Hellenic Journal of Psychology, Vol. 8 (2011), pp. 22-43

FACTORIAL AND CONSTRUCT VALIDITY OF THE GREEK VERSION OF THE STUDENTS' EVALUATIONS OF EDUCATIONAL QUALITY SCALE

Angelos Rodafinos[1], Vassilis Barkoukis[2], & Haralambos Tsorbatzoudis[2]

[1]City College, Affiliated Institution to the University of Sheffield, Thessaloniki, Greece, & [2]Aristotle University of Thessaloniki, Greece

Abstract: The validity and the applicability of a Greek translation of the Students' Evaluations of Educational Quality (SEEQ) scale were examined in two studies. Study 1 assessed the factorial and construct validity of the Greek SEEQ version with 377 students of the Aristotle University of Thessaloniki. The patterns of relationships between the subscales and background items were similar to those reported in related past studies. Confirmatory factor analyses of responses to the SEEQ scale supported a correlated 9 first-order factors model. In Study 2, 317 students selected and rated a 'good' and a 'poor' teacher. The scores on the SEEQ subscales (except for the Workload/Difficulty subscale) differed between good and poor teachers. Multigroup confirmatory factor analyses supported the invariance of the factor structure across ratings of good and poor teachers. Results provided initial support for the validity and the applicability of the SEEQ scale in the Greek university context.

Key words: Educational quality, Evaluations of teaching, Teacher characteristics.

INTRODUCTION

More than 2,000 articles, reviews, and books have been written in the past four decades on the subject of students' ratings of teaching. Although student evaluations of teaching effectiveness (SET) are not and should not be the single and only source of data to make a valid judgment on teacher performance (Cashin, 1988),

Address: Angelos Rodafinos, Department of Psychology, City College, Affiliated Institution to the University of Sheffield, 24 P. Koromila Street, 54 622 Thessaloniki, Greece. Phone/Fax: +30-2310-269065. E-mail: rodafinos@city.academic.gr

SETs are certainly a useful tool, as they may serve a number of functions, including offering feedback to lecturers, guiding student decisions regarding course selection, guiding administrative decisions regarding personnel issues, directing research on teaching, etc. Thus, today SETs are used by universities in most of the developed countries. In fact, the majority of western educational institutions require student evaluations of staff in all courses. Even students themselves favor the use of evaluations and consider teachers who use them to be more responsible, committed to teaching, and motivated (Brown, 2008).

Several instruments have been developed to record students' evaluations of teaching effectiveness, including the Student Description of Teaching questionnaire, the Michigan State Student Instructional Rating System, the Endeavor Instructional Rating Form, the Course Experience Questionnaire, the Students' Evaluations of Educational Quality and others (see Cashin & Downey, 1992; Marsh, 1987, 2001; Tagomori & Bishop, 1995). For their development, most of these instruments used a systematic approach and came up with similar, albeit not identical, factors.

Important questions in the course of the development of these questionnaires included whether they would be biased by demographic and background variables (i.e., instructor, student, marks, and class characteristics). According to Cashin (1988), SETs are, in general, unrelated to a number of variables associated with the instructor (sex, age, teaching experience, personality, and research productivity), the student (sex, age, level, GPA, and personality), the course (class size, time of the day), and the administration (the time during the term when ratings are collected). On the other hand, there are controversial findings regarding whether ratings may be biased by a number of other variables related to the instructor (faculty rank and expressiveness), the student (prior interest or motivation to take the course and expected grade), the course (level and academic field), and the administration (presence of the instructor, the purpose of the rating, and anonymity) (also see Marsh, 2007a; Marsh & Roche, 1997; Maurer, 2006; Sprinkle, 2008; Wright & Palmer, 2006). For example, although Clayson (1999) suggested that instructors' experience and personality characteristics have positive effects on SET, Marsh (2007a) reported that university teachers do not become more effective with experience and that other characteristics of the individual teacher were unrelated to overall teaching effectiveness. In addition, no specific association with instructors' rank has been found, as rank is associated with age and experience (Sprinkle, 2008). Additionally, although variables such as prior interest in the course, grade (i.e., expected, past, difficulty to get a good grade) and administrative issues have been thought to influence SET, research so far has provided only marginal support to this

assumption (Sprinkle, 2008). Furthermore, regarding course issues, Marsh (2001, 2007b) compared the importance of the course being taught with the influence of the teacher, and concluded that it is rather *who* teaches the course than the nature of the course itself. Hence, research so far has indicated that these background variables do not consistently affect SET. As Sprinkle (2008) noted, often it is students' personal biases that affect SET rather than the teacher's quality of teaching.

One of the widely used scales is the Students' Evaluations of Educational Quality scale (SEEQ; Marsh, 1982a). The SEEQ scale is one of the most thoroughly developed and widely used instruments. It is a multidimensional measure, which comprises of nine factors, namely Learning/Academic Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Exams/Marking, Assignments/Readings, Workload/Difficulty (Marsh, 1983, 1984, 1987; Marsh & Hocevar, 1991). The SEEQ has shown high generalizability, stability (Marsh, 1984), and validity as assessed with a number of methods (over time, in different courses, by student learning, by instructors' self-ratings, ratings of administrators, colleagues, and alumni), and is relatively free from bias. Indeed, numerous studies analyzing a very large number of surveys completed by students and instructors in a number of English (Marsh, 1981, 1986, 1987; Marsh & Roche, 1992) or non-English speaking countries (Marsh, Hau, Chung, & Siu, 1997; Marsh, Touron, & Wheeler, 1985; Watkins, 1992; Watkins & Akande, 1992; Watkins & Gerong, 1992; Watkins & Regmi, 1992; Watkins & Thomas, 1991) have supported the construct, and specifically the convergent validity (Marsh, 1984; Marsh & Roche, 1997), and the applicability of the SEEQ scale (for an overview see also Watkins, 1994).

Evaluation of educational quality in Greece

Despite the universal trends to use SETs in educational institutions, the majority of Greek State university lecturers, perceive that their academic freedom is threatened (for a discussion see Haskell, 1997) and strongly oppose or, at best, are fairly skeptical about the idea of being evaluated either by students or by other sources, despite the fact that neither their salary nor their job depends upon their classroom teaching skills. Yet, for quality assurance reasons and following European Union directives, the evaluation of the quality of instruction has to be enforced in the Greek educational system as well. Thus, the need for a standardized measure is both apparent and urgent.

Characteristics of the education systems of certain countries may influence student evaluation (Husbands & Fosh, 1993). One of the differences in higher educa-

tion, for instance, is the fact that according to the Greek constitution, the only legal provider of tertiary education, until recently, has been the state. Students are accepted at the university on the basis of their marks on the national examinations. Once enrolled, a student may proceed to the next level without successfully completing all or any units, there is no maximum time period allowed for the completion of one's studies, while assessment is based on final exams on textbook material (assignments or coursework are rarely used). Furthermore, it is unclear how Greek students behave in the course evaluation process, since the concept of students evaluating teachers is completely novel to them. It is also possible that, much like Chinese students (Ting, 2000), in the evaluations Greeks may pay more attention to the personal qualities of their teachers.

The present study

Applying measures developed in one setting in another without examining their validity and applicability is certainly not safe by default (Marsh, 1981). Yet, using a widely used measure is clearly preferable to using an untried instrument or developing a new one. The SEEQ scale was selected in the present study, on the basis of its validity, the support it has received when used in different cultures, as well as on factors such as efficiency, cost-effectiveness, and practicality.

Hence, the aim of Study 1 was to examine the factorial (i.e., the degree to which the measure of a construct conforms to the theoretical definition of the construct; Messick, 1995) and construct validity (i.e., the subscales will be interrelated as previous research has shown) of a Greek translation of the SEEQ scale. Based on previous evidence, it was assumed that the SEEQ scale will show factorial and construct validity (Hypothesis 1). The relationships between SEEQ scales and background variables (expected mark, past average mark, perceived difficulty to get a good mark, and prior interest level) were also examined. Based on previous research, it was expected that SEEQ scores will be relatively independent of background variables (Hypothesis 2).

The factorial invariance of the SEEQ factors across students' evaluations for good and poor teachers was tested in Study 2, to examine whether the questionnaire is invariant between these two groups. A secondary aim of Study 2 was to test for differences between the two groups in the SEEQ subscales. Marsh (2007a) indicated that although there might be substantial individual differences between teachers in terms of their teaching effectiveness, these differences were also highly stable over time, suggesting that the factors of the questionnaire are invariant. Thus, it was hypothesized that the questionnaire factors will be invariant across the two groups

(Hypothesis 3). It was also expected that good teachers, compared to poor ones, will receive higher ratings in SEEQ subscales (Hypothesis 4).

STUDY 1

Sample – Procedure

A total of 377 undergraduate physical education students completed the Greek version of the SEEQ. Of them, 163 were male (42.1%), 186 female (48.1%), and 38 did not specify their gender. The sample represents approximately 30% of the enrolled and active[1] students in the specific department.

The Greek version of the SEEQ scale (Marsh, 1982a, 2001) was administered by the researchers shortly before the end of the academic term to 20 randomly selected classes via the lottery method, taught by nine different instructors at the Department of Physical Education and Sports of Aristotle University of Thessaloniki, Greece. Courses at this Department include theoretical (e.g., sport psychology, biochemistry, didactic, motor learning) and practical (e.g., basketball, track and field, gymnastics) subjects. The SEEQ scale was administered in practical subjects where attendance is compulsory.

Standard printed material included instructions, consent forms explaining the purpose of the study and the rights of the participants to withdraw. Students were also informed about the confidentiality of their responses and did not provide their own names. Due to the random selection of the practical courses, it is likely that some students have completed more forms than others. In this case, they have evaluated the different instructors in the different courses. The completed forms were collected and placed in a folder by the teacher or a student in class to be submitted to the researchers, and were subsequently processed and analyzed.

Measures

The SEEQ scale. The development of the SEEQ scale is described in Marsh (1982a). It is an instrument developed to examine university students' evaluations

[1] Students may be enrolled but not active; that is, students may be enrolled and choose not to attend any compulsory practical courses but be examined in theoretical, non compulsory, courses. On the other hand, active students participate in the required activities of the institution, such as practical, compulsory, and theoretical courses.

of the teaching process. A number of studies have supported the scale's nine-factor structure (e.g., Marsh, 1983, 1984, 1987; Marsh & Hocevar, 1991). Marsh (1984) provided evidence on the reliability of the scale with Cronbach's alphas ranging from .70 to .90. Specifically, the nine factors (and their alphas in parentheses) were the following: Learning/Academic Value (.83), Instructor Enthusiasm (.82), Organization/Clarity (.74), Group Interaction (.90), Individual Rapport (.82), Breadth of Coverage (.84), Exams/Marking (.76), Assignments/Readings (.70), and Workload/Difficulty (.70). Furthermore, Marsh (1984) provided evidence on the construct validity and temporal stability of the scale.

The majority of the factors are comprised of four items, apart from the Assignments/Readings factor, which has two items. A nine-point scale from 1 (strongly disagree) to 9 (strongly agree) measures responses, apart from the Workload/Difficulty factor that uses an idiosyncratic scale with items for Difficulty (from 1 = very easy to 9 = very hard), Workload (from 1 = very light to 9 = very heavy), Pace (from 1 = too slow to 9 = too fast), and Hours needed to study (from 0 = none to 9 = nine or more hours).

Example items[2] for each of the nine factors, in the order presented above, include: "You have learned and understood the subject materials in this class," "Lecturer was dynamic and energetic in conducting the class," "Lecturer's explanations were clear," "Students were encouraged to participate in classroom discussions," "Lecturer was friendly towards individual students," "Lecturer contrasted the implications of various theories," "Methods of evaluating student work were fair and appropriate," "Required readings/texts were valuable to students," and "Average number of hours per week required outside class was ...".

The SEEQ scale was translated to Greek following the International Test Commission (ITC) guidelines for the adaptation of educational and psychological tests for cross-cultural assessment (also see Hambleton, 2001). Accordingly, two bilingual speakers translated the SEEQ scale to Greek and another two bilingual speakers back-translated the scale to English. Both translations were subsequently examined for their linguistic equivalence by a panel of researchers, who were familiar with the literature and the procedures followed for the adaptation of questionnaires in other languages.

The final form of the Greek version of the SEEQ scale was pilot-tested with a group of 101 secondary education teachers attending continuing education programmes. The teachers completed the inventory and were asked to suggest modifications, if they thought that certain items were unclear or non-applicable. Teachers

[2] A copy of the complete scale in Greek may be obtained by contacting the first author.

confirmed the face and content validity of the SEEQ scale, and verified that the SEEQ scale can be read and understood by ninth graders. Minor changes were suggested to modify certain expressions and replace a few words in order to achieve linguistic equivalence.

Student measures. Demographic and background variables, namely gender, expected mark, past average mark, perceived degree of difficulty to get a good mark, prior interest level, and overall evaluation were also collected. Expected and past average mark were measured on a response scale from 1-4 (fail) to 10 (excellent). Perceived degree of difficulty was measured on a response scale from 1 (very easy) to 9 (very difficult). Prior interest was measured on a response scale from 1 (very low) to 9 (very high). Finally, overall evaluation of the course compared to other courses was given on a response scale from 1 (very bad) to 9 (very good).

Data analysis

The SPSS 12.0 software was used to examine descriptive statistics of the students' background variables and scale's items, and correlations between the subscales and the background variables. Confirmatory factor analyses were conducted using the EQS 6.1 statistical package (Bentler, 2004). The data were initially examined for normality. In most cases skewness values were lower than 2.0, suggesting normal distribution of the data (Bollen & Long, 1994). However, in several occasions kurtosis scores were above 2.0. Thus, the maximum likelihood robust technique was used to address the possibility of non-normal distribution (Cantoni & Ronchetti, 2006). Two models were tested: a measurement model, assuming that the factors of SEEQ were not correlated to each other, and a structural 9-factor correlated model, assuming that the SEEQ factors were correlated to each other. Both absolute and incremental fit indices were used to evaluate the models tested, such as the comparative fit index (CFI), the Bentler-Bonett normed (NFI) and nonnormed fit index (NNFI), the standardised root mean squared residual (SRMR), the root mean square error of approximation (RMSEA) and the chi square to degrees of freedom (χ^2/df) ratio. The Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA) were used as more focal indices to evaluate the adequacy of models as they are not influenced by sample size (Fan, Thompson, & Wang, 1999). A cut-off value of .90 or above for the CFI is typically considered an acceptable criterion for model fit, although a value greater than .95 shows excellent fit (Hu & Bentler, 1999). A cut-off value of .08 or below for the RMSEA was considered appropriate for satisfactory model fit (Hu & Bentler, 1999).

Results

Means, standard deviations, and Cronbach's alphas for each SEEQ subscale and background variables are presented in Table 1. Cronbach's alpha reliability estimates ranged from .71 to .85, apart from the Workload/Difficulty scale (Cronbach's $\alpha = .63$).

Table 1. Descriptive data of the Greek SEEQ version and background variables

	M	*SD*	Cronbach's α
SEEQ subscales			
Academic Value	7.38	1.00	.75
Instructor Enthusiasm	7.87	1.10	.85
Organization/Clarity	7.48	1.19	.84
Group Interaction	7.51	1.27	.88
Individual Rapport	8.08	.91	.80
Breadth of Coverage	7.35	1.23	.85
Exams/Marking	7.54	1.22	.85
Assignments/Readings	7.42	1.38	.71
Workload/Difficulty	5.27	1.34	.63
Background Variables			
Expected Mark	7.84	1.37	
Past Average Mark	7.49	.99	
Difficulty to get a good mark	5.51	1.65	
Prior Interest Level	6.76	1.81	
Overall evaluation	7.88	1.13	

The results of the confirmatory factor analysis indicated that the measurement model produced poor fit to the data, $\chi^2(323, N = 328) = 2073.326, p < .001$, CFI = .65, NNFI = .61, AGFI = .52, SRMR = .35, RMSEA = .13. On the other hand, the 9-factor correlated model offered a borderline model fit; although its χ^2 was significant, $\chi^2(302, N = 328) = 797.222, p < .001$, most incremental and absolute fit indices were at acceptable levels, CFI = .90, NNFI = .89, AGFI = .81, SRMR = .05, RMSEA = .07. Given that the chi-square is heavily influenced (adversely) by large samples (due to excessive power) the fit indices were solely used for evaluation of model fit. Thus the model fit was considered to be adequate, nevertheless borderline. These findings suggest that Hypothesis 1 regarding the factorial validity was not fully supported. However, the correlated 9-factor model is evidence for construct validity. The results of the analysis revealed that the eight factors measuring quality of education (i.e., Learning/Academic Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Exams/Marking, Assignments/Readings) were strongly related to each other (*r*

ranged from .46 to .79) and weakly with the last one (i.e., Workload/Difficulty) with *r* ranging from .01 to .18.

The correlated 9-factor structure was also supported by factor loadings all of which were statistically significant at .01 and all (but two) exceeded .40 in standardized values, thus contributing significantly to the assessment of the construct of interest. In addition, the Root Mean Squared Error of Approximation was .07, with 90% confidence intervals ranging between .065 and .077. The range of the standardized item loadings was .218 to .875. The decomposition of the confirmatory factor analysis, as well as the descriptive statistics of each item, is presented in Table 2. Apart from the residuals of items 12 and 20 (r = .26) no other residuals were correlated.

Table 3 presents the correlations among SEEQ subscales and background variables. The results of the correlation analysis revealed that the eight factors measuring quality of education (i.e., Learning/Academic Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Exams/Marking, Assignments/Readings) were moderately related to the Overall Evaluation (r ranged from .37 to .47). Correlations with the ninth factor, Workload/Difficulty, were low (r ranged from .11 to .23) and in the case of Exams/Marking nonsignificant. This is evidence that only partly confirmed Hypothesis 1.

In a similar manner, the correlations between SEEQ subscales and the background variables measuring Interest level, Expected, Past, and Difficult Marks were also low ($r < .23$), and in some cases nonsignificant, with the exception of Average Mark, which was moderately correlated to the Workload/Difficulty subscale (r = .52). These findings confirmed Hypothesis 2.

STUDY 2

Sample – Procedure

The Greek version of the SEEQ scale was administered again around the end of the first academic semester, to 317 undergraduate physical education students of the Department of Physical Education and Sports of Aristotle University of Thessaloniki, Greece. Of them, 180 were male (56.8%), 135 female (42.6%), and three did not specify their gender. Senior and sophomore (third and fourth level) classes were selected for the administration of questionnaires, to ensure that students would have substantial experience from having attended more courses and

Table 2. Descriptive statistics and item decomposition of the Greek SEEQ version in Study 1

SEEQ items per subscale	*M*	*SD*	Skewness	Kurtosis	Item residual
Learning/Value					
Challenged and stimulated	7.42	1.26	-.93	1.92	.800
Learned something valuable	7.67	1.24	-1.35	2.53	.795
Increased subject interest	7.27	1.40	-.77	.23	.694
Understood subject matter	7.13	1.39	-.84	1.30	.751
Instructor Enthusiasm					
Enthusiastic about teaching	7.82	1.33	-1.36	2.46	.623
Dynamic and energetic	8.06	1.21	-1.81	4.94	.644
Enhanced with humor	7.76	1.43	-1.44	2.63	.662
Held your interest	7.82	1.32	-1.70	4.71	.498
Organization/Clarity					
Teacher explanations clear	7.73	1.35	-1.38	3.00	.637
Materials explained and prepared	7.47	1.42	-1.20	1.84	.483
Objectives stated and pursued	7.51	1.33	-1.10	1.36	.675
Facilitated taking notes	7.26	1.55	-.93	.88	.740
Group Interaction					
Encouraged class discussion	7.42	1.48	-1.26	2.25	.648
Students shared ideas	7.34	1.65	-1.37	2.29	.608
Encouraged questions and answers	7.51	1.46	-1.23	1.91	.494
Encouraged expression	7.76	1.35	-1.49	3.08	.601
Individual Rapport					
Friendly to individuals	8.21	1.06	-1.66	3.80	.738
Welcomed seeking help	8.18	1.04	-1.47	2.36	.585
Interested in students	8.01	1.15	-1.43	2.74	.976
Accessible to students	7.83	1.30	-1.32	2.28	.792
Breadth of Coverage					
Contrasted implications	7.20	1.48	-.92	1.08	.670
Gave background of ideas	7.37	1.33	-.94	1.27	.535
Gave different views	7.46	1.49	-1.26	2.31	.631
Gave current developments	7.42	1.56	-1.07	.98	.970
Examinations/Marking					
Feedback valuable	7.49	1.41	-1.33	2.74	.646
Evaluation methods fair	7.60	1.43	-1.36	2.33	.532
Tested course as emphasized	7.50	1.34	-1.13	2.05	.552
Assignments/Readings					
Readings were valuable	7.19	1.51	-.83	.72	.672
Contributed understanding	7.62	1.46	-1.39	2.69	.703
Workload/Difficulty					
Difficulty (easy-hard)	5.61	1.77	-.52	.33	.672
Workload (light-heavy)	5.89	1.88	-.51	.20	.703
Pace (slow-fast)	5.68	1.32	.55	1.36	.878
Hours of study	3.87	2.56	.43	-.91	.916

Table 3. Correlations between the subscales of the Greek SEEQ version and background variables

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Learning/Value		.59[b]	.57[b]	.51[b]	.48[b]	.55[b]	.47[b]	.43[b]	.14[b]	.11[a]	.01	.09	.23[b]	.46[b]
2. Instructor Enthusiasm			.66[b]	.61[b]	.61[b]	.62[b]	.51[b]	.39[b]	.11a	.14[a]	.03	.08	.05	.37[b]
3. Organization/Clarity				.61[b]	.59[b]	.64[b]	.64[b]	.47[b]	.15[b]	.16[b]	.04	.10	.08	.45[b]
4. Group Interaction					.55[b]	.59[b]	.54[b]	.42[b]	.12a	.11[a]	.07	.12[a]	.05	.38[b]
5. Individual Rapport						.54[b]	.57[b]	.41[b]	.03	.15[b]	-.02	.01	.08	.41[b]
6. Breadth of Coverage							.64[b]	.57[b]	.23[b]	.10[a]	.10[a]	.14[b]	.20[b]	.47[b]
7. Examinations/Marking								.56[b]	.05	.19[b]	.05	.07	.12[a]	.47[b]
8. Assignments/Readings									.20[b]	.17[b]	.07	.19[b]	.22[a]	.45[b]
9. Workload/Difficulty										-.10[a]	.24[b]	.52[b]	.20[a]	.04
10. Expected Mark											.29[b]	-.10	.14[b]	.06
11. Average Mark												.14[b]	.20[b]	.03
12. Difficult Mark													.18[b]	.10
13. Interest Level														.28[b]
14. Overall Evaluation														

Note: [a] $p < .05$; [b] $p < .01$. Values without a superscript represent nonsignificant scores.

interacted with various instructors. The sample represented approximately half of the enrolled and active student population at these levels.

Following a similar methodology to that used in previous related studies (Marsh, 1986; Marsh et al., 1997; Watkins, 1994; Watkins & Thomas, 1991), this time students were requested to think of one good and one poor teacher whose course they had attended during their studies at the university. The students were requested to select teachers from practical courses, in which participation was compulsory, in order to ensure that the courses would have similar requirements and characteristics, and that students would have formed their opinion having attended a class during the whole semester, rather than just a couple of lectures[3]. Students were then asked to complete two copies of the SEEQ scale, one for each teacher. Demographic variables assessed in Study 2, both for descriptive purposes and to ensure that any differences in the two groups of good and poor teachers were not a result of differences in these variables, included Class size, Course grade, and Age and Gender of the teacher. An additional 'not appropriate' response allowed students to mark items that were deemed to be inappropriate. For example, if there were no presentations that facilitated taking notes (Item 12) the students could mark the 'not appropriate' response. Anonymity was ensured in that students provided neither their own names nor the names of teachers who they nominated as good and poor.

[3] Students may choose to attend one or two lectures only in non-compulsory theoretical courses.

Data analysis

As in Study 1, descriptive statistics of the students' demographic variables and differences in the evaluations of good and poor teachers were examined via the SPSS 12.0. A Mann-Whitney test was used to examine differences in the distribution of male and female teachers in the good compared to the poor teacher groups. Paired sample *t*-tests were conducted to test for differences between good and poor teachers in demographic variables and in their evaluations by students.

To investigate the extent to which the SEEQ scale fit the data equally for poor and good teachers a series of confirmatory factor analyses were computed (Byrne, 1994; Cheung & Rensvold, 2002). Confirmatory factor analyses were conducted to test invariance between good and poor teachers using the EQS 6.1 statistical package (Bentler, 2004). First, the model fit was examined for the good (Model 1) and poor teachers (Model 2) data separately. Then, the model fit was evaluated through a series of models that placed increasing equality constraints on the data. Model 3 was a baseline model in which the fit of the model to the data for good and poor teachers was evaluated simultaneously with no equality constraints. In Model 4 the factor loadings were constrained to be equal for good and poor teachers and in Model 5 we tested the equivalence of item intercepts of those items found to have invariant factor loadings in Model 4. Models' invariance was initially tested through differences in chi-squares. However, due to chi-square's sensitivity differences on CFI values (ΔCFI) between the multi-sample models were used to evaluate models' invariance. A ΔCFI value less than or equal to .01 is an index of model's invariance (Cheung & Rensvold, 2002).

Results

Overall, students perceived the majority of items to be appropriate in evaluating teacher effectiveness. Specifically, 'not appropriate' responses in the majority of the items consisted less than 2% of the total sample responses for the particular item, apart from a couple of items (e.g., value of feedback), which still did not exceed 6%.

Paired *t*-tests ensured that the two groups of good and poor teachers were not evaluated differently as a result of certain demographic variables. Specifically, there were no significant differences in the age of teachers of the two groups, $t(291) = .86$, *ns*, and the size of classes taught by good and poor teachers, $t(391) = -.92$, *ns*. Furthermore, a Mann-Whitney test indicated no significant gender differences in the distribution of males and females teachers in the good compared to the poor teacher groups, $\chi^2 = -1.57$, *ns*. Male teachers in both groups represented approximately 70-

Table 4. Differences between good teachers and poor teachers in the items of the Greek SEEQ version

	Good teachers			Poor teachers			
SEEQ items per subscale	α	*M*	*SD*	α	*M*	*SD*	*t*(df)
Learning/Value	.80	6.93	1.29	.83	4.45	1.90	21.44 (313)
Challenged and stimulated		7.08	1.59		4.74	2.30	15.59 (308)
Learned something valuable		7.34	1.51		4.82	2.28	18.12 (309)
Increased subject interest		6.72	1.79		3.82	2.31	18.69 (304)
Understood subject matter		6.59	1.72		4.32	2.26	15.80 (303)
Instructor Enthusiasm	.81	7.24	1.36	.86	3.57	1.87	26.86 (315)
Enthusiastic about teaching		7.20	1.75		3.76	2.18	21.61 (313)
Dynamic and energetic		7.48	1.59		4.21	2.25	20.41 (310)
Enhanced with humor		6.90	1.81		3.37	2.36	20.61 (309)
Held your interest		7.40	1.54		2.92	2.09	28.67 (306)
Organization/Clarity	.78	7.01	1.33	.86	3.96	1.88	23.23 (311)
Teacher explanations clear		7.48	1.42		4.05	2.19	22.82 (307)
Materials explained and prepared		7.02	1.75		4.08	2.21	18.01 (306)
Objectives stated and pursued		7.16	1.57		4.30	2.27	17.75 (298)
Facilitated taking notes		6.34	2.03		3.34	2.17	17.62 (280)
Group Interaction	.83	6.90	1.41	.89	3.72	1.82	23.95 (311)
Encouraged class discussion		6.71	1.74		3.64	2.07	19.74 (304)
Students shared ideas		6.64	1.78		3.74	2.10	18.20 (301)
Encouraged questions and answers		7.06	1.78		3.81	2.09	20.04 (305)
Encouraged expression		7.24	1.58		3.75	2.12	22.80 (303)
Individual Rapport	.84	7.27	1.31	.87	3.67	1.81	26.92 (312)
Friendly to individuals		7.44	1.55		3.90	2.15	22.91 (310)
Welcomed seeking help		7.56	1.47		3.74	2.04	24.64 (306)
Interested in students		7.20	1.60		3.58	2.13	23.12 (303)
Accessible to students		6.85	1.76		3.43	2.12	20.60 (286)
Breadth of Coverage	.77	6.55	1.34	.87	3.91	1.86	19.99 (300)
Contrasted implications		6.29	1.72		3.94	2.17	13.81 (280)
Gave background of ideas		6.60	1.54		4.03	2.09	16.62 (276)
Gave different views		6.75	1.83		4.11	2.18	15.42 (287)
Gave current developments		6.55	1.80		3.72	2.18	16.53 (271)
Examinations/Marking	.76	6.94	1.36	.82	3.84	1.96	21.98 (299)
Feedback valuable		6.83	1.58		3.84	2.18	18.23 (282)
Evaluation methods fair		6.90	1.84		3.79	2.31	16.98 (273)
Tested course as emphasized		7.04	1.55		3.89	2.24	19.21 (273)
Assignments/Readings	.77	6.86	1.58	.86	4.58	2.26	14.59 (295)
Readings were valuable		6.87	1.77		4.62	2.44	13.10 (294)
Contributed understanding		6.88	1.66		4.55	2.38	13.80 (274)
Workload/Difficulty	.60	4.96	1.46	.69	4.79	1.99	1.44 (311) *ns*
Difficulty (easy-hard)		5.62	2.08		5.37	2.51	1.10 (305) *ns*
Workload (light-heavy)		5.06	2.11		4.71	2.51	13.80 (303)
Pace (slow-fast)		5.54	1.37		5.16	2.85	13.10 (287)
Hours of study		3.33	2.59		3.93	2.82	13.80(182) *ns*
Overall rating items	.62	7.64	1.29	.68	3.88	1.91	27.80 (312)
Overall course rating		7.23	1.79		4.40	2.27	17.29 (308)
Overall teacher rating		8.04	1.31		3.38	2.13	31.03 (311)
	Demographic variables						
Class size (students' number)		27.86	18.44		29.20	21.08	-.92 (301) *ns*
Course grade		7.98	1.46		6.04	1.86	12.80 (254)
Teacher age (years)		46.84	6.43		46.20	7.94	.86 (291) *ns*
Teacher gender (% female)*		24.30			29.10		-1.57 *ns*

Note. Items are paraphrased (for the full text see Marsh, 1987). α = Cronbach's alpha. All *t* values were significant at $p < .001$, unless otherwise indicated (*ns*). Teacher's gender comparison was calculated with χ^2.

75% of the sample. However, students reported receiving significantly better grades in courses taught by the good teachers (M = 8.01, SD = 1.44) compared to those taught by poor teachers (M = 6.07, SD = 1.87), $t(251) = 12.80, p < .001$.

Cronbach's alpha for each SEEQ subscale were calculated for both good and poor teachers (see Table 4). These were adequate and ranged between .76 and .89, apart from the Workload/Difficulty factor for good (.60) and poor (.69) teachers.

When tested independently both Models 1 (good) and 2 (poor teachers) provided borderline but adequate fit to the data (see Table 5). The factor structure was equivalent across the two data sets. The chi square difference tests indicated that additional equality constraints posed on item factor loadings (Model 4) and the equivalence of item intercepts of those items found to have invariant factor loadings (Model 5) did not significantly decrease the model fit. The results of Model 4 indicated that all items were found to be invariant across the two groups. These findings satisfy the criteria regarding factorial invariance (Bollen, 1989), and indicate that the scale is invariant across good and poor teachers. The fit indices of the models are presented in Table 5. The decomposition of Model 4, examining the invariance between good and poor teachers, is presented in Table 6. The above findings confirmed Hypothesis 3.

Table 5. Goodness-of-fit indices for good and poor teacher invariance models of the Greek SEEQ version

χ^2	$p <$	NNFI	CFI	SRMR	RMSEA	$\Delta\chi^2$	Δdf	ΔCFI	$p <$
Model: Good teachers									
(398, N = 249) = 687.49	.001	.88	.912	.06	.05	--	--	--	--
Model: Poor teachers									
(398, N = 226) = 833.94	.001	.88	.900	.07	.05	--	--	--	--
Model: Good and poor teachers together									
(793, N = 200) = 1477.75	.001	.89	.909	.04	.05	--	--	--	--
Model: Factor loadings constrained to be equal across teachers									
(824, N = 200) = 1576.76	.001	.88	.900	.04	.06	99.02	31	.009	.001
Model: Equivalence of factor loadings of item intercepts for those items found to be invariant (all questionnaire items)									
(851, N = 200) = 1655.35	.001	.87	.895	.05	.06	78.76	27	.005	.001

Note. NNFI = Non Normed Fit Index; CFI = Comparative Fit Index; SRMR = Standardized Root Mean Residual; RMSEA = Residual Mean Square Error of Approximation

Paired t-tests were used to compare students' evaluations for good and poor teachers. As a large number of comparisons were performed, the Bonferroni adjustment procedure was used to adjust the significance level to reduce the possibility of Type 1 error (the likelihood of showing group differences of statistical significance when they are not actually there). Mean scores and standard deviations for each SEEQ item are presented in Table 4 together with the t values. Results are reported for differences at the Bonferroni adjusted level of .001 or lower. Good teachers,

Table 6. Decomposition of invariant model for good and poor teacher of the Greek SEEQ version

SEEQ items per subscale	Good teachers	Poor teachers
Learning/Value		
Challenged and stimulated	.789	.669
Learned something valuable	.869	.778
Increased subject interest	.756	.765
Understood subject matter	.723	.640
Instructor Enthusiasm		
Enthusiastic about teaching	.809	.781
Dynamic and energetic	.860	.797
Enhanced with humor	.745	.716
Held your interest	.839	.774
Organization/Clarity		
Teacher explanations clear	.820	.716
Materials explained and prepared	.869	.840
Objectives stated and pursued	.826	.760
Facilitated taking notes	.631	.663
Group Interaction		
Encouraged class discussion	.751	.726
Students shared ideas	.775	.791
Encouraged questions and answers	.887	.873
Encouraged expression	.852	.781
Individual Rapport		
Friendly to individuals	.841	.740
Welcomed seeking help	.880	.828
Interested in students	.859	.839
Accessible to students	.780	.736
Breadth of Coverage		
Contrasted implications	.727	.720
Gave background of ideas	.839	.784
Gave different views	.768	.751
Gave current developments	.796	.771
Examinations/Marking		
Feedback valuable	.820	.735
Evaluation methods fair	.795	.725
Tested course as emphasized	.822	.754
Assignments/Readings		
Readings were valuable	.865	.784
Contributed understanding	.906	.898
Workload/Difficulty		
Difficulty (easy-hard)	.676	.623
Workload (light-heavy)	.856	.780
Pace (slow-fast)	.841	.623
Hours of study	.834	.712

compared to poor teachers, received higher scores in nearly all SEEQ subscales (except difficulty and hours of study out of class). The above findings confirmed Hypothesis 4.

DISCUSSION

The factorial validity and the applicability of a Greek translation of the Students' Evaluations of Educational Quality were examined in the present studies. Following the recommendations of Marsh (1986), a confirmatory factor analysis was used to investigate the structure underlying responses to the SEEQ items. Results of the confirmatory factor analysis in Study 1 provided preliminary support for the correlated 9-factor structure of the Greek version of the SEEQ scale. It is important to note that almost all the factor loadings were substantial. Only two factor loadings (i.e., for items "Gave current developments" and "Interested in students") were below .40. This might be due to cultural or education related issues. Perhaps Greek students perceive these concepts differently, or the findings are related to the practical character of the courses taught. Students in the Greek undergraduate educational system are usually assessed by a final exam, on material based on a single textbook, while often no additional sources (e.g., articles) are provided by the lecturers. Thus, it could be argued that students are not aware of current scientific developments in order to evaluate whether their teachers keep up with research. Furthermore, because of the small and non-representative (from a single department) sample of students, the results of the confirmatory factor analysis may be a little dubious; further investigation of factorial validity may be needed in future studies.

Furthermore, the results supported the construct validity of the Greek version of the SEEQ scale. More specifically, inter-factor correlation coefficients indicated modest relationships among SEEQ subscales ranging between .39 and .66, suggesting that the SEEQ factors are independent, although interrelated, constructs. Similar interrelations among SEEQ subscales (r ranged from .17 to .57) were reported by Marsh (1982b), apart from the interrelations with the Workload/Difficulty factor, the majority of which were nonsignificant. In the present study the Workload/Difficulty factor was positively but less strongly (compared to others) correlated to most factors, except Individual Rapport and Exams/Marking. The Workload/Difficulty factor was also unrelated to the Overall Evaluation item.

In terms of background variables that assessed marks (expected, average past mark, degree of difficulty to get a good mark) and prior interest level, these were not particularly related to SEEQ subscales. Overall, the general patterns of relations among SEEQ subscales and background variables were in line with findings from past research using the SEEQ scale. Specifically, not surprisingly, and similar to research findings in Western and Chinese universities (Marsh et al., 1997; Marsh &

Roche, 1992), Expected Mark and Workload/Difficulty were negatively correlated (r = -.10), denoting that students tend to perceive courses in which they earned poorer grades as somewhat more difficult. Results also indicated that students considered it difficult to receive good marks in courses regarded as high in Workload/Difficulty. The 'bias' hypothesis states that low in workload or easy courses receive higher ratings. Past studies (e.g., Marsh et al., 1997) have found «small relations with potential biasing factors» (e.g., the workload or the difficulty of a course) and suggested that «these should not be interpreted as biases» (p. 569). The absence of a significant correlation between Workload/Difficulty and Overall course Evaluations in the present study indicates the absence of such bias in student evaluations. However, the results of the study were correlational in nature and as such they should be treated with caution. More sophisticated designs might provide more conclusive evidence regarding the 'bias' hypothesis. Finally, students' initial Interest Level correlated positively with Academic Value, Breadth of Coverage, Exams/Marking, and Workload/Difficulty, but not with the rest of the SEEQ factors.

Data from Study 2 indicated that, comparable to past research in various countries (e.g., Hong Kong, Nepal, India, New Zealand, Nigeria and Philippines; see Watkins, 1994), no item was seen to be inappropriate by more than 6% of the students. Specifically, in regard to appropriateness, Value of Feedback was an item that was considered not appropriate by a few students. As Marsh et al. (1997) suggested this may be due to the fact that, as mentioned above, most courses in Greek tertiary education are assessed by only a final examination that typically is not returned to students. It is also possible that some students selected a teacher that only taught a component of a course that was not examined.

Data from study two also examined the invariance of the Greek version of the SEEQ scale across good and poor teachers. Specifically, a series of confirmatory factor analyses were computed to examine the extent to which the SEEQ scale fit the data equally for poor and good teachers. Confirmatory factor analyses indicated that the scale is invariant across ratings of good and poor teachers providing evidence that the psychometric properties of the SEEQ scale are equal across these two groups. These findings suggest that students' ratings had equivalent meaning between good and poor teachers. More specifically, students identified the same factors of quality of teaching in both good and poor teachers' ratings. These findings imply that the SEEQ scale is a useful instrument as it is not affected by the level of the teaching process (good or poor). As such, it could be effectively used in several occasions, such as evaluation of existing curricula, interventions examining the effectiveness of different educational curricula, and monitoring a teacher's personal improvement.

Second, when ratings of good and poor teachers were compared in study two, not surprisingly, good teachers were rated more favorably than poor teachers in nearly all SEEQ subscales. Differences were quite large for factors such as Individual rapport, Enthusiasm, Group interaction, and Organization, while they were smaller for Assignments/Readings and Breadth of Coverage. The size of these differences may very well have been augmented due to the halo effect created by the selection of good/poor teacher procedure, a finding also observed in past studies (Marsh, 1986). On the other hand, good teachers did not differ from poor teachers in Workload/Difficulty items, such as Difficulty and Hours of study out of class, and in demographic variables, such as Class size, teacher Age, and teacher Gender. Thus, similar to previous findings, it appears that Workload/Difficulty is not one of the major factors that differentiate good and poor teachers. Past studies have indicated that good teachers tended to teach slightly smaller classes, be a little younger and more likely to be female (Marsh et al., 1997). The absence of an effect due to the age of the teacher and the class size in the present study may very well be the result of the small variance in the age of instructors and the number of students per class (no more than 20) in the particular department. Yet, a significant effect was evident due to another background variable, that is, Course grade. Students recalled receiving significantly better grades in courses taught by good teachers compared to those taught by poor teachers, a factor that may have affected the students' perceptions of teaching quality.

Limitations

One of the main limitations of both studies is that the sample of the students selected was not particularly large neither did it represent a broad cross-section of the students in the particular university. Thus, there is a possibility that these results may not generalize well to students studying in other departments or universities. Past studies have employed substantially larger samples, using data from SETs collected over years or even decades in thousands of classes in North American universities. However, this was not feasible as no such records exist in the Greek universities. In fact, this was probably one of the few data collection efforts using a standardized SET instrument. In any case, the present small-scale data collection may still serve as a useful pilot study for future larger studies.

Another limitation might be the borderline factorial validity of the SEEQ scale found in Study 1. This result could be, once again, due to the small size of the sample, which, as mentioned, was drawn from a single educational discipline. Clearly, future studies should expand these finding incorporating both practical and theo-

retical courses, as well as other than physical education and sport sciences disciplines. Also, future studies may wish to consider the halo effect, that is, the tendency to form an overall positive (or negative, in the case of a reverse halo effect) impression of a person based on a particular outstanding trait or quality.

Finally, a further potential limitation of the study might be that, although the related correlation appeared to be quite small, we did not control the effect of student grades on teaching evaluations (leniency hypothesis). Gump (2007), in an extensive review, reports contradictory findings concerning this hypothesis, that is, so far it is not clear whether receiving a good mark in a course will affect the perceptions of teaching quality in a certain way. Future research designs should control for grade effects by isolating related variables that may influence students' perceptions of teaching evaluations.

Conclusions

In summary, findings from the two studies were consistent with related studies with different nationalities and provide initial factorial and construct validity evidence for the Greek version of the SEEQ scale. However, given that validation is an ongoing process where evidence needs to be collected from a number of sources and samples to strengthen and support the validity of scale scores (Messick, 1995), more studies should be conducted to examine additional validity aspects of the scale. In regard to the background variables, SEEQ factors were not strongly associated with variables assessing Marks (Expected, Average past mark, degree of Difficulty to get a good mark) and prior Interest level. Finally, the SEEQ factors were found to be invariant across good and poor teachers, although mean differences in favour of good teachers were found.

REFERENCES

Bollen, K. A. (1989). *Structural equations with latent variables.* New York: Wiley.

Bollen, K., & Long, J. S. (1994). *Testing structural equation models.* Newbury Park, CA: Sage.

Brown, M. J. (2008). Student perceptions of teaching evaluations. *Journal of Instructional Psychology, 35*(2), 177-181.

Bentler, P. M. (2004). *EQS: A structural equations program.* Encino, CA: Multivariate Software, Inc.

Byrne, B. M. (1994). *Structural equation modelling with EQS and EQS/Windows: Basic concepts, applications, and programming.* London: Sage.

Cantoni, E., & Ronchetti, E. (2006). A robust approach for skewed and heavy-tailed out-

comes in the analysis of health care expenditures. *Journal of Health Economics, 25*(2), 198-213.

Cashin, W. E. (1988). *Student ratings of teaching: A summary of the research* (IDEA paper no. 20). Manhattan, KS: Kansas State University, Center for Faculty Evaluation and Development.

Cashin, W. E., & Downey, R. G. (1992). Using global student rating items for summative evaluation. *Journal of Educational Psychology, 84,* 563-572.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indices for testing measurement invariance. *Structural Equation Modeling, 9,* 233-255.

Clayson, D. E. (1999). Students' evaluation of teaching effectiveness: Some implications of stability. *Journal of Marketing Education, 21,* 68-75.

Fan, X., Thompson, B., & Wang, L. (1999). The effects of sample size, estimation methods, and model specification on SEM fit indices. *Structural Equation Modeling: A Multidisciplinary Journal, 6,* 56-83.

Gump, S. E. (2007). Student evaluations of teaching effectiveness and the Leniency hypothesis: A literature review. *Educational Research Quarterly, 30,* 55-68.

Haskell, R. E. (1997). Academic freedom, tenure, and student evaluation of faculty: Galloping polls in the 21st century [Electronic version]. *Education Policies Analysis Archive, 5*(6). Retrieved February 19, 2008, from http://olam.ed.asu.edu/epaa/v5n6.html

Hambleton, R. K. (2001). The next generation of the ITC test translation and adaptation guidelines. *European Journal of Psychological Assessment, 17*(3), 164-172.

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1-55.

Husbands, C. T., & Fosh, P. (1993). Students' Evaluation of Teaching in Higher Education: Experiences from four European countries and some implications of the practice. *Assessment and Evaluation in Higher Education, 18*(2), 95-114.

Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. *Australian Journal of Education, 25,* 177-192.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology, 52,* 77-95.

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology, 74*(2), 284-279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology, 75,* 150-166.

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

Marsh, H. W. (1986). Applicability paradigm: Students' evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology, 78,* 465-473.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings,

methodological issues, and directions for future research [Special Issue]. *International Journal of Educational Research, 11,* 253-388.

Marsh, H. W. (2001, June). *Students' evaluations of university teaching.* Workshop presentation at Minho University, Braga, Portugal. Retrieved June 15, 2007, from http://apps.uws.edu.au/uws/edc/seeq/SETs_HerbMarsh_presentation_2001.pdf

Marsh, H. W. (2007a). Do university teachers become more effective with experience? A multilevel growth model of students' evaluations of teaching over 13 years. *Journal of Educational Psychology, 99,* 775-790.

Marsh, H. W. (2007b). Students' evaluations of university teaching: A multidimensional perspective. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in higher education: An evidence based perspective* (pp. 319-384). New York: Springer.

Marsh, H. W., Hau, K. T., Chung, C. M., & Siu, T. (1997). Students' evaluation of university teaching: Chinese version of the Students' Evaluations of Educational Quality instrument. *Journal of Educational Psychology, 89,* 568-572.

Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education: An International Journal of Research and Studies, 7,* 9-18.

Marsh, H. W., & Roche, L. A. (1992). The use of student evaluations of university teaching in different settings: The applicability paradigm. *Australian Journal of Education, 36,* 278-300.

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52,* 1187-1197.

Marsh, H. W., Touron, J., & Wheeler, B. (1985). Students' evaluations of university instructors: The applicability of American instruments in a Spanish setting. *Teaching and Teacher Education, 1,* 123-138.

Maurer, T. W. (2006). Cognitive dissonance or revenge? Student grades and course evaluations. *Teaching of Psychology, 33,* 176-179.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50,* 741-749.

Sprinkle, J. E. (2008). Student perceptions of effectiveness: An examination of the influence of student biases. *College Student Journal, 42*(2), 276-293.

Tagomori, H. T., & Bishop, L. A. (1995). Student evaluation of teaching: Flaws in the instruments. *Thought and Action, 11*(1), 63-78.

Ting, K. (2000). Multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*(5), 637-661.

Watkins, D. (1992). Evaluating the effectiveness of tertiary teaching: A Hong Kong perspective. *Educational Research Journal, 7,* 60-67.

Watkins, D. (1994). Student evaluations of teaching effectiveness: A cross-cultural perspective. *Research in Higher Education, 35,* 251-266.

Watkins, D., & Akande, A. (1992). Student evaluations of teaching effectiveness: A Nigerian investigation. *Higher Education, 24,* 453-463.

Watkins, D., & Gerong, A. (1992). Evaluating tertiary teaching: A Filipino investigation. *Educational and Psychological Measurement, 52,* 727-734.

Watkins, D., & Regmi, M. (1992). Student evaluation of tertiary teaching: A Nepalese investigation. *Educational Psychology, 12,* 131-142.

Watkins, D., & Thomas, B. (1991). Assessing teaching effectiveness: An Indian perspective. *Assessment and Evaluation in Higher Education, 16,* 185-198.

Wright, R. E., & Palmer, J. C. (2006). A comparative analysis of different models explaining the relationship between instructor ratings and expected student grades. *Educational Research Quarterly, 30,* 3-18.