EVALUATION OF THE PSYCHOMETRIC PROPERTIES OF A TEACHING FEEDBACK SURVEY; FIRST AND SECOND-ORDER CONFIRMATORY FACTOR ANALYSIS

Dr. Mikail Ibrahim
Faculty of Major Language Studies
Universiti Sains Islam Malaysia

This study examines the structural validity and dimensionality of scores from Mahfooz Ansari and Mustafa Achoui Ansari's Teaching Feedback Survey (2000) using both exploratory and confirmatory factor analyses (EFA and CFA). A total of 1504, 3rd, 4th year and postgraduate students from four Malaysian institutions of higher learning voluntarily participated in the study. An exploratory factor analysis with Maximum Likelihood (ML) was used with a randomly selected half of the sample. Four teaching effectiveness factors emerged. Although the result of this study replicated the findings reported by Mahfooz Ansari and Mustafa Achoui Ansari in terms of the number of the extracted factors, the total variance explained and factor loadings were higher in the current study than reported by Mahfooz Ansari and Mustafa Achoui Ansari. Furthermore, the use of EFA with ML instead of PCA used by Mahfooz Ansari et al allowed the researcher to isolate unique and error variances from the analysis unlike the technique used by Ansari et al. Additionally, both orthogonal and direct oblimin rotation techniques were examined.

However, the EFA suggested that direct oblimin psychometrically fit the data compared to its orthogonal counterparts. In the orthogonal rotation technique, three items were found to be cross-loaded (factorial complexity), another two item loadings were below the predetermined cut-off ($\geq$ .40), while direct oblimin did not witness any cross loading but two items were also below .40; the threshold for accepting the item as a part of the factor. A Confirmatory Factor Analysis with a maximum likelihood was run with the remaining half of the sample. The first-order CFA was firstly performed to test the structural validity, reliability and dimensionality of the scale. However, the higher intercorrelation among the factors suggested second-order. The study found that compared with first-order, the second-order fit the data perfectly well. The fit indices based on the validation sample collectively indicate a very good fit. RMSEA, which indicates the discrepancy between the proposed and the perfect model, suggested that error is near to zero which means that the model perfectly fit the data. In addition, internal consistency estimates provided further evidence of the reliability of factor scores.

Many higher institutions all over the world adopt a system of students' evaluation of teaching effectiveness especially in Western countries (Marsh, 1984; 1987; Greenwald & Gillmore, 1997). As quoted by Marsh, Remmers (1927) initiated the first systematic research programme of students' evaluation of teaching effectiveness. In 1927, he published the first standardized, systematic, and multi-trait scale to measure teaching instruction. Remmers tried to standardize the scale through examination of its reliability, validity, norms, halo effects, biased nature of the scale and the relationship between the expected grades and the students' actual ratings.

Many scales have been developed after Remmers' scale from different cultural and background points of view due to the complexity of the classroom and teaching and the learning processes, such as Marsh, 1987; Fernandez and Mateo, 1992; Ramsden, Martin and Bowden, 1989; Remedios, Lieberman and Benton, 2000; Remedios and Lieberman, 2008; Richardson, 1994; Mahfooz Ansari and Mustafa Achoui Ansari, 2000 and many more. These scales, whether named Students' Evaluation of Educational Quality, SEEQ (Marsh, 1984; 1987), Course Experience Questionnaire (Richardson, 1994), Endeavour Instructional Rating and Students' Rating of Instruction (d'Apollonia & Abrami, 1997); Teaching Feedback Survey, (Mahfooz Ansari & Mustafa Achoui Ansari, 2000) were constructed to evaluate students' experiences of teaching within the classrooms.

The purposes of these scales, according to Marsh (1984; 1987), are varied and among which are 1) diagnostic feedback to a Faculty on the effectiveness of their teaching, 2) to measure teaching effectiveness and to use it in tenure and /promotion decisions, 3) information for students to use in the selection of courses and instructors 4) a measure of the quality of the course, to be used in course improvement and curriculum development, and 5) an outcome on a process description for research and teaching.

The researcher (Marsh) believes that students' evaluation of teaching should be subject-matter dependent, due to the fact that the nature of any subject can determine students' evaluation attitude and their critical position. In fact, most of these scales are either teacher-based or course-based scales. As Marsh (1987) pointed out, focus on a subject matter or a specific teacher would yield a considerable amount of information that is "useful for feedback to faculty, useful for personnel decisions, useful to students in selection of courses, and useful for the study of teaching" (p.369).

However, from the perspective of institutions of higher education, it might be useful to focus on the entire degree programme for the assessment of quality improvement and maintenance (Richardson, 1994). By focusing on degree programmes, tremendous opportunity would be given to benchmark the effectiveness of teaching experience across different domains.

Furthermore, the content of the scales varied drastically from one scale to another, which subsequently affected the conclusion drafted from them and the trustworthiness of the scales. One of the major problems of the scales is that the psychometric properties of the scales were not properly tested. According to Marsh,, "part of the problem lies in the fragmentary approach to the design of both students-evaluation instruments and the research based upon them" (1987, p.260).

In relation to the dimensionality of the scales (see details in the next section), the items used to evaluate teaching effectiveness yielded different dimensions depending upon the sample characteristics, the initial item pool and the method of analysis used. While a group of teaching effectiveness scales concentrated on some of teachers' characteristics such as empathy, facilitation, personal attention, teacher support, students involvement, negative effect, enthusiasm and rapport and interaction as being more conducive to teaching effectiveness (Marsh, 1987; Ramsden, 1991), another group of scales focused on academic competence, communication competence, professional maturity, presentation, and organization and clarity as indicative of teaching effectiveness (Harrison, Douglas, & Burdsal, 2004). Meanwhile, Mahfooz Ansari and Mustafa Achoui Ansari (2000), in addition to delivery of information, meaningful interaction, feedback and fair treatment, due to the different cultural aspect also included Islamic orientation. This suggested that teaching effectiveness is a multi-trait and multi-dimensional phenomenon in which many characteristics of the instructor are involved.

Thus, the researcher attempts to examine the psychometric properties of Teaching Feedback Surveys?? among the selected higher institutions in Malaysia and how efficient the scales are in measuring teaching effectiveness. Therefore, the researcher attempted to psychometrically evaluate the structure validity and dimensionality of Mahfooz Ansari & Mustafa Achoui Ansari's (2000) scale.

DIMENSIONALITY OF TEACHING EFFECTIVENESS SCALES

Among the most controversial issues in the teaching effectiveness scale or course experience questionnaire is the issue of dimensionality. Many researchers of teaching effectiveness scales consider the quality of teaching as multi-dimensional (March, 1987, 1984, 1991; Ramsden, 1991; Mahfooz Ansari & Mustafa Achoui Ansari, 2000). Marsh (1984, 1987), the chief advocate of multi-dimensionality of teaching-effectiveness scales contended that a teaching-effectiveness scale is multi-dimensional in nature due to the fact that a single measure cannot adequately summarize the quality of teaching performance and the teacher as a professional engages in many activities within the classroom to prove his/her ability in handling the class efficiently.

Marsh demonstrated this multi-dimensional aspect of the scale by employing Principal Component Analysis (PCA). The Principal Component Analysis was employed to test the scale- construct validity and to extract underlying dimensions of the pool items**.** Eventually, the analysis of PCA resulted in nine distinctive but correlated dimensions: Learning/Value, Enthusiasm, Organization, Group Interaction, Individual Rapport, Breadth of Coverage, Exams/Grades, Assignment and Workload (Marsh, 1984, 1987). Marsh advocated the use of multiple indicators and ratings because students' rating is multi-dimensional in nature based on the different characteristics of the instructor.

According to Marsh, some criteria were used to prove the multi-dimensionality of the teaching effectiveness scale; these are, (a) teaching effectiveness is multi-faceted; (b) there is no single criterion of effective teaching; (c) different dimensions or factors of students' ratings will correlate more highly with different indicators of effective teaching. Furthermore, Mahfooz Ansari and Mustafa Achoui Ansari (2000), in their Teaching Feedback Survey, maintained the perception of multi-dimensionality of the quality of teaching experience.

Consistent with Marsh, Mahfooz Ansari and Mustafa Achoui Ansari, (2000) employed Principal Component Analysis through varimax and eventually extracted four meaningful factors; delivery of information, meaningful interaction, feedback and fair treatment and Islamic orientation.

On the other hand, Abrahami (1985, 1988, 1989) and Abrahami and d'Apollnia (1990, 1991) rejected the multi-dimensional concept of teaching effectiveness. Abrahami (1989) then proposed "for summative purposes, I favor the use of global rating items… or a carefully weighted average of rating factors in lieu of separate factor scores" (p.222). His major concerns that made him unequivocally discard the Marsh notion, according to him are (a) that he did not believe that there is sufficient evidence to establish the dimensions of effective teaching or whether and how they are inter-related, (b) the serious concerns about the content validity of specific items and some of the dimensions they comprise, (c) qualitative review of the multi-section validity studies suggested that many rating dimensions have lower correlations with students' learning (d) the generalizability of the specific rating factors is better known than global rating. These points constitute the evidence that Abrahami used to oppose Marsh's notion of multi-dimensionality.

According to d'Apollonia and Abrami (1997), although teaching effectiveness might generally be multi-dimensional, the large proportion of variance indicated is more towards global instructional skills rather than specific skill components which provided evidence for one global component (factor) rather than several specific factors.

VALIDITY OF THE TEACHING EFFECTIVENESS SCALES

As the dimensionality of teaching effectiveness scales entailed a lot of arguments and controversies, the validity of the scales also drew attention, heated discussion and was intensively researched decades ago. As researchers discussed the issues of validity, many disagreements and controversies have emerged. Many researchers concluded that different opinions on what constitutes teaching effectiveness and its definition severely affected its validity (March, 1984, Abrami & d'Apollonia, 1990). Although, generally, many researchers concluded that student rating is reasonably valid, useful for teaching improvement and relatively unaffected by external factors such as grading leniency, class size, charisma and workload (Marsh, 1984, 1987, & Marsh & Roche, 2000; d'Apollonia & Abrami, 1997), there is some substantial body of knowledge which emphatically asserted that the scales are invalid because they are biased towards extraneous variables that have no relationship with teaching quality (Damron, 1996; Greenwald & Gillmore, 1997; Emery, Kramer & Tian, 2003).

In other words, the correlations between the teaching effectiveness scales, on the one hand, and grading, workload and class size, on the other hand, suggest that the ratings might be biased for teachers who reward or they might penalize an instructor, not because of teaching characteristics, but because of the way he or she grades the students – either leniently or otherwise, or due to overloading them with work. Damron (1996) concluded that it is likely that the factors contributing most to students' instructional ratings are unrelated to an instructor's ability to promote students' learning. Moreover, the students' responses to the instructional rating varied across the subject domains (arts and science) and even students' academic levels (freshmen, sophomores and advanced students) (Emery, Kramer & Tian, 2003). This variability suggested that perhaps scales are not measuring what they purport to measure. Hence, it can be concluded that "validation studies that do not properly control for biasing (e.g. student characteristics, instructor characteristics, class characteristics) yield internally invalid and interpretable estimates of rating validity" (Emery, Kramer & Tian, 2003, p.39). As earlier emphasized, the construct validity of the scales was mainly tested through the employment of sophisticated statistical methods such as factor and confirmatory factor analysis, path analysis, structural equation modelling, and multi-trait, multi-method analysis (Marsh, 1984, 1987; Marsh & Roche, 1993, 2000; Remedios & Lieberman, 2008). However, different factors were extracted from the scales due to different types of the initial pool items (manifests) and the definition of the effectiveness of teaching, which give critics a chance to bombard the scales as invalid and vague.

Methodology

Participants

The subjects of this study were selected from four Malaysian institutions of higher learning, namely the Islamic Science University of Malaysia (USIM), the University of Malaya (UM), University Putra Malaysia (UPM) and the International Islamic University Malaysia (IIUM) which are located in the Selangor, Kuala Lumpur, and Negeri Sembilan area.

The 30 Teaching Feedback Survey items with demographic variables were distributed to 1504, 3rd, 4th year and postgraduate students from a randomly selected sample taken from the above higher institutions who voluntary answered the questionnaire. The data was equally and randomly divided into two. The first half was used to perform EFA and the second half was used for Confirmatory Factor Analysis. The sample comprised 301 (40.0%) males and 451 (60%) females. The reason for choosing students from these institutions of higher learning as the subjects of this study was to examine the appropriateness of the scales used to measure teaching effectiveness and their suitability to cover all the teaching aspects according to the Malaysian concept of teaching.

Instrument

The Teaching Feedback Survey scale, developed by Mahfooz Ansari and Mustafa Achoui Ansari (2000), was used in this study. The scale consisted of 30 items after Principal Component Analysis was used to reduce the initial pools items, absorb the underlying dimensions and test the construct validity of the scale. Principal Component Analysis, according to the authors, yielded four meaningful factors; Delivery of information, Meaningful Interaction, Feedback and Fair Treatment, and Islamic Orientation. The authors, due to the local culture where the instrument is going to be employed, added religious factors. The internal consistencies of the scale were tested using Cronbach's alpha. It was found that reliability ranged between .81 and .91, which suggested that the scale was highly reliable and could be used for any meaningful empirical study. The instrument ranged was based on a 5-point Likert Scale, from (1) Never to (5) Always.

Procedure

The questionnaires were distributed to the target subjects and the researcher employed both exploratory and confirmatory factor analyses. The exploratory factor analysis was used because of the favourability of EFA compared to PCA among practitioners. Although the scale developers (Mahfooz Ansari & Mustafa Achoui Ansari, 2000) employed PCA to test the reliability and extract the underlying structure of the scale, EFA with a maximum likelihood was used to extract underlying common variance among items loaded on their respective factors. It was believed that construction and the use of PCA was situation demanding, when computers were slow and expensive to use. Hence, PCA was then a quicker and cheaper alternative to EFA (Costello & Osborne, 2005). Furthermore, PCA was computed without regard to any underlying structure caused by using all of the variance of the manifest variables and all of the variance appears in the solution (Ford, MacCallum, & Trait, 1986). In other words, the factor solution in PCA extraction was a combination of unique variance, common variance and error whereas, on the other hand, in EFA factor extraction, the common variance is partitioned from its unique variance and error variance to reveal the underlying factor structure. This indicated that only shared variance (common variance) among the items would only appear in the solution. Thus, the EFA was used basically to extract only shared variance (common variance) among the items.

Preliminary Analysis

One of fundamental requirements of quantitative research is testing the appropriateness of the data. The suitability of data used to carry out quantitative analysis can be tested through the internal consistencies of the instrument used and its validity. To ensure the appropriateness of the instrument, the reliability of data was checked through Cronbach's alpha. The Cronbach's alpha of each item ranged between .93 and .97, which suggested high reliability of the data. Moreover, the investigation of distributional indices (Skewness and Kurtosis) suggested that an assumption of normal distribution was held. No items showed skew or kurtosis that exceeded the cutoff of $\pm$ 2 indicating no problems with univariate normality, while *Mahalanobis* was used to check the multivariate assumption of normality. When a further test was performed using the Kolmogorov-Smirnov test, the result indicated that the test was statistically insignificant ($p > .05$), except for the minor cases, while $p > .05$ meant that the normality assumption was held. Moreover, the Shapiro-Wilk test also confirmed the assumption of normality. Based on these results, it could be concluded that normality assumptions were tenable and the parametric data analyses were justifiable.

Results

As earlier indicated, an exploratory factor analysis (EFA) was used to test the dimensionality of the Teaching Feedback Survey (TFS) from a randomly chosen half of the sample (N = 1504) and the remaining responses were saved for a Confirmatory Factor Analysis (CFA). For the EFA, maximum likelihood with both orthogonal and direct oblimin rotations was used.

Several researchers and statistical practitioners advocate the use of ML over other methods because of its sensitivity to model misspecification, since less sensitivity to model misspecification can lead to higher type II error rates (Olsson, Foss, Troye & Howell, 2000; Olsson, Troye & Howell, 1999). Moreover, ML estimates result in the best fit between the matrix of observed variances and covariances, **S,** and the corresponding reproduced matrix, **Σ(θ),** is more stable, and has higher accuracy in terms of theoretical and empirical fit when the data fulfilled adequately normality assumptions and data is reasonably large (Tate, 1998; Olsson, Foss, Troye & Howell, 2000; Rao & Sachs, 1999; Brown, 2006). Both orthogonal and direct oblimin were used with the aim of obtaining simple structure and considering best results. Since orthogonal oblimin constrains factors to be unrelated, while direct oblimin assumes the factors to be correlated, it is worth testing the rotation that is psychometrically apt for the data.

The analyses showed there were substantial differences between the result of orthogonal and direct oblimin rotations. In the initial EFA, four factors with eigenvalues greater than 1.0 were extracted, accounted for 65.236 of the variance of the original items. Delivery of information accounted for 55.73% of the variance, Meaningful Interaction 5.79%, Feedback and Fear treatment 4.16 and Islamic Orientation 3.60%. Use of EFA for the data on antecedent (χ^2 19830.963, p = .001) indicated that coefficients in the correlation matrix were different from zero and were not likely to occur as a function of chance. Furthermore, the total matrix sampling adequacy of all antecedent items, an index of the extent to which the matrix partial and multiple correlation coefficients confirm to zero was found to be higher than/as high as .98. In addition to adopting eigenvalues greater than one, the scree plot was examined to cross check the authenticity of the number of the factors extracted. Although the same numbers of factors were extracted using both orthogonal and direct oblimin rotation techniques, the positions of the factors were varied which consequently affected the magnitudes of the total variances explained for each factor.

It was evidenced from the result that the direct oblimin rotation technique is more suitable for the data compared to its orthogonal counterpart. It was found that the oblimin rotation generated high loading items compared to orthogonal while interfactor correlations ranged from .48 to .77 sufficient to justify using direct oblimin rotation. According to the analyses, there are no less than 3 items crossed loaded (factorial complexity) in the orthogonal rotation, with another two items below the threshold of .40, whereas direct oblimin rotation did not witness any cross loading items, although two items fell below the threshold of .40. Hence, it obvious that direct oblimin provided superior simple structure. Additionally, the results of interfactors correlation also suggested that direct oblimin ($\delta = 0$) is the appropriate rotation for this analysis as opposed to the orthogonal rotation used by the authors. In relation to factor loading, it is worth noting that Mahfooz Ansari et al. set .30 as a threshold for an item to be used to interpret a factor. However, the current study set a threshold of greater than .40 as recommended by Steven (2003) because the total variance explained in a factor accounted for by the variable could be 16%, whereas the total variance explained for .30 is only .09%. This means that the items which loaded on a factor at .40 would more significantly and meaningfully contribute to the factor than a threshold of .30.

However, the items' 7 and 20 loadings (.364, .304 and .371, .319) were found to be below the threshold of .40 for both orthogonal and direct oblimin, respectively, while items 14, 15 and 23 loaded significantly in more than one factor in the orthogonal rotation. Moreover, the result of EFA also showed a low level of communalities for both items 7 and 20 which consequently led to the low level of their factor loading. According to Fabrigar et al. (1999), low level of communalities might result from low item reliability or the item(s) being unrelated to the domain of interest and sharing little in common with other variables, and thus, the items should be avoided.

The results of this study are quite similar to those reported by Mahfooz Ansari and Mustafa Achoui Ansari (2000) in terms of the number of factors extracted. However, the total variance explained is significantly higher in this study (65.24%) compared to what was found in Mahfooz Ansari and Mustafa Achoui Ansari's results (48.5%), while the variances explained for all four factors were 55.7, 5.8, 4.2, 3.6 compared with 33.7, 5.3, 5.2 and 4.3 in Mahfooz Ansari and Mustafa Achoui Ansari's study for Delivery of information, Meaningful Interaction, Feedback and Fair Treatment and Islamic Orientation, respectively. The total variance explained was relatively low in their study (the factor accounts for only 48.5% of total variance) suggesting that factors do not sufficiently explain the relationships among the items.

Loading for Exploratory Factor Analysis using Varimax and Direct Oblimin Rotation

No	Item	Orthogonal Factor					Direct oblimin Factor				
		1	2	3	4	CM	1	2	3	4	CM
1	My lecturers have good knowledge of the subjects	.714	.340	.105	.214	.682	.797	-.064	.083	-.173	.682
2	My lecturers are systematic in delivering lectures	.702	.261	.270	.204	.676	.804	-.045	-.039	.036	.676
3	My lecturers have clear pronunciation and intonation	.651	.288	.287	.167	.617	.717	.003	.038	.074	.617
4	My lecturers finish classes on time	.611	.176	.308	.156	.523	.725	-.011	-.112	.130	.523
5	My lecturers use class time effectively	.731	.212	.241	.207	.680	.883	-.058	-.129	.002	.680
6	My lecturers have mastery of the subjects	.711	.348	.133	.248	.706	.774	-.104	.086	-.152	.706
7	My lecturers use non-verbal communication	.364	.138	.378	.122	.310	.371	-.013	-.039	.289	.310
8	My lecturers are clear in presentation	.666	.332	.273	.244	.687	.694	-.092	.075	.025	.687
9	My lecturers use clear and understandable language	.683	.378	.226	.149	.682	.723	.044	.171	-.014	.682
10	My lecturers follow the course outlines	.654	.327	.177	.139	.585	.722	.035	.114	-.052	.585
11	My lecturers' lectures are well organized	.663	.297	.333	.166	.665	.721	.014	.043	.123	.665
12	My lecturers come prepared to the class	.679	.376	.253	.205	.708	.700	-.030	.149	.003	.708
13	My lecturers have proper pace of teaching	.671	.356	.282	.166	.685	.705	.022	.134	.053	.685
14	My lecturers act as model teachers	.520	.507	.280	.259	.674	.357	-.105	.413	.037	.674
15	My lecturers welcome students' comments	.414	.626	.266	.206	.677	.140	-.031	.667	.042	.677
16	My lecturers use a variety of teaching methods	.397	.472	.420	.288	.640	.162	-.158	.401	.227	.640
17	My lecturers encourage students' opinions	.352	.753	.232	.209	.788	-.029	-.027	.894	-.002	.788
18	My lecturers encourage students' participation	.352	.744	.194	.192	.751	-.013	-.010	.889	-.040	.751
19	My lecturers make the students work hard	.364	.465	.351	.220	.520	.146	-.083	.434	.176	.520
20	My lecturers encourage critical thinking	.192	.304	.177	.116	.174	.042	-.032	.319	.074	.174
21	My lecturers encourage additional learning	.397	.632	.271	.236	.686	.102	-.073	.674	.043	.686
22	My lecturers encourage students to ask questions	.361	.672	.232	.219	.683	.032	-.054	.763	.005	.683
23	My lecturers are prompt in giving feedback on exams	.322	.437	.545	.215	.638	.070	-.062	.320	.416	.638
24	My lecturers give attention to the weak	.301	.330	.665	.233	.695	.079	-.097	.233	.572	.695
25	My lecturers are fair and just in grading	.418	.399	.504	.219	.636	.243	-.063	.295	.450	.636
26	My lecturers return assignments with comments	.241	.216	.699	.178	.625	.067	-.054	.102	.661	.625
27	My lecturers discuss test results in the class	.176	.160	.656	.275	.563	-.014	-.217	.022	.609	.563
28	My lecturers relate topics to Islamic teaching	.283	.288	.346	.776	.884	-.008	-.897	.024	.060	.884
29	My lecturers promote Islamic values	.292	.291	.264	.789	.861	.013	-.924	.026	-.041	.861
30	My lecturers use examples that are Islamic	.277	.255	.297	.796	.863	.006	-.939	-.026	.004	.863
	Eigenvalue	16.72	1.74	1.25	1.08		16.72	1.74	1.25	1.08	
	Total variance explained	55.73	5.79	4.16	3.60		55.73	5.79	4.16	3.60	
	Reliability	.85	.84	.84	.89		.88	.85	.95	.84	

Inter-factor Correlation

Factor	1	2	3	4
1				
2	-.631			
3	.771	-.638		
4	.517	-.525	.477	

Extraction Method: Maximum Likelihood. Rotation Method: Oblimin with Kaiser Normalization

Confirmatory Factor Analysis

Both first and second-order Confirmatory Factor Analysis was employed to test the hypothetical structure of the scale extracted from EFA. The covariance matrices submitted for analysis were produced using LISREL 8.54 (Joreskog & Sorbom, 2003). The second half of collected data (N = 752) was subjected to CFA to examine whether they represented the Teaching Feedback Survey. Specifically, a first-order CFA was performed and a number of indices were employed to check the fitness of the model. The result of this analysis showed that the model fit the data reasonably well with Chi-Square ($\chi^2$1372.531), *df* 391, p = .001. With the exception of GFI and AGFI, the results of the measurement model generated fit indices which exceeded the recommended critical value of .90. More specifically, the fit indices were IFI .99, NFI .99, CFI .99, GFI .89, AGFI .87 and RMSEA .060. The value of CMIN/DF was also 3.51 which indicated that the measurement model fit adequately since the figure fell below the maximum recommended value of 5 (Marsh & Hocevar, 1985).

However, the Chi-Square was statistically significant, indicating that the model did not fit the data exactly, but with a relatively large sample size as in the present study (N = 752) even minor differences between the observed and implied covariance matrix may result in statistical significance (Schumaker & Lomax, 2004). In other words, with large sample size, the test has excessive type I error rate (Bollen, 1989, p.268), hence using other indices to determine the appropriateness of the model was justifiable.

Model	X2	Df	CMIN/DF	GFI	AGFI	IFI	NFI	CFI	RMSEA
First-order	1372.531	391	3.51	.89	.87	.99	.99	.99	.06
Second-Order	1672	391	3.28	.96	.94	.99	.98	.99	.05

This finding supported theoretically Ansari and Ansari's claim that the Teaching Feedback Survey scale had four separate factors which were Delivery of Information, Meaningful Interaction, Feedback and Fair treatment and Islamic Orientation (Mahfooz Ansari & Mustafa Achoui Ansari). Thus, the result of the Confirmatory Factor Analysis or Measurement Model also provided theoretical and empirical support for the existence of four separate factors on the Teaching Feedback Survey. This outcome is strengthened by lack of evidence of any offending estimates, such as negative variance in the results and high goodness of fit indices.

The analysis of second-order was warranted as a result of the high correlations between the factors involved as earlier elaborated. The results of the analysis suggested chi-square of ($\chi^2$1672.121 with df of freedom of 391 at .01. Furthermore, although the significance of p value is considered a negative sign in the confirmatory factor analysis, due to the sensitivity of chi-square as was earlier explained, especially when sample size is relatively high, the other indices were used to determine the fit. The result of generated fit indices exceeded the recommended critical value of .90. More precisely, GFI (.96), AGFI (.94), IFI (.99), NFI (.98) and CFI (.99) while the Root Mean Square Error of Approximation that assesses the extent to which a model fits reasonably well in the population was found to be .05.

The value of CMIN/DF was also 3.28 which indicated that the measurement model fit adequately since the figure fell below the maximum recommended value of 5. The finding supported theoretically and empirically Ansari and Ansari's claim that the teaching Feedback Survey scale had four separate factors (Mahfooz Ansari & Mustafa Achoui Ansari, 2000) in second order. Thus, the result of the Confirmatory Factor Analysis or Measurement Model also provided support for the existence of four separate factors on the Teaching Feedback Survey. This outcome is strengthened by lack of evidence of any offending estimates, such as negative variance in the results and high goodness of fit indices. Comparing both first and second-order analyses, it was obvious that second-order analysis was more appropriate for the data than first-order. The analysis yielded high fit indices for the second-order analysis compared to first-order which implies that the second-order is better suited to the data.

Discussion

In the above analyses, the underlying structure of the Teaching Feedback Survey of Mahfooz Ansari and Mustafa Achoui Ansari (2000) was assessed, using both Exploratory and Confirmatory Factor Analysis, respectively. Results from the current study demonstrated the factorial validity of the 30 items of the TFS scale. Four hypothesized factors emerged in the EFA and this structure was supported in CFA. The factor analysis also confirmed the association of items with their hypothesized factors. Although the EFA results were similar to what was found in Mahfooz Ansari & Mustafa Achoui Ansari's 2000 study, two items were found to fall below the cut-off of .40, the predetermined criteria set for accepting item to be meaningfully contributing to and interpreting the factor.

However, perhaps the most salient difference between this study and Mahfooz Ansari & Mustafa Achoui Ansari's 2000 study was the use of exploratory factor analysis with the ML rotation technique (both orthogonal and direct oblimin) in the current study, while Mahfooz Ansari and Mustafa Achoui Ansari employed Principal Component Analysis with the orthogonal rotation technique. Contrary to Mahfooz Ansari and Mustafa Achoui Ansari (2000), the use of EFA purposely to sabotage the unique and error variances from the analysis and extract factors based on common variances only among the variables. The uniqueness of EFA, therefore, was the isolation of extraneous variances associating with common variance and obtaining authentic or real shared information from the scale. Furthermore, a comparison between orthogonal and direct oblimin rotations indicated that the latter was more suitable for this set of data than its orthogonal counterpart used by the authors.

This was evidenced by the lack of factorial complexity in direct oblimin, while orthogonal oblimin witnessed at least three factorial complexities (cross-loading) which spoilt the simple structure of the model. Additionally, as earlier highlighted, the total variance explained was greater in this study (65.24%) compared to Mahfooz Ansari and Mustafa Achoui Ansari's 2000 study (48.5%) which suggested that the model does not sufficiently explain the relationships among the items in the Ansari one. Moreover, another prominent difference was the use of maximum likelihood, while Mahfooz Ansari and Mustafa Achoui Ansari's 2000 study used principal component. Maximum likelihood was believed to provide more capabilities/potential for statistical inference compared to principal component analysis (Floyd & Widaman, 1995; Fabrigar et al., 1999). Nevertheless, using Principal Component Analysis did not allow the researchers to know the real common variance among the items while the adoption of the orthogonal rotation method also denied the researchers the opportunity of thoroughly verifying interfactor correlation among the factors (Fabrigar et al., 1999). Interestingly, the CFA results also demonstrated adequate fit for the first and second-order; four factors of the Teaching Feedback survey, as measured by Mahfooz Ansari & Mustafa Achoui Ansari (2000), provided theoretical support for Ansari's scale.

However, the second-order analysis was found to be more appropriate for the data compared to the first-order analysis due to high fit indices, which indicated that theoretically and empirically the TFS scale adequately measures teaching effectiveness concepts. Additionally, it is important to recognize that the four factors in this model are all relatively highly correlated with each other which supports the claim of the presence of a second-order factor. Thus, this instrument can be used to elevate the quality of teaching and, subsequently, the quality of learning outcomes. Nevertheless, the scale can be further improved by extensively reviewing all the encompassed components of teaching effectiveness. The significant differences should be drawn between learning, breadth of information and presentation, on the one hand, and various types of interaction such as group interaction, individual rapport, on the other hand, as was done by Marsh (1987, 1989). Moreover, instead of having a number of 14 items to measure delivery of information, parsimonious, targeted and yet inclusive items should be developed to cater for other components that were yet not treated by the scale. Therefore, the revised version of the scale should treat/overcome/address the shortcomings, so the scale could be standardized and used extensively/globally especially in similar situations.

References

Abrami, P. C. & d' Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall & J. Franklin (eds.), *Student ratings of instruction: issues for improving practice* (pp.97-111). San Francisco: Jossey Bass.

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of higher education,* 8, 211-228

Abrami, P. C. (1989). SEEQing the truth about student ratings of instruction. *Educational researcher*, 18, 43-45

Bollen, K. A. (1989). Structural equations with latent variables. John Wiley & Sons, New York

Brown, T. A. (2006). Confirmatory factor analysis for applied research. The Guilford Press, New York

Costello, A. B. & Osborne, J. W. (2005). Best practices in exploratory factor analysis: four recommendations for getting the most from your analysis. Practical assessment, research and evaluation, 10, (7), http://pareonline.net/getvn.asp?v=10&n=7

Damron, J. C. (1996). Instructor personality and the politics of the classroom. Available at: *www.mankato.msus.edu/dept/psych/Damro_politics.html*

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

Emery, C.R., Kramer, T.R, & Tian, R.G. (2003). Return to academic standards: a critique of student evaluations of teaching effectiveness, Quality Assurance in Education 11(1)*: 37 - 46*

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. Psychological methods, 4(3), 272-299

Fernandez, J. & Mateo, M. A. (1992). Students' evaluation of university teaching quality: analysis of a questionnaire for sample of university students in Spain. Educational and psychological measurement, 675-686

Ford, J. K., MacCallum & Trait, M. (1986). The application of exploratory factor analysis in applied psychology: a critical review and analysis. Personnel psychology, 39, 291-314

Greenwald, A. G. & Gillmore, G. M. (1997). No pain, no gain? The importance of measuring course workload in students' rating of instruction. *Journal of educational psychology*, 89(4), 743-751

Hair, F., Anderson, E., Tatham, L., & Black, C. (1998). Multivariate data Analysis. New Jersey: Prentice-Hall International, INC.

Joreskog KG and Sorbom D. (2003) LISREL 8: User's reference guide (Scientific Software International, Chicago, IL).

Mahroof A. Ansari & Mustapha Achoui Zafar Araf Ansari (2000). Development of a measure if teacher effectiveness for IIUM. *Intellectual discourse*, 8(2), 199-220.

Marsh, H. W. & Hocevar, D. (1985). Application of confirmatory factor analysis to the study of self-concept: first and higher-order factor models and their invariance across groups. *Psychological bulletin*, 97, 562-582

Marsh, H. W. & Roche, L. (1993). The use of students' evaluation and an individually structured intervention to enhance University teaching effectiveness. *American educational research journal*, 30(1), 217-251.

Marsh, H. W. & Roche, L. A. (2000). Effects of grading leniency and low workload on students' evaluation of teaching: a popular myth, bias, validity or innocent bystanders? *Journal of educational psychology*, 92(1), 202-228

Marsh, H. W. (1984). Students' evaluation of University teaching dimensionality, reliability, validity, potential bias and utility. *Journal of educational psychology*, 76(5), 707-754

Marsh, H. W. (1987). Students' evaluation of university teaching, research findings, methodological issues, and directions for future research. *International Journal of educational research*, 11, 253-388

Mikail Ibrahim & Siti Aishah Hassan (2007) Quality supervision of PhD program at the International Islamic University, Malaysia: a Rasch measurement analysis. *Paper presented in the International conference on Higher Education*, 12-14 Dec 2007, at Hotel Palace of Golden Horses, Seri Kembangan Selangor. Organized by Faulty of Education, Universiti Putra Malaysia, Monograph 4, pp. 34-50

Olsson, U. H., Foss, T., Troye, S. V. & Howell, R. D. (2000). The performance of ML, GLS, and WLS estimation in structural equation modeling under conditions of misspecification and nonnormality. Structural equation modeling, 7(4), 557-595

Olsson, U. H., Troye, S. V. & Howell, R. D. (1999). Theoretic fit and empirical fit: the performance of Maximum Likelihood versus Generalized Least Squares estimation in structural equation modeling. Multivariate behavioral research, 34(1), 31-58

Ramsden, P., Martin, E., & Bowden, J. (1989). School environment and sixth form pupils' approaches to learning, *British Journal of educational psychology*, 59, 129-142

Rao, N. & Sachs, J. (1999). Confirmatory factor analysis of the Chinese version of the motivated strategies for learning questionnaire. Educational and psychological measurement, 59(6), 1016-1029

Remedios, R., & Lieberman, D. A. (2008). I liked your course because you taught me well: the influence of grades, workload, expectation and goals on students' evaluations of teaching, *British educational research journal*, 34, 91-115

Remedios, R., Lieberman, D. A. & Benton, T. G. (2000). The effects of grades on course enjoyment: Did you get the grade you wanted? *British journal of educational psychology*, 70, 353-368

Richardson, J. T. E. (1994). A British evaluation of course experience questionnaire. *Studies in higher*, 19, 59-68

Steven, J. (2003). *Applied multivariate statistics for the social sciences.* New Jersey: Lawrence Erlbaum Associates, Publishers.

Tate, R. (1998). An introduction to modeling outcomes in the behavioral and social sciences (2nd ed), Burgess publishing

Schumacker, E. & Lomax, G. (2004). *A beginner's guide to structural equation modeling*. New Jersey: Lawrence Erlbaum.