Intellectual Discourse, 2000
Vol 8, No 2, 199-220

Development of a Measure of Teacher Effectiveness for IIUM

Mahfooz A. Ansari, Mustafa Achoui
Zafar Afaq Ansari*

Abstract: *The paper reports the development of a multidimensional measure to assess teaching effectiveness. The scale, known as Teaching Feedback Survey for the International Islamic University Malaysia (TFS-IIUM), was developed and tested on a large sample of students and lecturers. By employing a principal components analysis with varimax rotation, an instrument consisting of 30 items was obtained, with four factorially independent dimensions of teaching effectiveness: Delivery of Information, Meaningful Interaction, Feedback and Fair Treatment, and Islamic Orientation. It documented high internal consistency reliability coefficients and a substantial amount of content, convergent-discriminant, and criterion-related validity coefficients.*

Evaluation of the performance of lecturers has a long history. The earliest instrument for evaluation of teaching effectiveness is reported to be the Purdue Rating Scale, developed by Remmer in 1928.[1] Since then, a number of instruments have been developed to measure teaching behavior. Notable among them are the instruments developed by Costin and his colleagues in 1971, Feldman in 1977, Frey in 1978.[2] In more recent years, Marsh and his colleagues have developed an

[*] Professors in the Department of Psychology, International Islamic University Malaysia. The study has been made possible by a grant from the Research Center of the International Islamic University Malaysia. Our deepest gratitude goes to Saliza Zakaria for her help in collection and analysis of data, and to IIUM faculty and students for their cooperation in data collection. Correspondence concerning this article should be addressed to the first author, who is now Professor of Management at the Universiti Sains Malaysia <mahfooz@usm.my>.

instrument which has been used widely.[3] The studies carried out to develop these instruments have usually (but not always) shown that teaching effectiveness is a multi-dimensional construct: that is, it is possible for a teacher to excel in some aspects of teaching, but not in others. It has also been found that teaching effectiveness can be assessed with a high degree of reliability and validity.

MEASURES OF TEACHING EFFECTIVENESS

The earliest instrument–Purdue Rating Scale developed by Remmers—consisted of 10 traits related to effective teaching.[4] The 10 traits were later reduced through factor analysis to two dimensions: Empathy and Professional Maturity. Empathy meant those personality characteristics of the teacher, which enhanced his or her esteem in the eyes of the students. Professional Maturity meant confidence and good presentation of the subject matter.[5] Feldman, who began with 20 categories of effective teaching, ultimately reduced them to three clusters, related to the roles of the teacher as a Presenter (communicator), Facilitator (interactor), and Manager (regulator).[6] Frey developed a scale consisting of 21 items, which were reduced to seven dimensions: Organization, Clarity of Presentation, Student Accomplishment, Examining, Class Discussion, Personal Attention, and Workload Difficulty.[7] Braskamp and his colleagues proposed five dimensions of teaching behavior: Teacher Skill, Negative Effect, Student Involvement, Teacher Support, and Teacher Control.[8] Finally, Marsh developed an instrument called Students Evaluation of Educational Quality (SEEQ). He discovered nine dimensions of teaching effectiveness: Learning Value, Instructor Enthusiasm, Breadth of Coverage, Organization and Clarity, Assignments and Readings, Examinations and Grading, Group Interaction, Individual Rapport, and Workload Difficulty.[9]

The number of dimensions emerging in various studies have been different depending upon the sample characteristics, initial item pool and the method of analysis used, and the descriptive labels attached to these factors. Yet, there is a remarkable similarity in the findings. Most studies show that there are some personality characteristics of the teacher that are more conducive than other characteristics to effective teaching. This constitutes a major dimension of teaching effectiveness. Then, there is another large factor indicating competence in communication and management of class. The first group is indicated by items covering Empathy, Facilitation, Personal Attention, Teacher Support, Student Involvement, Negative Affect, Enthusiasm, and Rapport and Interaction. The second group is indicated by items measuring Professional Maturity, Presentation, Teacher Skill, Teacher Control, Instructional Presentation, and Organization and Clarity. Apart from these two major dimensions, there are a number of smaller dimensions that emerge in different studies.

Reliability

The instruments measuring various dimensions of teacher effectiveness have generally been found to be internally consistent. In general, the investigators have reported correlations in the high .70s to .90s.[10] It has also been found that the students' ratings of teachers are fairly stable over time. Moderate to high correlations have been found between mid-semester and end-of-semester ratings of teaching assistants in psychology, social sciences, humanities, physical sciences, and biological sciences. The correlations ranged between .70 and .87 for four of the dimensions measured.[11] In a longitudinal study of Students' Evaluation of Educational Quality (SEEQ), profiles of a cohort of 221 teachers who had been evaluated regularly for over a 13-year period were studied. The ratings on separate scales were found to be stable over time, and so were the multi-dimensional profiles of ratings.[12]

Validity

Validity of students' ratings of teaching has been intensely researched during the last three decades or so. Greenwald, who summarized research in this area was able to locate 172 studies between 1971 and 1995, most of them during 1976-85.[13] The basic questions that have been asked in validity studies include: are the measures meant for assessing teaching effectiveness actually assessing teaching effectiveness; or are they measuring something else, like a lecturer's popularity, his or her ability to create momentary enthusiasm and interest in his or her lecture, lecturers' grading leniency, or the difficulty of the course. These are wide-ranging concerns, which require multi-dimensional effort to demonstrate construct validity of the measures of teaching effectiveness. Consequently, a number of different approaches have been used to study this problem. A review undertaken in the early nineties shows validity studies using relationship with the extent of learning, ratings of former students, lecturers' self-evaluation of their own teaching effectiveness, affective (evaluative) course consequences (for instance, plans to pursue further study in a particular area), and factor analysis. [14] These studies can be grouped into four different approaches, as described below.

Multi-Section Studies. A large group of studies has attempted to demonstrate convergent validity of the measures of teaching effectiveness. It has been shown that when the same course is taught in a number of sections, the differences in the average achievement of students in various sections taught by different instructors are reflected in the students' ratings of the instructors. This has been by far the most common method of demonstrating the validity of students' ratings of teachers. By using random assignment or ability pretests, student's characteristics can be controlled, and by having a common curriculum, textbook, and examinations, the effect of other factors can be eliminated.

According to some researchers, this approach is the most promising one because "... it minimizes the extent to which the correlation between students' ratings and achievement can be explained by factors other than instructor influence." [15] After a careful review of 43 validity studies using this methodology, several problems with the quality of studies conducted so far were identified.[16] Later, a meta-analysis of these 43 studies found that the mean validity coefficient (correlation between students' ratings and achievement) after attenuation was .47, with a 95% confidence interval extending from .43 to .51.[17] This was taken as indicating moderate to large validity for the tests.

Marsh has criticized this design on several counts. He considers it an "inherently weak design." First, the sample sizes are almost always very small. Secondly, the variance in achievement scores is mostly attributable to student presage variables, for example students' ability before starting the coursework, and it is difficult to find any major effect that can be attributed to teachers. Marsh also believes that "grading-satisfaction hypothesis" may explain the rating-achievement correlation.[18]

Multi-Trait Multi-Method Studies. Marsh prefers a multi-trait multi-method design.[19] In a typical study, college instructors evaluated their own teaching effectiveness with the same 35 items rating form that was used by their students. The student-instructor agreement was quite high: for the undergraduate courses taught by teaching assistants it was .46; for the undergraduate courses taught by faculty the correlation was .41; and for the graduate level courses, the correlation was .39.

Separate factor analyses of the three sets of data led to the same factor structure. The correlation between students and faculty ratings on the same factors were significant (median r=.45), but correlation between their ratings on different factors were low (median r =.02).[20]

Factor-Analytic Studies. Factor analysis is yet another method of showing validity, and has been used by a number of researchers. [21] Marsh and his colleagues have conducted the most extensive work in this area. They have published more than 30 factor-analytic studies, and identified nine factors of the instrument developed by them.[22] This invariance in factorial structure has been taken as an indicator of factorial validity.

Experimental Studies. A number of studies have used experimental designs to study the effect of certain variables. The most notable have been the effects of expressiveness and leniency in marking/grading. In some experimental studies, the general paradigm employed a factorial design in which the expressiveness of the lecturer and the amount of content were systematically varied. The lectures were carefully prepared and delivered by a professional actor. After viewing the videotape, the students evaluated teaching effectiveness with a rating instrument.[23] Ware and Williams, who used a uni-dimensional measure of teacher effectiveness, concluded that the amount of variance in students ratings explained by differences in expressiveness was consistently larger than the amount of variance explained by differences in content.[24] However when a multi-dimensional instrument is used, the effect of teacher expressiveness is largely confined to some factors like Teacher Enthusiasm. [25]

The effect of grading leniency as a factor threatening the validity of teacher rating scale has been hotly debated. It is well known that students' evaluative ratings are positively correlated with expected course grade, but such correlation cannot be taken as indicating validity. Experimental studies that have sought to clarify this relationship are fraught with major weaknesses.[26] Some researchers are of the view that while grading leniency can lead to inflated ratings by the students; this is not a serious matter because it can be statistically corrected.[27]

Scherr and Scherr, who reviewed research on other factors responsible for bias in student's ratings, concluded that only a limited number of such factors actually influence students evaluation. These factors include prior subject interest, workload/difficulty, and class size. According to them, grading practices of the teacher do influence student's evaluation, but are related to perception of fairness in grading.[28]

Utility

Are the measures of teacher effectiveness of any use in the educational institutions? Costin and his co-researchers have stressed the value of ratings for the individual faculty member and for the department and college as a whole. Consequently, they mentioned the following positive aspects of teaching effectiveness ratings (TER): (i) TER could provide a feedback which the instructor might not be able to elicit from students on a face-to-face basis; (ii) TER could provide departmental and college-wide norms against which individual faculty ratings could be judged; (iii) TER could provide a way in which a faculty member could, if he or she so desires, demonstrate his or her teaching effectiveness to those who have expressed an interest in evaluating these parameters for salary increase; (iv) TER could provide information to the department and college on areas of relative strength or weakness in teaching, and suggest directions for the development of new courses or programmes, and provide evaluative information and norms on the various new programmes, which are implemented; (v) Finally, TER could provide the student with a source of information to aid him or her in the selection of courses.[29]

Research has shown that introduction of teacher effectiveness assessment benefits students and teachers both. It has also been shown that the feedback which a teacher gets from students, particularly coupled with a candid discussion with an external consultant, can lead to improved performance as shown in the form of better ratings from the students, and improved learning of the students.[30]

The Need for a New Instrument for IIUM

The International Islamic University Malaysia (IIUM) is a unique institution of learning in many ways. It has teaching staff drawn from all over the world, whose educational experiences and backgrounds are quite varied. It attempts to provide a new kind of education, which is characterized by integration of Islamics with modern human social science. There is an attempt to integrate education with overall personality development along Islamic lines. Since the goals of this University are quite different from others, it was considered important to develop a measure of Teaching Feedback Survey (TFS), which is specially designed to assess the performance of teachers in accordance with these objectives.

In order to provide feedback on the performance of lecturers, the first Teacher Effectiveness Rating (TER) instrument was introduced in the University in 1991. This was subsequently revised several times. The present exercise was initiated in 1996, with qualitative and quantitative analyses of the existing instrument.[31] On the basis of this review, development of a new instrument "International Islamic University Malaysia-Teaching Feedback Survey" (IIUM-TFS) was undertaken.

ITEM DEVELOPMENT PROCESS

The Pre-Pilot Run: Item Generation

We followed both deductive and inductive approaches for item development.[32] The *deductive* approach required that we make a careful review of the literature. Our literature search, taken as a whole, revealed five broad dimensions making up effective teaching: Mastery or Knowledge of the Subject, Preparation and Organization of Lectures, Clarity of Presentation/Communication, Enthusiasm, and Ability to Stimulate Students Thought and Interest. In addition, we introduced another component keeping in view the vision and mission of IIUM. We called this dimension, "Islamic Orientation." Thus, our pool of items was initially based on a total of six major dimensions of teaching effectiveness. Our literature search based on a thorough review of teaching effectiveness measures including the past TER measures of IIUM.

We also employed an *inductive* approach for item generation–an approach called critical incident-like technique. This required that we gather behavioral descriptions of a highly effective/ineffective teacher. We collected these descriptions by asking our students of Personnel Psychology classes (PSYC 4510, taught by two members of this research team) to conduct an empirical study for a portrayal of an effective/ineffective teacher. We also conducted several rounds of focused-group interviews with undergraduate and graduate students. In addition, we interviewed several faculty colleagues.

By employing the above two methods—deductive and inductive—we collected a pool of around 80 items.

Item Review and Content Validation

The items were reviewed and judged at this stage for content validity. The three researchers and a research assistant (a psychology graduate) served as judges to evaluate each item to be identified in eight dimensions: Communication, Islamic Orientation, Knowledge, Feedback, Organization, Motivation, Intellectual Stimulation, and Time Management. This exercise was also directed at recommending modifications—change or drop items—and identifying unclear items. It resulted in 42 items measuring seven dimensions.

Next, we conducted a pre-pilot run of the 42 items on about 500 students drawn largely from the Faculty of Islamics & Human Sciences (known as Kulliyyah of Islamic Revealed Knowledge and Human Sciences). Students were asked not only to fill-in the questionnaire but also to comment on the items for their suitability—clarity/ambiguity, redundancy, etc. In addition, we sent a draft copy of the TER to over 400 IIUM faculty for comments and suggestions. We received a number of helpful comments and suggestions from some of our colleagues.

The modified TER items were presented in the Faculty's DCM (Dean's Committee Meeting). Again we received some very insightful comments, some of which were incorporated in our instrument.

We named the revised measure of teaching effectiveness, "Teaching Feedback Survey" (TFS). The measure had the following composition of items: (a) 43 single-statement items rated on a 5-point scale; (b) Four interrogative-statement items with binary alternatives; (c) Two interrogative-statement items rated on a 7-point scale; and (d) Seven biographical items, with different anchor points.[33]

The Pilot Run

The pilot study was conducted in Semester III, 1996/97. Six hundred twenty-nine undergraduate and graduate students--representing three faculties—Islamics & Human Sciences, Laws, and Economics & Management—were selected using a stratified random sampling procedure. They responded to the survey items. Our survey included 23 lecturers to be rated by 629 students for teaching effectiveness.

Item Review

We calculated descriptive statistics (*M* and *SD*) for each item. The analvsis revealed that of the 43, 15 items did not reach unity (i.e., *SD*

=1.00), suggesting that these items were not significantly discriminating between more effective and less effective teachers. We also examined the 43 X 43 inter-correlation matrix. At this stage, we had just 28 items.

Factor Analysis

The responses to 28 items were then subjected to a varimax rotated principal components analysis, as a test of the construct validity. The items to be retained were selected on the basis of the following criteria: The solution was constrained using the criterion of eigenvalue greater than 1.00, and meeting the criterion of factor loading generally not less than .35 on the defining component and no cross-loading greater than .25. The analysis confined to three factors meeting the above criteria that explained a total of 38.1% of the variance.

The first factor (employing 11 items) appeared to be the strongest one that we named, *Organization and Preparation*. The second factor (consisting of just 3 items) was the neatest one that we labeled, *Islamic Orientation*. The third factor was composed of 4 items that was termed, *Feedback*. The three factors were only moderately correlated, thereby suggesting a great deal of independence (non-overlapping variances) among the sub-scales (r^2 = .09).

Reliability

To examine the internal consistency of the sub-scales, Cronbach's coefficients alpha were computed. The three sub-scales were found to be fairly reliable, with reliability coefficients ranging from .74 to .90.

Validity

We had included two criterion measure items in our TFS scale--one relating to the learning aspect of the course and the other relating to the overall teaching effectiveness. We correlated these two items with the three derived sub-scales, and found that the three factors were positively and significantly correlated ($p <$ .001) with the two criterion items—r ranged from .18 to .49 for the learning criterion item and from .21 to .51 for the overall effectiveness item. The first factor (i.e., *Organization and Preparation*) was strongly correlated with both criterion items.

A final testing of the TFS was conducted in Semester I, 1997/98. This round of testing had a larger sample. It should be noted that although our pilot study included three major faculties, our bulk of data came from Islamics & Human Sciences. The final testing results based on a much larger sample are reported below.

THE FINAL TESTING

Scale Development Process: Sample

Following a stratified random sampling procedure, 979 undergraduate and graduate students rated their forty-one lecturers (see Table 1). These lecturers represented the four major faculties of IIUM—Islamics & Human Sciences (66%), Economics & Management (15%), Laws (15%), and Engineering (4%). Out of 41 lecturers, 38 were males and 3 females. The majority of them (about 59%) were Assistant Professors, whereas 20% were Associate Professors and 21% were Professors. About two-thirds of the sampled lecturers were international (from different nationalities). An approximately equal proportion of the students was chosen from different levels of their program. Table 2 contains the details of the student respondents. As is evident, the majority (about 90%) of the students were undergraduates. Around 60% of them were female. A sizeable number of students (over 60%) represented the faculty of Islamics & Human Sciences. A bulk of them (over 70%) had their CGPA between 2.0 and 3.0.

Table 1
Faculty-wise Break-up of Lecturers and Students

Faculty	*Lecturers*	*Students*
Islamic & Human Sciences	27	643
	6	154
Economics & Management Sciences	6	161
Engineering	2	21
Total	41	979

Procedure

Two part-time female research assistants (majoring in IRK with a minor in psychology) were especially recruited for data collection. Under the supervision of a senior research assistant, they approached the sampled 41 lecturers during their office hours, and handed over to them a request letter signed by the researchers. The letter contained a clear objective of the survey and a formal request to permit the research assistants to administer the TFS in their respective classes. After obtaining permission, the assistants administered the TFS to the students. Before administering, they assured the students of complete anonymity of individual responses. Then they supplied a 2B pencil (a special requirement for the computer form) to fill-in the TFS. On an average, the student took about 15 minutes in filling out the TFS.

Measures

Teaching Feedback Survey (TFS). The TFS was the revised 43-item scale from the pilot study. The students were asked to indicate on a 5-point scale (1 = *never*; 5 = *always*) the *frequency* with which each item was applicable to the lecturer. Also, they were provided with an additional response category--"Not Applicable"--to indicate if an item was not descriptive of the lecturer.

In addition, 4 "yes"-"no"-type items were used. These items were meant only for feedback purposes to the lecturers. Yet another two questions were asked for validation purposes—one relating to the amount of learning from the course and one relating to the overall effectiveness of the lecturer.

Social Desirability. In order to examine if the items were free from social desirability effect, we used the well-known 17-item Crowne & Marlow scale. The subjects were asked to indicate whether the statements (concerning personal attitudes) were "true" or "false" for them.[34]

In addition to the above measures, we used several single-statement items to assess the respondent's personal-demographic characteristics such as CGPA, program, and level.[35]

Table 2

Details of Sample: Frequency Count and Percentage

	Frequency	*Percentage*
Gender		
Male	378	38.6
Female	585	59.8
Program		
Undergraduate	913	93.2
Postgraduate	47	4.8
Year Level		
First	148	15.1
Second	237	24.2
Third	286	29.2
Fourth	277	28.3
Fifth	9	0.9
CGPA		
Not yet available	37	3.8
Less than 2.0	13	1.3
2.0 and less than 3.0	713	72.8
3.0 and above	191	19.5
Faculty		
Islamic~~s~~ & Human Sciences	622	63.5
Laws	170	17.4
Economics and Management	131	13.4
Engineering	31	3.2

Table 3
TFS Items and their Correlation with Social Desirability

	TFS Items in brief	*M*	*SD*	
01.	Is available during consultation hours	4.18	1.30	0.02
02.	Welcomes students' comments	4.46	0.87	0.09
03.	Uses a variety of teaching methods.	3.55	1.15	0.04
04.	Links theory and applications.	4.18	1.00	0.03
05.	Encourages students' opinions.	4.26	0.92	0.04
06.	Has a good knowledge of the subject.	4.70	0.66	0.01
07.	Teaches with a good sense of humor.	3.80	1.14	0.01
08.	Shows interest in feedback.	4.02	1.19	0.05
09.	Relates topics to Islamic teachings.	4.02	1.26	0.05
10.	Is systematic in presentation.	4.10	1.12	0.02
	Has clear pronunciation.	4.13	0.97	0.03
12.	Finishes class on time.	4.49	0.91	-0.04
13.	Encourages students' participation.	4.27	0.95	0.03
	Starts class on time.	4.30	0.89	0.02
15.	Makes the students work hard.	4.14	1.05	0.02
16.	Is prompt in giving feedback on exams.	3.84	1.36	0.00
17.	Misses classes without make-up.	1.93	1.65	-0.01
18.	Uses class time effectively.	4.42	0.89	0.05
19.	Encourages critical thinking.	4.10	1.08	0.03
20.	Has mastery over the subject.	4.65	1.00	-0.02
21.	Uses non-verbal communication.	3.87	1.19	-0.05
22.	Encourages additional learning.	3.94	1.05	0.02
23.	Is enthusiastic about teaching.	4.32	1.13	0.04
24.	Encourages students to ask questions.	4.25	0.98	-0.03
25.	Explains objectives at the beginning	3.95	1.14	-0.04
26.	Is clear in presentation.	4.13	0.90	0.04
27.	Uses clear, understandable language.	4.31	0.97	0.03
28.	Promotes Islamic values.	4.02	1.22	0.04
29.	Praises the students for performance.	3.91	1.25	-0.01
30.	Provides attention to the weak.	3.27	1.44	0.03
31.	Encourages students to come prepared.	3.88	1.22	-0.01
32.	Is fair and just in grading.	4.02	1.23	0.03
33.	Keep the students attentive.	4.10	1.09	0.06
34.	Returns assignments with comments.	3.32	2.07	0.04
35.	Generates a sense of enthusiasm.	3.94	1.36	-0.00
36.	Discusses test results in the class.	3.47	1.69	0.02
37.	Follows the course outline.	4.57	1.07	-0.01
38.	Misses classes without informing.	1.65	1.35	-0.02
39.	His/her lectures are well organized.	4.22	1.09	0.03
40.	Uses examples that are Islamic.	3.94	1.39	0.06
41.	Comes prepared to the class.	4.53	0.83	0.09
42.	Has a proper pace of teaching.	4.11	0.92	0.05
43.	Acts as a model teacher.	4.16	0.98	0.02

SCALE TESTING PROCESS

Item Review

We used three criteria for the selection of TFS items at this stage. First, we calculated descriptive statistics (*M* and *SD*) on each of the 43 items (see Table 3). The analysis revealed that item means were generally around the median of the anchor points, and they had a great deal of dispersion. Secondly, we examined the intercorrelations among the items. The correlation matrix (not reported here) revealed that the items were meaningfully correlated with one another. Thirdly, we calculated correlation for each of the 43 items with the social desirability score (see the last column of Table 3). Almost all correlations were near zero, thereby showing the TFS responses free from social desirability effect.

Factor Analysis

The TFS measure was next subjected to a varimax rotated principal components analysis, as a partial test of the construct validity. The criteria for the selection of items were the same as we had set for the pilot study. The solution was constrained using the criterion of eigenvalue greater than 1.00, and meeting the criterion of factor loading generally not less than .35 on the defining component and no cross-loading greater than .25. However, if an item had a very high loading on the defining component, the criterion of cross loading greater than .25 was a bit relaxed. Conversely, if an item had a little less than the required loading but had very low cross loading on other factors was retained in the factor. We made this relaxation only when the items were forming a meaningful configuration. Table 4 reports the factor loadings obtained along with eigenvalues and percentage of the variance explained. The analysis confined to four neat and interpretable factors that explained a total of 48.5% of the variance. The four factors are operationally defined below.

The first factor emerged as the strongest one. It included 14 items, explaining a total of about 37% of the variance in the matrix. It was composed of such teaching effectiveness areas as knowledge of the subject, presentation, lecture organization, and pace of teaching. We named this factor "Delivery of Information."

The second strongest factor had 8 items that involved content areas like encouraging the students to express their views, and motivating the students for critical thinking. We labeled this factor "Meaningful Interaction."

The third factor included 5 items pertaining to the teaching areas like returning assignments/exams promptly with helpful comments, and treating the students fairly in grading. This extracted factor was named "Feedback and Fair Treatment."

The last factor appeared to be the neatest one. It involved just 3 items, and included such content areas as promoting Islamic values, relating the concept with Islamic teachings, and using Islamically relevant examples. We called this factor "Islamic Orientation."

Assessment of Scale Independence

Although the four dimensions of teaching effectiveness are distinct, they are implicitly oriented toward teaching effectiveness, which leads us to expect some interdependence among them. Table 5 provides the descriptive statistics and intercorelations among the four factors. As can be seen, there was just one correlation that was above the .50 level--that is, between delivery of information and meaningful interaction. Overall, however, the teaching effectiveness dimensions were only moderately interrelated (average r^2=.21), thereby showing a great deal of independence among the factors.

Reliability

Cronbach's coefficients alpha were computed to examine the internal consistency reliability of the teaching effectiveness measures (see Table 5). The TFS dimensions were found to be highly reliable—coefficients alpha ranging between .81 and .91.

Validity

Before examining the validity of the TFS measure, we sought to examine if the respondents operated on a social desirability factor. The analysis indicated that the four factors were completely unrelated to social desirability (see the last column of Table 6). This fact may be taken as evidence that the TFS measures are free from social desirability effect.

Table 4

Factor Analysis Results of TFS Measures

Items	Factor 1	Factor 2	Factor 3	Factor 4
V6	44	27	07	05
V10	56	29	19	06
V11	62	19	06	03
V12	31	05	17	14
V18	37	20	15	11
V20	40	22	08	02
V21	34	24	15	17
V26	69	26	18	16
V27	67	18	12	13
V37	48	08	15	15
V39	64	14	20	15
V41	53	21	12	13
V42	52	15	23	21
V43	49	28	32	29
V2	21	54	18	21
V3	24	49	25	14
V5	21	75	09	17
V13	16	74	13	08
V15	25	42	28	-07
V19	18	51	15	15
V22	27	34	19	29
V24	24	60	15	13
V16	17	11	60	05
V30	18	32	60	09
V32	28	10	55	14
V34	06	11	71	08
V36	12	14	73	-00
V9	21	24	06	81
V28	19	20	11	81
V40	23	13	13	83
Eigenvalue	14.5	2.27	2.22	1.86
% Variance	33.7	5.3	5.2	4.30

Note. N = 979; Decimal points in factor loadings are omitted; N = 803; Factor 1=Delivery of Information; Factor 2 = Meaningful Interaction; Factor 3=Feedback and Fair Treatment; Factor 4=Islamic Orientation; The underlined loading indicates inclusion of the item in that factor; for description of items, see Table 3.

It is also evident in Table 6 that the TFS measures are positively and significantly correlated with the validity items. Taken as a whole, the first two factors—delivery of information and meaningful interaction--correlate more strongly with both validity factors, amount of learning in the course and overall teaching effectiveness. Yet, the other two factors, feedback and fair treatment and Islamic orientation, are also positively and significantly correlated with both validity factors. These information suggest that the TFS measures do not only have high reliability but they also have high validity coefficients.

Some Additional Analyses

We further hypothesized that any measure that attempts to assess teaching effectiveness must distinguish among the different lecturers on the four TFS factors.

Table 5

Descriptive Statistics, Reliabilities, and Intercorrelations of Teaching Feedback Survey (TFS) Dimensions

Factor	Factor 1	Factor 2	Factor 3	Factor 4
Factor 1	90			
Factor 2	67*	85		
Factor 3	46*	47*	81	
Factor 4	47*	45*	25*	91
M	4.28	4.09	3.45	3.94
SD	0.56	0.64	1.04	1.10

Note. N = 979; Decimal points in correlation matrix and alpha are omitted; Factor 1= Delivery of Information; Factor 2= Meaningful Interaction; Factor 3= Feedback and Fair Treatment; Factor 4= Islamic Orientation; Diagonal entries indicate coefficients alpha; *p < .01.

Table 6

Correlations of TFS Measures with Validity Measures and Social Desirability Factor

TFS Factors	*Amount of Learning*	*Overall Effectiveness*	*Social Desirability*
Delivery of Information	.56*	.61*	.02
Meaningful Interaction	.46*	.56*	.03
Feedback and Fair Treatment	.34*	.39*	.02
Islamic Orientation	.30*	.36*	.05

$N = 979$; $^{*}p < .01$.

To present this evidence, we made additional analysis to examine if there is any significant difference among the IIUM lecturers on the four teaching effectiveness dimensions. For this purpose, we computed a significance of difference on TFS scores across the sampled 41 lecturers. Table 7 provides a summary of one-way ANOVA.[36]

Table 7

Analysis of Variance of TFS Measures

TFS Factor	*df**	*F*
Delivery of Information	40,916	7.33**
Meaningful Interaction	40,924	6.48**
Feedback and Fair Treatment	40,764	7.59**
Islamic Orientation	40,895	17.62**

Note. *df vary because of missing cases in the cell; $^{**}p < .001$.

An inspection of Table 7 clearly indicates that the TFS has the ability to discriminate among the relatively more effective and relatively less effective lecturers. That is, lecturers can be ranked ordered in terms of their effectiveness. Also, it is very clear that a lecturer may be good at delivery of information, but she or he may not be equally good at Islamic orientation. Similarly, a lecturer may be good at interacting with students, but she or he may not provide prompt feedback to the students or he or she may be perceived as unjust teacher.

We had also recorded a few other points that were considered important concerning teaching effectiveness. These are number of courses taken with the lecturer under evaluation, level of cumulative grade-point-average, and the year level. Table 8 contains Pearson correlation coefficients of these factors with the four factors of teaching effectiveness.

Table 8

Correlations Between TFS Factors and Other Factors

Factor	*# Courses Taken*	*GPA*	*Year Level*
Delivery of Information	-.03	02	06*
Meaningful Interaction	04	-.00	.09*
Feedback and Fair Treatment	06*	-.01	13**
Islamic Orientation	-.04	-.00	-.06*

Note. $N = 979$; $^{*}p < .05$; $^{**}p < .01$

One common assumption people hold is that the greater the number of courses the student takes with a particular lecturer, the better the rating he or she would assign to the lecturer. Contrary to this assumption, our results suggest that the number of courses taken is independent of students' effectiveness ratings of their lecturer. Although one correlation—that is with feedback and fair treatment dimension—is marginally significant, the magnitude is too low (r =.06). Another common assumption people hold that the good cumulative grade-point-average holders assign good rating to the lecturers. This view is also at variance with our data (see Table 8). However, the maturity level of students seems to be favorable to the lecturer on three factors—delivery of information, meaningful interaction, and feedback and fair treatment--but unfavorable on the Islamic orientation dimension of teaching effectiveness.

CONCLUSION

The main objective of this exercise was to develop a psychometrically sound, multi-dimensional scale of independent factors to assess teaching effectiveness. The final measure (called IIUM-TFS) that emerged was a 30-item scale, with four relatively independent factors. On the basis of the various analyses performed, it appears that the newly developed measure has high reliability coefficients. The scale is free from social desirability effect, has built-in content validity, and has reasonable amount of construct validity. In addition, it has fairly good criterion-related validity. It has also been found to discriminate among lecturers.

Since it is not possible to address every issue in a single piece of research, future research should focus on comparing this scale with other scales to measure teaching effectiveness to further examine convergent and discriminant validity of the TFS scale. That is, it certainly needs further improvement. Despite the need for additional research to validate the newly developed TFS, future researchers and university authorities may be advised to use it for feedback purposes (to the lecturers) as opposed to any ad hoc measures.

Notes

[1] See, L.A. Braskamp, D. Caulley, & Frank Costin, "Students Ratings and Instructor Self-ratings and their Relationship to Students Achievement," *American Educational Research Journal* 16 (1979): 295-306.

[2] Braskamp, Caulley, & Costin, "Students Ratings"; Frank Costin, W.T. Greenenough, & R.J. Menges, "Students Ratings of College Teaching: Reliability, Validity and Usefulness," *Review of Educational Research* 41 (1971): 511-535; K.A. Feldman, "Consistency and Variability Among College Students in Rating their Teachers and Courses: A Review and Analysis." *Research in Higher Education* 6 (1977): 223-274; P.W. Frey, "A Two-Dimensional Analysis of Students Ratings of Instruction." *Research in Higher Education 9* (1978): 69-91.

[3] H.W. Marsh, "Validity of Students Evaluation of College Teaching: A Multitrait-multimethod Study," *Journal of Educational Psychology* 2 (1982):

263-279; H.W. Marsh, "SEEQ: A Reliable, Valid, and Useful Instrument for Collecting Students' Evaluations of University Teaching," *British Journal of Educational Psychology* 52 (1982): 77-95; H.W. Marsh & L.A. Roche, "Making Students Evaluation of Teaching Effectiveness Effective: The Critical Issues of Validity, Bias, and Utility," *American Psychologist* 52 (1997): 1187-1197.

[4] Braskamp, Caulley, & Costin, "Students Ratings."

[5] Costin, Greenenough, & Menges, "Students Ratings of College Teaching."

[6] K.A. Feldman, "The Superior College Teacher from the Students' View," *Research in Higher Education* 21 (1976): 45-116.

[7] Frey, "A Two-dimensional Analysis of Students Ratings."

[8] Braskamp, Caulley, & Costin, "Students Ratings";

[9] H.W. Marsh, "Multidimensional Students Evaluations of Teaching Effectiveness: A Test of Alternative Higher-order Structures," *Journal of Educational Psychology* 83 (1991): 285-296.

[10] See, for example, Costin, Greenenough & Menges, "Students Ratings."

[11] Ibid.

[12] H.W. Marsh & M. Bailey, "Multidimensional Students' Evaluation of Teaching Effectiveness," *Journal of Higher Education 64* (1993): 1-18.

[13] A.G. Greenwald, "Validity Concerns and Usefulness of Students Ratings of Instruction," *American Psychologist 52* (1997): 1182-1197.

[14] H.W. Marsh, & D. Hocevar, "The Factorial Invariance of Student Evaluations of College Teaching," *American Educational Research Journal* 21 (1984): 341-366.

[15] See, or example, S. d'Apollonia, & P.C. Abrami, "Navigating Students Ratings," *American Psychologist* 52 (1997): 1201.

[16] P.C. Abrami, S. d'Apollonia, & P. A. Cohen "The Validity of Students' Ratings of Instruction: What We Know and What We Do Not" *Journal of Educational Psychology* 82 (1990): 217-231.

[17] d'Apollonia, & Abrami, "Navigating Students Ratings."

[18] H.W. Marsh, "Students Evaluations of University Teaching; Dimensionality, Reliability, Validity, Potential Biases, and Utility," *Journal of Educational Psychology* 76 (1984): 707-754.

[19] D.T. Campbell & D.W. Fiske, "Convergent and Discriminant Validation of Multi-trait Multi-method Matrix," *Psychological Bulletin* 56 (1959): 81-105.

[20] For instance, Marsh, "Validity of Students Evaluation of College Teaching."

[21] Frank Costin, "Measuring Lecturing Behavior of College Instructors," *Professional Psychology* 5 (1974): 106-108; P.W. Frey, D.W. Leonard, & W.W. Beatty, "Student Ratings of Instruction: Validation Research," *American Educational Research Journal* 12 (1975): 327-336;

[22] See H.W. Marsh, Kit-Tai Hau, Choi-Man Chung, & T. L. Siu, "Students Evaluation of University Teaching: Chinese Version of the Students'

Evaluation of Educational Quality Instrument" *Journal of Educational Psychology* 89 (1997):568-572. Also see Marsh "Multidimensional Students Evaluations," and Marsh & Hocevar. "The Factorial Invariance of Student Evaluations."

23. H.W. Marsh, & J.E. Ware, "Effects of Expressiveness, Content Coverage, and Incentive on Multi-dimensional Student Rating Scales: New Interpretations of Dr. Fox Effect," *Journal of Educational Psychology* 74 (1982): 126-134.

24. J.E. Ware, & R.G. Williams, "A Reanalysis of the Doctor Fox Experiments," *Instructional Evaluation* 4 (1980): 15-18.

25. Marsh & Ware, " Effects of Expressiveness."

26. H.W. Marsh, & L.A. Roche "The Use of Student Evaluations for Teaching in Different Settings: The Applicability Paradigm," *Australian Journal of Education* 36 (1992):278-300.

27. A.G. Greenwald, & G.M. Gillmore, "Grading Leniency is a Removable Contaminant of Student Ratings," *American Psychologist* 52 (1997): 1209-1217.

28. F.C. Scherr & S.S. Scherr, "Bias in Student Evaluation of Teachers," *Journal of Education for Business* 65 (1990): 356-358.

29. Costin, Greenenough, & Menges, "Students Ratings."

30. J.U. Overall & H.W. Marsh, "Midterm Feedback from Students: Its Relationship to Instructional Improvement and Students Cognitive and Affective Outcomes," *Journal of Educational Psychology* 71 (1979): 856-865.

31. For details see, M.A. Ansari, M. Achoui, & Z.A. Ansari, *Development of a Measure of Teaching Feedback Survey (TFS),* Project Report Submitted to the Research Center, International Islamic University Malaysia (Kuala Lumpur, 1998).

32. T. Hinkin, "Guidelines for Scale Development for Study of Behavior in Organizations," *Proceedings of the Annual Meeting of the Academy of Management*, Las Vegas, Nevada, August, 1992; and D.P. Schwab, "Construct Validity in Organizational Behavior," in B.M. Staw & L.L. Cummings (Eds.), *Research in Organizational Behavior* vol. 2 (Greenwich, CT: JAI Press, 1980), 3-43.

33. For details see, Ansari et al, *Development of a Measure.*

34. D.P. Crowne, & D. Marlow, "A New Scale for Social Desirability Independent of Psychopathology," *Journal of Counseling Psychology* 24 (1960): 349-354.

35. For further details see, Ansari et al, *Development of a Measure.*

36. Ibid.