教育研究學報 1992，第七卷， 60-67
Educational Research Journal 1992, Vol.7, pp. 60-67

Evaluating the Effectiveness of Tertiary Teaching: A Hong Kong Perspective*

David WATKINS
University of Hong Kong

An investigation is reported which tests the applicability of two American instruments designed to assess tertiary student's evaluations of teaching effectiveness with 87 Hong Kong graduate students. Each student was asked to complete an evaluation questionnaire for both a 'good' lecturer and a poor lecturer so much of the analysis actually was based on the analysis of 174 rating forms. The scales were found to have generally high internal consistency reliability coefficients, most of the items were seen to be appropriate and all but one item was considered of importance by at least some of the students. In addition, all but the Work/Difficulty items clearly differentiated between 'good' and 'poor' lecturers. Further analysis supported the convergent and discriminant validity of the scales hypothesized to measure similar or dissimilar components of effective teaching. Factors analysis showed both a strong main factor and three minor factors providing some further support for a multidimensional model of teaching effectiveness.

這個報告是有關一個調查研究，看看美國兩個用來評定專上學生如何評核教學效能的工具的適用程度。調查的大專畢業生達八十七人。每一位畢業生必須對心目中的一位「好」和一位「差」的講師加以評核，填妥問卷後交回分析。問卷達一百七十四分。調查研究發現測驗量表內部的相符度及可靠性係數十分高。問卷中的各測驗項目均合乎是次測試的目的，而學生對各測驗項目均感重要。所有測驗項目均能分別「好」與「差」的講師，只有「功課量」及「難度」兩項目例外。進一步的分析則顯示量表所量度各個高效率教學的組合的識別性及輻合性極高。因素分析更顯示一個主要因素及三個次要因素對高效率教學的重要性，此乃直接支持多層面教學效能的模式理論。

Systematic evaluations of tertiary teaching have been commonplace in North American universities and colleges for at least the last fifteen years. Amongst the different criteria of teaching effectiveness that have been suggested are: (a) evidence on student learning as shown by course grades or scores on standardized achievement tests (unfortunately, even if it can be show that grades are comparable across courses, neither students nor lecturers seem to believe such scores necessarily reflect what a student has learnt from a course let alone that it is the lecturer who has brought about about this learning (Astin, 1974); (b) evaluation by one's peers or by neutral observers (these tend now to be used more as indicators against which student evaluations are validated) and (c) student ratings. This latter method has a controversial history but has been by far the most widely used.

The literature on the value of student ratings of tertiary teaching is a massive one which provides evidence of considerable hostility and suspicion on behalf of some U.S. faculty. While the early American evidence tended to support the reliability but cast doubt on the validity of student ratings, several recent reviews of research in this area are supportive of value (Dunkin & Barnes, 1986; Marsh, 1987). Moreover, it appear that much of the inconsistency in research findings is due to inadequate measuring instruments (Frey, 1982; Marsh, 1987). In particular, it appears now that teaching effectiveness is multifaceted and that any instrument which focuses on a single overall score is likely to be inadequate. For example, a lecturer who is well organized may not be the the best of oral communicators. Failure to separate these different components of effective teactive teaching has led to conflicting research findings as well as inadequate information for diagnostic or decision-making purposes (e.g. some aspects of poor teaching may be subject to improvement through training, others may not).

Tertiary institutions in Hong Kong are starting to take serious notice of student assessment of their

*This research was supported by a research grant to the author from the Committee for Research and Conference Grants of the University of Hong Kong.

lecturers' teaching ability. However, as yet there has been no published research on the quality of the measuring instruments available for use with students here.

This research assesses the reliability and validity of two American-developed measuring instrument - the Students' Evaluations of Educational Quality (SEEQ) developed by Marsh (1981) and the Endeavor Instructional Rating Form devised by Frey (1978). Although these instruments have been shown to be applicable in Spain (Marsh, Touron & Wheeler, 1985), Australia (Marsh, 1981; Marsh & Roche, 1991), and New Zealand (Watkins, Marsh & Young, 1987), there is still doubt about their cross-cultural validity, particularly in less developed countries (Clarkson, 1984).

The SEEQ Instrument

The SEEQ and the research that led to its development have been described by Marsh (1984, 1987). Numerous factor analyses have identified the nine SEEQ factors in responses from different populations of students and also in lecturer self-evaluations of their own teaching effectiveness when they were asked to complete the same instrument as their as their students. The nine SEEQ factors are Learning/Value, Instructor Enthusiasm, Organization/Clarity, Group Interaction, Individual Rapport, Breadth of Coverage, Examinations/Grading, Assignments/Readings, and Workload/Difficulty.

Marsh (1987) argued, as did Frey, that student's evaluations, like the effective teaching they are designed to reflect, should be multidimenioal. He supported this common-sense assertion with empirical results and also demonstrated that the failure to recognize this muitidimennsionality has led to misinterpretation in student-evaluation research.

The reliablility of responses to the SEEQ, based upon correlations among items designed to measure the same factor and correlations among responses by students in the same course, is consistently high (Marsh, 1987). To test the long-term stability of responses to the SEEQ, students from 100 classes were asked to re-evaluate teaching effectiveness several years after their graduation from their university program, and their retrospective evaluations correlated 0.83 with those the same students had given at the end of each class. Ratings on the SEEQ have successfully been validated against the ratings of former students, student achievement as measured by an objective examination in multisection courses, lecturers' evaluations of their own teaching effectiveness and affective course consequences such as applications of course materials and plans to pursue the subject further. None of a set of 16 potential sources of bias (e.g. class size, expected grade, prior subject interest) could account for more than 5% of the variance in seeq ratings, and many of the relationships were inconsistent with a simple bias explanation (e.g. harder, more difficult courses were evaluated more favorably). SEEQ ratings are primarily a function of characteristics of the person who teaches a course, rather than of the particular course which he or she teaches. Finally, feedback from the SEEQ, particularly when coupled with a candid discussion with an external consultant, led to improved ratings and better student learning.

The Endeavor Instrument

The Endeavor Instrument measures seven components of effective teaching – components that have been identified through the use of factor analysis in different settings (Frey, Leonard, & Beatty, 1975). The seven factors are Presentation Clarity, Workload, Personal Attention, Class Discussions, Organization-Planning, Grading, and Student Accomplishment. In validating the ratings obtained from this instrument, Frey has shown that the ratings on the Endeavor are correlated with student learning (Frey, 1978; Frey, et al., 1975). In these studies, as well as in similar studies described below, student ratings are collected in large multisection courses (i.e., courses in which the large group of students is divided into smaller groups or sections and all instruction is delivered separately to each section). Each section of students in the same course is taught throughout by a different lecturer, but each is taught according to a similar course outline, has similar goals and objectives and, most important, is tested with the same standardized final examination at the end of the course (for further discussion see Cohen, 1981; Marsh, 1984). Frey concluded that those sections of students that rate teaching to be most effective are also the sections that learn the most as measured by performance on the final examination, thus supporting the validity of ratings on the Endeavor Instrument.

Frey (1878) further argued that it is important to recognized the multidimensionality of evaluations of effective teaching. In an examination of the relationships between students' evaluations and a variety of other variables, he demonstrated that the size, and even the direction, of the correlations varies with the particular component of effective teaching considered. The failure to recognize this multidimensionality is an important weakness in much of the American research.

Inspection of the item content supported by previous research (Marsh et al., 1985) revealed considerable overlap in the dimensions measured by these two instruments (see Table 1). One SEEQ factor, Organization/Clarity, seems to have been divided into two factors for Endeavor instrument (Presentation Clarity and Organization/Planning). On the basis of earlier studies, 32 of the 34 SEEQ items (M1-M32) and 21 Endeavor items were classified into one of 16 dimensions (see Table 2). Items A1-A6 were added to try to better identify the factors involved.

TABLE 1
Pairs of Corresponding Factors in SEEQ and Endeavor

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

Research Aims

The following aspects of the SEEQ and the Endeavor instruments were assessed here for Hong Kong students:

1. The internal consistency reliabilities of the scales;
2. The ability of items to discriminate between "good" and "poor" lecturers;
3. The perceived importance and appropriateness of the items;
4. The validity of the underlying factor model;
5. The convergent and discriminant validity of the scales;
6. The relationship of scale scores to factors such as the size of the class, the grade achieved in the course, whether the course was a student's major or minor, and the age of the lecturer concerned.

Method

Subjects

The evaluation survey was administered to a total of 87 students enrolled in graduate education courses at the University of Hong Kong. The subjects were guaranteed the confidentiality of their responses and were not asked to identify themselves in any way. Each subject was asked to consider the lecturers who had taught their undergraduate tertiary courses and select a GOOD and a POOR teacher. They were told to limit their choices to lecturers that took the class for at least one term and who taught mainly using a lecture/seminar format. As the subjects had varying academic backgrounds in terms of disciplinary area, institution attended, and years of study few of their previous teachers would have been in common.

Statistical Analysis

Each item was initially tested in terms of (1) its ability to discriminate among the good and poor lecturers; (b) its appropriateness (i.e) the lack of "not appropriate" responses); and (c) its importance (i.e. the number "of most important" nominations). Items were categorized as representing 10 dimensions on an a priori basis (support for these dimensions was found in the Australian study described by Marsh in 1981 and confirmed in the New Zealand investigation of Watkins et al. in 1987) and a factor analysis of responses to all items was used to test the ability of the responses to differentiate among these hypothesized components of teaching effectiveness formed on responses to items from SEEQ and the Endeavor instruments.

All the statistical analyses were conducted with the commercially available SPSS-X statistical package (Hull & Nie, 1984). The factor analyses were performed with iterated communality estimates, a Kaiser nomalization, and an oblique rotation, also using the SPSS-X procedure. The Scree test criterion was used as a guide to the number of factors to be rotated but a numer of solutions were examined to find the best fit to the hypothesised factor structure. For purposes of this study, blank and "not appropriate" responses were considered to be missing values. Each of the factor analyses was performed on correlation matrices constructed with "pair-wise deletion" for missing data. As no sex differences were detected only the combined analyses are reported here.

Results

Scale Reliabilities

From Table 2 it can be seen that the estimated coefficients of reliability, α, ranged from 0.72 to 0.95 (the median α 's were 0.92 and 0.93 for the SEEQ and Endeavor, respectively).

TABLE 2
Paraphrased items and hypothesized factors for Marsh's (M) SEEQ instrument and Frey's (F) Endeavor instrument plus obtained scale reliability estimates, mean item responses for each 'lecturer' type and number of 'not appropriate' and 'not important' responses.

	Scales and Items	Mean responses for lecturers chosen as:		Number of 'most important' nominations	Number of in-appropriate' responses
		Good	Poor		
1.	Group interaction (SEEQ) α = 0.92				
	M13 Encouraged class discussion	6.90	4.48	7	7
	M14 Students invited to share knowledge/ideas	7.08	4.06	6	13
	M15 Encouraged questions and gave answers	6.98	3.40	8	6
	M16 Encouraged questioning of teacher's ideas	7.02	3.82	4	9
2.	Learning (SEEQ) α = 0.95				
	M1 Course challenging and stimulating	7.22	3.15	14	5
	M2 Learned something valuable	7.64	3.85	37	5
	M3 Class increased subject interest	7.26	2.85	26	8
	M4 Learned and understood subject matter	7.23	3.87	19	6
3.	Workload/difficulty (SEEQ) α = 0.72				
	M32 Course difficulty (easy-hard)	5.54	5.36	4	5
	M33 Course workload (light-heavy)	5.48	5.07	6	4
	M34 Course pace (slow-fast)	5.53	4.51	1	5
4.	Examinations/grading (SEEQ) α = 0.87				
	M25 Examination feedback valuable	6.36	3.97	4	29
	M26 Evaluation methods fair/appropriate	7.07	4.28	14	27
	M27 Tested course content as emphasized	6.58	4.28	1	26
5.	Individual Rapport (SEEQ) α = 0.91				
	M17 Lecturer friendly to individual students	7.44	4.94	9	4
	M18 Lecturer welcomed students seeking advice	7.24	3.95	21	5
	M19 Lecturer interested in individual students	6.60	3.40	8	6
	M20 Lecturer accessible to individual students	6.51	3.85	4	14
6.	Organization/clarity (SEEQ) α = 0.94				
	M9 Lecturer explanations clear	7.74	3.37	42	4
	M10 Course materials well explained and prepared	7.61	3.26	25	6
	M11 Course objectives stated and pursued	7.21	3.39	10	9
	M12 Lectures facilitated taking notes	6.54	3.16	10	10
7.	Lecturer Enthusiasm (SEEQ) α = 0.93				
	M5 Enthusiastic about teaching	7.76	3.98	24	5
	M6 Dynamic and energetic	7.49	3.34	21	5
	M7 Enhanced presentation with humor	6.71	3.26	7	7
	M8 Teaching style held your interest	7.24	2.54	33	4
8.	Breadth of coverage (SEEQ) α = 0.93				
	M21 Contrasted various theories	6.71	3.83	9	13
	M22 Gave background of ideas/concepts	6.98	3.52	10	9
	M23 Gave different points of view	6.89	3.65	7	13
	M24 Discussed current developments	6.77	3.35	6	14

9.	Readings (SEEQ) α = 0.84					
	M28	Readings/texts were valuable	6.86	4.42	4	8
	M29	They contributed to understanding	7.15	4.23	13	7
	Overall Rating Items (SEEQ)					
	M30	Overall Course Rating	7.47	3.51	8	6
	M31	Overall Lecturer Rating	7.70	3.24	13	6
10.	Class discussion (Endeavor) α = 0.95					
	F10	Class discussion was welcome	7.18	4.58	9	9
	F11	Students encouraged to participate	6.85	4.28	3	9
	F12	Encouraged students to express ideas	7.13	4.04	7	9
11.	Student Accomplishments (Endeavor) α = 0.91					
	F19	Understood the advanced material	7.07	3.98	8	10
	F20	Course improved ability to analyze issues	7.29	3.39	23	7
	F21	Course increased knowledge and competence	7.33	3.45	46	6
12.	Workload (Endeavor) α = 0.85					
	F4	Students had to work hard	6.84	5.16	0	8
	F5	Course required a lot of work	6.57	4.90	1	9
	F6	Course workload was heavy	5.95	4.45	2	6
13.	Grading/examinations (Endeavor) α = 0.95					
	F16	Grading fair and impartial	7.00	4.23	13	28
	F17	Grading reflected student performance	6.86	4.30	5	31
	F18	Grading indicative of accomplishments	6.78	3.87	2	31
14.	Personal Attention (Endeavor) α = 0.94					
	F7	Lecturer listened and was willing to help	7.32	4.29	16	4
	F8	Students able to get personal attention	7.11	4.10	12	12
	F9	Lecturer concerned about student difficulties	6.93	3.46	18	5
15.	Presentation clarity (Endeavor) α = 0.95					
	F1	Presentation clarified materials	7.61	3.16	27	5
	F2	Presented clearly and summarized	7.49	3.28	39	5
	F3	Made good use of examples	7.34	3.55	20	5
16.	Planning/objectives (Endeavor) α = 0.90					
	F13	Presentations planned in advance	7.67	4.01	28	29
	F14	Provided detailed course schedule	7.01	4.07	6	7
	F15	Class activities orderly scheduled	6.93	3.34	14	7

Evaluation of Items

It can be seen from Table 2 that all items, except for those on the Workload/Difficulty scale, discriminated quite well between "good" and "poor" lecturers. The discrimination is particularly clear for items on the Lecturer Enthusiasm, Learning Value, and Organization/Clarity dimensions. It is not surprising that the workload/difficulty-related items do not differentiate between "good" and "poor" lecturers – after all, a lecturer is surely "poor" if he *or* she provides material which is much too easy or much to difficult (similarly, far too great or too little quantity of work).

Only seven of the items, six concerned with assessment, were judged to be inappropriate by more than 17 (10%) of the students. Thus it seems that the great majority of items, except for the grading related items, were perceived as relevant to the evaluation of teaching by the students.

Students selected as many as five items that they felt were most important in describing each "lecturer" type. Only one item received no "most important" nominations. Items assessing the degree to which the course increased the students' interest in and knowledge of the subject matter; whether the lecturing style was enthusiastic and held one's interest; the degree to which the presentation was clear; and whether anything of value was learnt were seen as being the most important features.

Factor Analysis

Both the eigen values greater than 1.00 and the Scree test supported a four factor solution which accounted for 75.0% of the variance. Factor I was a strong general factor combining a number of aspects of teaching effectiveness. Three minor factors reflecting the course workload, and satisfaction with grading and group interaction were also identifiable.

Convergent and Discriminant Validity

The convergent and discriminant validity of the nine SEEQ and seven Endeavor scales were assessed in a modified multitrait-multimethod (MTMM) matrix, where the scales of teaching effectiveness are the multiple traits and the two different instruments correspond to the multiple methods (see Table 3).

TABLE 3
'Multitrait-multimethod' Matrix of Correlations among SEEQ and Endeavor Scales (n=174)

SEEQ scales	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1. Group Interaction															
2. Learning/value	0.75														
3. Workload/ difficulty	0.05	0.21													
4. Examinations/ grading	0.73	0.64	0.16												
5. Individual rapport	0.81	0.72	0.13	0.77											
6. Organization/ clarity	0.74	0.89	0.25	0.66	0.74										
7. Enthusiasm	0.82	0.87	0.20	0.67	0.78	0.89									
8. Breadth of coverage	0.77	0.82	0.22	0.66	0.77	0.81	0.83								
9. Assignments/ readings	0.74	0.80	0.15	0.63	0.67	0.76	0.76	0.74							
Endeavor scales															
10. Class discussion	0.93	0.64	-0.01	0.68	0.80	0.64	0.73	0.68	0.67						
11. Student accomplish-ments	0.73	0.91	0.19	0.64	0.75	0.89	0.85	0.81	0.76	0.63					
12. Workload	0.44	0.50	0.50	0.43	0.45	0.48	0.47	0.50	0.53	0.43	0.48				
13. Grading	0.72	0.68	0.18	0.86	0.75	0.69	0.67	0.69	0.61	0.65	0.68	0.51			
14. Personal attention	0.84	0.75	0.13	0.76	0.88	0.77	0.79	0.75	0.68	0.80	0.75	0.45	0.76		
15. Presentation clarity	0.76	0.87	0.17	0.65	0.72	0.91	0.89	0.79	0.76	0.69	0.83	0.48	0.68	0.78	
16. Organization/ planning	0.76	0.82	0.20	0.63	0.73	0.85	0.82	0.77	0.75	0.69	0.81	0.50	0.61	0.68	0.84

Note: Convergent validities are underlined.

Convergent validity refers to the correlations between SEEQ and Endeavor scales that are hypothesized to measure the same construct (see Table 1), while discriminant validity refers to the distinctiveness of the different dimensions and provides a test of the multidimensionality of the ratings. With only minor modifications, the criteria developed by Campbell and Fiske (1959) can be applied to these data as follows:

1. Convergent validities, correlations between SEEQ and Endeavor scales that are hypothesized to match, should be substantial. Here convergent validities ranged from 0.50 to 0.91 (median 0.88). So this condition is quite well satisfied for all seven cases.
2. One criterion of discriminant validity is that correlations between these matching scales should be higher than the correlations between non-matching SEEQ and Endeavor scales in the same row or column of the rectangular submatrix. This test is met by all but 3 of the 96 comparisons (counting half when the correlations are equal).
3. Another criterion of discriminant validity is that correlations between each convergent validity shoul be higher than those SEEQ and Endeavor scales correlations with the other eight SEEQ and six Endeavor scales, respectively. This test is met for all but 3 1/2 of the 98 such comparisons and, once again, this criterion is quite well satisfied. Inspection of Table 3 indicates that it is the Planning, and Workload scales which fail to satisfy criteria 3 and 4.
4. The pattern of correlations among SEEQ scales should be similar to the pattern of correlations among Endeavor scales (e.g. because the two SEEQ scales of Group Interaction and Individual Rapport are highly correlated, so should be the two corresponding Endeavor factors of Class Discussion and Personal Attention). Inspection of the correlations in Table 3 generally indicates a similarity in pattern of correlations.

Factors Related to Ratings

The mean responses of a number of factors possibly distinguishing between "good" and "poor" lecturers are reported in Table 4. It would seem that there is a trend for students to rate as "good" lecturers those who taught courses in their major area and in which they obtained higher grades. Higher ratings were not significantly associated with larger class sizes, the students' perceptions of their teacher's ages, or the lecturers' gender.

TABLE 4
Means of Classroom Aspects Possibly Related to Ratings of Lecturers (N = 87)

Item	"Good" Lecturer	"Poor" Lecturer
Is this subject your major (Yes = 1, No = 2)	1.21	1.34*
Student's estimate of size of class	39.78	35.45
Grade student obtained in course (A+ = 1 to E = 11)	4.19	6.59*
Estimated age of lecturer in years	42.97	41.56
Gender of lecturer (Male = 1, Female = 2)	1.11	1.13

* means are statistically different at the .01 level

Conclusions

The findings of this study generally lend support to the reliability and convergent and discriminant validity but throw doubt on the factor structure of the SEEQ and Endeavor instruments for use with Hong Kong students.

The internal consistency estimates obtained were fairly impressive for such short scales. The scale items differentiated "good" and "poor" lecturers in the ways expected. That workload and difficulty items did not differentiate was also quire predictable (and supports U.S. findings that students do not simply give high ratings to easy courses where they don't have to do much work). All but the items related to assessment were considered appropriate by most of the students while all but one item was chosen by at least a few as being most important. However, the item factor analysis only provided some support to a multidimenstional model of teaching effectiveness. A much stronger than expected general factor covering evaluative ratings of supposedly different aspects of teaching performance was obtained. However, investigation of the convergent and discriminant validity of the instruments by modified MTMM analysis provided clear evidence of the validity of the scales. An even stronger general factor was found in similar investigations with Indian and Nepalese students (Watkins & Regmi, 1992; Watkins & Thomas, 1991). Further research will be required to determine whether this is due to an artefact of "halo" effect resulting from asking students to rate a "good" and a "poor" lecturer or to Hong Kong, Nepalese and Indian students perhaps having a more global perception of teaching effectiveness than Western students. Qualitative investigations which explore student conceptions of teaching may well be needed to settle this issue.

These findings indicate that evaluation instruments developed at American universities may well be reliable when used in Hong Kong but that the separate components that underlie evaluations of teaching effectiveness at American universities may not be as distinct in Hong Kong. Taken together with earlier findings in Australia, New Zealand, Spain, Nepal, and India, strong support is provided for the cross-cultural reliability of these instruments but doubt is cast on the cross-cultural validity of the underlying model of teching effectiveness. Moreover, some of the difficulties attempting to compare lecturers' teaching effectiveness in practice are suggested by the finding that student ratings of lecturers may be related to factors such as the grade the student achieves in that course. To what extent such ratings represent genuine indicators of teaching effectiveness rather than biased responding is still a matter of conjecture although the American research seems to support the former conclusion (Marsh, 1987).

References

Astin, A.W. (1974). Measuring the outcomes of higher education. In H. Bowen (Ed.), *New directions for institutional research: Evaluating institutions for accountability* (No. 1). San Francisco: Jossey-Bass.

Campbell, D.T., & Fiske, D.W. (1959). Convergent and discriminant validation by the multitrait-multimethod 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.

Cohen, P.A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multi-section validity studies. *Review of Educational Research, 51,* 281-309.

Dunkins, M., & Barnes, J. (1986). Research of teaching in higher education. In M.C. Wittrock (Ed.), *Handbook of Research on Teaching* (3rd ed.). New York: McMillan.

Frey, P.W. (1978). A two-dimensional analysis of student ratings of instruction. *Research in Higher Education,* 9, 69-91.

Frey, P.W. (1982). Components of teaching. *Instructional Evaluation, 7,* 3-10.

Frey, P.W., Leonard, D.W., & Beatty, W.W. (1975). Student ratings of instruction: Validation research. *American Educational Research Journal, 12,* 327-336.

Hull, C.H., & Nie, H.H. (1984). *SPSS-X.* New York: McGraw-Hill.

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

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

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

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

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

Watkins, D., & Regmi, M. (in press). Student evaluations of tertiary teaching: A Nepalese investigation. *Educational Psychology.*

Watkins, D., & Thomas B. (1991). Assessing teaching effectiveness: An Indian perspective. *Assessment and Evaluation in Higher Education, 16,* 186-198.

Author

David WATKINS, Reader in Education, University of Hong Kong, Pokfulam Road, Hong Kong.