on teaching (Marsh & Roche, 1999; Radmacher & Martin, 2001). There is an essential need of a reliable and valid instrument of evaluation for teacher's effectiveness in institutions of Pakistan, where the education system is still not providing students the opportunity to give their feedback about teachers (Riaz, 2000).

Though, now the situation, especially on higher education level, is changing and many teachers desired to know the student's demands from their teaching, and how can they improve their performance in order to satisfy the needs of students and advance the overall quality of education system. For that purpose, both students and teachers need some reliable mode of teacher's assessment by the students. The primary purpose of developing the present rating scale is formative; that is, facilitating faculty growth, development, and self-improvement.

The study of gender differences between students and teachers related to their evaluation of teaching style effectiveness have also been planned in the present research. Though the previous researches have not shown very consistent differences between the gender of students in the evaluation of teaching style effectiveness (Marsh, 1984; Watchtel, 1998) but Tatro (1995) found that female students generally gave higher ratings than males, while Koushki and Kuhn (1982) found evidence supporting the reverse. However, in case of teacher's gender the previous researches revealed that gender (Freeman, 1994; Morris, Gorham, Stanley, & Huffman, 1996) has been investigated as important teacher characteristic related to effective teaching. As it is noted, that both genders have very different attitudes toward each other, especially in our culture, certain type of roles are associated to particular gender which are quite different to the roles assigned to the opposite gender (Hassan, 1994). The demands from a female teacher may be quite different, seen by each gender, as compared to male teacher. Keeping these cultural expectations in mind, it has been planned to see the differences in evaluation of teaching style regarding the gender of teachers. Researches show that the tendency of students to rate same-sex

instructors slightly higher than opposite-sex instructors (Centra, 1993; Feldman, 1993). The present study also aimed at measuring perception of each gender student about teaching effectiveness of teachers of both genders.

Following above mentioned purposes of teaching effectiveness, it becomes very essential to assess these needs of teaching evaluation systematically and adopt proper strategies for teaching effectiveness. Therefore present study aimed at developing teacher's evaluation scale that may be applicable to all the universities of Pakistan with semester system, to measure the attitudes of students towards the teacher. This scale would be equally applicable to social and natural sciences, as it includes the items particularly suitable for the subject matter that both groups deal with. It is felt that even if the rating scales for measuring teacher's effectiveness are present in the educational institutions, however, these should be renewed after every two years, as the demands of students are changing very rapidly.

Though there is a reliable scale known as Peshawar University Teachers' Evaluation Scale (PUTRS) developed by Riaz (2000) but the need for development of a new scale is felt as there are differences in educational system for which both scales are developed. PUTRS is developed for evaluation of teachers in annual system and UTES is developed to be used in semester system. In annual system teachers have ample time to build rapport with students and cover course content while semester system provide the University teachers with the opportunity to comprehensively re-examine, redefine, and remodel curriculum as necessary to deliver breadth of information and depth of insight, as well as utility, in order to best meet the requirements of modern life and academic demands. Hence there emerged a need for separate teaching evaluation method for semester system as well. Moreover the differences in educational system between Pakistan and Western countries necessitate the development of an indigenous scale for semester system too (Riaz, 2000). The UTES is equally applicable for natural and social sciences as it included the items dealing with the subject matter of both sciences.

Method

In the context of the aforementioned purposes, the major objectives of the present study were:

1. To develop an indigenous scale to measure the effectiveness of university teaching.
2. To establish the psychometric properties, that is, reliability and

validity indices will be determined.

3. To explore the differences of male and female students in the evaluation of university teacher's effectiveness.
4. To explore the differences in the evaluation of male and female University teachers.

Study I: Development of the University Teacher's Evaluation Scale

The development of the scale was carried out in three phases. Study I intend to develop the University Teacher's Evaluation Scale, and comprised of three phases. Phase I aimed to generate item pool with the help of literature review, generating focussed group discussions with students and taking opinion of students with the help of open ended questionnaires. At the end of study I subject matter expert's opinion was taken for selection of final items. The selected items were factor analysed to determine the factorial structure of final scale. Details of each phase of Study I are as follows:

Phase I: Generation of item pool. The overall item pool was generated with the help of literature review, focus group discussions, and open ended questionnaires.

Literature review. The item pool for the scale was generated with the help of review of the existing literature. The identified categories through extensive literature review were (1) Command/ knowledge/Expertise; (2) Individual/Group communication/Interaction; (3) Clarity of presentation; (4) Breadth of coverage; (5) Teaching environment/Classroom management; (6) Fostering intellect/ Creativity; (7) Assignments/Readings/Handouts; (8) Fairness in grading/Examination; (9) Teaching styles; (10)Attitude toward students; (11) Personality factors; (12) Punctuality; (13) Conceptual clarity; (I4) Organization/Balance/Planning; (15) Work load; (16) Teacher's involvement; (17) Ethical standards; and (18) Collaboration with parents (see for example, Abrami, 1985; Froyen & Iverson, 1999; Frey, Leonard, Beatty, & Shrock, 1981; Marsh & Thomas, 1992).

Focus groups. To generate items from identified categories, a series of focus group discussions were conducted with students of Quaid-i-Azam University, the size of each focus group comprised of 6 to 8 students. The total number of students participated in focus group discussions were 30 (16 men, 14 women). Their age ranged from 20-27 years (M = 23, SD = 7.25) from both natural (n =12) and social

sciences (n = 18). Informed consent was taken from the participants selected through purposive sampling; and they were approached at University campuses. Special permission was acquired from the library staff to conduct focus group discussions in a peaceful room at library**.** On the basis of these three focus groups 342 items were generated. The number of items generated for each category were: (1) Command/knowledge/Expertise included 16 items; (2) Individual/ Group communication/Interaction included 43 items; (3) Clarity of presentation included 8 items; (4) Breadth of coverage included 23 items; (5) Teaching environment/Classroom management included 8 items; (6) Fostering intellect/Creativity included 14 items; (7) Assignments/Readings/Handouts included 13 items; (8) Fairness in grading/Examination included 23 items; (9) Teaching styles included 76 items; (10) Attitude toward students included 35 items: (11) Personality included 21 items; (12) Punctuality included 5 items; (13) Conceptual clarity included 15 items; (14) Organization/Balance/ Planning included 16 items; (15) Work load included 8 items; (16) Teacher's involvement included 11 items; (17) Ethical standards included 2 items; and (18) Collaboration with parents included 5 items.

***Open ended questionnaires*.** To generate additional items, open-ended questionnaires were also administered on 20 students (10 men, 10 women) 5 men and 5 women from social sciences and 5 men and 5 women from natural sciences with age range of 20-27 years (M = 22, SD = 4.13) enrolled in M.Sc. and M.Phil Programs. The teacher's questionnaire was also given to 10 male and 10 female teachers; age range of the teachers was 35-50 years (M = 44, SD = 9.05). Students of Psychology (n = 5), International Relations (n = 3), and MBA (n = 2) participated from social sciences while from natural sciences, students of Biology (n = 4), Mathematics (n = 3), and Chemistry (n = 3) departments took part in the study. A total of 153 items were added to the existing categories. (1) Command/knowledge/Expertise (6 items); (2) Individual/group communication/Interaction (21 items); (3) Clarity of presentation (7 items); (4) Breadth of coverage (12 items); (5) Teaching environment/Classroom management (11 items); (6) Fostering intellect/Creativity (2 items); (7) Assignments/Readings/ Handouts (6 items); (8) Assignments/Fairness in grading/Examination (8 items); (9) Teaching styles (13 items); (10) Attitude toward students (16 items); (11) Personality factors (23 items); (12) Punctuality (5 items); (13) Conceptual clarity (5 items); (14) Organization/Balance/Planning (7 items); (15) Work load (4 items); (16) Teacher's involvement (3 items); (17) Ethical standards (3 items); and (18) Collaboration with parents (1 item). After adding the items

from open-ended questionnaire, a total of 495 items were generated on each category identified. On the basis of these focus group discussions and open-ended questionnaire the already identified areas though literature reviews were further confirmed.

After extracting items from all these sources, these were transcribed in the form of statements and a questionnaire was prepared by combining all these statements. The items for overlapping and repetitive content were carefully checked and the redundant items were dropped; while remaining items were improved by rephrasing. Frequencies were also assigned to see which items were greatly emphasized and were common in all these different sources before deleting any repeated item. For the final confirmation of short listed 195 items, the list was given to the judges.

Phase II: Judges' opinion. The judges (6 men and 5 women) included M.Phil degree holders (n = 3), Ph.D Scholars (n = 4), and regular faculty members (n = 4) having an experience of at least five years in the teaching and completed their PhD with age range of 26 to 45 years (M = 35, SD 12.4). On the basis of the judges' opinion, over-lapping and redundant items were eliminated and categories having the same type of items were merged. As a consequence, number of total categories reduced from 18 to 8 and the number of items in each category were also reclined. Thus the remaining categories included: (1) Command on the subject (**11** item); (2) Communication skills (12 items); (3) Clarity of presentation, (13 items); (4) Breadth of coverage (9 items); (5) Teaching environment/Classroom management (**11** items); (6) Fostering intellect/Creativity (7 items); (7) Assignments and Fairness in grading (10 items); and (8) Attitude toward students (12 items). The total number of items was reduced to 85 and all these categories were included in the University Teacher's Evaluation Scale.

Phase III: Factor analysis on the items of University Teacher's Evaluation Scale. In order to select the final items and get a factor situation of the scale, a factor analysis was carried out on independent sample.

Sample. The size of the sample was decided keeping in view the requirement of the sample size for factor analytic study. Kline (1986) has mentioned that a ratio of 3: 1 gave loadings essentially identical to those with a ratio of 10: 1. Therefore the sample consisted of 300 students of Masters Level including men (n = 150) and women (n = 150). The sample of male students of natural and social sciences (n = 150) and female students of natural sciences (n = 75) was taken

from Quaid-i-Azam University Islamabad, and female students of social sciences (n = 75) were incorporated from Fatima Jinnah Women University, Rawalpindi. The age range of respondents was from 20-27 years (M = 23.0, SD = 1.40). Informed consent was taken from participants and they were ensured that their responses will be kept confidential and will be used only for research purposes.

Measure. The initial form of the University-Teacher's Evaluation comprising 85 items was used to collect the data. The scale consisted of five response categories reflecting the desirability of the quality to be present in effective university teacher. The five response categories ranged from *do not agree* (1)*, to strongly agree* (5)*.* The minimum possible score was 85 and maximum score could be 425. Demographic information was also obtained along with the questionnaire. The greater score on UTES indicates student's positive evaluation of teaching effectiveness and low score reflected student's negative rating of teaching effectiveness.

Results

Factor Analysis

Firstly, as all the items of the scale were empirically determined, therefore it has sufficient content validity. To determine the dimensionality and construct validity of the scale developed, the 85 items were factor analyzed through Principal Component Factor Analysis. The value of Kaiser-Meyer-Olkin Measure of Sampling Adequacy .86 showed that data is meritorious for factor analysis. The large value of (X^2 = 12526.69, p < .001) shows that correlation matrix is not an identity matrix and variables are positively correlated with each other.

Before the Barlett Test of Sphericity, item total correlation was also computed which showed all the items correlated significantly with each other ranging from .27 to .59 (p < .001). Following Guertin and Bailey (1970) while all the items are found highly correlated with each other and with the total, the Direct Oblimin Method of Principal Component factor analysis was applied. On the basis of .40 factor loading criteria of Factor Analysis, 49 items were retained. All these items are falling in one category, showing the unifactor structure of the scale.

Table 1

Factor Matrix of the 85 items of University Teacher's Evaluation through Principal Component Analysis using Direct Oblimin Method

Item No.	Factor I	Factor II	Factor III	Factor IV	Factor V
1	**.44**	11	-.19	-6.5	-.30
2	.36	14	-.36	.20	.20
3	.36	.66	-.22	.18	.17
4	.26	-.86	-.22	.10	-.58
5	.37	-.30	-.40	-.91	.16
6	.29	-.32	-.66	-.89	.14
7	**.41**	-.39	-.39	-.14	-.87
8	**.43**	.37	-.21	-.94	-.33
9	.35	-.10	-.22	-.12	-.18
10	.26	.31	-.30	- .27	.17
11	.27	.35	-.30	-.32	-.21
12	.36	-.81	-.19	-.13	-.11
13	.26	.29	.11	-.25	-.10
14	**.40**	.37	-.18	-.20	.15
15	**.42**	-.39	-.23	-.13	-.54
16	**.49**	.18	-.17	-.76	-.67
17	.32	.30	-.24	-.69	-.10
18	.32	.35	-.85	-.30	-.75
19	**.48**	-.32	-.13	-.99	.22
20	**.53**	.27	-.16	-.11	-.80
21	**.53**	-.23	-.82	-.16	-.55
22	**.46**	-.25	-.20	-.10	-.64
23	.29	.30	-.21	-.19	-.61
24	**.45**	.23	.21	-.34	.23
25	.25	.30	.18	-.10	.15
26	**.49**	.17	-.63	-.14	-.17
27	.34	.29	.22	-.25	-.21
28	**.47**	-.61	-.23	-.52	-.38
29	**.52**	-.43	.13	-.20	-.97
30	**.48**	.15	-.08	-.23	-.25
31	**.49**	.23	-.13	-.10	.20
32	.38	.30	-.93	-.12	-.14
33	**.49**	.20	-.13	-.38	-.81

Continued…

Item No.	Factor I	Factor II	Factor III	Factor IV	Factor V
34	**.53**	-.42	-.20	.34	.21
35	**.46**	-.45	-.15	.39	.37
36	**.42**	.25	-.81	-.16	-.74
37	**.44**	-.31	-.12	-.86	.17
38	**.58**	.29	.28	-.25	.15
39	**.41**	-.53	-.30	-.17	.23
40	**.47**	.20	.01	-.60	-.32
41	**.42**	.11	-.12	-.40	-.12
42	**.58**	.23	- 14	.13	.14
43	**.45**	-.82	-.80	-.40	-.68
44	**.52**	.29	-.40	-.16	-.31
45	.32	.16	-.17	-.50	-.54
46	**.51**	-.17	.22	-.89	-.40
47	**.46**	-.59	-.36	-.25	.12
48	**.41**	.13	.23	-.15	.13
49	**.50**	-23.	-.50	-.12	-.49
50	**.46**	.19	-.59	.24	.15
51	**.50**	-.45	-.12	-.88	-.98
52	**.41**	.14	.22	.14	-.11
53	**.48**	.15	-.16	.14	-.31
54	**.47**	-.34	.16	-.36	.19
55	**.44**	.03	-.86	.29	-.25
56	.36	.16	.39	-.10	-.12
57	**.51**	.14	-.88	.15	-.15
58	**.47**	-.20	-.54	.14	-.18
59	**.44**	-.50	-.68	-.42	-.98
60	.30	.15	.33	-.10	-.62
61	.31	-.52	.31	.14	-.78
62	**.43**	-.17	.21	.16	.16
63	**.46**	.14	.11	.16	-.90
64	.32	-.52	-.43	-.00	-.84
65	.39	-.10	-.72	.27	-.27
66	**.41**	.12	.01	.15	-.20
67	**.48**	-.35	.15	-.54	-.18
68	.39	.12	.45	-.80	-.18
69	.38	-.67	.12	.19	-.48

Continued…

Item No.	Factor I	Factor II	Factor III	Factor IV	Factor V
70	.46	.21	-.86	.11	-.13
71	.39	.21	.17	.19	.68
72	.46	.18	.11	.24	.29
73	.34	-.06	.14	-.50	-.24
74	.39	.13	.24	-.21	-.17
75	.35	-.70	.30	.12	-.51
76	.39	.09	-.07	.45	-.12
77	.33	.18	-.97	.27	-.13
78	.35	.08	.30	-.42	-.81
79	.47	-.31	.27	-.71	-.98
80	.33	.11	.17	.17	.30
81	.36	-.85	-.86	.29	-.31
82	.38	.12	.21	.29	.18
83	.38	-.30	-.73	.29	.28
84	**.43**	-.15	.00	.23	-.33
85	.30	-.30	-.52	.15	-.73
Eigen Values	18.17	6.4	2.8	2.6	2.1
% of Variance	18.9	6.7	2.9	2.7	2.2
Cumulative %	18.9	25.6	28.5	31.2	33.4

Note. Factor loadings ≥ .40 have been boldfaced.

Factor analysis yielded 49 final items and their corresponding loadings for the scale, which was named as University Teacher's Evaluation Scale (UTES). As evident from results, Factor 1 has an Eigen value of 18.17 which explains 18.9 percent of total variance. Other four extracted factors have the minimum acceptable Eigen values and explain very little amount of variance and hence the newly developed measure appeared to possess the quality of unidimensional scale measuring teaching effectiveness.

Alpha Reliability of UTES

After the final selection of 49 items of the scale UTES, the alpha reliability (.92) and split half reliability coefficients (.78) of the final items were significantly high, thereby indicating high internal consistency of the scale.

Study II: Determination of Psychometric Properties of University Teacher's Evaluation Scale and Gender Differences in Evaluation

Study II aimed in determining psychometric properties of University Teacher's Evaluation Scale, including reliability and validity of scale and checking for gender differences in evaluation of teachers by university students.

Sample. Independent sample consisted of 30 students (14 women and 16 men) of Masters in Business Administration, were taken from Iqra University, Islamabad with age range of 20 to 24 years (M = 22.0, SD = 1.40). The demographic information such as gender of the student, discipline, semester, and age was also obtained. The evaluation of male teacher and female teacher was measured separately from the same respondents. For evaluation, both male and female teachers (working at respective Department for at least 2 years) of same popularity were taken separately. For this purpose information was qualitatively obtained from both the management (including Director of the respective Department) and students through informal discussion and semi-structured interviews. The sample items included "to what extent the teacher presents the course in a well-organised manner?", "Does the teacher make sincere efforts to enhance student learning", and "to what extent the teacher has been confident in teaching the course?" Respondents were approached at their respective campuses. Moreover, informed consent was acquired from them, and was ensured that their responses will be kept confidential and will be used only for research purposes.

Measures

University Teacher's Evaluation Scale (UTES). The 49 items were retained in UTES after study I. UTES was found to be a single factor instrument measuring teaching effectiveness. Responses were obtained on a Likert type 5-point rating scale (*strongly disagree* = 1, *disagree* = 2, *neutral* = 3, *agree* = 4, *strongly agree* = 5) with minimum possible score of 49 and the maximum possible score was 245. The higher score on UTES indicates student's positive evaluation of teaching effectiveness and lower score signify student's negative evaluation of teaching effectiveness.

Peshawar University Teacher's Rating Scale (PUTRS). Peshawar University Teacher's Rating Scale (PUTRS) was developed

by Riaz (2000), to evaluate student's perception of university teaching quality. Responses could be marked along Likert type 5-point rating scale (*always* = 5, *mostly* = 4, *sometimes* = 3, *rarely* = 2; *never* = 1) consisting of 25 items. The 25 affirmative statements refer to teacher's mastery of the subject, ability to stimulate intellectual curiosity, assignments/fairness in grading, student-teacher communication and teacher's attitude towards the students. The minimum score on PUTRS could be 25 whereas the maximum possible score was 125. Low scores on the test indicate poor quality of teaching whereas high scores demonstrate high quality of teaching. The scale can safely be regarded as a one factor test for the assessment of teaching quality at the university level. Riaz (2000) found .95 alpha reliability for PUTRS.

Reliability of the UTES

The two values of alpha for UTES and PUTRS for male and female teachers are computed. The Cronbach alpha coefficients on UTES for female teachers and male teachers are .93 and .89 respectively. Similarly, alpha values on PUTRS for female teacher is .98 and for male teacher is .90.

Table 2

Test-retest Reliability Coefficients of UTES and PUTRS

EFT	No. of Items		Alpha Coefficients		Split-Half
	Part -I	Part -II	Part -I	Part -II	
UTES	25	24	.86	.91	.80
PUTRS	13	12	.97	.96	.95
EMT					
UTES	25	24	.83	.84	.79
PUTRS	13	12	.82	.86	.82

Note. EFT = Evaluation of Female Teacher; EMT = Evaluation of Male Teacher; UTES = University Teacher's Evaluation Scale; PUTRS = Peshawar University Teacher's Rating Scale.

Test-retest Reliability was also calculated for estimating the temporal stability of the test. There was an interval of 15 days between the administration of the test and the retest. The two sets of

scores obtained were used to calculate a coefficient of correlation indicating the test retest reliability of the scale. The correlation between the test and retest of UTES was significant at $p < .0l$ (Women = .81, Men = .80).

Table 3

Gender differences among students on UTES and PUTRS in evaluation of male and female teacher

Scales	Women (n = 14)	Men (n = 16)			95% CI		Cohen's
	M(SD)	*M(SD)*	*t*(28)	*p*	*LL*	*UL*	*d*
EFT							
UTES	19.28 (20.03)	18.87(24.25)	1.76	.09	-3.7	-.2	0.06
PUTRS	11.57 (18.14)	10.37 (28.45)	1.92	.06	-3.9	-1.1	0.09
EMT							
UTES	19.78(20.18)	19.18(16.77)	1.51	.14	-2.9	-1.3	0.09
PUTRS	11.28(7.46)	10.00(12.96)	2.11	.04	-3.2	-1.2	0.78

Note. CI = Confidence Interval; LL = Lower Limit; UL = Upper Limit; EFT = Evaluation of Female Teacher; EMT = Evaluation of Male Teacher. EFT= Evaluation of Female Teacher; EMT= Evaluation of Male Teacher.

Results indicated that in UTES the difference was non-significant between female and male students in the evaluation of male teacher; however, in PUTRS the difference was significant and the female students evaluated the male teacher more favorably as compared to male students.

Convergent Validity

Convergent validation was established by exploring the correlation between UTES and PUTRS. Results indicated that both scales have significant positive correlation with each other (r =.71, $p <$ .001).

Discussion

Universities are often assessed in terms of efficiency and effectiveness of its teaching faculty. Teachers have an important role in the socio economic development of their state in particular and the

nation in general. The assessment of teacher's performance is not incorporated in the prevailing educational system of our country. As a result, educational standards are falling at an alarming rate. Every successful educational enterprise requires optimum utilization of human capabilities available to the system. Consequently, every such enterprise or activity needs periodic assessment and review. This has to be followed by search for better conceptual understanding, implementation strategies and practices. It is now well understood and appreciated that the role of teachers shall continually change in the 21st century for obvious reasons. While it will be necessary for the teachers and the teacher preparation systems to ensure regular acquisition of new skills and upgradation of existing skills, the assessment of the performance of teachers shall also remain an essential pre-condition for enhancing the efficacy of educational processes (Rajput, 1996). The present research aimed at development and validation of an indigenous scale for evaluation of effectiveness of University teaching.

The items for the scale were generated in Phase I. While developing the item pool for University Teacher's Evaluation Scale (UTES) a systematic process of empirical generation and careful selection of items was employed. The elaborated process for this purpose was carried out because of the emphasis that has been placed by several researchers and theorists on careful writing and selection of the items for development of an instrument (McKeachie, 1990; Rice, Stewart, & Hujber, 2000; Wylie, 1979). The findings of the study in phase III showed that 'Teachers Evaluation' is a uni-dimensional construct. The UTES constructed is found to be internally consistent and reliable scale. Thus the findings are consistent with the findings of previous research which identified teaching effectiveness as one factor construct (e.g., Marsh, 1980, 1981; Marsh & Bailey, 1993; Marsh & Hocevar, 1991; Marsh & Overall, 1981; Marsh & Roche, 1999; Marsh & Thomas, 1992; Riaz, 2000). After the final selection of 49 items of the UTES, reliability coefficients were determined, thereby demonstrating UTES as an internally consistent single factor scale.

Part II of the study pertains with the validation of the scale developed. The convergent validity of the scale was determined with an already developed reliable scale of teaching evaluation known as Peshawar University Teacher's Rating Scale (Riaz, 2000). The scale was administered on the students when they had not received their semester grading and was re-administered after they received the grading, but it does not affect their evaluation of teachers (e.g., see Tata, 1999).

However, the present research shows that the students' evaluations are reliable measure for teachers' evaluation, and there is consistency in student's ratings of both male and female teacher's evaluation in re-test. The correlation between the two scales came out to be quite satisfactory and assured the convergent validity of the scale. Though the PUTRS is a reliable scale but it is felt that teacher's evaluation scales should be developed or renewed after every two or three years as the demands and needs of students as well as teachers are changing very rapidly, so a proper assessment of these demands should be made.

The differences of male and female students in the evaluation of male and female teachers were investigated. The review of diverse literature reveals little consistent evidence of gender bias (Watchtel, 1998). For example, Tatro (1995) found that women gave higher ratings than male students, while present findings revealed that the difference is not significant, which is also supported by Koushki and Kuhn (1982). Results also indicated that in UTES there is non-significant difference in men and women in the evaluation of male and female teacher. However, in PUTRS the female students evaluated male teacher more favorably as compared to male students, hence; endorsing the findings of Feldman (1992, 1993). Existing research reviewed by Feldman (1993) on student ratings of male and female teachers in both the laboratory and the classroom settings. In his review of laboratory studies, Feldman (1992) reported that the majority of studies reviewed showed no difference in the global evaluations of male and female teachers. On the other hand, few studies in which differences are found, indicating male instructors receiving higher overall ratings than female teachers, in case when evaluated by female students.

Moreover, women faculty received higher ratings on questions addressing grading of students, and women students rated women faculty even higher than did male students. Researches (Basow, 1998; Seldin & Silberg, 1995) suggested that on questions of communication style and intellect (such as, speaks in an appropriate manner, has a wisdom to teach the subject), male faculty tended to be rated higher than their female counterparts by their students of both genders. However, the researches are not consistent in case of the category of 'command on the subject' (Dukes & Victoria, 1989). To receive good evaluations, male professors simply must demonstrate their competence and knowledge; that is, they need to fulfill their stereotypical gender role expectations. But female professors bear a double burden: they must fulfill both their gender role by being nurturing and warm, as well as their professional role by being

competent and knowledgeable. For example, separate studies led by Bennett (1982), Statham, Richardson, and Cook (1991) found that female professors are judged more negatively than men if they are not more interested in and available to the students than male professors. But even when female professors are more available and more helpful, their overall ratings are no higher. In order to receive comparable ratings, female professors need to do more than their men counterparts. Thus, findings of no difference between male and female professors in overall ratings may mask the fact that different standards are being used to judge men and women faculty.

Conclusion

The present practice of scale development for teacher's evaluation aims at professional and educational improvement. If teacher effectiveness is assessed from time to time and incorporated into the system, it may motivate those who are brilliant, enthusiastic idealistic and professionally better equipped. An important function of this appraisal for many teachers is to improve staff communication and strengthen the channels of communication within the institution.

It is further hoped that the scale developed will provide administrators and teachers with extensive opportunities for training that underscore the complexity of the art of teaching. Expectantly, it will raise the level of discussion about good teaching, what it looks like and the connection between good teaching and student achievement.

Limitations and Suggestions

Although the UTES, on the basis of its psychometric characteristics, can be regarded as valid and reliable instrument to assess teaching effectiveness, there are few limitations. The findings of the present research provide a favorable evidence of and convergent validation of the UTES, but it should not be considered conclusive and the study is needed to be replicated. It is quite hard to achieve the construct validation in a single research as it is an ongoing and dynamic process of revising the definition and measurement of the construct. The convergent validity can also be further determined by using other scales developed for the evaluation of university teaching and by using scales of constructs which theoretically correlate with effective teaching. Moreover further studies can be conducted to determine the discriminant validity of UTES with

related measures. Though Quaid-i-Azam University (QAU) is a national university with quota from each province representation, but still the size of sample for the research is not large enough to be representative of all the universities of Pakistan, because of time constraint and some other practical problems. There are many other variables like student's expectations concerning the instructors, student's belief, student's that the evaluations will be truly used for teaching improvement, emotional state of students at the time of evaluation, prior subject interest etc can influence their ratings.

Implications

Student's evaluation of teaching effectiveness can provide the diagnostic feedback to faculty about the effectiveness of their teaching that will be useful for the improvement of the teaching. A measure of teaching effectiveness can be used in administrative decision making. The information may further be used by students in the selection of courses and instructors. It also the measure of the quality of the course, to be used in course improvement and curriculum development. Apart from class room teaching student's opinions can also contain several questions pertaining to non class room aspects of the teacher's integrity, students' teacher relation, efforts that goes in preparation and updating of lecture material. Getting aware of student's perceptions and taking them positively, not only benefits a teacher and student, rather the whole system, the whole country. It is hoped that this scale will offer opportunities to new teachers and especially those having some difficulty in teaching, to improve by knowing their own limitations. The vast majority of those teachers who meet or exceed good standards of teaching but need support for their continued growth, will also be benefited by evaluations of their own teaching effectiveness.

References

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of Higher Education*, *8*(3), 211-228.

Abrami, P. C., & d'Apollonia, S. (1997). The dimensionality of ratings and their use in personnel decisions. *American Psychologist, 52*(11), 1198-1208.

Basow, S. A. (1998). Student evaluations: The role of gender bias and teaching styles. In L. H. Collins, J. C. Chrisler, & K. Quina (Eds.), *Career strategies for women in academia: Arming Athena* (pp.135-156). Thousand Oaks, CA: Sage Publications.

Bennett, S. K. (1982). Student perceptions of and expectations for male and female instructors. *Journal of Educational Psychology, 74*, 170-179.

Bhatanagar, R. P., & Jain, N. K. (1994). Net examination: One more analysis. *University News, 32*(36), 3-10.

Blunt, A. (1991). The effects of anonymity and manipulated grades on student ratings of instructors. *Community College Review, 18*, 48-54.

Braskamp, L. A., & Ory, J. C. (1994). *Assessing faculty work.* San Francisco: Jossey Bass.

Buskist, W. (2001). *Ways of the master teacher: Student and faculty perspectives on effective college and university teaching.* Invited address given at Stephen F. Austin State University, Nacogdoches, TX.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research*. Center for Faculty Evaluation and Development, Kansas State University.Kansas, United States of America.

Cashin, W. E. (1992). Student ratings: The need for comparative data. *Instructional Evaluation and Faculty Development, 12*, 146-150.

Centra, J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness.* San Francisco: Jossey-Bass.

Chen, Y., & Hoshower, L, B. (2003). Student evaluation of teaching effectiveness: An assessment of student perception and motivation. *Assessment and Evaluation in Higher Education, 28*(1), 71-88.

Clift, J. (1981). Establishing the validity of a set of summative teaching performance scales. *Assessment and Evaluation in Higher Education, 4*(3), 193-206.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-Analysis of multi-section validity studies. *Review of Educational Research*, *51*(3), 281-309.

d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist,* 52(11), 1198-1208.

Dukes, R. L., & Victoria, G. (1989). The effects of gender, status, and effective teaching on the evaluation of college instruction. *Teaching Sociology, 17,* 447-457.

Dunkin, M. J., & Barnes, J. (1986). Research on teaching in higher education. In M. C. Wittrock (Ed.), *Handbook of research on teaching* (pp. 754-777). New York, NY: Macmillan.

Feldman, K. A. (1992). College students' views of male and female college teachers: Part I-Evidence from the social laboratory and experiments, *Research in Higher Education, 33*(3), 317-375.

Feldman, K. A. (1993). College students' views of male and female college teachers: Evidence from students' evaluations of their classroom teachers. *Research in Higher Education, 34*, 151-211.

Freeman, H. (1994). Student evaluations of college instructors: Effects of type of course, gender and gender role, and student gender. *Journal of*

Educational Psychology, 86(4), 627-630.

Frey, P. W., Leonard, D. W., Beatty, W. W., & Shrock, S. A. (1981). Student ratings of instruction: validation research. *American Educational Research Journal*, 12 (4), 435-447.

Froyen, L. A., & Iverson, A. M. (1999). *The teaching process.* New Jersey, USA: Prentice Hall Publication.

Gavlick, M. (2006). Triangulating the relationships among publication productivity, teaching effectiveness, and student achievement. *New Directions for Institutional Research*, *1996*(90), 49-56.

Greenwald, A. G., & Gilmore, G. (1997). Grading leniency is a removable contaminant of student ratings. *American Psychologist, 52*(11), 1209-1217.

Guertin, W. H., & Bailey, J. P. (1970). *Introduction to modern factor analysis.* Michigan, USA: Edwards Brothers, Inc

Hassan, I. N. (1994). *Education of women in Asia.* Pakistan Report: RETA: 5513: Asian Development Bank. Federal Ministry of Education, Govt of Pakistan.

Hofman, J., & Kremer, L. (1980). Attitudes toward higher education and course evaluation. *Journal of Educational Psychology*, *72*(5), 610-617.

Hong, W., Xuezhu, C., & Zhao, K., (2007). On the relationship between research productivity and teaching effectiveness at research universities. *Frontiers of Education in China, 2*(2), 298-306.

Khandelwal, K. A. (2009). Effective teaching behaviours in the college classroom: A critical incident technique from students' perspective. *International Journal of Teaching and Learning in Higher Education, 21*(3), 299-309.

Kline, P. (1986). *A handbook of test construction: Introduction to psychometric design.* Methuen Massachusetts, USA: Methuen & Co. Ltd.

Koon, J., & Murray, H. G. (1995). Using multiple outcomes to validate student ratings of overall teacher effectiveness. *Journal of Higher Education, 66*, 61-81.

Koushki, P. A., & Kuhn H. A. (1982). How reliable are student evaluations of teaching? *Engineering Education, 72*, 362-367.

Kwan, K. P. (1999). How fair are student ratings in assessing the teaching performance of university teachers? *Assessment and Evaluation in Higher Education, 24*(2), 181-195.

Lersch, K. M., & Greek, C. (2001). Exploring the beliefs surrounding student evaluations of instruction in criminology and criminal justice undergraduate courses. *Journal of Criminal Justice Education, 12*(2), 283-299.

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.

Marsh, H. W. (1980). Students' evaluations of university teaching: Research findings, methodological issues and directions for future research. *International Journal of Educational Research, 11*(3), 253-388.

Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects on student ratings of instrument effectiveness. *Applied Psychological Measurement, 6*, 47-60.

Marsh, H. W. (1984). Students' evaluation of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology, 76*, 707-754.

Marsh, H. W. (1990). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structure. *Journal of Educational Psychology, 83*(2), 285-296.

Marsh, H. W. (2007). Do university teachers become more effective with experience: A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology*, *99*(4) 775-790.

Marsh, H. W., & Bailey, M. (1993). Multidimensionality of students' evaluation of teaching effectiveness: A profile analysis. *Journal of Higher Education, 64*(1), 1-15.

Marsh, H. W., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A-multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143-233). New York, NY: Agathon Press.

Marsh, H. W., & Dunkin, M. J. (1997). Students' evaluation of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.). *Effective teaching in higher education: Research and practice* (pp. 241-320). New York, NY: Agathon Press.

Marsh, H. W., & Hocevar, D. (1991). The factorial invariance of students' evaluations of college teachers. *American Educational Research Journal, 21*, 341-366.

Marsh, H. W., & Overall, J. U. (1981). The relative influence of course level, course type and instructor on student's evaluation of college teaching. *American Educational Research Journal, 18*, 103-112.

Marsh, H. W., & Roche, L. A. (1999). Rely upon set research. *American Psychologist, 54*(7), 517-518.

Marsh, H. W., & Thomas, C. S. (1992). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67*, 833-839.

McKeachie, W. J. (1990). Research on college teaching: The historical background. *Journal of Educational Psychology, 82*(2), 189-200.

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*(11), 1218-1225.

Morris, T. L., Gorham, J., Stanley, H. C., & Huffman, D. (1996). Fashion in the classroom: Effects of attire on student perceptions of instructors in college classes. *Communication Education, 45*(2), 135-147.

Murray, H. G. (1983). Low inference classroom teaching behaviors arid student ratings of college teaching. *Journal of Educational Psychology, 71*, 856-865.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82*(2), 250-261.

Narula, M. (2000). *Effective teaching in higher education.* New Dehli, India: Efficient Offset Printers.

Perry, R. P. (1997). Teaching effectively: Which students? What methods? In R. P. Perry & J. C. Smart (Eds.), *Effective teaching in higher education: Research and practice* (pp. 154-168). New York, NY: Agathon Press.

Radmacher, S. A., & Martin, D. 1. (2001). Identifying significant predictors of student evaluations of faculty through hierarchical regression analysis. *Journal of Psychology, 135*, 259-268.

Rajput, J. S. (1996). *Search for futuristic curriculum framework in teacher education.* New Delhi, India: University News

Regel, O. (1992). *The academic credit system in higher education: Effectiveness and relevance in developing countries.* Education and Employment Division Population and Human Resources Department the World Bank Report.

Riaz, M. N. (2000). Student evaluation of university teaching quality: Analysis of a teacher's rating scale for a sample of university students. *Pakistan Journal of Psychological Research, 15*(3-4) 107-117.

Rice, R. E., Stewart, L. P., & Hujber. (2000). Extending the domain of instructional effectiveness assessment in student evaluations of communication courses. *Communication Education, 40,* 253-266.

Seldin, P. (1999). *Changing practices in evaluating teaching: A practical guide to improved faculty performance and promotion/tenure decisions.* Bolton, MA: Anker Publishing company, Inc.

Seldin, P., & Silberg, N. T. (1995). Student evaluation of college professors: Are female and male professors rated differently? *Journal of Educational Psychology, 79*(3), 308-314.

Sproule, R. (2002). The under determination of instructor performance by data from the student evaluation of teaching. *Economics of Education Review*, *21*(3), 287-294.

Stack, S. (2003). Research productivity and student evaluation of teaching in social science classes: A research note. *Research in Higher Education, 44*(5), 539-556.

Statham, A., Richardson, L., & Cook, J. (1991). *Gender and university teaching: A negotiated difference. American Political Science Association* Albany, NY: SUNY Press.

Tata, J. (1999). Grade distributions, grading procedures, and students' evaluation of instructors: A justice perspective. *Journal of Psychology, 133*(3), 263-271.

Tatro, C. N. (1995). Gender effects on student evaluations of faculty. *Journal of Research and Development in Education, 28*, 169-173.

Trout, P. (2000, Apr. 21). Teacher evaluations. *Commonwealth, 27*(8), 10-11.

Wachtel, H. K. (1998). Student evaluation of college teaching effectiveness: A brief review. *Assessment and Evaluation in Higher Education, 23*(2), 191-212.

Wheeler, P. Haertel., & Scriven, M. (1992). *Resource handbook on performance assessment and measurement: A tool for students, practitioners, and policymakers.* Berkeley, CA: The Owl Press.

White, J. C. (2011). On the evaluation of teaching and learning in higher education: A multicultural inquiry. *Assessment and Evaluation in Higher Education*, *36*(2), 643-656.

Wylie, R. C. (1979). *The self concept: Theory and research on selected topics.* (2nd ed.). Lincoln, USA: University of Nebraska Press.

Received December 05, 2010

Revision received April 15, 2013