CUHK Education Journal I <>/. 22 No.2 pp. 291-304

A Cross-Cultural Investigation into Students' Evaluation of University Teaching

Wen-Ying LIN
The Chinese University of Hong Kong

David WATKINS
The University of Hong Kong

Qing-Mao MENG
Beijing Normal University

This study uses the applicability paradigm of Marsh (1981) to examine the validity of two evaluation instruments and their underlying model of teaching effectiveness across seven countries with diverse cultures and higher education systems. The results from the seven studies support the reliability, appropriateness, and to some degree convergent and discriminant validities of the two instruments. Similar patterns of item salience and discrimination between good and poor lecturers are also obtained. Hence, the similarity of the results from diverse academic settings generally lends support to the applicability and the cross-cultural validity of these two instruments and their underlying model of teaching. In addition, the finding that Hong Kong, Taiwan, and China are each relatively more similar to the West than among themselves may reflect the fact that their higher education systems are to a certain extent modeled after those of the West.

Students' evaluations of lecturers, which have been common in North American universities and colleges over the last twenty years, are designed to measure teaching quality. The literature on students' evaluations of teaching effectiveness (SETE) is composed of thousands of studies, dating back to the 1920s and earlier. In their recent review, Marsh and Dunkin (1992) spelt out the characteristics of SETEs as follows: (a) SETEs are multifaceted; (b) SETEs are reliable and stable; (e) SETEs are mainly a function of the instructor who teaches the course rather than the course that is taught; (d) SETEs are valid indicators of effective teaching; (e) SETEs are relatively unaffected by a variety of variables hypothesized as potential biases to the ratings; and (f) SETEs are considered useful by lecturers as feedback for teaching improvement, by students as information for course selection, and by adiministrators as input for personnel decisions.

The fact that SETEs are multifaceted is well supported by numerous empirical studies (Marsh, 1987). Many of these studies contain a number of well-constructed. US-developed evaluation instruments with clearly defined factor structures which are both theoretically hypothesized and empirically tested to provide measures of distinctive multifacets of teaching effectiveness. As Marsh (1987) pointed out, most of these instruments were developed using

a systematic approach and their identified factors were quite similar, thus rendering further support for their construct validity.

However, limited attempts have been made to investigate the applicability of these evaluation instruments, or the generalizability of related research findings, to students of countries with cultures and educational contexts seemingly rather different from those of the West. In fact, researchers from Third World countries have long cast doubt on the assumption that Western educational and psychological theories and measuring instruments are appropriate for non-Western subjects (Enriquez, 1977). To their disappointment, all too often some unsophisticated researchers apply, rather blindly, a test or an instrument developed in one culture and administer it in another without demonstrating the relevance of the construct or the validity of the instrument for the latter culture. In addressing this issue, Hui and Triandis (1985) and Watkins (1994) argued that for many research purposes one needs to demonstrate that the construct being measured in one culture is embedded in the same network of constructs in the same way as it is in another culture. Tentative assumptions that the operationalized constructs and the instruments are cross-culturally applicable are made and subsequently tested. In Cronbach's (1977) terminology, this involves testing both within-construct and between-construct portions of the nomological network. The underlying rationale of the argument is that if a construct has the same meaning in different cultures, it should result in the same empirical relationships. In addition, if the networks in different cultures are similar, it can be claimed that instruments used in this validation process are cross-culturally applicable and thus equivalent.

Hence, this study uses the applicability paradigm of Marsh (1981) to investigate the within-construct and between-construct aspects of two well-known, US-developed evaluation questionnaires, the Students' Evaluation of Educational Quality (SEEQ; Marsh, 1981) and the Endeavor (Frey, Leonard, & Beatty, 1975), and their underlying model of teaching effectiveness for seven countries with diverse cultures and systems of higher education. According to Berry's review (1989), there are two fundamentally different approaches to cross-cultural research: the "emic" and the "etic." The former employs concepts that derive from within a particular culture, while the latter seeks to compare different cultures on what are thought to be universal. Thus, this study is etic in nature since it aims at, by working comparatively across cultures, seeking support for the cross-cultural validity of the two instruments and understanding the similarities and differences among the educational contexts involved.

The SEEQ Instrument

The Students' Evaluation of Educational Quality (SEEQ) instrument, developed and discussed at length by Marsh (1981, 1987), measures a broadly representative set of evaluation factors and has strong factor analytic support. The nine evaluation factors, which the SEEQ was designed to measure, are Group Interaction, Learning/Value, Workload/Difficulty, Examination/Grading, Individual Rapport, Organization/Clarity, Instructor Enthusiasm, Breadth of Coverage, and Assignments/Readings. These factors are supported by more than 40 factor analyses. See Marsh (1984, 1987) for a detailed review of research developing the SEEQ and demonstrating the reliability and validity of the responses to this instrument.

The Endeavor Instrument

The Endeavor questionnaire was developed by Frey (1973) to measure seven components of effective teaching. The seven factors, identified through the use of factor analysis in different settings (Frey, 1973; Frey, 1978; Frey, Leonard, & Beatty, 1975), are Class Discussion, Student Accomplishment, Workload, Grading/Examinations, Personal Attention, Presentation Clarity, and Organization/Planning. See Marsh (1984, 1987) for a review of research demonstrating the reliability and validity of the responses to this instrument.

The SEEQ and the Endeavor were independently developed and do not measure the same number of components of effective teaching. However, an examination of the item content indicates that there is considerable overlap in the dimensions measured by the two instruments (see Table 1). There seems to be a one-to-one matching between the first five SEEQ factors and the first five Endeavor components (or scales). On the other hand, Organization/Clarity from the SEEQ appears to combine Organization/Planning and Presentation Clarity from the Endeavor. The remaining three

components from the SEEQ do not seem to correspond to any components from the Endeavor. The overlapping pairs of scales serve as the basis of a multitrait-multimethod (MTMM) convergent/ discriminant analysis, which will be discussed later in this paper.

Table 1: Pairs of Corresponding Scales in SEEQ and Endeavor Instruments

SEEQ Scales	Endeavor Scales
1. Learning/Value	1. Student Accomplishments
2. Group Interaction	2. Class Discussion
3. Individual Rapport	3. Personal Attention
4. Examinations/Grading	4. Grading
5. Workload/Difficulty	5. Workload
6. Organization/Clarity	6. Presentation Clarity
	7. Organization/Planning

The Applicability Paradigm

The applicability paradigm (Marsh, 1981) is used here to assess the applicability of the SEEQ and the Endeavor to students from a range of teritary institutions in a number of countries. The paradigm was first employed by Marsh (1981) to study the applicability of the two instruments to students at the University of Sydney in Australia. Specifically, students from 25 departments were requested to choose one of the best and one of the worst lecturers who had taught them and evaluated each on an evaltotion survey combining the items from the two instruments. As part of the study, students were asked to point out inappropriate items and to choose up to five items that they "felt were most important in describing either positive or negative aspects of the overall learning experience in this instructional sequence." This paradigm was laterf used in 12 other studies. In each of the studies, the analysis of the responses involves finding out which items best differentiate between good and poor lecturers, determining which items are most important and inappropriate, conducting a multitrait-multimethod (MTMM) analysis of construct validity of SEEQ and Endeavor scales, and in some cases correlating the scale scores with lecturer/class characteristics (e.g., lecturer's age and class size).

While most of the 13 studies generally support the reliability and validity of the two instruments and the multidimensional model of teaching effectiveness on which they are based, none of these studies can justify a claim of cross-cultural validity. According to Marsh (1986), comparison of results of each study should not only tell about the cross-cultural validity of the instruments but also provide a basis for understanding similarities and differences among the educational contexts involved. As a result, a number of studies along this line have been conducted, such as Marsh (1986), Marsh and Roche (1991), and Watkins (1994). These three studies generally conclude that the two instruments and the multidimensional model on which they based are appropriate in a wide variety of educational settings, and reveal some overall similarity in perception of teaching effectiveness. In particular, the data patterns for the West, such as the studies for Australia (Marsh, 1981; Hayton, 1983; and Marsh & Roche, 1991), New Zealand (Watkins, Marsh, & Young, 1987), and Spain (Marsh, Touron, & Wheeler, 1985), are very similar to one another, which indicates the greater campus/cultural similarities in these studies. On the other hand, among those non-Western studies, the data patterns for Hong Kong (Watkins, 1992) and India (Watkins & Thomas, 1991) appear more similar to those of the West, whereas those for Nepal (Watkins & Regmi, 1992), Papua New Guinea (Clarkson, 1984), Nigeria (Watkins & Akande, 1992), and the Philippines (Watkins & Gerong, 1992) are less similar to those of the West.

However, none of the studies by Marsh (1986), Marsh and Roche (1991), and Watkins (1994) contained results from the applicability studies conducted in China (Lin, Watkins, & Meng, 1994) and Taiwan (Lin et al., 1994). As both China and Taiwan have just begun to put emphasis on student evaluation, few published researches on the quality of the measuring instruments are available for use with their students. Hence, it is interesting to find out, through comparisons of China's and Taiwan's data patterns with those of the West, the cross-cultural validity and the applicability of the two instruments to students in China and Taiwan. In addition, although Hong Kong, Taiwan, and China have a lot in common, each still possesses its own individuality in terms of, for example, its government and education system. Thus, of particular interest is the question of whether Hong

Kong, Taiwan, and China are more similar in data patterns among themselves or more similar to those of the West.

Aims of Research

Specifically, the aims of this study are to investigate the following:

(a) To compare the internal consistency reliability of SEEQ and Endeavor scales in the seven cultures;

(b) To compare the convergent and discriminant validities of SEEQ and Endeavor scales through the analysis of modified multitrait-multimethod matrices in the seven cultures. Support for the multidimensionality of student ratings will be provided if discriminant validity meets the usual MTMM criteria;

(e) The appropriateness of the items of the two instruments will be examined across all seven countries;

(d) The relative importance of the questionnaire items for evaluating teaching quality will be compared across the seven countries;

(e) The items which best differentiate between good and poor lecturers in each country will be compared;

(f) The relationships between perceived quality of teaching and lecturer/class characteristics will be compared across the seven countries.

The Countries Sampled

The seven countries from which college students were sampled represent a range of levels of educational and economic development and cultural heritages. Except those for Taiwan and China, all evaluation surveys were done in English. Their systems of higher education are briefly described below.

Australia

Higher education system in Australia closely resembles those of other advanced English-speaking countries, especially Britain, Canada, and New Zealand. Before 1989, postsecondary education consisted of three major sectors: universities, colleges of advanced education (CAEs), and technical and further education (TAPE) schools. Since 1989, the former two sectors have been unified under the national system of higher education, emphasizing research and postgraduate study.

The subjects were 158 undergraduates from the University of Sydney, 30 percent of whom were enrolled in a course on human growth and development and the remainder were recruited on an ad hoc basis in various campus libraries, student union, and departmental lounges (see Marsh, 1981).

New Zealand

In close proximity to Australia, New Zealand has a university system quite similar to that of Australia in terms of standard and educational philosophy. The 119 subjects sampled in Watkins, Marsh, and Young (1987) were social science undergraduates enrolled at the University of Canterbury, which emphasizes research and postgraduate study (see Watkins et al, 1987; Marsh and Roche, 1991).

Nepal

Nepal is one of the world's poorest countries with per capita GNP at about US$180 in 1990. In 1988-89, there were 94,662 students enrolled at Tribhuvan University's 64 constituents and 71 private or affiliated campuses. Because salaries are extremely low for teaching staff, they have little incentive to do research or improve their teaching.

The subjects sampled were 297 students enrolled in graduate courses in science and arts at the Kathmandu campuses of Tribhuvan University (Watkins and Regmi, 1992). By Western standards, these courses are barely equivalent to senior undergraduate level.

The Philippines

The Philippines has had a long tradition of higher education, modeled after first the Spanish and later the American systems. Its higher education is characterized by quantity rather than quality. With a population of nearly 60 million, about 1.5 million were enrolled at tertiary institutions in 1988-89. According to Gonzalez (1989), most Filipino college courses, except those offered by a few prestigious universities, are equivalent to senior

secondary level by Western standards. The 77 subjects were undergraduate psychology students at an elite Catholic university outside Manila (Watkins and Gerong, 1992).

Hong Kong

Unlike the Philippines, higher education in Hong Kong is characterized by quality. As a British colony, Hong Kong has a system of higher education modeled largely on British standards and experience. However, with the majority of its population being Chinese, its system reflects a complex mixture of Eastern and Western cultural traditions.

A total of 87 students enrolled in graduate education courses at the University of Hong Kong were the subjects for the applicability study (Watkins, 1992). They were asked to complete the questionnaire for two of their lecturers when they were undergraduate social science students.

Taiwan

With strong American links, universities and colleges in Taiwan are quite comparable with those in the United States in respect of system and standard. Each year, a large number of college students enter American universities for graduate studies. Further, more and more teaching staff in tertiary institutions are previously US-trained Taiwanese students, most holding PhDs. Nevertheless, its Chinese cultural heritage remains basically intact despite much American influence on its education.

A total of 371 students at Taipei Municipal Normal College were sampled for the study (Lin, Watkins, & Meng, 1994). Specifically, they were from Departments of Education in Mathematics and Sciences, Language Education, Primary Education, and Education in Sociology.

China

Powered by the economic and social reforms, higher education in China has undergone drastic changes since 1978. Over the past 15 years, tens of thousands of students have gone to study abroad and thousands of foreign scholars have visited China. Many universities have established special relationships with sister universities in Europe, North America, Australia, and Asia. In particular, of long-term significance have been the influx of books, journals, and other academic materials bought for university and research institute libraries.

A total of 367 students of years three and four at Beijing Normal University were sampled for the applicability study (Lin, Watkins, & Meng, 1994). Specifically, the subjects were from seven randomly chosen classes, one from each of seven randomly selected departments (out of 15 departments). The seven randomly selected departments were Mathematics, Physics, Chemistry, Economics, History, Chinese, and Education.

Results

Reliability

The internal consistency reliability estimates, coefficient alpha, for the seven countries are shown in Table 2. Except that for Nepal, the mean alphas for the other six countries are 0.85 or higher for both SEEQ and Endeavor scales. For Nepal, the mean alpha of 0.74 is on the medium to low side.

Convergent and Discriminant Validities

As shown in Table 1, there seems to be a one-to-one matching between the first five SEEQ scales and the first five Endeavor scales, and the sixth SEEQ scale appears to combine two Endeavor scales. By Marsh's (1981) approach, convergent and

Table 2: University Summaries of Convergent-Discriminant Validity Criteria and Reliability Estimates

	SU	NZ	Nep	Phi	HK	Twn	Chi
Convergent Validity Means	0.83	0.85	0.67	0.88	0.83	0.88	0.85
Non-Convergent Validity Means	0.47	0.48	0.54	0.80	0.64	0.76	0.66
Criterion 1 (Convergent Validity) Proportion of Statistically Significant Correlations (Out of 7 at 0.01 level)	1.00	1.00	1.00	i.OO	1.00	1.00	1.00
Criterion2(DiscriminantValidity) Proportion of Successful Comparison (out of 96 Pairs)	0.99	1.00	0.88	0.86	0.97	0.92	0.99
Criterion 3 (Discriminant Validity) Proportion of Successful Comparisons (out of 98 Pairs)	0.98	1.00	0.88	0.91	0.96	0.92	0.97
Coefficent a Reliability Estimates							
Means for SEEQ Scales	0.90	0.90	0.74	0.92	0.89	0.89	0.86
Means for Endeavor Scales	0.89	0.91	0.74	0.94	0.92	0.91	0.85
Means for SEEQ & Endeavor	0.90	0.90	0.74	0.93	0.90	0.89	0.86

Note: SU = Sydney University; NZ = New Zealand: Nep = Nepal; Phi = Philippines; HK = Hong Kong; Twn = Taiwan; Chi = China.

discriminant validities of the nine SEEQ and seven Endeavor scales can be assessed in a modified multitrait-multimethod (MTMM) correlation matrix, where the multiple traits correspond to the scales of effective teaching and the multiple methods correspond to the two different instruments.

According to Campbell and Fiske (1959), convergent validity refers to the correlation between SEEQ and Endeavor scales that are hypothesized to measure the same component (or construct), whereas discriminant validity refers to the distinctiveness of the different dimensions (or scales) of the two instruments and offers a test of the multidimensionality of the two instruments. In other words, convergent validities of the two instruments are the correlations between matching SEEQ and Endeavor scales. On the other hand, discriminant validities of the two instruments can be classified into the following two categories: the correlations between non-matching SEEQ and Endeavor scales (i.e., heterotrait-heteromethod coefficients) and the correlations among different scales within each of the two instruments (i.e., heterotrait-monomethod coefficients). With some minor changes, the criteria developed by Campbell and Fiske to validate instruments can be applied to the data for each of the seven countries. The results of these modified MTMM analyses, summarized in Table 2, are interpreted as follows:

(a) Convergent Validities: Convergent validities (i.e., the correlations between supposedly matching SEEQ and Endeavor scales) should be high. Table 2 shows that the mean convergent validities range from 0.67 for Nepal to 0.88 for the Philippines, with all but one (for Nepal) being greater than 0.83. Considering the low alpha for Nepal, its mean correlation seems fairly high.

(b) Discriminant Validities: The Convergent validity correlations for each scale should be greater than the non-convergent validities involving that scale. From Table 2, it can be seen that the convergent validity means are far greater than the non-convergent validity means for New Zealand, Australia, China, and Hong Kong. On the other hand, the non-convergent validity means of 0.80 for the Philippines and 0.76 for Taiwan are

unexpectedly high, although the convergent validity means are still greater than the non-convergent validity means for these two countries and Nepal. Further, the usual MTMM criteria that the convergent validities should be greater than the corresponding heterotrait-heteromethod and the heterotrait-monomethod correlations are satisfied by 0.86-1.00 and 0.88-1.00 of the pairwise comparisons for the seven countries, respectively. Again, Nepal, the Philippines, and Taiwan are least successful in terms of pairwise comparisons.

In sum, support for the convergent and discriminant validities is strongest for New Zealand and Australia and weakest for Nepal, the Philippines, and Taiwan. The differences in terms of support among the seven studies, however, are not too substantial. Hence, the convergent and discriminant validities of SEEQ and Endeavor responses in each of the studies seem to be generally supported.

Inappropriate Items

In each of the seven studies, an item was counted as inappropriate if it was either marked as such or left blank by the student. For these studies, the sample for Hong Kong has the highest mean percentage (5.8%) of inappropriate items, followed by those for Australia and Nepal (3.6% for both), New Zealand (3.4%), Taiwan (2.0%), the Philippines (0.8%), and China (0.2%). The frequency of inappropriate responses is similar across all seven studies in that every item in each study is deemed appropriate by 80% or more of the students (see Table 3). In addition, across all studies, the items most frequently judged inappropriate are the ones regarding whether examinations are fair or feedback from examinations is valuable. As commented by Watkins (1994), this probably reflects a difference in assessment style in different campus settings (e.g., either assignments with no exams or end-of-year exams with no feedback other than the grade achieved).

Most Important Items

In all seven studies, each item of the questionnaire was selected by at least one student as being among the five most important items describing their chosen lecturers' teaching. The two items most often selected as being important are whether lecturer's teaching style holds students' interest and whether lecturer's explanations are clear (see Table 3). In addition, students from different countries have rather different expectations from their lecturers. Specifically, whereas students in Australia, New Zealand, China, and Taiwan considered the lecturers' enthusiasm and the interest they generated are most important, students in Hong Kong and the Philippines cared more about the learning outcomes achieved. The Nepalese students tended to be more concerned about whether they were allowed to share ideas in the class.

Patterns of Most Important Items Across Campus Settings

According to Marsh (1986), a better way to understand similarities in perception of teaching effectiveness in different settings is to compare the patterns of most important responses across different campus settings. Such analysis helps determine whether the items perceived to be most important in one campus setting are the same items perceived correspondingly in other campus settings. However, due to the large number of values (importance indices for 55 items in each of the seven studies), an objective index of similarity is needed for indexing the similarity of two or more sets of scores (i.e., the sets of importance indices in each of the studies). In line with Marsh (1986), a matrix of similarity indices, which involves calculating correlation coefficients between the importance scores of the 55 items for the seven countries and their total (i.e., correlating the proportion of subjects considering the item as important or inappropriate between any two countries among the seven), was constructed (see Table 4) to index the similarity in patterns of the most important items in each of the seven applicability studies and the total across the studies.

The similarity index relating each study to the total based on the seven studies combined shows how well the pattern of importance scores in any one study is representative of the overall pattern. With a similarity index of 0.91, the pattern for China is most representative of the overall pattern. Studies

Table 3: Paraphrased Items and the Scales of the SEEQ (M) and the Endeavor (F) Instruments

	Proportion of "Most Important" responses for each stud)								Proportion of "not appropriate" responses for each study							
	Tot	SU	NZ	Nep	**Phi**	**HK**	**Twn**	**Chi**	**Tot**	SU	NZ	**Nep**	**Phi**	**HK**	**Twn**	**Chi**
SEEQ Instrument																
Learning																
M 1 Course challenging & stimulating.	.13	.20	.28	.05	.16	.OS	.11	.14	.01	.00	.01	.03	.01	.03	.02	.00
M2 Learned something valuable.	.13	.16	14	.o"	.16	:;	.14	.13	.01	.01	.00	.01	.03	.03	.02	.00
\13 Class increased subject interest.	.14	10	12	o°	ns		20	.n	.02	.01	.01	.03	.01	.05	.04	.00
M4 Learned & understood subject matter.	.05	.05	.05	.06	.05	11	.04	.02	.01	.00	.00	.02	.02	.04	.01	.00
Enthusiasm																
M5 Enthusiastic about leaching.	.16	.28	.22	.OS	.10	.14	.13	.IX	.01	.00	.0(1	.02	.01	.03	.01	.00
M6 Dynamic and energetic.	.14	.20	.15	.13	.05	.12	.08	.21	.01	.00	.00	.03	.01	.03	.01	.00
M7 Enhanced presentation with humor.	.15	.16	.10	.09	.11	.04	.19	.20	.02	.01	.00	.05	.01	.04	.03	.00
M8 Teaching style held your interest.	.23	.37	.28	.10	.26	.19	.20	.27	.02	.00	.01	.02	.02	.02	.03	.00
Organization/Clarity																
M9 Lecturer explanations clear.	.19	.21	.17	.29	.14	.24	.12	.20	.01	.01	.00	.04	.03	.02	.01	.00
M10 Course Materials well explained & prepared.	.10	.11	.13	.10	.03	.14	.06	.11	.02	.01	.01	.03	.03	.04	.02	.00
M11 Course objectives stated & pursued.	.05	.17	.09	.02	.02	.06	.07	.03	.02	.01	.02	.03	.02	.05	.03	.00
M12 Lectures facilitated taking notes.	.09	.10	.14	.06	.05	.06	.09	.08	.02	.02	.00	.03	.01	.06	.02	.00
Group Interaction																
M13 Encouraiyd class discussion.	.09	.11	.07	.08	.08	.04	.10	.08	.03	.10	.05	.04	.01	.04	.02	.00
M14 Students invited to share knowledge/idea.	.07	.04	.03	.15	.06	.04	.06	.04	.03	.08	.07	.03	.02	.08	.02	.00
M15 Encouraged questions & gave answers.	.10	.04	.07	.14	.06	.05	.14	.09	.02	.02	.04	.02	.01	.04	.01	.00
M16 Encouraged questioning of teacher's ideas.	.06	.03	.04	.06	.12	.02	.08	.06	.02	.04	.05	.03	.01	.05	.02	.00
Individual Rapport																
M17 Lecturer friendly to individual students.	.09	.07	.07	.14	.08	.05	.09	.07	.01	.03	.01	.03	.01	.02	.01	.00
M18 Lecturer welcomed students seeking advice.	.07	.08	.10	.06	.08	12	.09	.04	.02	.02	.02	.04	.03	.03	.02	.00
M19 Lecturer interested in individual students.	.07	.03	.06	.07	.03	.05	.09	.07	.02	.05	.02	.05	.02	.04	.02	.00
M20 Lecturer accessible to individual students.	.04	.03	.04	.06	.08	.02	.04	.02	.04	.07	.07	.06	.02	.08	.04	.01
Breadth of Coverage																
M21 Contrasted various theories.	.07	.04	.07	.08	.06	.05	.07	.09	.03	.06	.07	.03	.07	.08	.01	.00
M22 Gave background of ideas/concepts.	.05	.04	.03	.06	.03	.06	.04	.06	.02	.03	.03	.02	.02	.05	.02	.00
M23 Gave different points of view.	.06	.07	.06	.04	.07	.04	.07	.08	.04	.09	.09	.04	.01	.08	.02	.00
M24 Discussed current developments.	.07	.03	.07	.09	.10	04	.05	.JO	.03	.06	.05	.05	.01	.08	.03	.00
Examinations/Grading																
M25 Examination feedback valuable.	.04	.03	.08	.06	.05	"2	.05	.02	.07	.13	.11	.09	.03	.17	.05	.00
M26 Evaluation method fair/appropriate.	.05	.05	.06	.09	.07	.08	.04	.02	.05	.07	.06	.08	.01	.16	.03	.00
M27 Tested course content as emphasized.	.04	.04	.03	.02	.01	.01	.05	.04	.06	.08	.09	.10	.03	.15	.03	.00

Table 3: Paraphrased Items and the Scales of the SEEQ (M) and the Endeavor (F) Instruments (Continue)

	Proportion of "Most Important" responses for each study								Proportion of "not appropriate" responses for each study							
	Tot	**SU**	NZ	**Nep**	**Phi**	**HK**	**Twn**	**Chi**	**Tot**	**SI'**	**NZ**	**Nep**	**Phi**	**HK**	**Twn**	**Chi**
Readings/Assignments																
M28 Readings/texts were valuable.	.05	.03	.06	.05	.07	.02	.05	.04	.03	.04	.06	.03	.01	.05	.05	.01
M2^ They contributed to understanding.	.05	.06	.05	.05	.(15	(IS	.05	.03	.03	.01	.05	.06	.01	.04	.03	.00
Overall Rating Items																
M30 Overall course rating.	.06	.10	.08	.06	.03	.05	.05	.04	.01	.00	.00	.02	.01	.04	.01	.00
M31 Overall lecturer rating.	.11	.12	.04	.08	.05	.08	.07	.24	.01	.00	.01	.02	.03	.04	.01	.00
Workload/Difficulty																
M32 Course difficulty (easy-hard).	.04	.03	.07	.06	.07	.02	.05	.01	.01	.01	.00	.01	.01	.03	.00	.00
M33 Course workload (light-heavy).	.04	.06	.08	.02	.07	.04	.06	.02	.01	.00	.00	.02	.01	.02	.01	.01
M34 Course Pace (slow-fast).	.05	.07	.07	.04	.05	.01	.08	.02	.01	.00	.00	.02	.02	.03	.00	.(X)
Endeavor Instrument																
Presentation Clarity																
Fl Presentations clarified materials.	.12	.20	.10	.11	.13	.16	.07	.13	.01	.00	.01	.02	.02	.03	.01	.00
F2 Presented clearly & summarized.	.14	.29	.20	.12	.07	.22	.10	.10	.01	.00	.00	.01	.03	.03	.01	.00
F3 Make good use of examples.	.10	.07	.07	.17	.08	.12	.09	.10	.01	.00	.01	.0?	.01	.03	.02	.00
Workload																
F4 Students had to work hard.	.05	.04	.02	.11	.07	.00	.06	.03	.02	.00	.01	.03	.03	.05	.02	.00
F5 Course required a lot of work.	.03	.02	.03	.07	.03	.01	.05	.01	.02	.00	.01	.02	.03	.05	.02	.00
F6 Course workload was heavy.	.04	.01	.03	.08	.07	.01	.06	.02	.01	.00	.01	.02	.03	.04	.02	.00
Personal Attention																
F7 Lecturer listened and was willing to help.	.07	.05	.05	.12	.07	.09	.06	.05	.03	.07	.07	.02	.02	.02	.02	.00
F8 Students able to get personal attention.	.05	.06	.04	.08	.03	.07	.04	.04	.05	.18	.09	.05	.03	.07	.02	.00
F9 Lecturer concerned about student difficulties.	.05	.07	.03	.08	.05	.10	.05	.03	.02	.03	.03	.02	.01	.03	.02	.00
Class Discussion																
F10 Class discussion was welcome.	.06	.02	.03	.14	.06	.05	.05	.04	.03	.10	.05	.04	.01	.05	.02	.00
F11 Students encouraged to participate.	.06	.11	.05	.07	.08	.02	.05	.04	.03	.10	.07	.03	.01	.05	.02	.00
F12 Encouraged students to express ideas.	.05	.01	.02	.07	.08	.04	.07	.05	.02	.05	.05	.02	.01	.05	.02	.00
Planning/Objectives																
F13 Presentations planned in advanced.	-09	.13	.09	.11	.01	.16	.09	.06	.02	.01	.00	.03	.01	.17	.02	.00
F14 Provided detailed course schedule.	.05	.06	.09	.08	.03	.04	.04	.02	.02	.01	.01	.03	.03	.04	.04	.01
Fl5 Class activities orderly scheduled.	.07	.01	.02	.11	.03	.08	.09	.09	.04	.09	.09	.05	.02	.04	.02	.00
Grading/Examinations																
F16 Grading fair and impartial.	.06	.05	.04	.06	.12	.08	.08	.05	.06	.07	.09	.12	.01	.16	.03	.01
F17 Gradubg reflected student performance.	.04	.03	.04	.03	.06	.03	.05	.03	.06	.10	.09	.10	.01	.18	.03	.00
FIK Grading indicative of accomplishments.	.03	.02	.03	.02	.03	.01	.07	.03	.06	.09	.08	.11	.01	.18	.02	.00

Table 3: Paraphrased Items and the Scales of the SEEQ (M) and the Endeavor (F) Instruments (Continue)

	Proportion of "Most Important" responses for each study-								Proportion of "not appropriate" responses for each study							
	Tot	**SU**	NZ	**Nep**	**Phi**	**HK**	**Twn**	**Chi**	**Tot**	**SU**	NZ	**Nep**	**Phi**	**HK**	**Twn**	**Chi**
Student Accomplishments																
HI 9 Understood the advanced material.	.06	.10	.04	.05	.05	.05	.05	.09	.03	.04	.06	.02	.03	.06	.03	.00
1 20 Course improved ability lo anal)/e issues.	,(W	.H)	.10	.00	.14	.13	.OS	.11	.02	.01	.03	.02	.02	.04	.02	.00
F21 Course increased knowledge and competence.	.14	.08	.13	.09	.28	.26	.13	.14	.01	.01	.01	.02	.01	.04	.02	.00

Note: SU = Sydney University; NZ = New Zealand; Nep = Nepal; Phi = Philippines; HK = Hong Kong; Twn = Taiwan; Chi = China.

for Australia, New Zealand, and Hong Kong result in similarity indices of 0.76 and above. The patterns for Nepal and the Philippines least resemble the overall pattern, with similarity indices of 0.54 and 0.64, respectively.

From Table 4, it can be observed that the most important items in one setting also appear to be the most important in other settings (mean r = 0.53). Hence, a question of interest is how well the pattern for each study is comparable with that of each of the other studies. In particular, whether Hong Kong, Taiwan, and China are more similar in data patterns among themselves or more similar to those of the West. As expected, Australia and New Zealand are most similar to each other in data patterns, with a high similarity index of 0.86. Surprisingly, the similarity indices among the three Chinese-speaking areas are moderate to low — 0.52 between Hong Kong and Taiwan, and 0.60 between Hong Kong and China. On the other hand, the similarity indices between them and the Western countries are relatively high — 0.72 between China and Australia, 0.66 between China and New Zealand, and 0.65 between Hong Kong and Australia. In comparison, Taiwan data are slightly less similar to those of the West. In addition, data for Nepal and the Philippines are least similar to those of the other five countries.

Differentiating Between good and poor Lecturers

In all seven studies, all but the Workload/ Difficulty items differentiate clearly between good and poor lecturers. Specifically, items regarding whether the teaching style holds students' interest, whether the lecturer's explanations are clear, and whether the lecture increases students' interest in the subject are among the items consistently differentiating teaching quality.

To determine if students in each sample used similar bases to evaluate their lecturers, another set of similarity indices were computed. In each study, the mean difference between the ratings of good and poor lecturers was computed for each of the 55 items and then each item was ranked based on the magnitude of that difference. The Spearman rank order correlations between the ranks of these 55 items for the seven countries were calculated and are shown in Table 5.

It can be seen from Table 5 that the correlations range from 0.62 to 0.90 (all statistically different from zero at 1% significance level with 53 degrees of freedom) and the mean correlation is 0.79, indicating strong agreement across the seven studies in respect of the factors that students used to assess teaching quality. Moreover, a close examination of Table 5 indicates that the bases used by students to

Table 4: Cross-Campus Similarity in Patterns of Items Judged to be Most Important in the Seven Studies

	Tot	SU	NZ	Nep	Phi	HK	Twn	Chi
Tot								
SU	.84	—						
NZ	.82	.86	—					
Nep	.54	.26	.21	—				
Phi	.64	.46	.57	.21	—			
Hk	.76	.65	.63	.47	.56	—		
Twn	.82	.58	.63	.29	.58	.52	—	
Chi	.91	.72	.66	.39	.53	.60	.71	—

Note: Tot = total; SU = Sydney University; NZ = New Zealand; Nep = Nepal; Phi = Philippines; HK = Hong Kong; Twn = Taiwan; Chi = China.

Table 5: Correlations between the 55 Items Ranked According to their Degree of Differentiation between "Good" and "Poor" Teachers across seven Studies

	SU	NZ	Nep	Phi	HK	Twn	Chi
SU							
NZ	.88	—					
Nep	.80	.71	—				
Phi	.73	.78	.62	—			
Hk	.89	.90	.80	.83	—		
Twn	.80	.84	.70	.80	.89	—	
Chi	.77	.76	.65	.72	.82	.83	—

Note: Tot = total; SU = Sydney University; NZ = New Zealand; Nep = Nepal; Phi = Philippines; HK = Hong Kong; Twn = Taiwan; Chi = China.

evaluate teaching quality generally appear to be that Hong Kong is more similar to New Zealand (0.90), Australia (0.89), and Taiwan (0.89) than to China (0.82); Taiwan is as similar to New Zealand (0.84) and Australia (0.80) as to China (0.83); and China is less similar to New Zealand (0.76) and Australia (0.77) than to Hong Kong (0.82) and Taiwan (0.83).

Characteristics of Good and Poor Lecturers and Their Classes

Further investigation on cross-cultural trends in students' evaluations was done to compare the characteristics of the lecturers and the classes concerned (Australia is excluded because its study does not contain this type of investigation). From Table 6, it can be observed that in each of the studies good lecturers tend to give higher grades and, except for Taiwan and China, they are perceived by students to be of the same ages as poor lecturers. Good lecturers tend to be perceived as younger by students in Taiwan but as older by students in China. Nepal and New Zealand are the only countries where students' ratings appear to correlate

with class size. Surprisingly in these two countries, ratings received in large classes are on the average higher than those received in small classes. Finally, for Hong Kong, New Zealand, and the Philippines, good lecturers are more likely to be teaching one of the students' major subjects.

Conclusion

Before making any conclusions, it must be realized that the interpretation of the results should be made with some reservation because apparent differences in the seven studies may be not only due to cross-cultural differences but also due to campus/ sampling differences. Ideally, to make comparisons on a common basis across different cultures, subjects should be chosen from similar majors and universities representative of each of the countries. This was achieved fairly well as students in all but the Nepalese studies were education or social science majors. This may partly explain the relatively lower level of similarity index for Nepal. In addition, since only one campus was sampled for each country and sample size is small in some of the studies, the generalizability of the results to other campuses may be limited.

However, the results provide some encouraging evidence which supports the applicability of both SEEQ and Endeavor instruments as well as the cross-cultural validity of the two instruments and their underlying model of effective teaching. Specifically, the results indicate that, except the study for Nepal, the other six studies generally have high internal consistency reliability coefficients for SEEQ and Endeavor scales. Four (Australia, New Zealand, Hong Kong, and China) of the seven samples show clear evidence of convergent and discriminant validities. In the seven studies, all but the items relating to assessment are considered appropriate. Furthermore, the pattern of importance ratings of the questionnaire items from subjects of the seven countries reflects some overall similarity in perception of teaching quality. Comparison of items discriminating between good and poor lecturers suggests some degree of agreement regarding the most salient items across all seven studies. It appears that in all the studies students' evaluations are quite related to the grades awarded, but relatively not consistently related to other class/lecture characteristics such as class size and lecturer's age.

The fact that Australia and New Zealand are most similar in data patterns may not only reflect, as commented by Watkins (1994), the greater campus/ cultural similarities between the two countries but also be a result of the higher reliabilities of the scales for these two studies, since students in these two studies responded to the survey questions in their mother tongue. On the contrary, for students in Hong Kong, Nepal, and the Philippines, the evaluation surveys were done in English, which is at best their second or third language; for students in China and Taiwan, the surveys were translated into Chinese, which may cause slight distortion in meaning from the original English version. In addition, the fact that Hong Kong, Taiwan, and China are each relatively more similar to the West than among themselves may suggest that they each have been influenced more or less by the West and, accordingly, each of their higher education systems may reflect a similar but unique blend of Chinese

Table 6: Means of Characteristics of Good and Poor Lecturers and their Classes by Country

Characteristic	New Zealand		Nepal		Philippines		Hong Kong		Taiwan		China	
	Good	Poor	Good	Poor	Good	Poor	Good	Poor	Good	Poor	Good	Poor
Major Subject	1.22	1.40*	1.00	1.02	1.05	1.59*	1.21	1.34*	1.16	1.12	1.10	1.16
Class Grade	8.20	7.15*	1.16	2.05*	3.38	4.63*	4.19	6.59*	84.73	79.52*	84.32	76.97*
Class Size	106.50	90.71*	55.05	35.25*	42.79	48.00	39.78	35.45	30.23	31.36	73.24	74.97
Lecturer's Age	40.14	42.84	44.49	41.14	34.38	38.70	42.97	41.56	42.03	46.38*	45.08	39.48

Note: An Asterisk * indicates within-country means are significantly different at 0.01 level.
Major subject was coded Major = 1, Minor = 2. Class grades were coded according to the grading system at that campus: New Zealand from A (highest) = 10 to E = 0; Nepal first division = 1 (highest), second division = 2, third division = 3; Philippines from 1.0 (highest) to 5.0; Hong Kong from A + = 1 (highest) to E = 11.

and Western cultural traditions. Moreover, the result that China and Taiwan are slightly more similar to each other than to Hong Kong may suggest that, besides methodological differences (i.e., survey was done in Chinese for China and Taiwan and in English for Hong Kong), the former two are more alike in culture and education system, whereas the latter has modeled its education system largely on British standards and experience for more than a century.

Synthesizing the results for the seven countries, this study generally supports the within-construct and between-construct validities of the two instruments and their underlying model of teaching quality for use in a variety of cultures. However, it must be realized that the approach for this study is an etic one. According to Watkins (1994), it is possible that culture-specific aspects of effective teaching may still exist, which can only be determined by emic-type research, probably of a qualitative nature, within each of the cultures concerned. In particular, the discrepancy in data patterns (especially in their lack of differentiation among the multiple components of effective teaching) between the studies for Nepal and the Philippines and those for the other five countries may suggest a need for an emic-type qualitative research to determine if such discrepancy is due to whether students in Nepal and the Philippines have a more global perception of effective teaching than their counterparts in the other five countries or whether there is any halo effect caused by the "good-and-poor" selection procedure.

References

Berry, J. (1989). Imposed-etics, derived-emics: The operationalization of a compelling idea. *International Journal of Psychology,* 24:721-735.

Campbell, D.T.. & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-muitimethod matrix. *Psychological Bulletin,* 56:81-105.

Clarkson, P.C. (1984). Papua New Guinea students perceptions of mathematics lecturers. *Journal of Educational Psychology,* 76:1386-1395.

Cronbach, L.J. (1971). Test validation. In R.L. Thorndike (Ed.), *Educational Measurement.* Washington, D. C: American Council of Education.

Enriquez, V.G. (1977). Filipino psychology in the third world. *Philippine Journal of Psychology, 10:3*-17.

Prey, P.W. (1973). Student ratings of teaching: Validity of several rating factors. *Science,* /<92:83-85.

Frey, P.W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education,* 9:69-71.

Frey. P.W,, Leonard, D.W., & Beatty. W.W. (1975), Student ratings of instruction: Validation research. *Research in Higher Education,* 9:69-71.

Hayton, G.E, (1983). An investigation of the applicability in Technical and Further Education of a student evaluation of teaching instrument. Unpublished thesis. Faculty of Education, University of Sydney.

Hui, C.H., & Triandis. H. (1985). Measurement in cross-cultural psychology: A review and comparison of strategies. *Journal of Cross-Cultural Psychology, 16:*131-152.

Lin, W-Y., Watkins, D., & Meng, Q-M. (1994). Students" evaluations of tertiary teaching effectiveness: A Taiwan investigation. A manuscript submitted for publication.

Lin, W-Y., Watkins, D.. & Meng, Q-M. (1994). Students' evaluations of tertiary teaching: A China perspective. A manuscript submitted for publication.

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

Marsh, H.W. (1984), Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychology,* 7(5:707-754.

Marsh, H.W. (1986). Applicability paradigm: Students" evaluations of teaching effectiveness in different countries. *Journal of Educational Psychology,* 75:465-473.

Marsh, H.W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research. *International Journal of Educational Research, 17*:253-388.

Marsh, H.W, & Dunkin, M.J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (Ed.), *Higher Education: Handbook of Theory and Reseach* (Vol. 9). New York: Agathon.

Marsh, H.W., & Roche, L.A. (1991). The use of student evaluations of university teaching in different settings: The applicability paradigm. As yet unpublished research. University of Western Sydney.

Marsh, H.W., Touron, J., & Wheeler, B. (1985). Students' evaluation of university instructors: The applicability of American instruments in a Spanish setting. *Teaching and Teacher Education: An International Journal of Research and*

Studies, 7:123-138.

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 university teaching: 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,* 727-734.

Watkins, D., Marsh, H.W., & Young, D. (1987). Evaluating tertiary teaching: A New Zealand perspective. *Teaching and Teacher Education: An international Journal of Research and Studies,* 2:41-53.

Watkins, D., & Regmi, M. (1992). Student evaluations 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,* 7(5:185-198

Dr. Wen-Ying LIN is Lecturer of Department of Educational Psychology, The Chinese University of Hong Kong.
Dr. David WATKINS is Reader of the Faculty of Education, The University of Hong Kong.
Dr. Qing-Mao MEMO is Professor of Department of Psychology, Beijing Normal University.