Bulletin of Education & Research
December 2007, Vol. 29, No. 2, pp. 21-32

Are Students' Evaluation of Teaching Valid? Evidence Form an Iranian Higher Education Institution

Naser Shirbagi*

Abstract

Because of the importance of assessing teaching effectiveness based on reliable and valid instruments, a study was conducted to identify dimensions of teaching effectiveness from the undergraduate students' perspectives. The study adopted an exploratory descriptive design. Respondents to participate in the research were 250 undergraduate students at an Iranian higher education institution. A twelve-item format questionnaire was the main instrument for data collection. The data collected was then subjected to factor analysis, and a model produced of teaching effectiveness and charisma. The findings suggest that how much students perceive their lectures as a charismatic teacher is an important predictor of Student Evaluation of Teaching (SET) scores. The study presents a challenge to the use of SET in higher education system in Iran and, in particular, raises questions of fairness if such ratings are to be used in decisions relating to employment issues. The finding suggests that SET be applied cautiously in faculty performance evaluation.

Keywords: *Student Evaluation of Teaching, Lecture Attributes, Charisma, Course Attributes*

Introduction

There are thousands of articles and books dealing with the extremely complex research on student evaluation of teaching (SET). Research in this area began as early as 1936 with Heilman and Armentrout work and continued unabated. The quantity of research is indicative of the importance of SET in higher education. While some studies have found that using SET is generally a valid method of assessment, others have found it to be flawed (Steiner *et al*., 2006). In spite of these inconclusive results and concern about them shared by many in academia, SET is still the most common measure of teaching effectiveness used today in Iran.

The practice of SET in universities is ubiquitous in many countries such as, the US and the UK. In the US, information from SET can be used for faculty decisions about conditions of employment such as salary and promotion. In the UK, information from SET is considered as important

*Department of Education, Faculty of Humanities, University of Kurdistan, Sanandaj, Iran

evaluative information, but also as a guide for potential changes in course material and method of delivery. In short, SET is an integral part of higher education practices in most countries.

Literature Review

Effort to isolate the variables explaining SET scores have been made for more than 40 years. Unfortunately, early efforts suffered from one or more serious shortcomings in the statistical methods used, and all research has been hampered to some extent by the unavailability of data from more than two or three consecutive semester (McPherson, 2006).

Shevlin *et al.*, (2000) pointed out that there are three ways to recognize good teachers. Firstly, might value teachers by their ability to effect personal development in their students. This is a long-term outcome and also attempting to quantify it is problematic. Secondly one might value teachers by their effectiveness in facilitating good academic work in their students. A third way of evaluating teachers is to ask their students to rate them. This is the most immediate and the most widely used of the three strategies and is commonly measured by questionnaire at the end of courses. One of the issues to consider is whether administrators are measuring the most important variables of teaching effectiveness or whether some variables are becoming more important just because they are measurable. Another important issue to consider is the validity of scores of teaching effectiveness gathered from student evaluations.

One of the most important benefits of student ratings worth mentioning here is that the process of designing or filling out the forms encourages teachers and students to reflect on their educational experiences, and as a result, develop clearer conceptions about what efforts they must make in order to achieve better teaching and learning results. There are basically two types of evaluation: *summative* and *formative*. Summative evaluations occur usually at the end of the teaching of a course and are used to calculate a final assessment. Formative evaluations, on the other hand, are nearly always feedback immediately to bring about changes while a course is being taught.

Despite the perceived importance of SET there are some concerns related to the assessment of teaching effectiveness that are yet unresolved. The primary concern about use of SET is the issue of whether or not they actually measure teaching effectiveness. There is little agreement as to what constitutes effective teaching. For example, Swartz *et al.*, (1990) identify two factors of effective teaching as (1) clear instructional presentation, and (2) management of student behavior. Further studies identify more and different factors of teaching effectiveness. For example, Patrick and Smart (1998) identify three factors of teaching effectiveness as (1) respect for students, (2) organization and presentation skills, and (3) ability to challenge students. Other researchers have suggested as many as seven factors

(Ramsden, 1991) or nine factors of effective teaching (Marsh & Dunkin, 1992).

The second major concern entails bias. A number of variables unrelated to teaching skills have been shown to affect SET in some studies (Cashin, 1995, Marsh & Roche, 1997, McKeachie, 1997). The relationships between ratings of teaching effectiveness and variables related to student characteristics, lecturer behavior, and the course administration have been examined (d'Apollonia & Abrami, 1997). For example, in relation to student characteristics, Marsh (1987) and Feldman (1976) reported a positive association between expected grades and ratings of teaching effectiveness. Further to this, Marsh & Roche (1997) reported similar relationships between ratings and the prior subject interest of the student and the reason for taking the course. The variable related to the lecturer behavior that has received the greatest research interest is that of grading leniency. Greenwald and Gillmore (1997) demonstrated that grading leniency had a strong positive relationship with ratings of teaching effectiveness. A further problem concerns the validity of the conclusions that are drawn from SET data due to the lack of statistical sophistication in the personnel committees that may use the information (McKeachie, 1997). Overall, research on the effects of extraneous variables on the validity of SET suggests the need for caution in the interpretation of this data. It would appear, then, that consensus on the characteristics of effective teaching is low, and there are a number of factors that challenge the validity of the data (Marsh & Roche, 1997). Also if students have a positive personal and/or social view of the lecturer this may lead to more positive ratings irrespective of the actual level of teaching effectiveness (Shevlin *et al.*, 2000)

Students may respond to central quality of leadership that then influences their evaluations of teachers. One approach to leadership that offers parallels to teaching is charismatic leadership. For example, House's (1976) theory of charismatic leadership emphasizes the relationship between the leader and the follower. Weber (1947) provided the most well-known definition of charisma as the special personality characteristics that gives a person superhuman or exceptional powers and is reserved for a few, is of divine origin, and result in the person being treated as a leader. According to charismatic approach the personal characteristics of a charismatic leader include being dominant, having a strong desire to influence other, being self-confidence, and having a strong sense of one's own moral values. In addition to displaying certain personality characteristics, charismatic leaders also demonstrate specific types of behavior that are set strong role model, show competence, articulate goals, communicate high expectations, express confidence and arouse motives. Bass (1985) provided a more expanded and refined version of transformational leadership that was based on, but not fully consistent with, the prior works of Burns (1978) and House (1976).

This model has four factors that includes: 1) idealized influence, 2) inspirational motivation, intellectual stimulation, and 4) individualized consideration. Factor 1 is called charisma. It describes leader who act as strong role model for followers; follower identity with these leaders and want very much to emulate them. These leaders are deeply respected by followers (Avolio, 1999). The distinctions between transformational leadership and charismatic leadership are not clear (Shackleton, 1995), and even if we make the distinction between these two, then the feature of Bass's model that has been found to have the greatest effect on satisfaction ratings is idealized influence or charisma (Bryman, 1992). The features of charismatic leadership and transformational leadership resemble the features of teaching effectiveness identified above (Patrick & Smart, 1998). Charisma has been shown to affect voter judgments of politicians (Pillai *et al.,* 1997), as well as leadership at work (Fuller *et al*., 1996). Because of the special features of the teacher's role, in which, they challenge, assess and motivate students; the impact of charisma in SET is further intensified (Woods, 1993).

Charisma is such a salient trait in students' perceptions of teachers that it affects assessment of teacher effectiveness. After reviewing relevant literature, a study was devised to examine the relationship between charisma and teaching effectiveness. It was supposed that the student's perception of the lecturer would significantly predict teaching effectiveness ratings. The main aim of this study was to determine whether a *halo effect*[1] occurs in the completion of SET ratings and to estimate the magnitude of this effect.

Method

Sample

The sample consisted of 250 undergraduate students at an Iranian mid-sized university. Students were all enrolled full-time on courses within *Faculty of Humanities Science*. Because of the anonymous and secrecy nature of the evaluation no details of demographic variables are available. The participants were required to rate their lecturer. In total, ten lecturers (eight males and two females) were rated during this study.

[1] The halo effect refers to a cognitive bias whereby the perception of a particular trait is influenced by the perception of the former traits in a sequence of interpretations. The halo effect is involved in Kelley's (1950) implicit personality theory, where the first traits we recognize in other people then influence the interpretation and perception of latter ones. Attractive people are often judged as having a more desirable personality and more skills than someone of average appearance.

Instrument

An eleven-item teaching effectiveness self-report scale[2] (Shevlin *et al.*, 2000) was administered to students by a member of administrating staff. The scale was designed to measure two dimensions of teaching effectiveness. Six items related to lecturer's attributes and measured the *lecturer ability* factor, and five items related to aspects of the particular module or course and measured the *module attributes* factor. Responses to the items were made on a five-point Likert scale anchored with *strongly agree* and *strongly disagree*. An addition item was included, *'The lecturer has charisma'*, which used the same response format as the other items. Descriptive statistics and coefficient alphas for measures after translation are shown in Table 1.

Table 1

Mean, standard diviations and coefficient Alphas of Measure

Measures	M	SD	No of Items	Alpha
Lecturer attributes	23.37	5.96	6	0.907
Course attributes	18.45	4.48	5	0.813
0verall	45.77	11.01	12	0.932

Table 1 shows that coefficient alphas of measure in all cases were in satisfactory level and the translated version of questionnaire has defensible reliability to use within Iranian Higher Education context.

Analyses and Results

Table 2

Intercorelation coefficients among charisma and teaching effectiveness factors

Item	M	SD	1	2	3	4	5	6	7	8	9	10	11
1	4.1	1.1											
2	3.8	1.1	.70										
3	3.8	1.2	.59	.62									
4	3.9	1.2	.60	.64	.64								
5	3.7	1.3	.66	.65	.68	.69							
6	3.7	1.2	.59	.55	.54	.56	.61						
7	3.5	1.1	.33	.42	.37	.36	.33	.33					
8	3.7	1.2	.34	.35	.36	.37	.31	.35	.42				
9	3.7	1.2	.66	.63	.69	.63	.69	.58	.42	.46			
10	3.7	1.2	.44	.51	.53	.56	.58	.52	.46	.42	.66		
11	3.9	1.3	.61	.60	.59	.51	.57	.53	.36	.41	.58	.49	
12	3.9	1.4	.60	.66	.64	.56	.64	.55	.39	.41	.63	.54	.80

Note: All coefficients are significant at 0.01 level, (N=250)

[2] With permission of authors, the original scale (English) was translated to Persian language using back translation approach. Comparison was made of the original and back translated version and no any important discrepancies in the translations were seen.

Table 2 reports means, standard deviation of items as well as correlation coefficients among items. In addition to correlation coefficients magnitudes reported, the significance of each was tested. In all instances, the coefficients were statistically significant at 0.01 level. The correlation coefficients (*r*) between item 12 and other items in all case were high and greater than 0.50 (except for item 12 with 7, and 12 with 8).

The model depicted in figure 1 is a graphical representation of the expected loading of our eleven-item teaching effectiveness scale. The model was drawn to indicate that we have two expected factors, lecturer attributes with six and course attributes with five expected factor loadings.

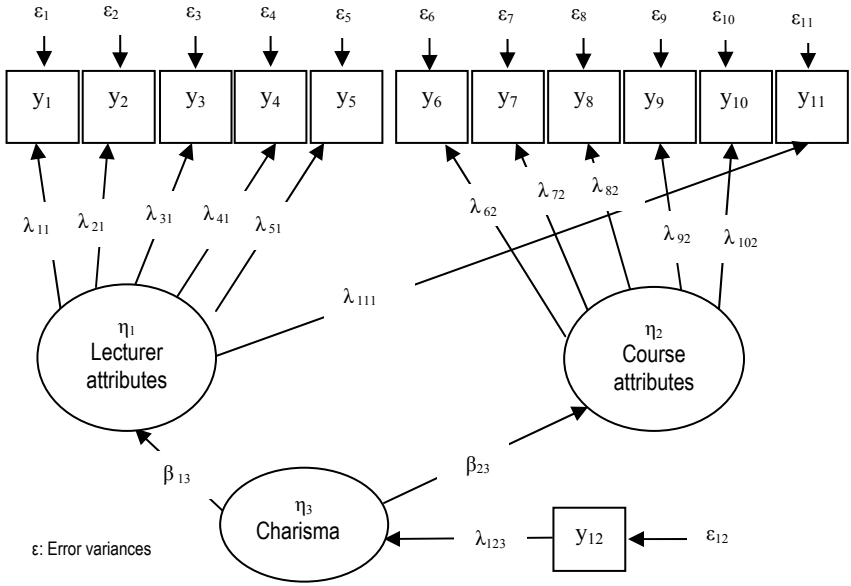

Figure 1: Path Diagram of Charisma and Teaching Effectiveness Factors Model

The model presented in Figure 1 (adapted with permission from Shevlin *et al*., 2000), was built in LISREL8.3 (Joreskog & Sorbom, 1999a). From the sample data a covariance matrix was computed using PRELIS2.3 (Joreskog & Sorbom, 1999b) using maximum likelihood. Figure 1 specifies a two-factor measurement model for the eleven items (y1 -y11) measuring student evaluations. The two factors, lecturer ability (η1) and module attributes (η2) are measured by their respective items in the self-report teaching evaluation scale. The factor loadings are given the symbol λ, and the error variances for each item the symbol ε. The lecturer ability (η1) and module attributes (η2) factors are regressed on the charisma factor (η3). The regression coefficients are symbolized as β. The charisma factor is measured

by a single item (y12). The model estimates can be used to determine the percentage of variation in the lecturer ability and module attributes factor that is attributable to the charisma factor.

The fit indices regarding the model of charisma and teaching effectiveness factors from the LISREL analyses are shown in table 3.

Table 3
Fit indices for charisma and teaching effectiveness factors model

Items	Values in present study	Values in Shevlin's *et al.* (2000) study
χ^2	185.65	114
Df	52	52
P	0.00	<0.05
90% Confidence Interval for NCP	(95.8 to 179.1)	-
Root Mean Squire Error of Approximation (RMSEA)	0.09	0.075
Goodness-of-Fit Index (GFI)	0.90	0.92
Comparative Fit Index (CFI)	0.93	0.94
Root Mean Squire Residual (RMSR)	0.066	-
Normal Fit Index (NFI)	0.91	-
Standardized Root Mean Square Residual (SRMR)	0.049	0.044
Incremental Fit Index (IFI)	0.93	0.94

Table 3 shows fit indices[3] for charisma and teaching effectiveness factors model. It should be noted that, good model fit is indicated by a non-significant chi-squire. Although the chi-square value for the model was large relative to the degree of freedom, and statistically significant, this should not lead to the rejection of the models as the large sample size increases the power of the test (Tanaka, 1987). Therefore, other fit indices must be considered. One of the popular fit index, refereed to as goodness-of-fit index (GFI), compare the relationship between the variable obtained from the sample with those hypothesized in the model. Good model fit is indicated by GFI greater than 90. Another fit index is comparative fit index (CFI). Obtained CFI values above 0.90 indicate acceptable model fit. While the χ^2 value suggests the model may not adequately fit the data, the CFI, GFI, and SRMR would suggest good model fit. That is, overall, the model is a reasonable description of the data.

The completely standardized solutions for path coefficients are reported in Table 4. The factor loading indicates that the items used in the teaching effectiveness self-report scale are good indicators of the lecturer ability and module attributes factors. All the factor loadings are positive, high and statistically significant.

[3] It is important to keep in mind that several fit indices are typically reported for any given model. Each available index of fit addresses a slightly different issue and therefore no index of fit is considered to be prefect.

Table 4
Standardized Parameter Estimates for Teaching Effectiveness Model

Parameter	Estimate in Present Study	Estimate in Shevlin's *et al.* (2000) Study
β_{13}	0.96*	0.83*
β_{23}	0.94*	0.61*
λ_{11}	0.78*	0.60*
λ_{21}	0.83*	0.76*
λ_{31}	0.82*	0.82*
λ_{41}	0.76*	0.77*
λ_{51}	0.77*	0.77*
λ_{62}	0.71*	0.53*
λ_{72}	0.63*	0.57*
λ_{82}	0.60*	0.54*
λ_{92}	0.86*	0.56*
λ_{102}	0.75*	0.74*
λ_{111}	0.79*	0.85*
λ_{123}	0.79*	0.67*

Note: *$p < 0.05$.

The standardized regression coefficients from the charisma factor to the lecturer ability (β_{13}) and module attributes factors (β_{23}) are 0.98 and 0.94, respectively. These effects are statistically significant ($P < 0.05$). Therefore, the charisma factor accounts for 92% of the variation of the lecturer ability factor and 88% of the module attributes factor. These results are consistent with the previous findings of Shevlin's *et al*., (2000) study.

Conclusions

The results of this research raise issues concerning the interpretation and utility of SET ratings. The SET ratings were demonstrated to be significantly affected by the students' perception of the lecturer as a result of that questioning the validity of this particular scale. Further, they raise questions about how the effect of confounding variables can be minimized by that means increasing the validity of SET ratings. However, the findings could be argued to be likely to generalize to most teaching assessment instruments on the basis of the prevalence of the halo effect.

The results indicate that a halo effect does indeed operate during the measurement of teaching effectiveness as the relationships between the charisma factor and the lecturer ability and module attributes were statistically significant. In fact, the effect is large with the charisma factor accounting for 92% and 88% of the variation in the lecturer ability and module attributes factors respectively[4]. That is, a significant proportion of the scale's variation is reflecting a personal view of the lecturer in terms of

[4] These values were 61% and 37% respectively, in Shevlin's *et al.,* (2000) study.

their charisma rather than lecturing ability and module attributes. In other words, the findings suggest that how much students perceive their lectures as a charismatic teacher is an important predictor of SET scores.

The two-factor structure of the scale, with high factor loadings, would appear to suggest that meaningful variables related to teaching quality were being measured. However, the important point is that the two factors are reflecting a positive halo effect as well as variance attributable to teaching quality. This raises questions in respect of the utility of using information from such scales since the attribute of charisma is having a central trait effect on student evaluations.

In short, this study presents a challenge to the use of SET in higher education system in Iran and, in particular, raises questions of fairness if such ratings are to be used in decisions relating to employment issues. This suggests that these results be applied cautiously in faculty performance evaluation. In this area several suggestions may prove helpful. It is clear that SET scores should not be used to make fine distinctions between faculty members, nor should they be used to rank-order them. The potential for bias based on a number of factors makes using SET in this way unfair. They should instead be used as a general guide to assessing teaching. Programs can consider using a combination of evaluation tools and averaging their results. Use of a variety of evaluation tools (e.g. self, peer) rather than relying solely on SET is necessary. Comprehensive and usable information may be provided for effective teaching. Universities should provide clear policy guidelines on quality control for faculties to develop multiple teaching effectiveness evaluation instruments. Programs might consider conducting research on their specific units to periodically assess what inappropriate variables seem to be influencing SET scores. If, for example, a given variable is appears to be negatively biased, then programs can statistically adjust scores to address the bias. It should be kept in mind that the activity of teaching is essentially one of the human interaction, and as such is inextricably tied to the student's perception of lecturer's personality. An evaluation of teaching effectiveness, however, must be based on outcomes.

Additional research on SET validity issue in higher education certainly would be useful. For example, more interesting alternative models can specify that may be examined in future research.

References

Avolio, B.J. (1999). *Full leadership development: Building the vital forces in organization*. Thousand Oaks, CA: Sage

Bass, B. M. (1985). *leadership performance beyond expectations, New York*: Free Press.

Bryman, A. (1992). *Charisma and Leadership in Organizations* (London, Sage)

Burns, J. M. (1978). *Leadership.* New York; Harper & Row

Cashin, W. E. (1995). Student rating of teaching: The research revisited/ IDEA Paper no.32.] Manhattan, KS: *Center for faculty Evaluation and development in higher education*, Kansas State University.

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

Feldman, K. A. (1976). Grades and college students' evaluations of their courses and teachers, *Research in Higher Education*, 18(1), pp. 3-124.

Fuller, J. B., Patterson, C. E. P., Hester, K. & Stringer, D. Y. (1996). A quantitative review of research on charismatic leadership, *Psychological Reports*, 78(1), pp. 271-287.

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

Heilman, J. D. & Armentrout, W.D. (1936). Are student rating of teachers affected by grades? *Journal of Educational Psychology*, 27, (March):197-216.

House, R. J. (1976). A 1976 theory of charismatic leadership, in: J. G. Hunt & L. L. Larson (Eds) *Leadership: The Cutting Edge*, pp. 189-207 Carbondale, IL, Southern Illinois University Press.

Jöreskog, K., & Sörbom, D. (1999a). *LISREL 8.30*, Chicago: Scientific Software Inc.

Jöreskog, K., & Sörbom, D. (1999b). *PRELIS 2.30*, Chicago: Scientific Software Inc.

Kelley, H. H. (1950). The warm-cold variable in first impressions of persons, *Journal of Personality and Social Psychology*, 18(3), pp. 431-439.

Lowman, J. & Mathie, V. A. (1993). What should graduate teaching assistants know about teaching? *Teaching Of Psychology*, 20(2), pp. 84-88

Marsh, H. W. (1987). Students' evaluations of university teaching: research findings, methodological issues, and directions for future research, *International Journal of Educational Resea*rch, 11(3), pp. 253-388.

Marsh, H. W. & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective, *American Psychologist*, 52(11), pp. 1187-1197

Marsh, H. W. & Dunkin C. M. (1992). Students' evaluations of university teaching: a multidimensional perspective, in: J. C. Smart (Ed.) *Higher Education: handbook on theory and research,* Vol. 8, pp. 143-234 (New York, Agathon Press).

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

McPherson, M. A. (2006). Determinants of how Students Evaluated Teachers, *Journal of Economic Education*: Winter, 37, 1:13-20.

Patrick, J. & Smart, R. M. (1998). An empirical evaluation of teacher effectiveness: the emergence of three critical factors, *Assessment and Evaluation in Higher Education*, 23(2), pp. 165-178.

Pillai, R., Stites, D., S., Grewal, D. & Meindl, J. R. (1997). Winning charisma and losing the presidential election, *Journal of Applied Social Psychology,* 27(19), pp. 1716-1726

Ramsden, P. (1991). A performance indicator of teaching quality in higher education: the course experience questionnaire, *Studies in Higher Education,* 16(2), pp. 129-150.

Shackleton, V. (1995). Business Leadership, (London, Routledge).

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 and Evaluation in Higher Education,* 25:397-405.

Steiner, S. Gerdes, K., Holley, L.C., & Campbell, H. E. (2006). Evaluation teaching: listening to students while acknowledging bias, *Journal of Social Work Educatio*n, 42, 2. 355- 376.

Swartz, C. W., White, K. P. & Stuck, G. B. (1990) The factorial structure of the North Carolina Teacher Performance Appraisal Instrument, *Educational and Psychological Measurement*, 50(1), pp. 175-185.

Tanaka, J. S. (1987). "How big is big enough?" Sample size and goodness of fit in structural.

Weber, M. (1947). *The theory of social and economic organizations*, (T. Parsons, Trans), New York: Free Press.

Woods, P. (1993). The charisma of the critical other: enhancing the role of the teacher, *Teaching and Teacher Education*, 9(8), pp. 545-557.