

A Psychometric Evaluation of Two Teaching Effectiveness Scales

Mikail Ibrahim¹

¹ Faculty of Major Language Studies, Universiti Sains Islam Malaysia, Nilai, Malaysia

Correspondence: Mikail Ibrahim, Faculty of Major Language Studies, Universiti Sains Islam Malaysia, Nilai, Malaysia. Tel: 17-28-20-605. E-mail: ibidun18@yahoo.com

Received: April 12, 2012 Accepted: May 28, 2012 Online Published: June 25, 2012

doi:10.5539/jsd.v5n7p91

URL: <http://dx.doi.org/10.5539/jsd.v5n7p91>

Abstract

The call for teaching accountability in higher education initiated teaching effectiveness research and its scales development. Attention in many institutions of higher learning has been diverted recently to the improvement of teaching performance as another way besides academic research to promote the higher institutions. The diversity of attention is a response to external calls for accountability in teaching as a result of the under-estimation of the significance of the teaching process compared to research activities. As research on teaching effectiveness has increased, so has the number of different measures of teaching effectiveness. Hence, in this article, the researchers examined the psychometric properties of two teaching effectiveness scales, namely the Marsh Student Evaluation of Educational Quality (1987) and Mahfooz Ansari and Mustafa Achoui Ansari Teaching Feedback Survey (2000) in terms of their factorial and construct validity. A total of 1504 3rd and 4th year and postgraduate students were selected from four renowned Malaysian public Universities, namely USIM, UM, UPM and IIUM. The study found that although the two scales were constructed to assess teaching effectiveness in higher institutions, the Marsh scale was extensively used in the literature and more comprehensive in relation to the numbers of factors. The study found that although there is room for improvement for both scales, the Marsh's scale is psychometrically more sound, and theoretically more comprehensive than Ansari and Ansari's scale.

Keywords: teaching effectiveness scale, psychometric properties, confirmatory factor analysis, Principal Component Analysis (PCA)

1. Introduction

The widespread demand from both institutional authorities and stakeholders for greater accountability in higher education has sparked research on what constitutes teaching effectiveness. These demands, coupled with those of the quality improvement movement, have resulted in a need for valid, reliable and comparable performance data on teaching quality. Recently, higher institutions of learning all over the world have begun adopting evaluation systems to promote their academic standards and the quality of the experience learned during the students' life. Historically, this kind of evaluation is normally carried out by the authorities of the higher institutions. However, due to the significance of students' experience and assessment ability, the institutions of higher learning are involving students' assessments in policy design.

In Malaysia, the authorities and the stakeholders are asking institutions of higher learning to promote the students' learning experience and accountability. The accountability of an institution of higher learning would not be accomplished or completed without direct comments from students on their experience and their evaluation of the teaching qualities of that particular higher institution. The students' evaluations are not intended to penalize the instructors but rather to improve the process of teaching and learning since the meaningful interpretation of evaluation is judging the worth or merit of something (Worthen et al., 1997). Research has suggested that 22% and 20% of University Sains Malaysia (USM) students had negative perceptions of their lecturers' ability to comment on their work and give helpful feedback on how students are doing, respectively (Sarjik Kuar, 2003)

The findings indicated that there is a portion of students in some of the Malaysian institutions of higher learning who are not totally satisfied with the performance of their instructors and the quality of the teaching experience they are receiving from their respective institutions. The teacher evaluation system should, therefore, be rooted in two broad purposes: (1) it should be outcome-oriented, (2) it should be improvement oriented. According to

Colby, Bradshaw and Joyner (2002), evaluation is used generally for the purpose of instructional improvement and evidence of accountability which are consequently believed to have positive effects upon academic performance, teachers' instructional improvement and the overall image of the institution. Since Malaysian institutions of higher learning have aspired to be globally competitive through sound research activities, the quality of the teaching experience, the quality of programmes on offer, the infrastructure, and the percentage of international lecturers and students within each particular establishment, students' opinions cannot be ignored or underestimated.

Although many Islamic countries -especially Malaysia- are now encouraging student-evaluation systems, many teachers (lecturers) still consider inside classroom activities as being a personal issue in which outsiders are not allowed to be involved.

Moreover, even though students' evaluation scales have been used extensively to measure teaching effectiveness, the opinions, according to Marsh (1987), vary from reliable, valid, and useful to unreliable, invalid and useless. In addition to that, Marsh alleged that the previous studies had been superficial, and unsound both methodologically and from the perspective of the research paradigm (Marsh, 1987). It was also reported that the scales also contained inconsistencies in domain definition, lack of correlation with students' achievement, and even item quality such as double-barreled, unclear and spurious items (Donald et al., 2008). Thus, amid these serious allegations about the quality of teaching effectiveness scales, further investigation is warranted. In support of this thought, Cashin and Downey (1992) contended that the main objective of the teaching effectiveness assessment is to make an objective description and accurate judgment of effective teaching as accurately as possible and to control all irrelevant variables other than the instructor's teaching effectiveness. Hence, the researcher's task is not to evaluate why an instructor is effective or ineffective or how the performance of instructors can be improved, but rather to follow a diagnostic process of the situation under study.

Furthermore, since the findings from Western research settings may not necessarily hold true for subjects in other cultures, there is a serious need for an empirical study that will confirm and validate the previous studies on students' evaluation scales and add some new perspective to the previous attempts by those researchers. An awareness of the inadequate data from non-Western research areas, and possible problems of generalizability of Western findings to subjects in other settings, partly motivated this study. More precisely, as little has been done in some of Malaysia's institutions of higher learning to academically examine students' evaluation of teaching effectiveness, this study aims to determine whether a questionnaire derived from many Western studies could be generalized for use in the Malaysian context or whether the scale can be replicable in a different culture and ideology. The present study attempts to investigate the psychometric properties of the Students' Evaluation of Educational Quality (SEEQ) scale constructed by Marsh and the Teaching Feedback Survey of Mahfooz Ansari and Mustafa Achoui Ansari (2000) among third and fourth year university and postgraduate students, in order to examine which of the two scales is more appropriate to assess teaching effectiveness in the Malaysian context.

2. Historical Background of the Study

The call for teaching accountability in higher education started more than two decades ago (Barrie et al., 2005), when stakeholders and educational specialists started advocating more transparency in higher education institutions (Mikail & Siti Aishah, 2007; Mikail Ibrahim, 2012). Numerous institutions in many Western countries, especially in Australia and the United Kingdom, diverted their attention to the improvement of teaching performance as another way besides academic research to promote the institutions of higher learning. The diversity of attention is a response to external calls for accountability in teaching as a result of an under-estimation of the significance of the teaching process compared to research activities. This development was not necessarily intended to undervalue the importance of research activities but was rather a consideration of both the teaching exercise and the research activities complementing each other for the betterment of the institutions and students alike. In 1996, the Institute of Teaching and Learning of the University of Sydney developed a new policy on the quality assurance of teaching and learning to foster students' standards (Barrie et al., 2005). The policy also involved the development of a system to enact the policy in ways that would align policy and management directives, faculty strategic initiatives, and teaching and learning practices at the level of actual subjects.

The quest for effective teaching remains a demanding, complex and daunting task. In spite of the wealth of research evidence on the nature of effective and ineffective teaching, there are still problems about spelling out what effective teaching really is (Centra, 1993). Most research efforts aim at investigating teacher effectiveness

by probing the following dimensions: special characteristics of the teacher, which include cognitive dimensions, personality dimensions, perceptions of self and others, instructional procedures and interaction styles.

An important aspect emanating from the last of these (interaction styles), is the fact that teachers are the ones who contribute most to the educational enterprise and, therefore, need to ensure that the learner is engaged appropriately with the instructional material. In this regard, it is important that teachers are able to link teaching to learning functions in order to facilitate the optimal realization of learning outcomes (Grosser, 2007). "Teacher enriches interactions within the learning environment by providing students with relevant experiences and with the substantial scaffolding that permits them to construct meaningful interpretations and assimilate new understandings. Therefore, the more teachers know about their students and the domains and tasks under study, the more likely they are to establish a learning environment that benefits those students" (Alexander et al., 1994).

According to Stringer and Irwing (1998), teaching effectiveness could be figured out through the extent that students' performance improves after a period of instruction, in a manner consistent with the objectives of the instruction and in accordance with prior set goal(s). Thus, the effectiveness of teaching can be measured through the changes which have occurred in the students' knowledge (declarative and procedural), their level of motivation, ability to cope with constant changes in life, and effective management of stress after they had been given the instruction. Due to this uniqueness of effective teaching, many post-secondary institutions adopted students' rating of instruction as one measure of instructional effectiveness (d'Apollonia & Abrami, 1997; Griffin, 2004).

Thus, one of the ultimate goals of any institution of higher learning is to be a custodian, dispensing knowledge and transferring culture to their students which will produce capable citizens who can carry out their responsibilities adequately. Obviously, knowledge is not merely the memorization of facts and figures but rather an ability to digest, accumulate and assimilate the information obtained for the betterment of the mind, society and human beings. Thus, an institution of higher learning should be willing and able to judge to what extent it succeeds in: (1) giving students the confidence and ability to take responsibility for their own continuing personal and professional development; (2) preparing students to be personally effective within the circumstances of their lives and works and (3) promoting the pursuit of excellence in the development, acquisition and application of knowledge and skills (Stephenson, 1992).

Furthermore, the features of effectiveness are argued to be the characteristics of both teachers and students alike. This simply means that teaching effectiveness should not be erroneously attributed to teachers alone due to the fact that students' characteristics are fundamental requirements of any teaching and learning success.

In Malaysia, government and stakeholders have been challenging higher institutions of learning in recent years to improve the standard of education especially after a global rating system of higher institutions started. The government has firmly stated that no compromise must be made on the quality of education. Thus, the then Prime Minister, Yab Dato' Seri Abdullah Ahmad Badawi, emphatically concluded that quality is the only way to compete and be among the best in the world (New Straits Times, 2nd Oct 2007; New Straits Times, 8th April, 2008; New Straits Times, 6th August, 2008). Hence, to be among the best, the Malaysian institutions of higher learning must fulfil the fundamental requirements that will imbue and enhance the quality of teaching, research and publication, administrative effectiveness and transparency, amenities and infrastructure. These qualities will not only contribute to the effective pursuit of knowledge but will also facilitate recruitment because of the soundness of their academic reputation among international lecturers and students alike. Teaching effectively might not be directly examined as an independent dimension of ranking, but as an independent measure, through the alumni and staff winning major international awards, highly cited researchers in major fields, articles published in selected top journal, articles indexed by major citation indexes, and performance per capita.

It is also believed that the other objectives of the institutions of higher learning would hardly be achieved if in-depth knowledge is not meaningfully transferred from capable and well-equipped instructors to their students or subordinates. "Perhaps what we need is a succinct model of good university teaching which powerfully conveys the main requirements for good teaching virtually anywhere and which includes a built-in respect for diversity and otherness" (Badley, 2000). Researchers (Ramsden, 1991; Eiszler, 2002) studied the impact of effective teaching on students' performance, as opposed to over-estimating and solely judging the quality of lecturers based on their research and publications.

3. Teaching Effectiveness Scales

Many institutions of higher learning all over the world have adopted a system of students' evaluation of teaching effectiveness especially in Western countries (Marsh, 1984; 1987; Greenwald & Gillmore, 1997). As reported by Marsh, Remmers (1927) initiated the first systematic research programme of students' evaluation of teaching

effectiveness. In 1927, he first published the first standardized, systematic, and multi-trait scale to measure teaching instruction. Remmers tried to standardize the scale through the examination of its reliability, validity, norms, halo effects, the 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 et al. (1989), Remedios et al. (2000), Remedios and Lieberman (2008), Richardson (1994), Mahfooz Ansari and Mustafa Achoui Ansari (2000), Harrison and Douglas (2004), Meyer and Eley (2006) and many more. These scales, whether they were called student evaluations of educational quality (Marsh, 1984; 1987), course experience questionnaires (Richardson, 1994), endeavour instructional rating and students' rating of instruction (d'Apollonia & Abrami, 1997); or teaching feedback survey, (Mahfooz Ansari & Mustafa Achoui Ansari, 2000), were all constructed to evaluate students' experiences of teaching within the classroom.

The purposes of these scales, according to Marsh (1984; 1987), are varied. They include: 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 courses-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 the institution 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, a tremendous opportunity would be given to benchmark the effectiveness of the teaching experience across different domains.

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

In relation to the dimensionality of the scales, 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 teacher characteristics such as empathy, facilitation, personal attention, teacher support, student involvement, negative affect, enthusiasm and rapport and interaction as 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 et al., 2004). Meanwhile, Mahfooz Ansari and Mustafa Achoui Ansari (2000), in addition to the delivery of information, meaningful interaction, feedback and fair treatment and due to the different cultural aspects, 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 purposive students' evaluation of teaching effectiveness among the selected institutions of higher learning in Malaysia and how efficient the scales are in measuring teaching effectiveness.

4. Method

4.1 Participants

A total of 1504 3rd, 4th year, and postgraduate students from four Malaysian public institutions of higher learning were selected and voluntarily participated in this study. The subjects of this study were selected from the 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. There were initially 1550

participants but data from 46 participants were discarded due to many reasons. These data were discarded because (1) participants skipped the majority of items and (2) the participants were deleted because of a pattern in their responses. The data were 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 602 (40.0%) males and 902 (60%) females.

4.2 Measures

Participants completed two types of teaching effectiveness scales, namely the Marsh Student Evaluations of Educational Quality (Marsh, 1984; 1987) and the Teaching Feedback Survey, (Mahfooz Ansari & Mustafa Achoui Ansari, 2000). The first scale consisted of 29 items and was categorized into eight dimensions; Learning/Value, Enthusiasm, Organization, Group Interaction, Individual Rapport, Breadth of Coverage, Exams/Grades, and Assignment. The ninth factor was not included (workload) due to the scale of measurement used for the factor (nominal scale). The reliability of the scale was established and ranged from .86 to .94 across many studies. Four items were used each to measure learning/value; enthusiasm, organization, group interaction, individual rapport, and breath of coverage factors while three items were used to measure exam/grade, and two items were assigned to measure assignment, respectively. The second instrument was developed by Mahfooz Ansari and Mustafa Achoui Ansari (2000) and was used to investigate teaching effectiveness in the International Islamic University (IIUM). The scale consisted of 30 items and Principal Component Analysis (PCA) was used to categorize it into latent factors which are delivery of information, meaningful interaction, feedback and fair treatment and Islamic orientation. Interestingly, the internal consistencies of the scale were tested and they were between .81 and .91. This figure indicated that the instrument is psychometrically sound to be used in academic research activities. Both scales were then joined together with short demographic variables. All items were measured on a 7-point scale (1 =Not true to 7= Always true).

4.3 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's distributional characteristics (Skewness & Kurtosis value) suggested that an assumption of normal distribution was held. No items showed skew or kurtosis that exceeded the cutoff of ± 2 indicating no problems with univariate normality, while Mahalanobis was used to check 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. Then CFA was conducted on each teaching effectiveness scale to investigate if the scales' factors adequately described the data. In these analyses, each item was assigned to its designed latent factor and the factor pattern coefficients and correlation between the latent factors were estimated. Multiple fit indices were used to assess the adequacy of these measurement models and the chi-square statistics divided by its degrees of freedom was calculated. Carmine and McIver (1981) recommended that a ratio of X^2 divided by df should be less than 3.0 for adequate fit but Marsh and Hocevar (1985) suggested that a ratio as high as 5.0 is considered acceptable. Chi-Square and other fit indices such as GFI, AGFI, NFI, CFI and TLI were examined and the measures should be $\geq .90$ to be considered good fit. RMSEA was also tested and should be less than .08 for an adequate model.

It is worth mentioning that although the two scales were generally measuring the same concepts (teaching effectiveness), the number of their factors varied extensively from eight and four factors for the Marsh and Ansari scales, respectively.

5. Result and Discussion of Measurement Model

Two different CFAs were performed for Marsh and Mahfooz Ansari & Mustafa Achoui Ansari to test the underlying structure of the scales. The first measurement model was run on Marsh's teaching effectiveness using the LISREL programme version 8.3 (Jöreskog & Sörbom, 2004), a famous statistical package extensively used for multivariate analyses such as CFA and SEM with a maximum likelihood estimation. The scale consisted of 29 items, with eight latent variables, namely Learning/Value, Enthusiasm, Organization, Group Interaction, Individual Rapport, Breadth of Information, Examination, and Assignment. The first-order analysis allowed the factors to correlate together, while six items were freely correlated.

Given the potential limitations of an exclusive reliance on the X^2 especially when sample size is relatively large, other indices were used such as GFI, AGFI, CFI, TLI and RMSEA to determine the fitness of the model. The model was considered fit when GFI, AGFI, CFI, and TLI were greater than .90 and when the RMSEA value was about .05 but not greater than .08. According to the researchers, GFI is considered to be like R^2 in regression analysis. The GFI is a ratio of the squared difference between the observed and reproduced matrices to the observed variance, not adjusted for the degree of freedom while AGFI adjusts the GFI index for the degree of freedom of a model relative to the number of variables (Schumacker & Lomax, 1996; Hair et al., 1998; Brown, 2006; Stevens, 2007; Whitman et al., 2009; Donnellan, 2008). The Comparative Fit Index (CFI) is also known as the Bentler comparative fit index. It is another method to conceptualize goodness of fit by comparing an existing model with a null model which assumes the latent variables in the model are uncorrelated (Tabachnick & Fidell, 2007; Johnson et al., 2007). The RMSEA is similar to the RMR, but it assesses the discrepancy between the model implied covariance (correlation) matrix and the observed covariance (correlation) matrix by taking into account the degree of freedom or number of free parameters required to achieve a given level of fit (Rao & Sachs, 1999). Practitioners and statistical experts have different thresholds for RMSEA but most salient ideas are that the value of RMSEA must be between .05 and .08 and between .05 and 1.00 (Rao & Sach, 1999; Hulpia et al., 2009; Cadiz et al., 2009). RMSEA was first constructed by Steiger and Lind (1980) and later was expanded by Browne and Cudeck (1993) to test the adequacy of the model through the examination of how well the parameters of the factor model are able to reproduce the sample correlations. The equation of RMSEA is as follows:

$$\text{Estimated RMSEA} = \sqrt{\frac{\hat{F}_o}{df_{model}}} = \sqrt{\frac{T - df}{df(N - 1)}}$$

Where T represents the likelihood ratio test statistic, df represents degrees-of-freedom from the hypothesized model and N represents sample size. The numerator, therefore, represents the sample estimate of the noncentrality parameter λ and it is an estimate of the degree of model misspecification (see Steiger & Lind, 1980; Browne & Cudeck, 1993; Curran et al., 2002; 2003). Where $\hat{F}_o = \frac{x^2_{model} - df_{model}}{N}$ when the model perfectly fits the data the $\hat{F}_o = 0$. The greater the model misspecification the larger the \hat{F}_o .

A CFA was performed on eight factors of the Marsh Student Evaluation of Educational Quality. The model had the following latent structure: items 1-4 were indicators of the latent variable Learning/Value, items 5-8 were indicators/manifest of latent Enthusiasm, items 9-12 were indicators of Organization, items 13-16 were indicators of Group Interaction, items 17-20 were indicators of latent Individual Rapport, items 21-24 were indicators of latent Breadth of Coverage, items 25-27 indicated latent Exams/Grades, and items 28 -29 were indicative of Assignment and Workload. It is worth mentioning that some residuals were allowed to be correlated but within the factor.

The hypothesized measurement model was supported by a Confirmatory Factor Analysis. The fit of the model to the data was excellent. Even though the Chi-Square statistics with a degree of freedom of 346 was $X^2 964.82$ ($p = .001$) resulted in the conclusion to reject the null-hypothesis that the model was incorrect, the other indices proved otherwise. The reported root-mean-square residual of approximation (RMSEA) was approximately zero (i.e. much smaller than the 0.05 value often considered to reflect a good fit) and the p-value of the null hypothesis is .001, and $RMSEA < 0.05$ was .99. Furthermore, the goodness of fit index (GFI) was .92, the adjusted goodness of fit index (AGFI) was .90, the Normed fit index (NFI) was .99, and the Comparative Fit Index (CFI) was .99 which indicated that the model fit since the indices were larger than the recommended .90. Finally, the largest standardized covariance residual was 3.0 in magnitude, a value which indicated that the model fit the data.

Marsh's Student Evaluation of Educational Quality (second-Order)

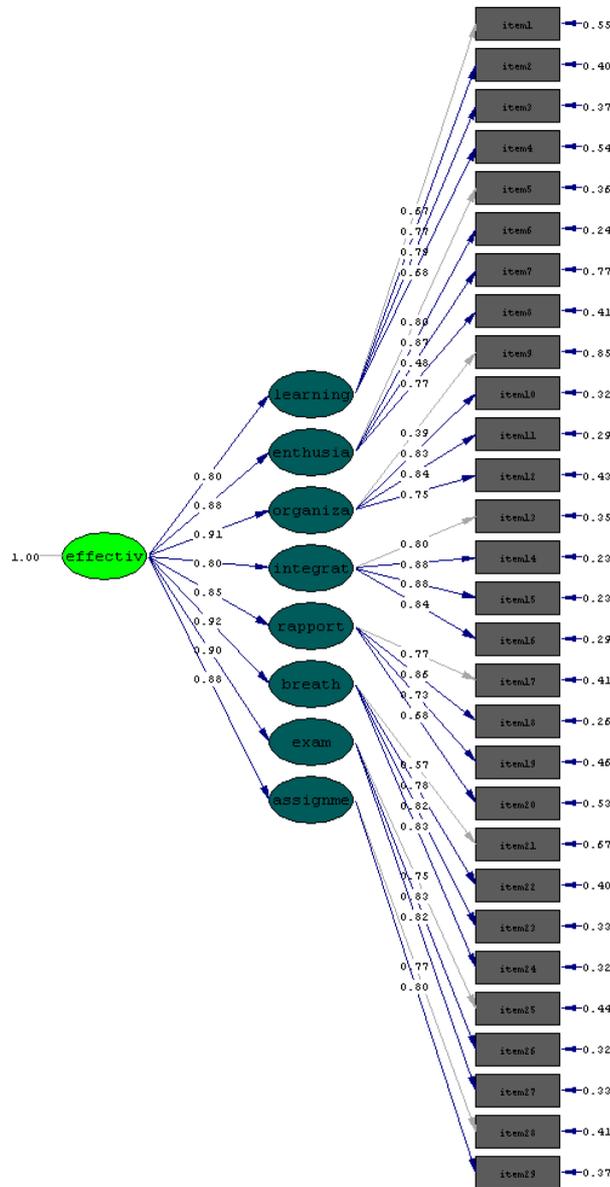


Figure 1. Marsh's student evaluation of education quality scale, 1987 (Second-Order)

The value of CMIN/DF was also 2.79 which indicated that the measurement model fit adequately since the figure fell below the maximum recommended value of 5. According to many statistical practitioners (Marsh & Hocevar, 1985), the ration of CMIN/df should be less than 5.0 to judge the fitness of the model (for more review, see Marsh & Hocevar, 1985; Carmines & McIver, 1981). This finding supported theoretically the Marsh Student Evaluation of Educational Quality Scale (SEEQ, 1987) which claimed that the scale had eight separate factors. Thus, the result of the Confirmatory Factor Analysis or Measurement Model also provided support for the existence of eight separate factors for the SEEQ scale. 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 estimated loadings, reflecting the validity of each observed variable as a measure of the latent variable, were generally very high, ranging from .40 to .88. Finally, the estimated correlations among the latent variables were all positive with magnitudes of .64 to .90 consistent with the expectations based on the literature. Therefore, it can be concluded based on multiple indicators of first-order of Confirmatory Factor Analysis (CFA) that the model fit and the hypothesized measurement model was acceptable.

As expected, the interfactor correlations were mostly moderate and suggested the existence of second-order factors. The second order is requested and warranted when first-order factors are explained by some higher order factor structures (Schumacker & Lomax, 2004). In second-order or higher-order CFA, the focus is on the intercorrelations represented among the factors (Brown, 2006). According to him, a goal of higher-order analysis is to provide a more parsimonious account of the correlations among the lower-order factors. These specifications emphasize that higher-order factors have direct effects on lower-order factors; these direct effects and the correlations among the higher-order factors are responsible for the covariation of the lower-order factors (Brown, 2006).

Thus, the same 29 items were again analyzed but using second-order CFA. The analysis revealed a slight increase of Chi-Square X^2 1301.43, ($df = 369$), at .01. The p-value also suggested that data does not fit the model, which indicated the existence of a discrepancy between the data and the model. Furthermore, although the significance of p- value is considered a negative sign in the measurement model, due to the sensitivity of chi-square especially when the sample size is high, as was previously elaborated, other indices were used to determine the model fit. The other indices suggest that the model fits perfectly. More precisely, the GFI .96, IFI .99, AGFI .94, CFI .96. Interestingly, the Root Mean Square Error of Approximation that assesses the extent to which a model fits reasonably well in the population was (.06), a value that was considered to be within the recommended value below .100, with the p-value of RMSEA of .01.

The result of the Confirmatory Factor Analysis (CFA) provided support that the model fit perfectly. Thus, a comparison of both first and second orders suggested that both were statistically fit, although the second-order was slightly better than the first-order in terms of fitness indices especially for GFI and AGFI. Moreover, the reliability of each item for the second-order analysis was also slightly higher compared to the first-order analysis.

Mahfooz Ansari & Mustafa Achoui Ansari Teaching Feedback Survey (Second-Order)

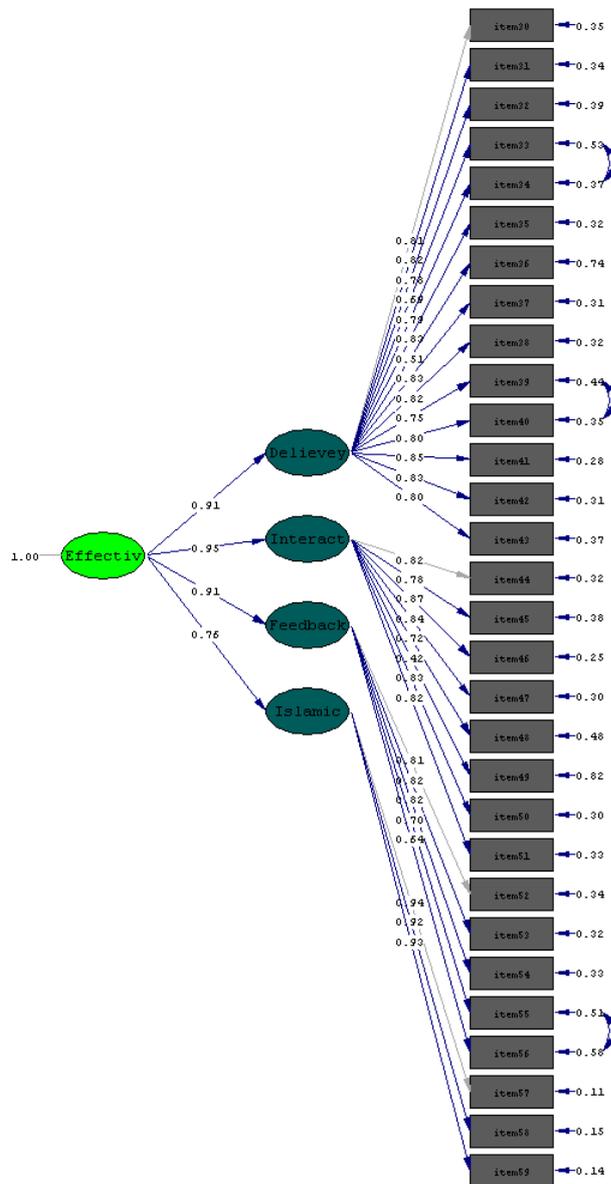


Figure 2. Ansari & Mustafa achoui teaching feedback survey, 2000 (Second-Order)

On the other hand, a CFA was conducted on Mahfooz Ansari & Mustafa Achoui Ansari (2000) based on the results of their Principal Component Analysis. The model found items 1-14 were indicators of latent delivery of information, items 15-22 were indicative of latent meaningful interaction, items 23-27 were indicated latent Feedback and fair treatment, and items 28-30 were indicators of latent Islamic orientation. The analysis suggested generally that the model was *infit* with Chi-Square (X^2 1372.531), df 391, $p = .001$. Nevertheless, other fit indices with the exception of GFI and AGFI, indicated that the model *fit* reasonably. More specifically, the fit indices were GFI .89, AGFI .87, IFI .99, NFI .99, CFI .99 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 & Vocevar, 1985). This finding supported theoretically Mahfooz Ansari & Mustafa Achoui 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, 2000). 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 1672.121 with df of freedom of 391 at .01.

However, since the chi-square is very sensitive to the sample size, as earlier narrated, other indices were used to investigate and determine the appropriateness of the model. The result of most generated fit indices exceeded the recommended critical value of .90. The analysis revealed IFI .99, NFI .98 and CFI .99. However, the GFI and AGFI levels were slightly lower in the second-order analysis than in the first-order analysis with .86 and .83 for GFI and AGFI respectively. Interestingly, the Root Mean Square Error of Approximation that assesses the extent to which a model fits reasonably well in the population was also within the recommended value below 1.00 with a value of .06.

The value of CMIN/DF was also 4.28 which indicated that the measurement model fit adequately since the figure fell below the maximum recommended value of 5. The finding supported theoretically Ansari and Ansari's claim that the teaching Feedback Survey scale had four separate factors (Ansari & Ansari, 2000). 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.

Table 1. Descriptive information for the Marsh (1987) and Mahfooz Ansari et al. (2000) scales

| Scale | Item Numbers | Mean | SD | α | Average Inter-item Correlation |
|-------------------------------------|--------------|------|-------|----------|--------------------------------|
| Marsh Scale SEEQ 1987 | | | | | |
| Learning/value | 1-4 | 5.62 | .94 | .921 | .70 |
| Enthusiasm | 5-8 | 5.45 | 1.09 | .919 | .73 |
| Organization | 9-12 | 5.30 | 1.08 | .918 | .75 |
| Group Interaction | 13-16 | 5.64 | 1.04 | .918 | .74 |
| Individual Rapport | 17-20 | 5.15 | 1.058 | .917 | .75 |
| Breath | 21-24 | 5.26 | 1.00 | .913 | .81 |
| Examination | 25-27 | 5.35 | 1.03 | .914 | .79 |
| Assignment | 28-29 | 5.64 | 1.04 | .919 | .73 |
| Ansari & Ansari 2000 TFS | | | | | |
| Delivery of Information | 1-14 | 5.60 | .95 | .851 | .806 |
| Meaningful Interaction | 15-22 | 5.66 | .99 | .839 | .833 |
| Feedback & fair Treatment | 23-27 | 4.27 | .98 | .855 | .788 |
| Islamic Orientation | 28-30 | 5.34 | 1.45 | .907 | .725 |

6. Conclusion

There is increasing interest in what connotes teaching effectiveness in institutions of higher learning due to the effect of teaching effectiveness on learning outcomes. Hence, since research on teaching effectiveness is on the increase, the inventories to measure it are also increasing. The primary aim of the current study was to compare the psychometric properties of the selected two scales of teaching effectiveness. In general, it appears that Marsh's scale held up the better of the two although Mahroof Ansari et al.'s also demonstrated good construct validity and scale characteristics. More precisely, the result of this study suggested that broadly, Marsh's scale is a much better scale than Mahfooz Ansari and Mustafa Achoui Ansari's on both psychometric and practical grounds. Theoretically, Marsh's scale is comprehensive and encompasses broad factors that were considered parts of teaching effectiveness elements. Although Marsh's scale is relatively shorter compared to Mahfooz Ansari and Mustafa Achoui Ansari's one, however, it assesses a broader range of constructs and differentiated among learning, enthusiasm, organization, and breath of information. It also distinguished between different types of interaction such as individual interaction and group interaction, while Mahfooz Ansari and Mustafa Achoui Ansari did not differentiate between them. Although both scales tapped breath of teaching, enthusiasm, and learning/value, an advantage of Marsh's scale is that it had a separate factor for each of these domains while Mahfooz Ansari and Mustafa Achoui Ansari had them as items under delivery of information factors. Thus, it can be concluded that Marsh's scale is based on a more completing theoretical framework than that of Mahfooz Ansari and Mustafa Achoui Ansari. Psychometrically, although RMSEA (which indicates the discrepancy of the

obtained model from the actual model) was within the recommended value for both scales, it was closer to zero in Marsh's scale compared to the Mahfooz Ansari and Mustafa Achoui Ansari scale. Furthermore, while other fitness indices such as GFI, AGFI, IFI, NFI, and CFI were above the recommended value of .90, these indices appeared to be higher in Marsh's scale compared to Mahfooz Ansari and Mustafa Achoui Ansari's scale and also better in second-order than in the first-order. Thus, it can be psychometrically concluded that both scales were more fit in second-order than in first-order. Although both scales were found to be psychometrically sound, caution should be exercised when researchers are looking for an appropriate scale to be used for their studies because a non comprehensive scale would not be able to capture all the angles that a researcher attempts to study and then might not be able to contribute to the theoretical underpinnings.

References

- Alexander, P. A., Kulikowich, J. M., & Jotton, T. L. (1994). The role of subject-matter knowledge and interest in the proceeding of linear and nonlinear texts. *Review of Educational Research, 4*(2), 201-252.
- Badley. (2000). Developing globally-competent university teachers. *Innovation in Education and Teaching International, 37*(3), 244-253.
- Barrie, S., Ginns, P., & Posser, M. (2005). Early impact and outcomes of an instructionally aligned, student focused learning perspective on teaching quality assurance. *Assessment and Evaluation in Higher Education, 30*(6), 641-656.
- Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, the Guilford Press.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen, & J. S. Long (Eds.), *Testing Structural Equation Models*. Newbury Park, CA: Sage.
- Cadiz, D., Sawyer, J. E., & Griffith, T. L. (2009). Developing and validating field measurement scales for absorptive capacity and experienced community of practice. *Educational and Psychological Measurement, 69*(6), 1035-1058.
- Carmines, E. G., & McIver, J. P. (1981). Analyzing models with unobserved variables: analysis of covariance structures. In G. W. Bohnstedt & E. F Borgatta (Eds.), *Social measurement: current issues* (pp. 65-115). Beverly Hill, CA: Sage.
- Cashin, W. E., & Downey, R. G. (1992). Using global students' items for summative evaluation. *Journal of Educational Psychology, 84*(4), 563-572.
- Centra, J. A. (1993). *Reflective faculty evaluation; enhancing teaching and determining faculty effectiveness*. San Francisco: Jossy-Bass Publishers.
- Colby, S. A., Bradshaw, L. K., & Joyner, R. L. (2002). *Teacher evaluation: A review of the literature*. A paper presented at the annual meeting of the American Educational Research Association. New Orleans.
- Curran, P. J., Bollen, K. A., Chen, F., Paxton, P., & Kirby, J. (2003). Finite sampling properties of point estimates and confidence intervals of the RMSEA. *Sociological Methods & Research, 32*(2), 208-252.
- Curran, P. J., Bollen, K. A., Paxton, P., Kirby, J., & Chen, F. (2002). The noncentral chi-square distribution in misspecified structural equation models: finite sample results from a Monte Carlo simulation. *Multivariate Behavioral Research, 37*, 1-36.
- d'Apollonia, S., & Abrami, P. C. (1997). Navigating student rating of instruction. *American Psychologist, 52*(11), 1198-1208.
- Donald, B., Brian, E., Curtis, M., William, M., Craig, P., Kirk, J., ... Zachary, W. (2008). *Developing a psychometrically sound measure of collegiate teaching proficiency*. EBSCOhost Research database.
- Donnellan, M. B. (2008). A psychometric evaluation of two achievement goal inventories. *Educational and Psychological Measurement, 68*(4), 643-658.
- Eiszler, C. F. (2002). College students' evaluation of teaching and grade inflation. *Research in Higher Education, 43*(4), 483-501.
- 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, 67*(5), 675-686.
- 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.

- Griffin, B. W. (2004). Grading leniency, grade discrepancy and student ratings of instruction. *Contemporary Educational Psychology, 29*, 410-425.
- Grösser, M. (2007). Effective teaching: Linking teaching to learning functions. *South African Journal of Education, 27*(1), 37-52.
- Hair, F., Anderson, E., Tatham, L., & Black, C. (1998). *Multivariate data Analysis*. New Jersey: Prentice-Hall International, INC.
- Harrison, P. D., & Douglas, D. K. (2004). The relative merits of different types of overall evaluations of teaching effectiveness. *Research in Higher Education, 45*(3), 311-323.
- Hulpia, H., Devos, G., & Rosseel, Y. (2009). Development and validation of scores on the distributed leadership inventory. *Educational and Psychological Measurement, 69*(6), 1013-1034.
- Johnson, B., Stevens, J. J., & Zvoch, K. (2007). Teachers' perceptions of school climate; a validity study of scores from the revised school level environment questionnaire. *Educational and Psychological Measurement, 67*(5), 833-844.
- Joreskog, K., & Sorbom, D. (2003). LISREL (version 8.54) computer software. Chicago: Scientific Software international.
- Mahroof A., Ansar, & Mustapha Achoui Zafar Araf Ansari. (2000). Development of a measure of 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. (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.
- Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87*(4), 666-679.
- Meyer, J. H., & Eley, M. G. (2006). The approaches to teaching inventory: a critique of its development and applicability. *British journal of Educational Psychology, 76*, 633-649.
- 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 Faculty of Education, Universiti Putra Malaysia, Monograph 4, pp. 34-50.
- Mikail Ibrahim. (2012). Evaluation of the psychometric properties of a teaching feedback survey; first and second-order Confirmatory Factor Analysis. *American International Journal of Contemporary Research, 2*(2), 134-142.
- New Straits Times. (2007, October 2), p. 2.
- New Straits Times. (2008 April 8), p. 9.
- New Straits Times. (2008, August 6), p. 11.
- Ramsden, P. (1991). A performance indicator of teaching quality in higher education; the course experience questionnaire. *Studies in Higher Education, 16*, 127-150.
- 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, & Sachs. (1999). Confirmatory factor analysis of the Chinese version of the motivated strategies for learning questionnaire. *Educational and Psychology 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 Education, 19*, 59-68.
- Sarjit K. (2003). *Evaluating teaching effectiveness in higher education: A case study of USM*. Retrieved from <http://WWW.usm.my/ipptn/HER2>
- Schumacker, E., & Lomax, G. (1996). *A beginner's guide to structural equation modeling*. New Jersey: Lawrence Erlbaum.
- Steiger, J. H., & Lingd, J. M. (1980). *Statistically based tests for the number of factors*. Paper presented at annual meeting of psychometric society, Iowa City, IA.
- Stephenson, J. (1992). Capability and quality in higher education. In Stephenson, J., & Weil, S. (Eds). *Quality in learning*. Kogan Page, UK.
- Stevens, J. J., & Zvoch, K. (2007). Confirmatory factor analysis of the TerraNova comprehensive tests of basic skills/5. *Educational and Psychological Measurement, 67*(6), 976-989.
- Stringer, M., & Irwing, P. (1998). Students' evaluations of teaching effectiveness: a structural equation modeling. *British Journal of Educational Psychology, 68*, 409-426.
- Tabachnick, G., & Fidell, S. (2007). *Using multivariate statistics*. Boston: Allyn & Bacon.
- Whitman, D. S., Van Rooy, D. L., Viswesvaran, C., & Kraus, E. (2009). Testing the second-order factor structure and measurement equivalence of the Wong and Law emotional intelligence scale across gender and ethnicity. *Educational and Psychological Measurement, 69*(6), 1059-1074.
- Worthen, B. R., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Program evaluation, alternative approaches and practical guidelines*. New York: Longman.