The Validity and Reliability of the Student Evaluation of Teaching: A case in a Private Higher Educational Institution in Malaysia.

Tan Lay Khong
Taylor's Business School
Taylor's University
Malaysia.
Email: welman.tan@taylors.edu.my

Abstract

Most universities are using the Student Evaluation of Teaching (SET) as an instrument for students to assess a lecturer's teaching performance. It is an essential instrument to reflect the feedback in enhancing the quality of teaching and learning. The purpose of this paper is to examine the validity and reliability of the SET as a valid instrument in evaluating teaching effectiveness in a private higher education institution in Malaysia. Exploratory Factor Analysis and Confirmatory Factor Analysis have validated all 10 items of SET whereby all items indicated high reliability and internal consistency. A Confirmatory Factor Analysis using AMOS software also confirmed that a single factor model was used to evaluate teaching effectiveness. The single factor model was further validated using 1000 repeated samples of Bootstrap method in AMOS.

Keywords: Student Evaluation of Teaching, validity, reliability

1. Introduction

Most universities use the Student Evaluation of Teaching (SET) to assess a lecturer's teaching performance. It is extensively used in higher educational institution as an instrument to measure the quality of teaching effectiveness. In some higher institutions, SET is used as a tool for academic staff appraisal and promotion. In the UK, SET is renowned and used by the Quality Assurance Agency for Higher Education (QAA) as a documentation for subject reviewing (Shevlin et. al). In Australia, Course Experience Questionnaire (CEQ) is widely used as a performance indicator of teaching quality. Thus, it is an important instrument used to gauge the teaching effectiveness in universities (d'Apollonia & Abrami, 1997). It also helps to improve future teaching performance of the instructors (Sheehan and Duprey, 1999). Consequently, it is essential to evaluate the validity and reliability of SET in measuring teaching effectiveness.

Many literatures have clearly defined the SET constructs that measures the teaching effectiveness. One of a comprehensive review by Marsh (1991) concluded that effective teaching is a multidimensional construct. Ramsden (1991), Trigwell & Prosser (1991), Richardson (1994) and Wilson et al. (1997) had further developed the evaluation instrument and identified five scales of effective teaching in higher education. They are good teaching, clear goals and standards, appropriate workload, appropriate assessment and student independence. Ramsden and Wilson suggested a two-factor model of students' evaluation teaching. Trigwell & Prosser had also identified two-factor model: one construct reflecting good teaching, clear goals and student independence and the other reflecting appropriate assessment and workload. However, Richardson had extracted a single factor model with the same five scales.

Marsh (1987) stated that the first SET was published as far back as in 1915. Most SET at that time emphasized on the factors that influence the rating of teaching performance such as gender, age, course type, level of subject, class size, grade expectations, rank and experience of instructor, etc.(Marsh, 1987). Since then, not many studies were done on validity and reliability of the SET.

Studies on validity and reliability of SET were only examined, beginning in the 1970s (Centra, 1993). Kerlinger (1986) defined reliability as "relative absence of errors of measurement in a measuring instrument" and the items measured in the SET are internally consistent. On the other hand, validity is a measurement used to quantify the construction of a teaching evaluation (Messick 1989). Kerlinger (1986) stated that validity refers to the items being measured and whether the instrument measures what we want to measure. Since then, there have been numerous questionings and doubts on the validity and reliability of student's perceptions of teaching (Sproule, 2002). As such, the opinions of student evaluations vary from "valid, reliable and useful" to "invalid, unreliable and useless" (Marsh, 1984).

As teaching evaluation becomes more essential in higher educational institutions in assessing teaching effectiveness, the extent of the validity of SET instrument has important implications to the various stakeholders of an institution or university. Thus there is a need to have a valid, reliable and comparable performance data for the teaching quality improvement (Wilson et al., 1997). As a result, this study is aiming to examine the validity and reliability of the SET administered in the institution. In addition, this study also aims to explore a further statistical analysis in validating the SET instrument.

2. Instrument Validity

Most teaching evaluation questionnaires have not presented sufficient evidence of validity. If an instrument provides a measure of what it actually measures, validity is established. If the students give good ratings to effective teachers, the ratings are valid. This is done by determining the size of correlation between ratings and effectiveness of teachers.

There are two types of validity that are commonly used – content validity and construct validity. Content validity refers to whether the content of the questions or items measured in the SET are representative and adequate. It is arguable that in some testing, some questions in the questionnaires are not related to the intended subject of testing. The content validity would become a trivial issue if the questionnaires contain sufficient questions to address the construct to be tested.
(Fox 1994).

Construct validity, on the other hand, is considered as the most important aspect of validity studies. Marsh (1984) defines construct validity as items measured in student evaluations must be related to variables that are indicative of teaching effectiveness but if they are not related, it will lead to potential biases in the construct. Construct validity is further classified into two types – convergent validity and divergent validity. Convergent validity examines if the constructs of the items measured in student evaluations are highly related with each other whilst divergent validity tests of items in the individual construct are only related with the items in the constructs.

3. Historical Trends in Establishing Validity

Over the past several decades since 1970s, many researches have been focused on issues related to the validity of course evaluation or teaching evaluation questionnaires. Some studies appeared to have statistical evidence of validity. However, to date many researchers are still exploring a valid course evaluation questionnaire to measure the quality of teaching.

Prior to 1970, research published raised the concerns about the validity of student evaluation but there were no established statistical methods to validate the questionnaire. Exploration of validity started in the 1980s whereby there were numerous approaches used to validate a course evaluation questionnaire. Most researchers in 1980s started the validity by exploring the correlational construct validity (Greenwald, 1997). Marsh (1987), the pioneer in validity studies claimed that

construct validity is the best method in validating the course evaluation questionnaire. Again in 1991, Marsh commented that there were few well-constructed instruments that used factor analysis to support the construct validity in measuring the teaching effectiveness.

Beginning in the 1990s, numerous researchers started to use Exploratory Factor Analysis (EFA) to validate the course questionnaires. Kremer (1990) used EFA to discover set of constructs in teaching, research and service. Marsh (1991) recommended the usage of advance methodology and demonstrated the application of Confirmatory Factor Analysis (CFA) and Hierarchical CFA (HCFA) using LISREL V (Joreskog & Sorbom, 1988). Shevlin et al. (2000) used LISREL 8 (Joreskog & Sorbom, 1993) for CFA to develop a two-factor confirmatory factor model of teaching effectiveness, namely "lecturer attributes" and "module attributes'.

Wilson et al., (1997) further developed the Course Experience Questionnaire (CEQ) by using the application of Structural Equation Modelling (SEM). Their results confirmed the validity of the CEQ as a performance indicator of university teaching quality. Kember et al. (2008) also used CFA in assessing the nine constructs in the Exemplary Teacher Course Questionnaire (ETCQ).

However, no researcher has validated their instrument using the Bootstrap method. Bootstrap is a resampling technique, whereby numerous samples are taken from the original sample. Using this technique to validate a questionnaire does not need a huge sample as was relied upon in the past. This paper will therefore demonstrate the application of Bootstrap method in validating the SET.

4. Methods

A sample of 200 undergraduate students was taken from a private higher education institution, namely Taylor's University in Malaysia. The SET, a 10-item questionnaire, was administered among these students to evaluate their teaching and learning experiences in the programme. The items were measured on a 5 point Likert scale, where 1 indicated 'Strong Disagreement' and 5 indicated 'Strongly Agreement' to the statement.

The data were analyzed using statistical software, SPSS, version 11.5. First, the Cronbach's alpha coefficient was used to test the internal consistency. A Cronbach's alpha of more than 0.7 indicates that SET is reliable. Next, construct validity was tested to determine whether each item correlated adequately with at least

one of other item in the construct This was done by studying the correlation matrix among the 10 items. Exploratory Factor Analysis (EFA) was used to check if all 10 items can be reduced into a smaller dimension. Factor loading of 0.5 was used as the cut-off point. Confirmatory factor analysis (CFA) using AMOS was performed to test how well the measured items represent the number of constructs (Hair et al. 2010). In CFA, Chi-square/df, GFI, AGFI, TLI, CFI and RMSEA were used to test model acceptability.

Finally, the model was validated using 1000 repeated samples using the Bootstrap method in AMOS. The Bootstrap method, first introduced by Efron (1979), is a resampling technique whereby numerous subsamples are taken from the original sample, with replacement. The model is fitted within each sub-sample and the single model fit is tested against the bootstrap samples model fit summary. In this study, the Boolen-Stein p-value of more than 0.05 was taken as an indicator of "correctness" of model fit. In AMOS, the default model estimation procedure uses the maximum likelihood estimation method. This method requires the assumption of multivariate normality among the variables. It has been generally accepted that if the normalized kurtosis value (or the critical ratio) is more than 5, then there is a violation of the assumption of multivariate normality. In such situation, Yung & Bentler (1996) suggested the usage of bootstrap estimates rather than simply relying on the spurious estimates from the maximum likelihood method.

5. Results

The means and standard deviations (SD) for each item in the SET questionnaire are presented in Table 1. The mean values for all 10-items in the SET questionnaire are more than 3, indicating a general agreement to the statements in the questionnaire. The reliability analysis on the items in the SET questionnaire gave a Cronbach's alpha value of 0.908, which is much higher than the usual required value of 0.70 (Hair et. al., 2010). Thus, the 10-items SET questionnaire exhibit a very high internal consistency. The inter-item correlation values are presented in Table 1. The highest correlation coefficient for each item with at least one other item in the construct ranges from 0.3 to 0.9, indicating that the items in the construct correlated adequately.

Table 1: Mean scores and SDs of the 10-items SET questionnaire

	Descriptive statistics		Inter-item Correlation Matrix										Factor loadings
Item	Mean	SD	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	
Q1	3.81	0.792	1.000										.799
Q2	3.87	0.746	0.698	1.000									.773
Q3	4.1	0.796	0.597	0.615	1.000								.757
Q4	3.75	0.857	0.565	0.601	0.546	1.000							.766
Q5	3.93	0.832	0.405	0.381	0.467	0.516	1.000						.667
Q6	3.91	0.806	0.523	0.490	0.493	0.380	0.461	1.000					.717
Q7	3.43	0.922	0.333	0.308	0.318	0.401	0.330	0.448	1.000				.576
Q8	3.95	0.852	0.603	0.543	0.572	0.559	0.490	0.556	0.389	1.000			.801
Q9	3.76	0.881	0.539	0.526	0.529	0.538	0.455	0.513	0.456	0.632	1.000		.782
Q10	3.58	0.942	0.579	0.494	0.432	0.545	0.491	0.477	0.446	0.541	0.580	1.000	.754

In Exploratory Factor Analysis (EFA), the Kaiser-Meyer-Olkin (KMO) value was 0.921, which is considered to be excellent (Foulger, 2010). The Principal Component method was used in the extraction, with a default Eigen value of 1. Using this method, a single factor was extracted that explained 55% of the total variance in the 10 items. This is more than the minimum acceptance of 50%. The factor loadings are shown in the last column of Table 1. The minimum factor loading was 0.576, which is more than the minimum set value of 0.5 and this shows that the SET is constructed based on single factor model. The composite reliability value was 0.924, much higher than the desired minimum value of 0.7, shows that the SET construct is a reliable instrument.

The validity of a single construct 10-items model was further tested using Confirmatory Factor Analysis (CFA) in AMOS software. A single factor model (Figure 1) was found to be acceptable [Chi-square/df <3, GFI, AGFI, TLI and CFI all more than 0.9 and RMSEA less than 0.08]. The average variance extracted was 50.3% and the composite reliability was 0.909.

This confirms that the SET is a single factor model. The multivariate kurtosis value was 19.14 with a critical ratio value of 8.74. This condition necessitates the need for model validation using bootstrap resampling technique. In this study, the Bollen-Stine bootstrap p-value for 1000 bootstrap samples was 0.088, which is more than 0.05. Thus the model "correctness" is acceptable at 5% level of significance. The standardized regression weights, the 95% confidence limits and the p-values for the items in the construct are provided in Table 2.

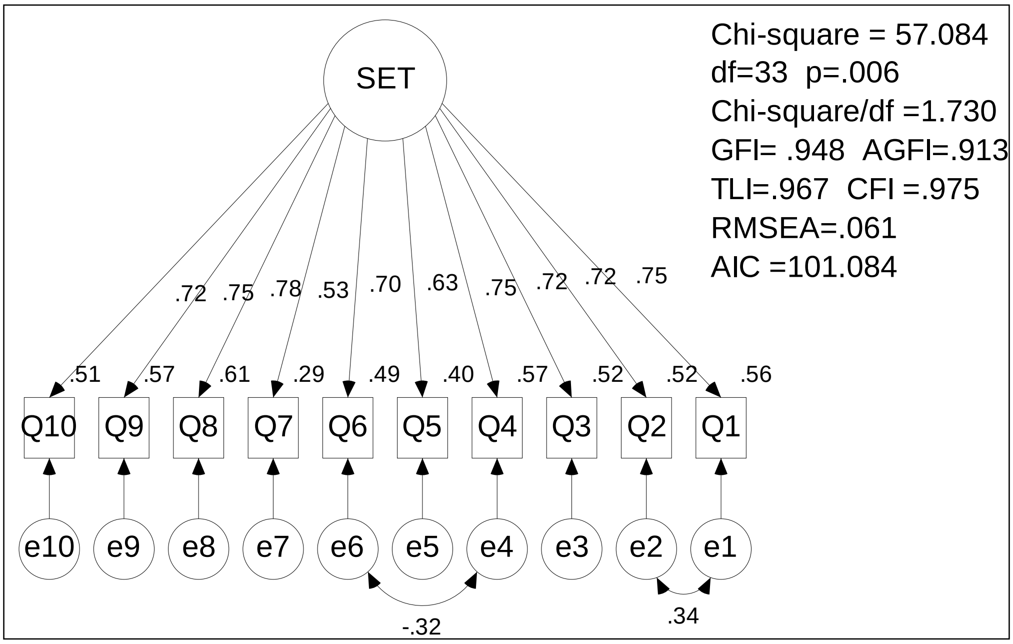

Figure 1 The AMOS output path diagram

Table 2: Standardized regression weights

Parameter			Estimate	95% CI		P
				Lower	Upper	
Q1	<---	SET	.751	.684	.811	.001
Q2	<---	SET	.723	.653	.784	.001
Q3	<---	SET	.720	.650	.784	.001
Q4	<---	SET	.755	.688	.821	.001
Q5	<---	SET	.631	.552	.700	.001
Q6	<---	SET	.700	.610	.773	.001
Q7	<---	SET	.534	.431	.623	.001
Q8	<---	SET	.779	.722	.828	.001
Q9	<---	SET	.752	.685	.817	.001
Q10	<---	SET	.715	.648	.773	.001

Generally, all the factor loadings are positive, high and statistically significant, indicating that the items used in the SET instrument are good indicators of assessing teaching effectiveness. All items except item Q7 have reasonable percentage of variation that explained the variation of the SET.

6. Conclusion

The conclusion of this study showed that the SET is a valid instrument in evaluating teaching effectiveness. All 10-items of the SET questionnaires were validated and exhibited a very high internal consistency and items were correlated adequately. Exploratory Factor Analysis and Confirmatory Factor Analysis confirmed that a single factor model was used in the SET. This model was further validated using Bootstrap method and it was confirmed that this model “correctness” was acceptable.

Though the single model factor was validated, the model could be argued that the SET has construct validity but lacks content validity as some of items (item 6, 7, 9, and 10) do not seem to measure teaching effectiveness. It is therefore suggested that a further study, particularly on content validity to be conducted so as to develop a more comprehensive SET instrument as a valid tool to measure teaching effectiveness in the institution.

7. References

[1] D'Apollinia, S. &. Abrami, P.C. (1997). Navigating student ratings of instruction, *American Psychologist*, 64, 431 - 441.

[2] Efron, B. (1979). Bootstrap methods: Another look at the jackknife. Annals of Statistics, 7, 1-26.

[3] Feldman, K.A. (1977) Consistency and variability among college students in rating their teachers and courses. *Research in Higher Education*, 10, 137 – 194.

[4] Foulger, D. and other participants. (June 1, 2010). Statistical Methods. MediaSpaceWiki. Retrieved on from http://evolutionarymedia.com/cgi-bin/wiki.cgi?StatisticalMethods.

[5] Fox, R. (1994). Validating lecturer effectiveness questionnaires in accounting. *Accounting Education*, 3 (3), 249 – 258.

[6] Greenwald, A. G. (1997). Validity Concerns and Usefulness of Student Ratings of Instruction. *American Psychologist*, 52(11), 1182 - 1186.

[7] Kember, D. and Leung, D.Y.P. (2008). Establishing the validity and reliability of course evaluation questionnaires. *Assessment & Evaluation in Higher Education*, 33(4), 341 - 353.

[8] Kerlinger, Fred N. (1986). *Foundations of Behavioural Research: Educational and Psychology Inquiry*, 3rd.edition. New York: Holt, Rinehard and Winston.

[9] Kremer, J.F. (1990). Construct Validity of Multiple Measures in Teaching, Research, and Service and Reliability of Peer Ratings. *Journal of Educational Psychology*, 82(2), 213 - 218.

[10] Marsh, H.W.(1984). Students, evaluations of university teaching: dimensionality, reliability, validity, potential biases and utility, *Journal of Educational Psychology*, 76(5), 707 - 754.

[11] Marsh, H.W..(1987). Students' evaluations of of university teaching: research findings, methodological issues, and directions for further research, *International Journal of Educational Research,* 11(3), 253 – 288.

[12] Messick, S. (1989). Validity In .R.Linn (Ed.) *Educational measurement* (4th ed., 13 - 103), New York: Macmillan Publishing.

[13] Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the course experience questionnaire, *Studies in Higher Education*, 16, 129 - 150

[14] Richardson, J.T.E. (1994). A British evaluation of the course experience questionnaire. *Studies in Higher Education*, 19, 59 - 68.

[15] Sheehan, E.P. & DuPrey,T. (1999). Student evaluations of university teaching. *Journal of Instructional Psychology*, 26, 188.

[16] Shevlin, M.; Banyard, P.; Davies, M. & Griffiths, M. (2000). The Validity of Student Evaluation of Teaching in Higher Education: love me, love my lectures? *Assessment & Evaluation in Higher Education*, 25(4), 397 - 405.

[17] Sproule, R. (2002). The underdetermination of instructor performance by data from the student evaluation of teaching, *Economics of Education Review*, 21, 287 – 294

[18] Trigwell, K. & Prosser, M. (1991). Improving the quality of student learning: the influence of learning context and student approaches to learning on learning outcomes. *Higher Education*, 22, 251 - 266.

[19] Wilson, K.L.; Lizzio, A. and Ramsden, P. (1997). The development, validation and application of the Course Experience Questionnaire. *Studies in Higher Education*, 22(1), 33 - 53.

[20] Yung, K.H., Bentler, P.M. (1996). Bootstrapping techniques in analysis of mean and covariance structures. In G.A. Marcoulides & R.E. Schumacker. *Advanced structural equation modeling: Issues and techniques.* Erlbaum N.J.

Appendix: Student Evaluation of Teaching Questionnaire (SET)

1. The outline and expectations for this course as supplied by the teacher were clear.
2. The lessons were organised and prepared.
3. The teacher was knowledgeable about the course content.
4. The course content was effectively presented.
5. Opportunities were provided for student participation.
6. The homework and classroom assignments were helpful.
7. The textbooks and/or recommended materials were useful.
8. The teacher was available for consultation and was helpful.
9. The assessment was fair.
10. This course met my needs and goals for future study and/or employment.