

Educational and Psychological Measurement

<http://epm.sagepub.com>

A Multilevel Factor Analysis of Students' Evaluations of Teaching

Michael D. Toland and R. J. De Ayala

Educational and Psychological Measurement 2005; 65; 272

DOI: 10.1177/0013164404268667

The online version of this article can be found at:
<http://epm.sagepub.com/cgi/content/abstract/65/2/272>

Published by:

 SAGE Publications

<http://www.sagepublications.com>

Additional services and information for *Educational and Psychological Measurement* can be found at:

Email Alerts: <http://epm.sagepub.com/cgi/alerts>

Subscriptions: <http://epm.sagepub.com/subscriptions>

Reprints: <http://www.sagepub.com/journalsReprints.nav>

Permissions: <http://www.sagepub.com/journalsPermissions.nav>

Citations (this article cites 15 articles hosted on the SAGE Journals Online and HighWire Press platforms):
<http://epm.sagepub.com/cgi/content/refs/65/2/272>

A MULTILEVEL FACTOR ANALYSIS
OF STUDENTS' EVALUATIONS OF TEACHING

MICHAEL D. TOLAND
R. J. DE AYALA
University of Nebraska–Lincoln

This study examined the factorial validity of scores on the newly developed Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS) through a series of confirmatory and multilevel structures. Conventional confirmatory factor analyses using the total covariance and pooled within-covariance matrices from two midwestern universities indicated that a reduced 25-item SETERS fit the data better than the original 34-item SETERS. Furthermore, multilevel factor analysis was conducted on the combined samples. This analysis suggested that one or three factors at the between and within levels were a plausible representation of SETERS scores. Pearson's correlations between individual scores on the SETERS and the Students' Evaluation of Educational Quality questionnaire provided additional validity evidence for the two measures. The need for additional empirical research on the SETERS before widespread use is discussed.

Keywords: *teaching effectiveness; factor analysis; multilevel; validity*

Previous studies examining the structural validity of scores from a measure of students' evaluations of teaching (SET) have relied on individual responses or class-average responses as the unit of analysis. However, when using individual (student) responses as the unit of analysis, the standard statistical assumption of independence is violated. Specifically, conventional factor analysis assumes that students within classes share no common perceptions when rating their respective teachers. When this assumption is violated, any substantial similarities within groups lead to inaccurate estimates

Correspondence concerning this article should be addressed to Michael D. Toland, Department of Educational Psychology, University of Nebraska–Lincoln, Lincoln, NE 68588-0345; e-mail: tolandmd@unlserve.unl.edu.

Educational and Psychological Measurement, Vol. 65 No. 2, April 2005 272-296
DOI: 10.1177/0013164404268667
© 2005 Sage Publications

of a model's parameters, standard errors, and fit indices (Heck, 2001). For example, Marsh (1983) discussed how "the class-average response is nearly always appropriate, and any findings based upon individual students as the unit of analysis must also be demonstrated at the class-average level" (p. 152).

Since Marsh's (1983) statement, most research examining the dimensionality of students' ratings of university teaching have used class-average responses as the unit of analysis for single-level analyses and have commonly found these ratings to be multidimensional in nature (d'Apollonia & Abrami, 1997). One of the better supported measures of SET has been an instrument originally developed by Marsh and his colleagues (Marsh, 1977, 1983; Marsh, Fleiner, & Thomas, 1975), the Students' Evaluation of Educational Quality (SEEQ) questionnaire. The SEEQ questionnaire consists of 35 items that make up nine dimensions. These nine evaluation dimensions are learning/value, enthusiasm, organization, group interaction, individual rapport, breadth of coverage, examination/grading, assignments, and workload/difficulty. Numerous studies conducted with Marsh as lead researcher have supported the nine dimensions that constitute the SEEQ using class-average responses (e.g., Marsh, 1983, 1991; Marsh & Hocevar, 1991).

Although researchers have most often used class-average responses in confirming the factor structure of SET, there still remains a problem with single-level analysis. Specifically, results found at the group-average level appear stronger than they would be if within-group variation were also incorporated into the analysis, because all the variability present within each group unit (e.g., all students' ratings within a class) is collapsed to a single mean (Kaplan & Elliott, 1997). Once again, the class-average response ignores the individual variability that occurs within each class. In most cases, using individual responses and class-average responses as the primary units of analysis presents problems for single-level analyses.

Recently, Ting (2000) acknowledged the problem with data analyzed at the individual or class-average level by applying hierarchical linear modeling (HLM) to data he collected. Ting examined students' ratings as outcomes (e.g., lecturing performance, course design, and overall evaluation) of an interaction process among different types of student-, teacher-, class-, and course-level variables in the Chinese context. This study's outcomes showed that like their Western counterparts, Chinese students were able to evaluate their learning conditions according to different dimensions of teaching effectiveness. Also, the three overall measures of teaching effectiveness correlated positively with one another.

A more recent study by Marsh and Hattie (2002) also used HLM. As part of a larger study, Marsh and Hattie investigated the extent to which the relationship between teaching effectiveness and research productivity varied as a function of academic department. For the purposes of their study, teaching

effectiveness was assessed using the standard teaching evaluation form, which included an overall course rating composed of students' ratings of overall course value, course materials, and teacher presentation. HLM analyses from this study indicated that the teaching-research relationship was near zero and did not vary as a function of academic department.

Marsh and Hattie's (2002) and Ting's (2000) studies are two of the only studies found in a review of the literature to have applied multilevel modeling to SET. As Muthén (1991) suggested, this may be because statistical methodology and software development have lagged behind research needs. Although software development has advanced to deal with the problem of what constitutes a proper unit of analysis, no research studies were found that examined the dimensionality of SET using multilevel factor analysis (MFA). MFA is able to take into account the nested observations as well as shed light on within- and between-class variance components (Muthén, 1991). Therefore, the present study applied MFA modeling to Abrami, d'Apollonia, and Rosenfield's (1996) proposed SET model.

To better understand Abrami et al.'s (1996) model, some description of their research is necessary. Abrami et al. used multivariate meta-analysis techniques to identify the common factors across SET instruments. In their study, they coded more than 458 items on 17 scales of SET into common factor categories. Then, they used a subset of the items that had consistent intercategory correlation coefficients across student rating forms and combined the intercategory correlation coefficients to produce an aggregate correlation matrix. Finally, they factor-analyzed the correlation matrix using principal components (PC) analysis. Their results indicated a common PC across the SET forms that explained about 63% of the variance in teaching effectiveness. However, for reasons that are unclear, they subsequently conducted an oblique rotation with a δ value of 0.2 and found four first-order components. The first component was labeled "instructor's role in delivering information" and included statements about overall course and instructor ratings, clarity, presentation, organization, monitoring, and the instructor's enthusiasm. The second component was labeled "instructor's role in facilitating a social learning environment," which consisted of statements about concern for students, the availability of the instructor, respect for students and diversity, and the friendliness of the classroom atmosphere. The third component was labeled "instructor's role in regulating student learning," which consisted of statements about teacher feedback and evaluation. The fourth component, which was not interpretable, included a mixture of statements about the instructor's knowledge of course content and class materials and the use of instructional objectives.

d'Apollonia and Abrami (1997) extended Abrami et al.'s (1996) research by discussing how the first-order component structure found by Abrami et al. was similar to the three roles of the instructor found by Feldman (1976). That

is, when the third and fourth components were combined, they more closely resembled the three factors found by Feldman. Abrami et al. subsequently conducted a second-order component analysis, which indicated that students' ratings of teaching effectiveness did indeed measure a general instructional skill (GIS), therefore confirming the PC analysis that there was a global measure of instructional skill. The GIS consisted of three instructional skills: the delivery of instruction, the facilitation of interactions, and the evaluation of student learning.

The purpose of the present study was to build on the previous research of Abrami et al. (1996) and d'Apollonia and Abrami (1997). Specifically, we tested the theoretical predictions of d'Apollonia and Abrami's three-factor model by developing a new measure, the Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS). The SETERS quantifies the level of teaching effectiveness on three factors: Instructor's Delivery of Course Information (ID), Teacher's Role in Facilitating Instructor/Student Interactions (I/S), and Instructor's Role in Regulating Students' Learning (RL). Additionally, the present study sought to summarize and explore the three-factor model within and between variations and to establish if the same model holds at the individual and class levels.

First, we hypothesized that a three-factor solution, as proposed by d'Apollonia and Abrami (1997), would perform better than a global measure of teaching effectiveness. This is because the majority of previous studies have found SET measures to be multidimensional in nature. Second, it was expected that scores on all three of the factors of the SETERS would be positively related to the nine factors of Marsh's (1983) SEEQ questionnaire. Third, it was expected that scores on the SEEQ factor Examination/Grading would have the strongest positive relationship with the factor RL on the SETERS, because the items on these specific factors are the most closely matched of all factors. Finally, scores on all SETERS factors were predicted to have the weakest relationship with the SEEQ factor Workload/Difficulty because the items that constitute this SEEQ factor are the least closely matched of any items on the SETERS.

Method

Measures

SETERS. Teaching effectiveness was defined by the following three factors: (a) Instructor's Delivery of Course Information (e.g., enthusiasm, organization, presentation, clarity), (2) Teacher's Role in Facilitating Instructor/Student Interactions (e.g., group interaction; rapport; understanding learners' backgrounds, ethnicities, and attitudes), and (3) Instructor's Role in Regulating Students' Learning (e.g., exams, assignments, readings, quizzes).

A pool of 45 items (15 items per factor) were written and developed by the primary author on the basis of the definition of each factor outlined above and in accordance with Abrami et al.'s (1996) and d'Apollonia and Abrami's (1997) three-factor model, a review of the literature on SET (e.g., Jackson et al., 1999; Kindsvatter, Wilen, & Ishler, 1988; Marsh & Dunkin, 1992; Young & Shaw, 1999), and forms in current usage. Also, some general principles of teaching and learning put forth by Fincher (1985) and summarized by Marsh and Dunkin (1992) helped aid the development of the items used on the SETERS.

Four full-time graduate students, familiar with item development, within the Department of Educational Psychology at a moderately sized, midwestern public university assessed items for clarity, redundancy, and ambiguity. Of the original 45 items, 11 were deleted on the basis of the graduate students' recommendations that these items were redundant or poorly worded.

To provide content validity evidence, six preselected full-time faculty members from a small, midwestern private university who were familiar with the content area of students' evaluations of university teaching were recruited to review the 34 newly revised items. They were informed that they were to evaluate each of the items for clarity and then match individual items with appropriate factors for content validity. Faculty members' suggestions on item clarity and appropriate factors were considered, and modifications to items were made. SETERS items were then assigned a 5-point, Likert-type rating scale, with higher scores suggesting that an instructor had a superior teaching skill on that item or factor.

A demographics page was added to the beginning of the SETERS form. This page contained questions on gender, class standing (freshman, sophomore, junior, or senior), grade point average prior to the course, level of interest in the subject matter prior to the course (1 = *very low* to 5 = *very high*), and expected course grade.

A preliminary version of the teaching effectiveness rating scale was pilot tested on approximately 27 volunteer undergraduate students enrolled in a section of introductory statistical methods. Undergraduate students were asked to evaluate their instructor using the SETERS while being timed for approximately 15 minutes and to further discuss any ambiguous questions or problems they may have had with the instrument. Students' feedback on these items was summarized and used to adjust some of the existing items for wording. Additional psychometric evaluation of the SETERS was conducted and is described in detail later. The final version of the 25-item SETERS is given in the Appendix.

SEEQ questionnaire. To establish evidence of convergent validity for scores on the SETERS, data were collected on one of the best known developed student evaluation questionnaires, the SEEQ questionnaire (Marsh,

1987). The SEEQ is a 35-item (scored on a 5-point, Likert-type scale) questionnaire designed for use with undergraduate- and graduate-level students. The nine SEEQ factors are Learning/Value (4 items), Instructor Enthusiasm (4 items), Organization/Clarity (4 items), Group Interaction (4 items), Individual Rapport (4 items), Breadth of Coverage (3 items), Examination/Grading (4 items), Assignments/Readings (4 items), and Workload/Difficulty (4 items). The consistency of student's ratings on the basis of previous research has shown subscale reliability estimates (coefficient α) ranging from .87 to .98 (Marsh, 1983) and subscale interrater reliability estimates for class-average responses ranging from .90 to .95 (Marsh & Hocevar, 1984).

Participants

The target population was undergraduate students. All students voluntarily participated in this study for extra credit or to receive the partial fulfillment of research participation required for their particular courses. Because data were collected from multiple courses in this study, it was possible for students to respond more than once. Therefore, one of any multiple occurrences was randomly retained, and all remaining occurrences were removed from the data set.

The remaining data consisted of SET from 35 undergraduate classes at a small, midwestern private university (Sample 1) and 19 undergraduate classes at a moderately sized, midwestern public university (Sample 2). Across the two samples, there were 828 undergraduate students (387 in Sample 1 and 441 in Sample 2), 35 courses (26 in Sample 1 and 9 in Sample 2), and 33 teachers (21 in Sample 1 and 11 in Sample 2). Students from Sample 1 ranged in age from 19 to 62 years ($n = 383$), with a mean age of 21.29 years ($SD = 4.97$ years), and students from Sample 2 ranged in age from 17 to 55 years ($n = 439$), with an average age of 20.77 years ($SD = 3.53$ years). Of the 54 classes, 35.2% were from Sample 2, and 18 classes sampled at this university were from the college of education and one was from arts and sciences. Ten of the instructors in Sample 2 were graduate student instructors, and 1 was a professor. Of the teachers sampled in Sample 1, 5 were professors, 5 were associate professors, 4 were assistant professors, and 7 were instructors. Of those courses sampled at this university (Sample 1), 9 were from the biblical studies division, 13 were from the general education division, and 4 were from the professional studies educational division.

Procedure

Data were collected 3 weeks before the end of the semester from students enrolled in classes at either university. Students enrolled in any of the classes had the opportunity to complete a packet that consisted of an informed con-

sent form, the demographic sheet, and two measures of SET, the SETERS and the SEEQ questionnaire. At the small, midwestern private university, during the classes' regularly scheduled times, a student volunteer followed an outline of study directions for administering and collecting packets from students. At the moderately sized, midwestern public university, at a prearranged time in a room other than the assigned classroom, the primary author followed the same outline of directions for administering and collecting packets from students.

At the beginning of each administration, students were told of the study's purpose and given the packet described above. Once all students had completed the packet, they were reminded of the study's purpose and provided with mailing and e-mail addresses to allow them the opportunity to request additional information pertaining to the study findings.

Data Analysis

MFA was applied to these data to allow simultaneous examination of the within- and between-class structures while taking measurement error into account. Muthén (1994) suggested that an MFA should involve four steps: (a) conventional confirmatory factor analysis (CFA), (b) the estimation of between variation, (c) the estimation of within structure, and (d) MFA.

The present study implemented the above strategies for each sample and combined samples. First, the students' responses in Sample 1 were used to assess the fit of the three- and one-factor models through conventional CFA. Then, Muthén's (1994) second and third steps were used with this sample to determine the removal of previously identified redundant items from the 34-item instrument. After the removal of these redundant items, the reduced model was tested. Similarly, these same three steps were conducted with the respondents in Sample 2 to provide a cross-validation of the measurement model established in Sample 1. Because Sample 2 was independent of Sample 1, it gave some protection against capitalization on chance and specification errors that were internal to the model (Mueller, 1997).

Because multilevel analysis usually requires at least 50 to 100 groups to look at the between-class variation and draw full conclusions (Muthén, 1994), the data collected in both samples were combined for multilevel analysis. The two samples could have been kept separate, but then it would not have been possible to use MFA to explore the three-factor model, because there was an insufficient number of classes to look at the between portion of the model (R. Heck, personal communication, February 17, 2002; Muthén, 1994). Moreover, results from the combined samples should be interpreted with caution, because each sample was previously tested for significance, and by examining the combined samples, we have capitalized on chance results by inflating Type I error.

Six indices were used to assess the measurement model's fit to the data with the CFAs. These indices included the χ^2 index, the goodness-of-fit index (GFI), the nonnormed fit index (NNFI), the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean residual (SRMR). The MFA models were tested with Muthén and Muthén's (1998) quasi-maximum likelihood fitting function (MLM), which includes robust standard errors and adjustment to the χ^2 test statistic due to unbalanced group sizes. Currently, there are a few indices available in Mplus (Muthén & Muthén, 2001) for testing multilevel models with unbalanced group sizes and MLM estimation. Three of these indices are the CFI, the RMSEA, and the SRMR. Parameter estimates were also used to assess the measurement models' fit during all these analyses.

The six above-mentioned fit indices were chosen for this study because no single fit index is considered to be the definitive marker of a model with "good" fit; each index serves a different purpose and should be interpreted in combination with the other indices. The χ^2 index is an absolute index that tests for lack of fit resulting from overidentifying restrictions placed on a model. A nonsignificant p value (e.g., $p > .05$) is desired, but the χ^2 index is usually inflated by the number of restrictions imposed on a model and sample size. Values of 1 for the GFI and the NNFI indicate perfect model fit; however, some researchers have suggested cutoff values greater than .95 to indicate good model fit. The following fit index cutoff values suggested by Hu and Bentler (1999) were used for determining goodness of fit: CFI $> .95$, RMSEA $< .06$, and SRMR $< .08$.

The consistency of students' ratings was estimated for the total scale and each factor on the SETERS and SEEQ by computing Cronbach's coefficient α and the interrater reliability estimates for class-average responses (see Winer, 1971). Validity evidence for the SETERS was obtained by estimating the relationship between the factors of the SETERS and those of the SEEQ questionnaire (Marsh, 1987) by using Pearson's correlation coefficient. Specifically, convergent evidence was established when a specifically hypothesized high positive correlation existed between scores on the SETERS and SEEQ factors.

Results

Because the accuracy of the maximum likelihood (ML) ratio χ^2 statistic is based to some extent on the degree of multivariate normality of the observed data, we examined the tenability of this assumption for our data. As a preliminary step, all items were examined for univariate normality by the inspection of each item's skewness and kurtosis. Kline (1998) suggested that all absolute index values (i.e., skewness and kurtosis indices) less than 2 should be considered to reflect a fairly normal distribution. According to this criterion,

the skew and kurtosis for 32 of the 34 items in Sample 1 were reflective of a normal distribution, whereas 31 of the 34 items in Sample 2 were fairly normal. In both samples, all items had skew values less than 2.0, but kurtosis values for three items indicated a potential problem. The normalized values of Mardia's index of multivariate kurtosis were 26.93 and 28.07 for each sample. These large and statistically significant values indicate that the items collectively do not have a multivariate normal distribution. However, these normalized values may be statistically significant in a large sample, similar to the sample sizes used even with small deviations from multivariate normality (Kline, 1998).

In conventional CFA, the total covariance matrix could be corrected for any lack of multivariate normality by using Satorra and Bentler's (1994) scaled χ^2 statistic and robust standard errors. However, there is not a method for correcting any lack of multivariate normality in the pooled within-group covariance matrix for individual-level analysis. Therefore, ML estimation will be used throughout individual-level analyses for comparative purposes, and results should be interpreted with some caution.

Prior to analyses, Item Pairs 10 and 11, 12 and 14, 29 and 31, 33 and 34, 18 to 20, and 21 to 23 were identified as being redundant items. For instance, Items 10 and 11 both reflected the teacher's role in the presentation of course material, were worded identically, and differed only with respect to the time referent. Because one of the goals of the study was to retain items that contributed most to the three-factor model, these redundant items were retained in the initial CFAs for the SETERS. Therefore, item coefficients, error variances, reliabilities, interrater reliability estimates for class averages, corrected item total correlations, means, and standard deviations were inspected to help delineate which of the redundant items were not contributing to the performance of the model and should be dropped from future analysis.

Initial CFAs for the SETERS

Initial descriptive statistics about the 34 items scored on the SETERS indicated that students in both Sample 1 ($n = 381$) and Sample 2 ($n = 405$) believed that their teachers were relatively high in teaching effectiveness (with item means ranging from 3.56 to 4.42 in Sample 1 and from 3.67 to 4.59 in Sample 2); items had relatively small standard deviations, ranging from 0.82 to 1.15 in Sample 1 and from 0.58 to 1.06 in Sample 2. The findings of scale analyses are summarized in Table 1. Sample 1 α coefficients ranged from .91 to .94 across the factors and total scores, with item total correlations ranging from .53 to .77. Sample 2 α coefficients ranged from .88 to .93, and item total correlations ranged from .38 to .73. The interrater reliability estimates for items in both Samples 1 and 2 varied from .42 to .82 and

Table 1
Descriptive Information for the Subscales and Total Scale on the Students' Evaluation of Teaching Effectiveness Rating Scale in Samples 1 and 2

	Sample 1				Sample 2			
	ID	I/S	RL	Total	ID	I/S	RL	Total
34-item scale								
Number of items on original scale	12	10	12	34	12	10	12	34
<i>M</i>	46.8	40.3	45.5	133	49.4	41.6	47.2	138.39
Adjusted <i>M</i> ^a	3.90	4.03	3.79	3.90	4.12	4.16	3.94	4.07
<i>SD</i>	9.18	7.18	8.43	23.1	7.75	6.20	7.15	19.22
α	.94	.93	.91	.97	.93	.89	.88	.96
Interrater reliability estimates for class averages								
ICC	.84	.61	.78	.79	.94	.90	.89	.92
ICC	.32	.13	.25	.26	.47	.28	.27	.37
25-item scale								
Number of items on reduced scale	9	6	10	25	9	6	10	25
<i>M</i>	35.2	24.3	38.2	97.80	37.20	25.6	39.7	102.66
Adjusted <i>M</i> ^a	3.91	4.06	3.82	3.91	4.13	4.27	3.97	4.11
<i>SD</i>	6.98	4.16	6.81	16.8	5.78	3.52	5.72	13.72
α	.92	.86	.89	.96	.9	.82	.84	.94
Interrater reliability estimates for class averages								
ICC	.84	.78	.65	.8	.95	.88	.88	.92
ICC	.32	.14	.25	.27	.48	.22	.26	.37

Note. ID = Instructor's Delivery of Course Information; I/S = Teacher's Role in Facilitating Instructor/Student Interactions; RL = Instructor's Role in Regulating Students' Learning; ICC = intraclass correlation.

a. The adjusted mean was determined by dividing the subscale mean by the number of items on the subscale.

from .65 to .93, whereas interrater reliability estimates for the factors ranged from .61 to .84 and from .89 to .94.

Fit statistics for the 34-item one- and three-factor models using the total covariance matrix (S_T) are presented in Table 2. The single-level CFA using ML estimation was performed to determine the fit of three- and one-factor models in each sample. Analyses showed a poor fit for the three- and one-factor models, but all item parameters estimates were statistically significant at the .01 level. Chi-square difference tests were also conducted, for instance, to compare the χ^2 values for the one-factor model (Model 1) and the three-factor model (Model 2; Table 2). Tests indicated that the three-factor model (Model 2) fit these data statistically significantly better than the one-factor model (Model 1) in Sample 1, $\chi^2(3) = 807.89$, $p < .001$. Similarly, the three-factor model (Model 8) fit these data better than the one-factor model (Model 7) in Sample 2, $\chi^2(3) = 514.23$, $p < .001$ (see Table 2).

Although the single-level models using S_T did not fit either sample very well, the intraclass correlation (ICC) for the observed indicators indicated

Table 2
Chi-Square and Fit Indices for the Students' Evaluation of Teaching Effectiveness Rating Scale for Samples 1 and 2

Method	Model	χ^2	df	Probability	RMSEA	GFI	CFI	NNFI	SRMR
Sample 1									
Original items									
S_T	1 (34 items, one factor)	3,132.25	527	.00	.140	.60	.74	.72	.070
	2 (34 items, three factors)	2,324.36	524	.00	.100	.71	.82	.81	.079
S_{PW}	3 (34 items, one factor)	2,531.78	527	.00	.130	.63	.74	.72	.071
	4 (34 items, three factors)	1,899.29	524	.00	.095	.73	.82	.81	.078
Respecification									
	5 (25 items, one factor)	967.19	275	.00	.096	.79	.85	.83	.059
	6 (25 items, three factors)	820.02	272	.00	.082	.83	.88	.86	.057
Sample 2									
Original items									
S_T	7 (34 items, one factor)	2,555.10	527	.00	.120	.67	.75	.73	.069
	8 (34 items, three factors)	2,040.87	524	.00	.051	.74	.81	.80	.073
S_{PW}	9 (34 items, one factor)	2,049.92	527	.00	.099	.72	.73	.71	.069
	10 (34 items, three factors)	1,749.95	524	.00	.085	.77	.78	.77	.071
Respecification									
	11 (25 items, one factor)	761.32	275	.00	.075	.85	.85	.84	.056
	12 (25 items, three factor)	679.87	272	.00	.067	.87	.88	.87	.054

Note. RMSEA = root mean square error of approximation; GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = nonnormed fit index; SRMR = standardized root mean residual.

that there was considerable variation ($ICC > 5\%$) between classes, with ICCs ranging from 6% to 29% for Sample 1 and from 8% to 43% in Sample 2. Table 1 indicates that there was about the same level of variation between groups for the three factors of effective teaching, with ICCs ranging from 13% to 32%. In Sample 2, the variation between groups was slightly higher for the three factors (Table 1). Because individual-level measurement error contributes to the within variances, it probably deflated the ICCs (Muthén, 1994).

Given that there was sufficient between-class variation in the observed variables (i.e., $ICC > 5\%$), it was reasonable to proceed to the estimation of the within structure (Muthén, 1994). The fit of the two models in each sample on the basis of the pooled-within covariance matrix (S_{PW}) was examined. Even though all item parameter estimates were statistically significant, at the .01 level for both the three- and one-factor models, the conventional ML analysis and indicators of model fit gave a poorer fit for S_T than for S_{PW} in each sample for both the one- and three-factor models (Table 2). Because the difference in the number of observations between the two approaches (S_T vs. S_{PW}) was negligible ($n = 381$ vs. $n - G = 346$ in Sample 1, and $n = 405$ vs. $n - G = 386$ in Sample 2, where G is the number of classes), sample size differences cannot in and of themselves explain the results. Heck (2001) and Muthén (1994) pointed out that when conventional factor analysis is used with hierarchical data, the model test of fit is inflated because of nonzero ICCs. Therefore, the poor fit in the conventional factor analyses might have been expected given the nonzero ICCs. Although five of the six indicators for the within-structure models using S_{PW} indicated poor-fitting models, the three-factor model fit statistically significantly better than the one-factor model in both samples. For instance, looking at Models 3 and 4 for Sample 1 in Table 2 shows that the χ^2 values using S_{PW} for the 34-item one- and three-factor models were 2,531.78 ($df = 527$) and 1,899.29 ($df = 524$). The χ^2 difference test shows the superiority of the three-factor model over the one-factor model in Sample 1, $\chi^2(3) = 632.49$, $p < .001$, and in Sample 2, $\chi^2(3) = 299.97$, $p < .001$.

No attempt was made to modify the structure of the 34-item model, because the lack of fit was primarily seen as a result of the redundant items that were described earlier. These same items also exhibited highly correlated errors, as indicated by the sizes of modification indices and standardized residuals. This analysis, in conjunction with the item descriptive statistics, theory, and content appropriateness, was used to determine which of the duplicate items were not contributing to the performance of the model.

Modification of the SETERS

Items 11, 14, 18, 19, 21, 22, 29, and 34 were dropped from the SETERS because the content appropriateness and/or phrasing were already being met by other related items on each respective factor. Additionally, Item 7 was dropped from the scale because its ICC, interrater reliability estimate for class average, and corrected item total correlation values were consistently lower relative to other items on the ID factor.

The reduced three-factor model consisted of 25 items with 9, 6, and 10 items per ID, I/S, and RL factor, respectively. Subscale coefficient α values were slightly less than those from the 34-item scale. Means, standard deviations, and interrater reliability estimates of class average for factors remained about the same after the deletion of items from the original 34-item scale. In general, the original 34-item scale and reduced 25-item scale results are fairly similar across both samples (Table 1).

The reduced one- and three-factor models estimated with ML using S_{PW} indicated improved model fit compared with the 34-item scale. For example, in Sample 1, the 34-item three-factor model χ^2 was 1,899.29, $df = 524$, and SRMR = .078, whereas the reduced model χ^2 was 820.02, $df = 272$, and SRMR = .057. Similarly, in Sample 2, the reduced model indicated improvement in model fit. In general, the reduced model fit these data better from a descriptive view, but still, five of the six indicators of model fit suggested moderate-fitting models. For instance, a χ^2 difference test using S_{PW} for the 25-item scale was conducted to compare the three-factor model, $\chi^2(272) = 820.02$, with the one-factor model, $\chi^2(275) = 967.19$ (i.e., Model 6 vs. Model 5, Table 2) in Sample 1. Tests indicated that the three-factor model fit significantly better than the one-factor model in both Sample 1, $\chi^2(3) = 147.17$, $p < .001$, and Sample 2, $\chi^2(3) = 81.45$, $p < .001$.

The standardized factor pattern and structure coefficients, item error variances, and item reliabilities for the three-factor model using the pooled within-group covariance matrix are presented in Table 3. The t values for all standardized factor pattern coefficients were statistically significant ($p < .01$), whereas Sample 2 standardized factor pattern coefficients tended to be slightly smaller in magnitude than Sample 1 coefficients. Factor intercorrelations using S_{PW} for ID and I/S, I/S and RL, and ID and RL, respectively, were .87, .92, and .86 in Sample 1 and .86, .88, and .89 in Sample 2. These indicated strong positive relationships among the factors without the between-class variability being imposed on these relationships. In general, the correlations among factors and overall model fit indices for both samples were descriptively similar in magnitude.

Table 3
 Standardized Parameter Estimates From Pooled Covariance Matrix for the 25-Item, Three-Factor Students' Evaluation of Teaching Effectiveness Rating Scale Model in Samples 1 and 2

Item No.	ID		I/S		RL		Item Error Variance	Item R^2
	Pattern	r_{st}	Pattern	r_{st}	Pattern	r_{st}		
1	.61 (.55)	.61 (.55)	—	.53 (.47)	—	.53 (.49)	.39 (.24)	.38 (.30)
3	.73 (.66)	.73 (.66)	—	.64 (.57)	—	.63 (.59)	.30 (.27)	.54 (.43)
4	.70 (.62)	.70 (.62)	—	.61 (.53)	—	.60 (.55)	.38 (.27)	.48 (.38)
5	.68 (.50)	.68 (.50)	—	.59 (.43)	—	.58 (.45)	.52 (.55)	.46 (.25)
8	.67 (.60)	.67 (.60)	—	.58 (.52)	—	.58 (.53)	.38 (.33)	.44 (.35)
9	.75 (.61)	.75 (.61)	—	.65 (.52)	—	.65 (.54)	.39 (.31)	.56 (.37)
10	.69 (.63)	.69 (.63)	—	.60 (.54)	—	.59 (.56)	.41 (.38)	.47 (.40)
12	.72 (.60)	.72 (.60)	—	.63 (.52)	—	.62 (.53)	.42 (.49)	.52 (.35)
13	.76 (.75)	.76 (.75)	—	.66 (.65)	—	.65 (.67)	.27 (.20)	.58 (.56)
20	—	.60 (.48)	.69 (.56)	.69 (.56)	—	.63 (.49)	.36 (.40)	.47 (.31)
23	—	.67 (.52)	.77 (.60)	.77 (.60)	—	.71 (.53)	.30 (.38)	.59 (.36)
24	—	.62 (.52)	.71 (.61)	.71 (.61)	—	.65 (.54)	.39 (.31)	.50 (.37)
25	—	.64 (.64)	.73 (.74)	.73 (.74)	—	.67 (.65)	.30(.20)	.54 (.50)
26	—	.49 (.46)	.56 (.53)	.56 (.53)	—	.52 (.47)	.52 (.54)	.31 (.28)
30	—	.67 (.61)	.77 (.71)	.77 (.71)	—	.71 (.62)	.28 (.25)	.60 (.48)
2	—	.48 (.46)	—	.51 (.46)	.56 (.52)	.56 (.52)	.40 (.21)	.32 (.27)
6	—	.49 (.42)	—	.52 (.41)	.57 (.47)	.57 (.47)	.58 (.47)	.33 (.22)
15	—	.44 (.39)	—	.47 (.39)	.51 (.44)	.51 (.44)	.70 (.67)	.26 (.20)
16	—	.55 (.54)	—	.59 (.54)	.64 (.61)	.64 (.61)	.50 (.47)	.41 (.37)
17	—	.59 (.51)	—	.63 (.50)	.69 (.57)	.69 (.57)	.45 (.56)	.48 (.32)
27	—	.58 (.57)	—	.63 (.56)	.68 (.64)	.68 (.64)	.40 (.46)	.46 (.39)
28	—	.57 (.49)	—	.61 (.48)	.66 (.55)	.66 (.55)	.34 (.38)	.43 (.30)
31	—	.60 (.58)	—	.64 (.57)	.70 (.65)	.70 (.65)	.44 (.38)	.49 (.43)
32	—	.57 (.51)	—	.61 (.50)	.66 (.57)	.66 (.57)	.34 (.34)	.44 (.32)
33	—	.60 (.51)	—	.64 (.50)	.70 (.57)	.70 (.57)	.43 (.61)	.49 (.32)

Note. The first values are from Sample 1, and the values in parentheses are from Sample 2. ID = Instructor's Delivery of Course Information; I/S = Teacher's Role in Facilitating Instructor/Student Interactions; RL = Instructor's Role in Regulating Students' Learning; r_{st} = structure coefficient. Dashes represent pattern coefficients constrained and not estimated in the model.

MFA for the SETERS

The three initial analysis steps suggested an MFA model with three factors for both within and between. Because multilevel analysis usually needs at least 50 to 100 groups to examine the between-class variation and draw full conclusions (Muthén, 1994), the initial MFA would not converge for either sample. As a result, Samples 1 and 2 were combined to explore the multilevel structure of the SETERS in the combined sample. Prior to combining, the inspection of both samples' covariance matrices indicated that they both had similar item variances and covariances. Therefore, these samples were con-

Table 4
Chi-Square and Fit Indices for the 21-Item, Three-Factor Students' Evaluation of Teaching Effectiveness Rating Scale Model in the Combined Sample

Method	Model Tests							
	χ^2	<i>df</i>	Probability	RMSEA	GFI	CFI	NNFI	SRMR
S_T	1,050.44	186	.00	.083	.87	.91	.89	.048
S_{PW}	802.46	186	.00	.072	.90	.90	.89	.049
Multilevel factor analysis								
MLM	1,517.46	376	.00	.062		.877		
Multilevel factor analysis								
Between								.077
Within								.050

Note. RMSEA = root mean square error of approximation; GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = nonnormed fit index; SRMR = standardized root mean residual; MLM = Muthén and Muthén's (1998) quasi-maximum likelihood fitting function, a quasi-maximum likelihood χ^2 value estimated by Mplus for unbalanced clusters.

sidered homogeneous and were treated as such in this exploratory MFA. Initial MFA on the 25-item three-factor model using the combined sample would not converge. Items 28, 31, 32, and 33 (instructor's regulation of student learning) had ICC values below 5%, indicating that they were not adequate for multilevel analysis, and were subsequently dropped from further analysis. All MFA results are based on the remaining 21 items. Although these 21 items could still be considered to be representative of the first two factors of the SETERS, there was potentially some construct underrepresentation for the factor measuring instructor's regulation of student learning. The average class size was about 15, and there was an unusually large range of class sizes (1 to 66) in this analysis, indicating that these data are far from balanced.

Inspection of the 21 items' univariate skew and kurtosis indices showed that all items had an absolute skew index less than 2.0, but only 19 items had an absolute kurtosis index less than 2. The normalized version of Mardia's index of multivariate kurtosis was 32.23. To protect against nonmultivariate normality, the MLM quasi- χ^2 statistic and robust standard errors were used for multilevel analysis. Even though the MLM statistic and standard errors are expected to give protection against nonnormality, this issue has not been exhaustively studied (Muthén, 2003).

The MLM quasi- χ^2 test of model fit was 1,517.46 for the 21-item three-factor model, with 376 degrees of freedom ($n = 792$), while RMSEA = .062 and the SRMR values for between and within were .077 and .050, respectively (Table 4). Therefore, compared with the χ^2 value calculated using S_{PW} , the addition of the between-class structure increased the quasi- χ^2 value by

Table 5
Item Characteristics for the 21-Item, Three-Factor Students' Evaluation of Teaching Effectiveness Rating Scale Model in the Combined Sample

Factor	Item No.	Item Characteristics					
		ICC	S_T	SPW	Reliability		
					Within	MFA Between	
ID	1	.07	.48	.35	.13	1.00	
	3	.23	.58	.49	.52	.89	
	4	.20	.50	.44	.42	.81	
	5	.34	.48	.36	.36	.77	
	8	.18	.50	.40	.41	.95	
	9	.34	.63	.48	.47	.93	
	10	.28	.57	.43	.43	.96	
	12	.29	.56	.44	.45	.92	
	13	.20	.64	.57	.58	.97	
	I/S	20	.21	.45	.40	.41	.69
		23	.11	.54	.49	.49	.95
		24	.19	.47	.44	.44	.59
		25	.13	.58	.55	.55	.87
26		.11	.28	.28	.29	.39	
30		.16	.61	.53	.35	.99	
RL	2	.18	.38	.33	.31	.63	
	6	.14	.29	.30	.32	.28	
	15	.24	.30	.25	.25	.54	
	16	.14	.52	.42	.41	1.00	
	17	.17	.47	.40	.41	.82	
	27	.15	.49	.41	.41	.95	

Note. MFA = multilevel factor analysis; ICC = intraclass correlation; ID = Instructor's Delivery of Course Information; I/S = Teacher's Role in Facilitating Instructor/Student Interactions; RL = Instructor's Role in Regulating Students' Learning.

about 715, with an additional 190 degrees of freedom. It is also interesting to note the close agreement in the estimated within reliabilities for MFA and S_{PW} , except for Item 1 (Table 5). These results seem to indicate that we could accept the model as a plausible representation of the sample data.

Because the fit of the model was determined to be adequate, we inspected the size of the parameter estimates. On the individual (within) level, the standardized pattern coefficients for the ID factor ranged from .36 to .76, for the I/S factor pattern from .54 to .74, and for the RL factor from .50 to .64 (Table 6). All pattern coefficients were considered to be substantial and were statistically significant (i.e., tested as the ratio of the unstandardized estimate to its standard error) at the .01 level (Heck, 2001). The correlations among factors within classes were .85, .89, and .87, for ID and I/S, ID and RL, and I/S and RL, respectively.

Table 6
Standardized Parameter Estimates for the 21-Item, Three-Factor Students' Evaluation of Teaching Effectiveness Rating Scale Model in the Combined Sample

Item No.	MFA															
	S_T				S_{PW}				Within				Between			
	Pattern		r_s		Pattern		r_s		Pattern		r_s		Pattern		r_s	
	ID	RL	ID	RL	ID	RL	ID	RL	ID	RL	ID	RL	ID	RL	ID	RL
1	.69	—	—	.61	.62	.59	—	—	.59	.51	.53	.36	—	—	1.0	.32
3	.76	—	—	.68	.68	.70	—	—	.70	.61	.62	.72	—	—	.94	.64
4	.71	—	—	.63	.64	.66	—	—	.66	.57	.59	.65	—	—	.90	.59
5	.69	—	—	.62	.62	.60	—	—	.60	.52	.53	.60	—	—	.88	.53
8	.71	—	—	.63	.64	.63	—	—	.63	.55	.56	.64	—	—	.98	.57
9	.79	—	—	.79	.70	.71	.69	—	.69	.60	.61	.68	—	—	.97	.61
10	.75	—	—	.67	.68	.66	—	—	.66	.57	.59	.66	—	—	.98	.59
12	.75	—	—	.67	.68	.66	—	—	.66	.57	.59	.67	—	—	.96	.60
13	.80	—	—	.71	.72	.75	—	—	.75	.65	.67	.76	—	—	.98	.68
20	.67	—	—	.67	.60	—	.63	—	.55	.63	.54	.64	—	—	.83	.56
23	.73	—	—	.73	.66	—	.70	—	.61	.70	.60	.70	—	—	.97	.61
24	.68	—	—	.68	.61	—	.67	—	.58	.67	.57	.67	—	—	.77	.58
25	.76	—	—	.76	.68	—	.74	—	.64	.74	.63	.74	—	—	.93	.64
26	.53	—	—	.47	.53	.48	—	.53	.46	.53	.45	.54	—	—	.63	.47
30	.78	—	—	.69	.78	.70	—	.73	.64	.73	.62	.59	—	—	1.0	.51
2	.62	—	—	.56	.56	.62	—	—	.57	.51	.50	.57	—	—	.79	.48
6	.54	—	—	.49	.49	.54	—	—	.55	.49	.48	.55	—	—	.53	.49
15	.55	—	—	.50	.50	.55	—	—	.50	.45	.44	.50	—	—	.74	.44
16	.72	—	—	.65	.65	.72	—	—	.65	.58	.57	.64	—	—	1.0	.56
17	.69	—	—	.62	.62	.69	—	—	.63	.56	.55	.63	—	—	.91	.64
27	.70	—	—	.63	.63	.70	—	—	.64	.57	.56	.64	—	—	.98	.64

Note. MFA = multilevel factor analysis; r_s = structure coefficient; ID = Instructor's Delivery of Course Information; I/S = Teacher's Role in Facilitating Instructor/Student Interactions; RL = Instructor's Role in Regulating Students' Learning. Dashes represent pattern coefficients constrained and not estimated in the model.

At the class level, the standardized pattern coefficients for ID ranged from .88 to 1.00, for I/S from .63 to 1.00, and for RL from .53 to 1.00 (Table 6). All coefficients were considered to be substantial and were statistically significant at the .01 level. There were substantial correlations among the three factors: .95 for ID and I/S, .94 for ID and RL, and .95 for I/S and RL. These correlations and the within factor correlations suggested that perhaps one teaching effectiveness factor would be enough to capture both within-class and between-class variation. The results of the one-factor model at both levels gave an MLM quasi- χ^2 test of model fit of 1,761.05 ($df = 380$), RMSEA = .068, and SRMR values of .069 and .052 for between and within, respectively. A comparison of overall model fit indices for the three-factor model at both levels and the one-factor model at both levels descriptively suggested that the models fit about the same.

The amount of variance in the three factors that existed among the groups was also determined. The ICCs for these three factors correspond to what Muthén (1991, 1994) called true differences in these factors across the teachers, after adjusting for other sources of variability such as clustering effects and measurement error associated with each factor's indicators (e.g., differences in how students evaluated their own teachers). The actual differences across the sample of classes were about 40% for the ID factor, 20% for I/S, and 26% for RL. For instance, the ICC for ID represents the adjustment for the individual-level unreliability associated with the 10 items used to measure this factor (e.g., ICCs that ranged from .07 to .34; Table 5). Accordingly, the size of the ICCs suggested that the difference in teaching effectiveness, as measured by the three factors, compared across the set of classes was relatively moderate.

Analyses of the SEEQ Questionnaire

Reliability estimates (coefficient α) for Sample 1 ranged from .76 to .90, whereas Sample 2 reliability estimates ranged from .64 to .92. The interrater reliability estimates for the SEEQ factors ranged from .72 to .92 and .65 to .95. Table 6 shows the results of the nine-factor SEEQ χ^2 test and other measures of model fit for Samples 1 and 2. The analysis of S_T showed that the nine-factor model exhibited a poor fit for five of the six fit indices for both samples.

The estimation of the proportion of between-group variation (ICC) for the nine factors ranged from .11 to .52 for Sample 1 and from .11 to .58 for Sample 2. In general, the ICCs are similar across both samples with respect to each factor.

Because of the relatively high levels of between variation, it was reasonable to estimate the within structure. For each sample, the conventional ML analysis gave a worse fit for the S_T than for S_{pw} for the nine-factor model

(Table 7). This difference in fit was expected given the moderate-size ICCs and the moderate average class size of about 15. Interestingly, the S_{PW} analysis showed that Sample 2 fit the within part of the model better than Sample 1. Even though the ICCs suggested that MFA was warranted, MFA was not conducted on the SEEQ, because there was an insufficient number of classes to adequately estimate the between portion of the model.

Validity Evidence

Table 8 shows the Pearson correlations (r) for scores on the three SETERS factors with scores on the nine SEEQ factors. Because it was expected that scores on all three of the factors of the SETERS would be positively related to scores on the nine factors of Marsh's (1983) SEEQ questionnaire, one-tailed tests were performed. Using individual-level student responses, all of the following correlations were tested at the 1% significance level; the critical value was $r \geq .23$. Correlations for scores on the three SETERS factors with the nine SEEQ factors ranged from .38 to .76 in Sample 1 and from .13 to .73 in Sample 2. As hypothesized, scores on the SEEQ factor Examination/Grading had the strongest positive relationship with the RL factor on the SETERS, whereas scores on all three SETERS factors had the weakest relationship with the SEEQ factor Workload/Difficulty.

Discussion

A great deal of research exists in support of SET instruments as measures of teaching effectiveness. However, prior research has not looked at the multilevel structure of SET instruments. This study was concerned with exploring the within and between variation of a three-factor model of teaching effectiveness using MFA.

The 21-item SETERS was developed for use in MFA. Initially, the SETERS consisted of 25 items, with the RL factor measured by 10 items. However, 4 of the items (Items 28, 31, 32, and 33) did not fit the criteria for MFA and were subsequently dropped from this analysis. Because these 4 items were commonly represented on similar SET instruments, we did not appropriately represent the construct of teaching effectiveness. Therefore, it is suggested that inferences drawn from the MFA be treated as preliminary validity evidence for the SETERS scores. Subsequently, the three-factor model of d'Apollonia and Abrami (1997) was tested, and the results at both the within- and between-class levels appeared to indicate a reasonable fit to the combined samples (Table 4). It appeared that the three SETERS factors could be adequately represented at the within- and between-class levels.

The first hypothesis predicted that a three-factor solution, as proposed by d'Apollonia and Abrami (1997), would perform better than a global measure

Table 7
Chi-Square and Fit Indices for the Students' Evaluation of Educational Quality Questionnaire in Samples 1 and 2

Method	Model	χ^2	df	Probability	RMSEA	GFI	CFI	NNFI	SRMR
Sample 1									
S_T	1 (35 items, nine factors)	2,154.98	524	.00	.11	.69	.83	.81	.073
S_{PW}	2 (35 items, nine factors)	1,365.22	524	.00	.074	.79	.87	.85	.056
Sample 2									
S_T	3 (35 items, nine factors)	1,344.22	524	.00	.065	.83	.90	.89	.041
S_{PW}	4 (35 items, nine factors)	1,041.28	524	.00	.053	.86	.91	.90	.049

Note: RMSEA = root mean square error of approximation; GFI = goodness-of-fit index; CFI = comparative fit index; NNFI = nonnormed fit index; SRMR = standardized root mean residual.

Table 8

Pearson Correlations for the 25-Item Students' Evaluation of Teaching Effectiveness Rating Scale (SETERS) With the Students' Evaluation of Educational Quality (SEEQ) Questionnaire in Samples 1 and 2

SEEQ Subscale	SETERS Subscale		
	ID	I/S	RL
Learning/Value	.75 (.73)	.69 (.56)	.74 (.59)
Instructor Enthusiasm	.73 (.73)	.68 (.61)	.70 (.57)
Organization/Clarity	.76 (.68)	.62 (.59)	.66 (.63)
Group Interaction	.68 (.64)	.70 (.69)	.66 (.59)
Individual Rapport	.62 (.50)	.71 (.64)	.65 (.60)
Breadth of Coverage	.60 (.65)	.67 (.58)	.56 (.60)
Examination/Grading	.61 (.51)	.63 (.51)	.70 (.65)
Assignments/Readings	.70 (.67)	.63 (.54)	.68 (.55)
Workload/Difficulty	.45 (.13)	.38 (.14)	.45 (.19)

Note. Correlations were calculated using pairwise deletion. The first values are from Sample 1, and the values in parentheses are from Sample 2. ID = Instructor's Delivery of Course Information; I/S = Teacher's Role in Facilitating Instructor/Student Interactions; RL = Instructor's Role in Regulating Students' Learning. Sample 1 *n* ranged from 320 to 382; Sample 2 *n* ranged from 394 to 439.

of teaching effectiveness using the SETERS. On the basis of research and theory, teaching effectiveness as measured by students' ratings of teaching is multidimensional in nature. The results from both samples showed that the three-factor model consistently had marginally higher values for GFI, CFI, and NNFI and lower values for RMSEA and SRMR than did the one-factor model for the SETERS (Table 2). The consistencies in these values existed regardless of whether one was using the total covariance matrix or the pooled within-covariance matrix. However, MFA results showed that a one-factor model at both levels performed about as well as a three-factor model at both levels. These results contradict previous work that has found that SET measures are multidimensional. It is suggested that a factor score divided by the number of items on each factor could be used to help individual teachers ascertain problematic areas of their teaching effectiveness.

The results also supported our second hypothesis by showing that there was a strong positive relationship between scores on the SETERS factors and each SEEQ factor (Table 7). The results also supported our third hypothesis by showing that scores on the SEEQ factor Examinations/Grading was most related with scores on the SETERS factor RL. Additionally, scores on the SEEQ factor Workload/Difficulty had the weakest relationship with the three SETERS factor scores and supported our final hypothesis. These results support the study prediction and supplied convergent validity evidence for scores on the SETERS.

In general, using current cutoff criteria, all models indicated a lack of good fit or moderate fit with these data. This might be attributed to several

aspects of the study. For example, if data had followed a multivariate normal distribution, the pooled within-group and scaled between-group covariances could have been viewed as observed values with an observed sampling distribution. Typically, this distribution is used by structural equation modeling programs to estimate the χ^2 model test and standard errors of the parameter estimates. However, it is unknown how well the sampling distribution of the multivariate multilevel covariance estimates follows the sampling distribution of the observed covariances (R. Heck, personal communication, March 21, 2002; Hox, 2003). As a result, our interpretation proceeded with caution because we did not know how well the MLM fitting function and standard errors performed under nonnormal multivariate distributions with unbalanced group sizes.

One could plausibly suggest that the moderate levels of model fit to these data are a result of the models being incorrect representations of teaching effectiveness as measured by the SETERS. Other models that could potentially be tested with the SETERS are Burdsal and Bardo's (1986) six-factor model or the four-factor model initially proposed by Abrami et al. (1996). In addition, the models tested in this study could be used with additional samples using more sophisticated sampling methods and possibly much larger sample sizes at the between level.

In addition to questions regarding the lack of fit, it should be noted that these results cannot be generalized beyond these two samples, because convenience sampling was used to collect these data. Also, comparison of parameter estimates across samples (Sample 1 vs. Sample 2) may be problematic because of scale issues. An additional limitation of this study is the relatively small number of classes. Recall that only 54 classes were used at the between-group level of the MFA, and this is a relatively small sample size for MFA. Potentially, MFA results may have provided more support for the three-factor or one-factor model at both levels if the number of classes had been larger.

Appendix

Students' Evaluation of Teaching Effectiveness Rating Scale

For the following statements we would like you to respond by indicating your level of agreement with each statement about this course instructor, using the number codes provided below. Please omit or leave blank any of the items that do not pertain to the instructor of this course.

Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly Agree
1	2	3	4	5

Instructor's Delivery of Course Information

- 1. The instructor presented the course material with enthusiasm.
- 3. The instructor's explanations of course content were clear.
- 4. The instructor clarified course material by reviewing concepts.
- 5. The instructor used a variety of approaches/strategies when presenting material.
- 7. The instructor presented the course material at an appropriately paced sequence.
- 8. The instructor explained the situations under which course content could be applied.
- 9. The instructor was an effective instructor compared with other college or university instructors I have had.
- 10. The instructor related course material to my present needs.
- 11. The instructor related course material to my future needs.
- 12. The instructor increased my interest in the course subject matter.
- 13. The instructor helped me understand the course content.
- 14. I was motivated to learn the course material.

Instructor/Student Interactions

- 18. The instructor encouraged me to ask questions during class.
- 19. The instructor encouraged me to participate in class discussions.
- 20. The instructor encouraged class discussion.
- 21. The instructor encouraged me to express my opinions about course material.
- 22. The instructor encouraged me to share my knowledge about course content.
- 23. The instructor respected my opinions about course content.
- 24. The instructor encouraged me to interact with other students in class.
- 25. The instructor was informative when responding to students questions in class.
- 26. The instructor could be contacted outside of class time.
- 30. The instructor promoted a comfortable learning atmosphere.

Regulating Student Learning

- 2. The instructor appeared knowledgeable in the course content area.
- 6. The instructor clearly outlined the direction of the course through a syllabus.
- 15. The instructor provided me with hands on activities with the course subject matter.
- 16. The instructor understood my pace/rate for learning course content.
- 17. The instructor acknowledged my individual learning achievements/accomplishments.
- 27. The instructor helped me with my individual learning needs.
- 28. The instructor graded my assignments/examinations according to instructor prescribed standards for grading.

- ___29. The instructor provided me with feedback on my learning progress.
- ___31. The instructor's feedback on my learning progress was valuable.
- ___32. The assignments/examinations covered course content as emphasized by the course instructor.
- ___33. The instructor provided feedback on my assignments/examinations that helped me learn from my mistakes.
- ___34. The instructor's feedback on my assignments/examinations let me know what I knew and did not know.

Note. Items in boldface type constitute the final instrument.

References

- Abrami, P. C., d'Apollonia, S., & Rosenfield, S. (1996). The dimensionality of student ratings of instruction: What we know and what we do not. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 11, pp. 213-264). New York: Agathon.
- Burdsal, C. A., & Bardo, J. W. (1986). Measuring students' perceptions of teaching: Dimensions of evaluation. *Educational and Psychological Measurement*, 56, 63-79.
- d'Apollonia, S., & Abrami, P. C. (1997). Navigating student ratings of instruction. *American Psychologist*, 52, 1198-1208.
- Feldman, K. A. (1976). The superior college teacher from the student's view. *Research in Higher Education*, 5, 243-288.
- Fincher, C. (1985). Learning theory and research. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 1, pp. 63-96). New York: Agathon.
- Heck, R. (2001). Multilevel modeling with SEM. In G. Marcoulides & R. Schumacher (Eds.), *New developments and techniques in structural equation modeling* (pp. 89-127). Mahwah, NJ: Lawrence Erlbaum.
- Hox, J. J. (2003). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Lawrence Erlbaum.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indices in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, 6, 1-55.
- Jackson, D. L., Teal, C. R., Raines, S. J., Nansel, T. R., Force, R. C., & Burdsal, C. A. (1999). The dimensions of students' perceptions of teaching effectiveness. *Educational and Psychological Measurement*, 59, 580-596.
- Kaplan, D., & Elliott, P. R. (1997). A didactic example of multilevel structural equation modeling applicable to the study of organizations. *Structural Equation Modeling*, 4, 1-23.
- Kindsvatter, R., Wilen, W., & Ishler, M. (1988). *Dynamics of effective teaching*. White Plains, NY: Longman.
- Kline, R. B. (1998). *Principles and practice of structural equation modeling*. New York: Guilford.
- Marsh, H. W. (1977). The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal*, 14, 441-447.
- Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. *Journal of Educational Psychology*, 75, 150-166.
- Marsh, H. W. (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., & Dunkin, M. J. (1992). Students' evaluations of university teaching: A multidimensional perspective. In J. C. Smart (Ed.), *Higher education: Handbook of theory and research* (Vol. 8, pp. 143-233). New York: Agathon.
- Marsh, H. W., Fleiner, H., & Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology, 67*, 833-839.
- Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness. *Journal of Higher Education, 73*, 603-641.
- Marsh, H. W., & Hocevar, D. (1984). The factorial invariance of student evaluations of college teaching. *American Educational Research Journal, 21*, 341-366.
- Marsh, H. W., & Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. *Teaching and Teacher Education, 7*, 9-18.
- Mueller, R. O. (1997). Structural equation modeling: Back to basics. *Structural Equation Modeling, 4*, 353-369.
- Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.
- Muthén, B. O. (1994). Multilevel covariance structure analysis. *Sociological Methods & Research, 22*, 376-398.
- Muthén, B. O. (2003, January 22). Mplus discussion: Multilevel data/complex sample: Multilevel factor analysis. Message posted to <http://www.statmodel.com/discussion/messages/12/101.html?1043216963>
- Muthén, L. K., & Muthén, B. O. (1998). Mplus user's guide. Los Angeles: Authors.
- Muthén, L. K., & Muthén, B. O. (2001). Mplus for Windows (Version 2.01) [Computer software]. Los Angeles: Authors.
- Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variable analysis: Applications for developmental research* (pp. 399-419). Thousand Oaks, CA: Sage.
- Ting, K. F. (2000). A multilevel perspective on student ratings of instruction: Lessons from the Chinese experience. *Research in Higher Education, 41*, 637-661.
- Winer, B. J. (1971). *Statistical principles in experimental design*. New York: McGraw-Hill.
- Young, S., & Shaw, D. G. (1999). Profiles of effective college and university teachers. *The Journal of Higher Education, 70*, 670-686.