教育研究學報 1997，第十二卷第二期，205-213

Educational Research Journal 1997 Vol. 12, No. 2, pp. 205-213

Weighting the Dimensions of Student Ratings of Teaching Effectiveness

Derek Cheung

Hong Kong Baptist University

An important unresolved issue in student evaluation research is how different rating dimensions should be weighted for summative evaluation. This study demonstrated the use of hierarchical confirmatory factor analysis to determine the weights. A model for evaluating the teaching effectiveness of a B.Ed. programme delivered by distance education was hypothesized, which consisted of four first-order factors and one second-order factor. The four first-order factors represented four separate dimensions of the construct Teaching Effectiveness: Student Development; Assessment; Learning Materials; and Face-to-face Component. The second-order factor subsumed all the four dimensions and was treated as a general factor of effective teaching. End-of-semester student ratings were collected by a multifactor rating form. Hierarchical confirmatory factor analysis of the student ratings provided clear support for the a priori model. Ways to compute the weights for the four dimensions based on second-order factor loadings were demonstrated.

Introduction

Student evaluations of teaching are the predominant method used to evaluate the teaching effectiveness of university courses. However, administrators within tertiary education often complain about research that tells them that student evaluations of teaching effectiveness are multidimensional, saying that they are not informed about how they should weight the dimensions so as to compute an aggregate score for summative purposes or for making personnel decisions.

The purposes of student evaluations of teaching in tertiary education are well documented (e.g., see Marsh, 1987; Marsh & Roche, 1993). Teaching effectiveness is a multidimensional construct (Marsh, 1987). For formative purposes, researchers tend to agree that a multidimensional profile of student ratings is more useful than an average. For summative purposes or personnel decisions such as promotion, tenure, contract renewal, salary adjustment, resource allocation and merit awards, a carefully weighted average of specific dimensions seems to be superior to an unweighted average or an array of separate factor scores (Abrami, 1989; Marsh, 1987, 1991).

Although Rich (1976) found that 75% of the faculty respondents from a sample of colleges and universities agreed that student ratings should be used in tenure decisions, some researchers doubt whether administrators can properly utilize student evaluations for summative purposes or personnel decisions. For example, Franklin and Theall (1990) reported that departmental administrators were unable to interpret commonly used descriptive statistics and to select valid indicators of teaching effectiveness. They gave the following examples of poor practice:

> We saw that some departmental administrators, who routinely used ratings to make decisions about personnel, evaluation policy, and resource allocation, were not familiar enough with important ratings issues to make well-informed decisions. We regularly heard of personnel decisions that were made on the basis of a single course's ratings and of cases in which workload or difficulty were the deciding factors in such decisions, or in which mean scores, separated by tenths or even hundredths of a point, were accepted as valid indicators of individual differences in teaching performance. (Franklin & Theall, 1990, p. 78)

Key Words: Structural equation modeling; Student evaluation of teaching; Teaching effectiveness.

The author would like to thank Prof. John Hattie and Mr. Raymond Lam for their advice on running the LISREL program, and Mrs. Tammy Lai and two anonymous reviewers for their constructive comments on an earlier draft of this paper.

Correspondence concerning this article should be addressed to Derek Cheung, School of Continuing Education, Hong Kong Baptist University, 4/F Bank Center, 636 Nathan Road, Hong Kong. Fax: (852) 2304 6572 Email: spcheung@hkbu.edu.hk

Abrami and d'Apollonia (1990) also pointed out that administrators cannot be expected to have the expertise of synthesizing the information from various dimensions of teaching effectiveness. They explained that administrators tend to weight factor scores equally or focus on particularly strong or weak areas of teaching.

There are at least three reasons why it is important to solve the above weighting problem: (1) to understand how different dimensions of the construct "teaching effectiveness" are related to each other; (2) to prevent misuse of student ratings; and (3) to summarize student ratings as a weighted average which can then be used as a variable for other research studies. Furthermore, since the teaching effectiveness of university staff does not correlate with their research output (Hattie & Marsh, in press) and administrators need to make informed decisions based on student ratings of courses, sensible use of the student data is critically important. It is unfortunate that little research has been done to investigate how weight should be assigned to each dimension of teaching effectiveness. As Abrami and d'Apollonia (1990) and Marsh (1987) emphasized, precise and defensible procedures for determining the weights are not available. The situation in distance education is even worse than that in traditional on-campus courses because the knowledge about the design and organization of student feedback for courses taught by distance education is still in its infancy (Calder, 1994). To validly evaluate the teaching effectiveness of distance learning courses in tertiary education, measuring instruments with clearly defined dimensions are required. No such instruments with confirmed factor structure, however, can be found in the literature.

This paper reports the development of a rating form which is based on a conceptual framework for evaluating the teaching effectiveness of courses taught by distance education, and demonstrates how the weights of specific dimensions of the construct "teaching effectiveness" can be determined by means of hierarchical confirmatory factor analysis.

Literature Review

Methods Used to Determine the Weights in Past Studies

One of the important unresolved issues surrounding student evaluation research is the determination of the weights of individual dimensions of the construct "teaching effectiveness." Marsh (1987) suggested that the weight assigned to each dimension should be determined by logical and empirical analyses. Few research studies in this area, however, have appeared in the literature. A few researchers attempted to compute a weighted average, but they did not determine the weights of individual dimensions. For example, Cashin and Downey (1992) asked the instructor, a faculty committee or the department head to weight the ten course objectives on the Instructional Development and Effective Assessment rating form for each specific course. To compute a weighted composite criterion measure of teaching effectiveness, essential, important, and minor important objectives (not dimensions) were given double, single, and zero weights, respectively. Thus, the weighted average was not calculated by combining the multidimensional scores together. Marsh (1994, 1995) also revealed other weaknesses of this approach to determination of weights.

Marsh (1987) proposed that individual dimensions can be weighted in relative importance by the lecturer being evaluated, by the department head or a promotions committee. For example, Marsh and Roche (1993) asked lecturers to rate the importance of each dimension. The importance ratings were then ipsatized (Marsh & Roche, 1993, p. 229). The importance weighted total score was computed by taking the mean cross-product of each ipsatized importance rating multiplied by the student rating of the corresponding dimension. The weights were different from lecturer to lecturer. The drawback of this approach to determining the weights is obvious; a lecturer might heavily weight those dimensions on which he/she is effective and minimally weight other dimensions. Thus, the assignment of weights might be biased.

Kwan (1993) compared the rankings of quality of courses based on students' global ratings, total score ratings, and weighted multidimensional ratings. He constructed 12 items to measure 12 dimensions (i.e., one item per dimension), namely, relevance to study/profession, interest and challenge in presentation, enthusiasm about teaching and subject, organization in presentation, relationship with other subjects, concern for students, encouraging active learning, promoting independent learning, up-to-date knowledge, meaningful assignments, clarity of presentation, and perceived learning. Principal component analysis of student responses to the 12 items yielded two factors. Factor 1 contained all dimensions except "Relevance to study/profession" and "Relationship with other subjects" which loaded on factor 2. A weighted multidimensional score was computed as a linear combination of the standardized values of ratings on the 10 dimensions weighted by their respective factor score coefficients on the first factor. The second factor was discarded in the calculation of the weighted multidimensional score.

Broder and Dorfman (1994) assigned weights to attributes of teaching quality by ordinary least squares regression. The beta coefficients of attributes were determined. They reported that approximately 81 percent of the variation of the global ratings on instructor performance could be explained by four attributes: enthusiasm (24%), knowledge of subject (23%), tying information together (20%), and ability to stimulate thinking (14%). Similarly, Ryan and Harrison (1995) determined the beta weights for individual dimensions by multiple regression. However, the study was conducted by asking students to respond to hypothetical instructors in imagined classrooms and there was only one single item per dimension. Students were required to make overall evaluations of hypothetical instructors based on a manipulation of the nine factors in Marsh's Students' Evaluations of Educational Quality (SEEQ) Form. The beta weights for the nine SEEQ's factors were found by multiple regression. Ryan and Harrison suggested that a composite weighting scheme may be determined by asking the faculty in an academic unit to participate in a similar experiment.

Although multiple regression has been commonly used to determine the weights of dimensions, it is well known that regression parameters are faulty if the observed variables contain measurement errors or there is interdependence or simultaneous causation among the observed variables (Goldberger, 1973). These faults can be corrected by use of statistical methods such as structural equation modeling, but no such studies are reported in the literature.

The above review of literature has identified some of the major limitations of previous research studies, such as the use of single-item scales, analysis of student ratings at non-dimensional level, determination of the weights as regression parameters, and possible biases in the assignment of weights by instructors themselves. The present research incorporated some modifications on most past designs. It was believed that the specific dimensions must be weighted in proportion to the validity of data on the dimensions. Because the ratings are collected from students, not from the instructors being evaluated, the department head or a promotions committee, the weights should be empirically determined by analyzing the validity of student ratings. In this study, student ratings of courses were collected by multi-item scales rather than single-item scales, and the construct validity of student ratings at the dimension level was tested by hierarchical confirmatory factor analysis (HCFA) through the LISREL program (Joreskog and Sorbom, 1993). In order to determine the weights of individual dimensions, a model with four first-order factors and one second-order factor was hypothesized in the present research. The specific dimensions of the construct "teaching effectiveness" formed the four first-order factors, while the second-order factor represented a general factor of effective teaching. Issues surrounding the higher-order structure of student ratings of university courses are reviewed in the next section.

Higher-order Structures

For student evaluation research, there has been considerable debate on the number of higher-order factors in student ratings. Feldman (1976) inferred the presence of three higher-order factors based on the pattern of correlations among 19 different categories of student responses. The three higher-order factors related to instructor's presentation of material, facilitation of learning, and regulation of students.

A few researchers have attempted to find out the higher-order structure of student ratings by exploratory factor analysis (EFA). For example, Smalzried and Remmers (1943) inferred the existence of two second-order factors by analyzing their ten first-order factors through EFA. The two second-order factors were labelled as Empathy, and Professional Maturity. Similarly, Frey (1978) found two second-order factors based on his seven first-order factors, with labels Pedagogical Skill and Rapport. As Marsh (1991) pointed out, the methodological aspects of each of the above studies as an investigation of higher-order factors are weak—exploratory rather than confirmatory factor analysis was employed; each scale or first-order factor was measured by a single item only; and many items loaded substantially on more than one higher-order factor and thus findings were not easily interpreted.

The research of Marsh (1991) is probably the first published study of using HCFA to test higher-order factors of student ratings. He demonstrated that the nine first-order SEEQ factors were subsumed by four second-order factors, namely, Presenter, Rapport, Course Materials, and Workload. He also pointed out some necessary conditions under which HCFA can be applied to research on student ratings (e.g., the existence of a well-defined first-order factor structure, the upper limit for the fit of higher-order model), but the focus of his study was not on the determination of the weights of dimensions. In this study, HCFA was employed to investigate not only the first and second-order factor structures of students' ratings of teaching effectiveness, but also the weighting scheme for the first-order factors.

It is worth emphasizing that none of the past studies reported the existence of a single higher-order factor. Abrami (1985) proposed a single higher-order factor model, but no empirical evidence was shown. In the present research, the possibility of the existence of a single higher-order factor of student ratings was empirically tested by HCFA.

Context for the Study

The Bachelor of Education (B.Ed.) programme offered by Hong Kong Baptist University was used as the context for this research. The B.Ed. programme is a part-time in-service programme to upgrade the qualification of primary teachers as well as to enhance their professional competencies. It is mainly delivered by distance education and students can select their own route through an array of courses.

Like other distance education programmes, the B.Ed. programme provides students with study guides and books of reading. Students are separated from the lecturers or writers of the study guides, and tutorials are the only face-to-face component of the programme. The number of tutorials varies from 8 to 14, depending on the nature of a particular course. Students' performance in a course is usually assessed by written assignments, a terminal examination, and the extent of participation in tutorials.

A total of 16 courses were offered in the B.Ed. programme in the semester Spring 1996. Examples of courses are Issues in Human Development, Curriculum Development, Assessment Practices, Research Methods, School Counselling and Guidance, Language in Education, and Special Needs Education. Each student was allowed to take at most two courses in a semester. In Spring 1996, 1705 students enrolled in the B.Ed. programme. To assure and improve the quality of the courses offered by the B.Ed. programme, end-of-semester student evaluations of courses were conducted for both formative and summative purposes. Mean ratings of items were produced and distributed to members of the examination board, the course leader, lecturers, and tutors. Lecturers used the mean ratings to judge the quality of their courses and to make decisions on contract renewal of a particular tutor. They also wished to receive a weighted average to facilitate a final judgment about a course's effectiveness.

Methodology

Development of the Rating Form

One of the important factors affecting the quality assessment in distance teaching organizations is the quality of the courses offered to students (Calder, 1995). For student evaluation of the effectiveness of courses taught by distance education, no well-constructed rating forms with clearly defined factor structures can be found in the literature. This is undesirable because student ratings serve both summative and formative purposes, such as the provision of comparative data across different distance learning courses, monitoring of the consistency of standards, diagnosis of the need of improvement, appraisal of tutor performance, and identification of problematic areas (Calder, 1994). Although many tertiary institutions that offer distance education programmes routinely collect student ratings of their courses, little attention has been paid to the development of the rating forms, resulting in invalid and unreliable student data.

Teaching effectiveness is context-dependent (Murray, Rushton & Paunonen, 1990), and thus rating forms (e.g., Marsh's SEEQ, Frey's Endeavor instrument) which have been designed for evaluating traditional on-campus courses were considered not appropriate for distance education. Instead, the author has developed a conceptual framework for evaluating not only the B.Ed. courses offered by Hong Kong Baptist University, but also a variety of similar distance learning courses in other settings. The conceptual framework consists of four dimensions, and each dimension subsumes a number of attributes. Based on this conceptual framework, at least one item was constructed to evaluate each attribute. As a result, a total of 35 items were constructed: 7 items on student development, 7 items on assessment, 7 items on learning materials, and 14 items on face-to-face component. All items were written in Chinese and positively phrased. Students were required to rate the items on a 5-point scale, varying from "strongly agree" to "strongly disagree." In order to keep the focus solely on the determination of the weights of dimensions, this paper does not delineate the theoretical justifications for inclusion of the four dimensions in the conceptual framework and the process of development of the instrument. Interested readers can refer to Cheung (in press) for details. The nature of the four dimensions are summarized below:

1. *Student Development.* This dimension evaluates how students perceive their progress on relevant learning objectives of a particular distance education course. Three attributes are subsumed under this dimension, namely, cognitive, psychomotor, and affective learning outcomes. Sample items are: "After taking this course, I developed skills and points of view needed by professionals in the field most closely related to this course" and "after taking this course, I learned to apply the knowledge and concepts in new situations."
2. *Assessment.* The Assessment dimension emphasizes the quality of the process of assessment of student performance in a distance education course. Four attributes are included, which deal with guidance given to students on how to complete the assessment requirements, appropriateness of the assessment tools, accuracy and fairness of assessment, and difficulty and workload. Sample items are: "The written assignments had clear and specific instructions" and "the written assignments were relevant to and integrated with what was presented in the course."
3. *Learning Materials.* This dimension concerns the quality of the pre-packaged self-instructional written learning materials given to students in a distance education course. Ideally, the learning materials developed in-house should meet specialized needs of a particular course and should be professionally presented and produced. A total of four attributes are subsumed under this dimension, which focus on the design and production of the study guides, integration of the study guides with the selected readings, quality of the selected readings, and students' workload. Sample items are: "The selected readings were generally interesting" and "the questions in the study guide helped me to learn and achieve the objectives."
4. *Face-to-face Component.* This dimension relates to how students perceive the quality of tutorials (and lectures if any) incorporated in a distance education course. The effectiveness of this face-to-face component is governed by tutor performance. Ten essential attributes have been hypothesized, covering feedback to students, presentation, organization, tutor's knowledge base, students' belief of the value of tutorials, individual rapport, group interaction, tutor's expectation of student performance, enthusiasm, and breadth. Sample items are: "The tutor gave clear and understandable explanations" and "the tutor used the tutorial time appropriately and effectively."

Data Collection and Analysis

In Spring 1996, student ratings of each B.Ed. course were obtained in a regular tutorial session during the last two weeks of the semester. Students responded anonymously to the items on the rating form, and the tutor was asked to take a short break outside the classroom during the evaluation period. A total of 2121 completed rating forms were returned by students.

The student ratings were first coded by optical mark reading machine. Using the SPSS program, student data on each of the four dimensions were then separately analyzed based on the value of item-total correlation and Cronbach's alpha. Items with unsatisfactory item-total correlations were deleted. In addition, the student data on each dimension were checked by exploratory factor analysis with principal axis factor extraction. The number of factors was restricted to one and only those items with satisfactory factor loadings were retained to form a scale.

Having done the item analyses, all the remaining items were subjected to a first-order confirmatory factor analysis (CFA) using the

LISREL program (Joreskog & Sorbom, 1993). Maximum likelihood was selected as the method of parameter estimation and listwise deletion of cases with missing data was used. In the a priori model, each item was allowed to load on only one factor (i.e., the dimension the item was constructed to measure), and the errors associated with all items were posited to be uncorrelated. The ability of the four-dimension model to fit student responses was judged by the value of such fit statistics as the root mean square error of approximation, adjusted goodness of fit index and comparative fit index. The correlations among the four factors were also examined to see whether it was worthwhile to subject the student responses to a second-order CFA.

To determine the weights of the four dimensions, a single second-order factor was hypothesized, which incorporated the four first-order factors. HCFA was employed using the LISREL program, testing the first-order and second-order factors simultaneously in a single analysis. As in the first-order CFA, each item was allowed to load on only one of the four factors. The HCFA model was intended to explain the covariation among the four first-order factors. The second-order factor loadings (i.e., the gamma values generated by the LISREL program) were taken as the weights of specific dimensions of the construct "teaching effectiveness."

Results and Discussion

A total of six items (i.e., items Q7, 17, 20, 23, 28 and 33) were deleted because of their relatively low item-total correlations from reliability tests and low factor loadings from EFA. It was found that the variance of student responses to these six items was small. Perhaps they failed to show satisfactory item-total correlations and factor loadings because of lack of discriminating power. For example, 89.5% of students agreed or strongly agreed that "The tutor had a friendly attitude towards students" (item Q17), and only 12.4% of students disagreed or strongly disagreed that "The written assignments were challenging" (item Q23). To improve the rating form, further research is needed to construct new items which can generate more widely spread scores, particularly for the dimension Assessment.

The remaining 29 items are shown in Table 1. They constituted four scales which correspond to the four dimensions of teaching effectiveness conceptualized in this study. The values of Cronbach's alpha, item-total correlation, and factor loading from EFA indicate that the student ratings collected by these four scales are valid and reliable.

When the student ratings were subjected to first-order CFA, the model was found to fit the student responses quite well (see Table 2). Different factors measured distinct dimensions of the construct "teaching effectiveness." It is important to note that the model is very restrictive; there were a total of 116 factor loadings (i.e., 29

Table 1
Results of Item Analyses

Items		Item-total Correlation	Factor Loading from EFA
Student Development ($\alpha = .87$)			
Q1	Understand concepts	.59	.64
Q6	Apply knowledge	.65	.71
Q11	Communicate ideas	.65	.71
Q16	Stimulate interest	.63	.69
Q21	Value new viewpoints	.61	.67
Q26	Contribute to discussion	.63	.68
Q31	Develop skills	.68	.74
Assessment ($\alpha = .76$)			
Q3	Clear instructions	.52	.60
Q8	Questions well designed	.63	.77
Q13	Make students think	.59	.71
Q18	Relevant to the course	.50	.59
Learning Materials ($\alpha = .82$)			
Q4	Study Guide well designed	.53	.59
Q9	Integrated with readings	.51	.57
Q14	Interesting readings	.62	.69
Q19	Easy to understand	.58	.65
Q24	Help students to learn	.58	.65
Q29	Questions arouse interest	.57	.65
Q34	Amount of readings	.54	.61
Face-to-face component ($\alpha = .93$)			
Q2	Constructive comments	.64	.66
Q5	Interesting tutorials	.65	.70
Q10	Clear explanations	.81	.85
Q12	Knowledge base	.75	.78
Q15	Usefulness of tutorials	.70	.72
Q22	Encourage students	.60	.63
Q25	Realistic expectation	.63	.65
Q27	A variety of strategies	.75	.78
Q30	Well prepared	.77	.80
Q32	Skillful in observation	.71	.74
Q35	Time management	.79	.82

Note. The questionnaire items are paraphrased.

items x 4 factors), 87 of them were fixed at zero. So, the results were consistent with the *a priori* model and students could differentiate the four dimensions. Furthermore, the four first-order factors were found to be substantially correlated (see Table 3), indicating that a second-order CFA should be tested. The dimension Face-to-face Component was least closely related to the other three dimensions.

Table 2
Standardized First-Order Factor Loadings and Fit Statistics

Item No.	First-order CFA	Hierarchical CFA
Student Development		
Q1	.66	.66
Q6	.70	.70
Q11	.72	.72
Q16	.70	.71
Q21	.66	.66
Q26	.69	.69
Q31	.73	.74
Assessment		
Q3	.62	.63
Q8	.73	.73
Q13	.74	.74
Q18	.58	.58
Learning Materials		
Q4	.59	.59
Q9	.57	.57
Q14	.68	.68
Q19	.64	.64
Q24	.67	.66
Q29	.68	.68
Q34	.59	.58
Face-to-face Component		
Q2	.66	.66
Q5	.68	.68
Q10	.85	.85
Q12	.78	.78
Q15	.73	.73
Q22	.62	.62
Q25	.65	.65
Q27	.77	.77
Q30	.80	.80
Q32	.73	.73
Q35	.82	.82
Fit Statistics:		
Chi-square	2515.96(df=371)	2677.63(df=373)
Root mean square error of approximation	.056	.058
Root mean square residual	.034	.040
Goodness of fix index	.90	.90
Adjusted goodness of fit index	.89	.88
Comparative fit index	.92	.91

Table 3
Correlation among Factors

	First-order CFA				Hierarchical CFA			
	1	2	3	4	1	2	3	4
1	1.00				1.00			
2	.69	1.00			.73	1.00		
3	.72	.79	1.00		.74	.70	1.00	
4	.65	.48	.45	1.00	.56	.53	.53	1.00

Table 4
Standardized Second-order Factor Loadings and Error Variances

Factors	Second-order Factor Loadings	Error Variances
1	.88	.23
2	.84	.30
3	.84	.29
4	.64	.60

Tables 2 and 3 also display the results from the HCFA. It was found that the fit statistics were just minimally reduced by adding a single second-order factor to the a priori model. Hence, in contrast to all past studies, the existence of a single general second-order factor of the construct "teaching effectiveness" was confirmed.

The second-order factor loadings of individual dimensions and error variances in the prediction of the first-order factors from the second-order factor are shown in Table 4. There is no necessary connection between number of items constructed to measure a first-order factor and loading on a second-order factor in HCFA (E. Rigdon, personal communication); the sum of the standardized error variance and the square of second-order factor loading is actually equal to the factor variance. Factor 1 (i.e., the dimension Student Development) obtained the highest second-order factor loading. This implies that the amount learned by students in the course was the most important factor affecting student ratings of distance education courses. This finding is consistent with the result reported by Ryan and Harrison's (1995) study even though they did not employ distance education as their research context. Many years ago, Tyler (1958), and Cohen and Brawer (1969) also argued that student gains in a course should be the most direct and ultimate criterion for evaluating teaching performance.

Factor 2 (the dimension Assessment) and Factor 3 (the dimension Learning Materials) got the same size of second-order factor loading, but its value was slightly lower than that of Factor 1.

The second-order factor loading of Factor 4 (the dimension Face-to-face Component) was surprisingly low; it was the least important factor affecting student ratings of the distance learning programme. In other words, the tutors' influence was tangential to the overall teaching effectiveness. The reasons why students perceived the face-to-face component of the B.Ed. programme least effective were not investigated in the study. It seems that student ratings related more to those factors that they had more control over, such as the amount learned in the course, preparing the written assignments and examinations, and studying the pre-packaged self-instructional written learning materials. Although good tutors would help, students might have perceived that the quality of tutorials was largely out of their control. The B.Ed. programme required students to attend at least 75% of the tutorial sessions, but students were not allowed to select or change their tutors. Some students might have felt helpless if they had encountered less competent tutors. Some tertiary institutions, which offer distance learning courses, even do not require students' compulsory attendance at tutorials. Recently, Roberts (1996) asked distance education students to rank the helpfulness of six learning processes. Attendance at tutorials was just ranked fifth and was considered by students to be less helpful than studying the course materials, completing assignments, attending compulsory summer school, and reading and using the feedback on assignments. Kember and Murphy (1992) also pointed out that the tutor's effect is indirect in both distance education and open learning.

Clearly, the face-to-face component is an important human dimension of effective teaching in distance education. One of the major purposes of tutorials is to help deepen students' understanding of the topics covered in the written learning materials through activities such as group discussion, presentation and role-play. During tutorials, students are also provided with opportunities to listen to other students' problems and to build their confidence. In their study of the delivery formats of a variety of distance education courses offered in Hong Kong, Kember et al. (1992) interviewed 60 students and only one of them preferred telephone counselling to direct face-to-face contact with tutors. This implies that distance education students treasure the provision of tutorials. However, as the production of quality learning packages is vital to distance education, allocation of human and other resources by distance education institutions are usually biased towards the development of the learning materials. Less attention has been paid to the planning of tutorials (Jennings & Ottewill, 1996). Evidence from this study suggests that the potential benefits of this human dimension of effective teaching should be targeted for further investigation.

With the availability of the gamma values, procedures can then be formulated to compute the weighted average. One of the possible ways to calculate the weighted multidimensional average is to divide the gammas by the total gamma and use them as weights. Hence, if this procedure is used for the present study, the weights for the dimensions Student Development, Assessment, Learning Materials, and Face-to-face Component will be equal to 28%, 26%, 26% and 20%, respectively. Since the square of gamma is equal to the variance of the first-order factor that can be explained by the second-order factor, another possible way to compute the weighted average is to square the four gammas, and then divide each gamma squared by the total gamma squared. With this approach, the weights for the dimensions Student Development, Assessment, Learning Materials, and Face-to-face Component will be 30%, 27%, 27% and 16%, respectively. However, an empirical comparison of the validity of different weighting schemes was outside the scope of the present study. Future research should be planned to compare the validity and usefulness of a simple unweighted average, a weighted average with weights based on students' or lecturers' relative importance ratings, and a weighted average with weights empirically determined by HCFA. Furthermore, there is a major limitation of the present study. That is, the same student data set was used to establish the scales and to confirm the model. Researchers should try to improve this aspect of the research in their future design.

Conclusion

Although student evaluation of courses is central to quality assessment of distance education programmes, most evaluations were conducted on an ad hoc basis. This paper has reported how a rating form with good psychometric properties can be developed. Furthermore, it is believed that both multidimensional profiles of student ratings and a carefully weighted average are useful. If administrators are not familiar with multidimensional profiles or intend to use a total score for making a final judgment about a course's effectiveness, the weights assigned to individual dimensions of the construct Teaching Effectiveness should be pre-determined by evaluators, and then a weighted average can be given to administrators to avoid any misuse of student ratings. As Franklin and Theall (1990, p. 75) emphasized, "Administrators should be able to use ratings fairly and efficiently in performance appraisal without exposing themselves and their institutions to liability for misuse." It is difficult, if not impossible, to ask administrators themselves to employ a defensible empirical method to compute the weighted average. This paper has demonstrated that evaluators may help administrators solve the weighting problem by HCFA and compute the weighted average on the basis of the second-order factor loadings.

In this study, determination of the weights depends upon the one single second-order factor in the *a priori* model. I suggest that high-inference dimensions be conceptualized in order to increase the probability of success. Specific aspects of effective teaching can be treated as subdimensions or attributes. Provided that the first-order factors are moderately correlated, it is worth testing the second-order factor.

References

Abrami, P. C. (1985). Dimensions of effective college instruction. *Review of Higher Education, 8,* 211-228.

Abrami, P. C. (1989). How should we use student ratings to evaluate teaching? *Research in Higher Education, 30,* 221-227.

Abrami, P. C., & d'Apollonia, S. (1990). The dimensionality of ratings and their use in personnel decisions. In M. Theall, and J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice.* San Francisco: Jossey-Bass.

Broder, J. M., & Dorfman, J. H. (1994). Determinants of teaching quality: *What's important to students? Research in Higher Education, 35,* 235-249.

Calder, J. (1994). Course feedback: Its costs and benefits; its limitations and potential. In G. Dhanarajan, P. K. Ip, K. S. Yuen, & C. Swales (Eds.), *Economics of distance education: Recent experience.* Hong Kong: Open Learning Institute Press.

Calder, J. (1995). Evaluation and self-improving systems. In F. Lockwood (Ed.), *Open and distance learning today.* London: Routledge.

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

Cheung, D. (in press). Developing a student evaluation instrument for distance teaching. *Distance Education.*

Cohen, A. M., & Brawer, F. B. (1969). Measuring faculty performance. Washington, DC: ERIC Clearinghouse for Junior College Information / American Association of Junior Colleges.

Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education, 5,* 243-288.

Franklin, J., & Theall, M. (1990). Communicating student ratings to decision makers: Design for good practice. In M. Theall, & J. Franklin (Eds.), *Student ratings of instruction: Issues for improving practice.* San Francisco: Jossey-Bass.

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

Goldberger, A. S. (1973). Structural equation model: an overview. In A. S. Goldberger and O. D. Duncan (Eds.), *Structural equation models in the social sciences.* New York: Seminar Press.

Hattie, J. A., & Marsh, H. W. (in press). The relationship between research and teaching: a meta-analysis. *Review of Educational Research.*

Jennings, P. L., & Ottewill, R. (1996). Integrating open learning with face-to-face tuition: a strategy for competitive advantage. *Open Learning, 11*(2), 13-19.

Joreskog, K. G., & Sorbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Chicago: Scientific Software International, Inc.

Kember, D., Lai, T., Murphy, D., Siaw, I., & Yuen, K. S. (1992). A synthesis of evaluations of distance education courses. *British Journal of Educational Technology, 23,* 122-135.

Kember, D., & Murphy, D. (1992). *Tutoring distance education and open learning courses.* Campbelltown, NSW: Higher Education Research and Development Society of Australia Inc.

Kwan, K. P. (1993). Using student ratings to evaluate teaching performance. Paper presented at the annual conference of Hong Kong Educational Research Association, Hong Kong.

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. (1991). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. *Journal of Educational Psychology, 83,* 285-296.

Marsh, H. W. (1994). Weighting for the right criteria in the instructional development and effectiveness assessment (IDEA) system: Global and specific ratings of teaching effectiveness and their relation to course objectives. *Journal of Educational Psychology, 86,* 631-648.

Marsh, H. W. (1995). Still weighting for the right criteria to validate student evaluations of teaching in the IDEA system. *Journal of Educational Psychology, 87,* 666-679.

Marsh, H. W., & Roche, L. (1993). The use of students' evaluations and an individually structured intervention to enhance university teaching effectiveness. *American Educational Research Journal, 30,* 217-251.

Murray, H. G., Rushton, J. P., & Paunonen, S. V. (1990). Teacher personality traits and student instructional ratings in six types of university courses. *Journal of Educational Psychology, 82,* 250-261.

Rich, H. E. (1976). Attitudes of college and university faculty toward the use of student evaluation. *Educational Research Quarterly, 3,* 17-28.

Roberts, D. (1996). Feedback on assignments. *Distance Education, 17,* 95-116.

Ryan, J. M., & Harrison, P. D. (1995). The relationship between individual instructional characteristics and the overall assessment of teaching effectiveness across different instructional contexts. *Research in Higher Education, 36,* 577-594.

Smalzried, N. T., & Remmers, H. H. (1943). A factor analysis of the Purdue Rating Scale for instructors. *Journal of Educational Psychology, 34,* 367-369.

Tyler, R. W. (1958). The evaluation of teaching. In R. M. Cooper (Ed.), *The two sides of the log.* Minneapolis: University of Minnesota Press.