Teaching & Teacher Education, Vol. 7, No. 4, pp. 303–314, 1991
Printed in Great Britain

0742–051X/91 $3.00+0.00
Pergamon Press plc

STUDENTS' EVALUATIONS OF TEACHING EFFECTIVENESS: THE STABILITY OF MEAN RATINGS OF THE SAME TEACHERS OVER A 13-YEAR PERIOD

HERBERT W. MARSH

University of Western Sydney, Macarthur, Australia

and

DENNIS HOCEVAR

University of Southern California, U.S.A.

Abstract—Students' evaluations of teaching effectiveness (SETEs) are weakly related—negatively—to teaching experience and age according to Feldman's (1983) comprehensive review of cross-sectional studies. Cross-sectional studies, however, provide a weak basis for inferring the future ratings of less experienced teachers or the past ratings of more experienced teachers. Considered here are ratings of 6024 classes taught by a diverse cohort of 195 teachers representing 31 academic departments who were evaluated continuously over a 13-year period using the same multidimensional Students' Evaluations of Educational Quality instrument. For both undergraduate and graduate level courses, there were almost no changes over time for any of the nine content-specific dimensions, the overall course rating, or the overall instructor rating. The findings were consistent for teachers who had little, moderate, or substantial amounts of teaching experience at the start of the study. These results are important because this is apparently the only study to examine the stability of faculty ratings using a longitudinal design with a large and diverse group of teachers over such a long period of time.

Students' evaluations of teaching effectiveness (SETEs) are widely collected and used for a variety of purposes such as personnel decisions, feedback to faculty on the effectiveness of their teaching, input into students' course selection, and research on teaching. An enormous amount of research has demonstrated that SETEs are multidimensional with a well-defined factor structure, internally consistent, reasonably valid when compared to a variety of other indicators of effective teaching, and relatively unaffected by potential biases to the ratings (see Marsh, 1987, for an overview of this research). Nevertheless, most of this research has considered ratings collected in one specific course on a single occasion and there is surprisingly little research on the stability of mean ratings received by the same instructor over an extended period of time. The purpose of the present investigation is to examine changes in ratings of a large number of teachers who have been evaluated continuously over a 13-year period with the same multidimensional Students' Evaluations of Educational Quality (SEEQ) instrument.

The Stability of Students' Evaluations of Teaching

There are many approaches to the study of stability and change (Goldstein, 1979; Plewis, 1985; Rogosa, 1979; Rogosa, Floden, & Willett, 1984; Willett, 1988). The two most common, however, refer to the stability of means over time (mean stability) and to the

stability of individual differences over time (covariance stability). The present investigation emphasizes the mean stability over time, but it is useful to review both approaches. In each case it is desirable to have longitudinal data in which the same individuals are evaluated on many different occasions.

A number of researchers (e.g., Bausell & Bausell, 1979; Kulik & Kulik, 1974; Marsh, 1981; Gilmore, Kane, & Naccarato, 1978) have examined correlations between ratings of the same instructor in different offerings of the same course, the same teacher in different courses, and different teachers teaching the same course in an attempt to disentangle the relative influence of the course and the teacher. Using a path analytic approach, Marsh (1981) found that students' evaluations were primarily a function of the teacher rather than the course. For overall ratings of the instructor and of the course, the correlations between ratings of different instructors teaching the same course (one estimate of the course effect) were −.05 and −.01, respectively, whereas correlations between ratings for the same instructor in two different classes (.61 and .59) and in two different offerings of the same course (.72 and .71) were much larger. Based on these findings, Marsh (1987) concluded that SETEs are primarily a function of the teacher who teaches a course rather than the course that is being taught. These studies of covariance stability support the practice of aggregating ratings across different courses and suggest that individual differences in teaching effectiveness are stable, but do not address the issue of mean stability.

Overall and Marsh (1980) examined stability in a longitudinal study in which the same students evaluated teachers at the end of the course and retrospectively several years after finishing the course and at least one year after graduation from the program. They showed that mean ratings were nearly the same at both points in time and that class-average responses for the end-of-term responses correlated .83 with the retrospective ratings. The study demonstrates that the perspectives of the same students do not change over time and counters the claim that students would evaluate instructors differently after being called upon to apply course materials in further coursework or after graduation. These results, however, address the stability of responses by the same students about teaching effectiveness from a single occasion and not the stability of the teaching effectiveness over multiple occasions.

Most studies of the mean stability of teaching effectiveness are based on cross-sectional instead of longitudinal designs, and much of this research has been at the primary and secondary school level. In this research, teaching experience or age is related to various indicators of effective teaching—though typically not SETEs. In an early review of this research, Ryans (1960) reported an overall negative relation between teaching experience and teaching effectiveness. He suggested, however, that there was an initial increase in effectiveness during the first 5 years, a leveling out period, and then a period of gradual decline. Barnes (1985) reviewed research since the early 1960s and also found that after the first few years teaching experience was negatively related to measures of student achievement and teaching effectiveness. She further reported that teaching experience beyond the first few years was associated with a tendency for teachers to reject innovations and changes in educational policy. At the university level, Feldman's (1983) comprehensive review of studies examining relations between seniority and SETEs provides an important basis for the present investigation.

Feldman's 1983 Review

Feldman (1983) conducted the most extensive review of studies relating overall and content-specific dimensions of SETEs to teacher age, teaching experience, and academic rank. In earlier research, Feldman (1976) devised a set of categories reflecting content-specific components of teaching effectiveness from the students' perspective (see Table 1) and used these categories to facilitate his 1983 review. In Table 1 are presented Feldman's categories, the SEEQ factor mostly nearly related to each category, and a summary of the relation of ratings in each category to his seniority measures.

Feldman (1983) reported that SETEs were only weakly related to the three measures of seniority (age, years experience, and academic rank). He also argued, however, that distinct

patterns were evident (see Table 1). Overall evaluations tended to be negatively correlated with age and—to a lesser extent—years of teaching experience, but tended to be positively correlated with academic rank. Thus, younger teachers, teachers with less teaching experience, and teachers with higher academic ranks tended to receive somewhat higher evaluations. Age and teaching experience showed reasonably similar patterns of correlations with overall and content-specific dimensions (Feldman, 1983). Academic rank, however, had a more varied pattern of relations with the content-specific dimensions. Academic rank tended to be positively correlated with some characteristics such as subject knowledge, intellectual expansiveness, and value of course materials, but negatively correlated with other characteristics such as class discussion, respect for students, helpfulness and availability to students (see Table 1).

Feldman (1983) suggested that the strength of relations of ratings with teaching experience and, perhaps, age might be underestimated in studies that considered only linear relations. In a few studies that specifically examined nonlinear relations, there was some suggestion of an inverted U-shaped relation in which ratings improved initially, peaked at some early point, and then declined slowly thereafter. Although too few studies examined this nonlinearity to say precisely the point at which

Table 1

Categories of Effective Teaching and Their Relation[a] to Teacher Age, Experience, and Rank (from Feldman, 1983, Tables 1–4) and the Corresponding SEEQ Factor Most Nearly Matching Each of Feldman's Categories

	Seniority measures									
	Age			Experience			Rank			
Feldman's categories	Pos	0	Neg	Pos	0	Neg	Pos	0	Neg	SEEQ factor most similar to category
1. Stimulation of interest	0.0	1.0	2.0	0.0	2.7	3.5	1.0	8.2	2.5	Instructor enthusiasm
2. Enthusiasm	0.0	0.5	2.0	1.0	2.0	1.5	1.0	3.5	1.5	Instructor enthusiasm
3. Subject knowledge	0.0	1.0	1.0	1.0	2.0	0.0	5.0	0.5	0.0	Breadth of coverage[b]
4. Intellectual expansiveness	0.0	0.0	1.0	0.0	3.0	0.0	4.0	0.5	0.0	Breadth of coverage
5. Preparation and organization	0.0	3.5	2.0	0.5	5.8	1.5	1.5	8.2	0.8	Organization/Clarity
6. Clarity and understandableness	0.0	0.5	2.0	0.0	3.2	3.0	1.5	7.5	3.3	Organization/Clarity
7. Elocutionary skills	0.0	0.0	1.0	0.0	1.0	1.0	0.0	3.0	1.0	None
8. Sensitivity to class progress	0.0	0.0	0.5	0.0	2.2	0.0	0.0	1.5	0.2	None
9. Clarity of objectives	0.0	0.0	0.0	0.5	2.0	1.5	0.5	3.3	1.3	Organization/Clarity
10. Value of course materials	0.0	2.0	1.0	0.0	4.3	2.5	3.0	5.8	1.5	Assignments/Readings
11. Supplementary materials	0.0	0.0	0.0	0.0	1.3	0.0	0.6	0.8	0.0	Assignments/Readings
12. Perceived outcome/impact	0.0	0.0	1.5	1.0	2.0	0.0	0.5	1.0	0.5	Learning/Value
13. Fairness, impartiality	0.0	0.0	1.0	0.0	4.8	0.0	0.0	5.3	4.8	Examinations/Grading
14. Classroom management	—	—	—	—	—	—	—	—	—	None
15. Feedback to students	0.0	0.0	0.0	0.0	4.0	0.0	0.0	3.0	1.5	Examinations/Grading
16. Class discussion	0.0	1.0	1.0	0.0	3.6	3.0	0.0	3.3	5.9	Group interaction
17. Intellectual challenge	0.0	1.0	1.5	0.0	2.5	1.0	0.5	2.8	2.1	Learning/Value
18. Respect for students	0.0	2.5	0.5	0.0	4.8	1.5	0.0	3.8	3.3	Individual rapport
19. Availability/helpfulness	0.0	0.5	1.0	0.0	3.6	0.5	1.0	2.8	3.6	Individual rapport
20. Difficulty/workload	—	—	—	—	—	—	—	—	—	Workload/Difficulty
Overall evaluation	0	6	6	2	8	5	10	21	1	Overall instructor & course

Note. The actual categories used by Feldman in different studies (e.g., Feldman, 1976, 1983, 1984, 1986, 1987, 1988, 1989a, 1989b) varied somewhat. Categories 14 and 20 were not included by Feldman (1983).

[a]The numbers are the number of studies in Feldman's 1983 review reporting significantly positive (Pos), nonsignificant (0), and significantly negative (Neg) relations. Values for the specific categories can be nonintegers because Feldman (1983) weighted results by taking into account the number of categories into which each association could be coded.

[b]Whereas this factor most closely matches the corresponding category, the match is apparently not particularly close.

ratings peaked, Feldman's review suggested that it occurred somewhere between 3 and 12 years of teaching experience.

Feldman (1983) noted that many different interpretations of the results existed. He suggested, for example, that seniority may covary with other variables that affect ratings, that academic rank may product different expectations in students, that teachers may change as they grow older, or that different cohorts of students may change in what they expect of a good teacher. Feldman also noted that studies in his review were based on cross-sectional rather than longitudinal data so that the associations might just reflect differences in the cohorts of teachers being compared. Surprisingly, however, Feldman gave much less weight to this possibility, emphasizing instead the possibility that teachers had actually changed over the years. The inconsistent patterns of relations involving age and experience on the one hand, and academic rank on the other, suggest that differential selectivity may play a role in the interpretation of these relations. It must be emphasized that cross-sectional studies provide a poor basis for inferring what ratings younger, less-experienced teachers will receive later in their careers or what ratings older, more-experienced teachers would have received if evaluated earlier in their careers. A better basis for these inferences are longitudinal data such as those considered here.

Methods

Sample and Procedures

During the period 1976 to 1988 nearly one million SEEQ instruments (Marsh, 1982, 1984, 1987) were completed in the evaluation of almost 50,000 classes at a large private U.S. university. Typically SEEQ instruments were distributed to faculty shortly before the end of each academic term, administered by a student in the class or by administrative staff according to standardized written instructions, and taken to a central office where they were processed. Although an academic unit's participation in this program was voluntary, the university required that all units systematically collect some form of students' evaluations and did not consider tenure/promotion recommendations that did not include such documentation. Thus, most academic units that used SEEQ required that all faculty be evaluated in all courses. A normative archive consisting of class-average responses for all classes evaluated during this 13-year period served as the basis for the present investigation.

The SEEQ instrument consists of 33 specific rating items, two overall rating items, and several additional background/demographic items. It is designed to measure nine evaluation factors that have been supported by more than 30 published factor analyses (e.g., Marsh, 1983, 1984, 1987; Marsh & Hocevar, 1991). As is typical in SEEQ research, the nine SEEQ scales are summarized with empirically derived factor scores (e.g., Marsh, 1983). For present purposes, a factor analysis (Marsh & Hocevar, 1991) was conducted that clearly identified all nine SEEQ factors, and factor scores based on this analysis were used in subsequent analyses. The results of this factor analysis are presented in Marsh and Hocevar (1991) and are very similar to early factor analyses presented by Marsh (1983, 1984, 1987).

For purposes of the present investigation all teachers who were evaluated at least once during each of 10 different years between 1976 and 1988 were selected. This process identified 195 different teachers who had been evaluated in a total of 6024 different courses, an average of 30.9 classes per teacher. These teachers came from a total of 31 different academic departments representing social sciences, business, safety and systems management, engineering and humanities.

Statistical Analysis

The statistical analyses consisted of a series of multiple regressions in which ratings by each instructor were related to linear and nonlinear components of the year the ratings were collected (1976−1988), the course level (2 = graduate, 1 = undergraduate), and their interactions. Separate analyses were conducted for each of the nine SEEQ factors, the overall instructor rating, and the overall course rating. In one set of analyses all 6024 courses were considered separately. In a second set of analyses,

all courses taught by the same instructor at the same level (graduate or undergraduate), offered in the same year were averaged. This aggregation resulted in a total sample of 3135 unique combinations of instructor, year, and course level. Because both sets of analyses resulted in nearly identical conclusions, only results of the second set of analyses are considered.

Results

The purpose of the present investigation was to determine changes in SETEs of the same instructors over a 13-year period. Hence, this was a study of the mean stability of the ratings of the same teachers over time. A multiple regression approach to ANOVA was used in which the linear and nonlinear effects of year, course level, and their interaction were evaluated. In the first set of analyses (Table 2) only the effects of course level were consistently significant. As previously demonstrated (Marsh, 1984, 1987), instructors are evaluated more favourably in graduate level courses than in undergraduate level courses. This effect of level, however, does not interact with either the linear or quadratic components of year. Whereas there was a very weak, linear decline in evaluations, the sizes of this effect were small—never explaining more than half of 1% of the variance in any of the evaluation scores—and failed to reach statistical significance for 7 of the 11 evaluation scores.

The most important influence in the students' evaluations of teaching is the instructor. In order to evaluate the influence of the instructor the mean rating of each instructor over all undergraduate classes and over all graduate classes was computed. For each of the student evaluation scores, ratings of the same instructor were averaged over time separately for undergraduate and graduate level courses. In a second set of regression models (Table 3), these instructor mean ratings were added to the predictor variables considered in the first set of analyses (Table 2). Hence, the effects of the individual instructor were controlled in evaluating the effects of the other variables (see Pedhazur, 1982, for a discussion of criterion coding). As has been found previously, the individual instructor accounts for most of the variance in

Table 2

Changes in Multiple Dimensions of Students' Evaluations Over Time for Ratings of the Same Instructors: The Effects of Year (1976–1988), Level (Undergraduate and Graduate), and Their Interaction (N = 3135)

Component	*r* with Year	Year	Standardized beta weights for: Year 2	Level	Yr. × Lev.	Yr. 2 × Lev.	Mult. R.
Factor scores							
Learning/Value	.000	.000	−.035	.168**	.015	.032	.173**
Enthusiasm	−.018	−.025	−.022	.118**	.011	.014	.132**
Organization	−.043	−.046	−.026	−.026	.029	−.006	.065
Group interact	−.044	−.049*	.023	.260**	−.021	.032	.289**
Indiv. rapport	−.066*	−.067**	.042	.179**	−.007	.006	.198**
Breadth	−.038	−.038	.035	.161**	.015	−.035	.148**
Exams	−.021	−.032	−.077**	.145**	−.019	.033	.151**
Assignments	−.022	−.028	.008	.174**	.011	.003	.178**
Workload	.016	.009	−.034	.068*	.013	−.002	.078*
Overall ratings							
Course	−.041	−.051*	−.016	.168**	.008	.026	.193**
Instructor	−.062*	−.071*	−.010	.144**	.012	.024	.174**

Note. Multiple regression was used to predict factor scores and overall ratings from year (linear and quadratic components), course level (1 = undergraduate, 2 = graduate), and the year × level interaction.
*$p < .05$; **$p < .01$.

Table 3

Changes in Multiple Dimensions of Students' Evaluations Over Time for Ratings of the Same Instructors: The Effects of Year (1976–1988), Level (Undergraduate and Graduate), and Their Interaction (N = 3135)

Component	*r* for Instr.	Standardized beta weights for: Instr.	Year	Year^2	Level	Yr. × Lev.	Yr.^{2} × Lev.	Mult. R.
Factor scores								
Learning/Value	.701**	.703**	.001	−.045**	−.023	.018	.025	.703**
Enthusiasm	.822**	.822**	−.016	−.019	−.003	.010	.006	.822**
Organization	.770**	.770**	−.048**	−.025	.000	.017	−.004	.772**
Group interact	.814**	.815**	−.012	−.020	−.009	−.013	.010	.815**
Indiv. rapport	.747**	.746**	−.026	.016	.006	.006	−.009	.748**
Breadth	.735**	.735**	.005	−.011	.000	.009	−.007	.736**
Exams	.678**	.678**	−.028	−.017	.006	−.008	−.014	.678**
Assignments	.704**	.704**	−.004	−.024	−.008	.012	.006	.704**
Workload	.797**	.797**	−.020	−.009	.010	−.009	.007	.797**
Overall ratings								
Course	.725**	.725**	−.031	−.028	−.013	−.031	.019	.726**
Instructor	.756**	.755**	−.048**	−.020	−.010	.009	.015	.758**

Note. The Instructor (Instr.) component was obtained by taking the mean of the instructor ratings for undergraduate classes and for graduate classes, and then including these means in the prediction of ratings. Because these means were computed separately for graduate and undergraduate level courses, it has the effect of eliminating variance due to course level.
*$p < .05$; **$p < .01$.

each of the different SEEQ scores. Because this approach controls for individual differences in teaching effectiveness and so much variance is explained by the instructor, this analysis provides a much more powerful test of changes in ratings of the same teacher over time. The results again show that there are almost no systematic changes in ratings over time. Year accounts for no more than one quarter of 1% in any of the evaluation scores, and only reaches statistical significance for 2 of 11 scores.

Controls for Changing Standards

The results summarized above indicate that there are almost no linear or quadratic trends in the evaluations of the same teachers over this 13-year period. We interpret this to mean that the teaching effectiveness of instructors in our longitudinal sample was stable over time from the perspective of students. However, because the students who evaluated the teachers each year differed, the ratings may reflect a combination of changes in the teachers and changes in the standards used by students to evaluate the teachers. Hence, it is possible that the lack of change in the ratings reflects counter-balancing effects of changes in teacher effectiveness and changes in standards students use to evaluate teaching effectiveness.

In order to examine this potential problem, correlations between the ratings and the year they were collected (1976–1988) are presented for 6024 classes taught by teachers included in our longitudinal sample and for the remaining 9641 classes offered in the same 31 academic departments by other teachers (assistant, associate, and full professors) not otherwise considered here. Correlations based on teachers not included in our study provide a possible control for changes in the standards students use to evaluate teaching effectiveness. If, for example, ratings steadily improve over time for ratings based on the remaining 9641 classes, then it might be argued that the lack of change for the 6024 classes considered here really reflects a relative decline. Correlations between year and the 11 SEEQ scores (Table 4) vary from −.061 to .032 for the teachers in the longitudinal sample, from −.067 to .097 for teachers not in the longitudinal sample, and from −.056 to .070 for the combined sample. The sizes of these relations are very small and do not differ substantially in the two samples. Whereas alternative explanations may exist, the results suggest that the overall level of SETEs in these departments has not varied substantially over this 13-year period and supports our interpretation of the results.

Linear and Quadratic Effects

Feldman (1983) suggested that there may be a nonlinear relation between years of teaching experience and student ratings in which ratings initially increase, peak, and then decline slowly. Such a nonlinear effect would complicate our study, because the teachers varied substantially in the amount of teaching experience they had at the start of the study. If this nonlinear trend exists: (a) ratings of the least experienced

Table 4

Correlations Between Students' Evaluations and Year: Total Sample, Longitudinal Sample, Non-Longitudinal Sample

	Total sample (n = 15,665)	Long. sample[a] (n = 6024)	Non-long. sample (n = 9641)
Factor scores			
Learning/Value	.010	.005	.018
Enthusiasm	.020**	−.020	.045***
Organization	−.010	−.044***	.013
Group interact	−.041***	−.042**	−.047***
Indiv. rapport	−.056***	−.049***	−.067***
Breadth	−.052***	−.036**	−.059***
Exams	−.052***	−.036**	−.063***
Assignments	−.021**	−.009	−.028**
Workload	.070***	.032*	.097**
Overall ratings			
Course	−.005	−.042***	.017
Instructor	−.027***	−.061***	−.004

[a]Because ratings of classes taught by the same teacher in the same year were not averaged in the non-longitudinal comparison group, the 6024 (unaveraged) classes were evaluated in the longitudinal group instead of the 3125 averaged responses considered in earlier analyses of the longitudinal group. The comparison of correlations presented here for the longitudinal group are, however, very similar to those presented in Table 2.

*p < .05; **p < .01; ***p < .001.

teachers in our study should increase during at least the first few years of the study; (b) ratings of teachers with intermediate amounts of experience should be stable during the first few years of the study, followed by a gradual decline; and (c) ratings of the most experienced teachers should decline throughout our study. It is possible that such a trend is lost by considering results averaged across teachers with varying amount of experience.

In order to examine this potential problem, we limited consideration to instructors who had the same academic rank (assistant, associate or full professor) during the first 3 years that they were evaluated. Assistant professors are typically promoted within 5 or 6 years—sometimes sooner. Because assistant professors who do not receive tenure can not stay at the university more than 7 years, they are automatically excluded from our 13-year longitudinal study. Thus, teachers who were assistant professors during the first 3 years of our study were typically inexperienced teachers. Conversely, it typically takes at least 7 years—often much longer—to become a full professor. Thus, teachers who were full professors during the first 3 years of our study were typically experienced teachers. Based on academic rank we formed three groups who had relatively little experience (assistant professors), an intermediate amount of experience (associate professors), and considerable experience (full professors). A subsequent set of multiple regressions was conducted in which the effects of year (linear and quadratic components), level (graduate vs. undergraduate), initial rank, and all possible interactions were estimated. Of critical concern were the interactions of initial rank with linear and quadratic components of year and, perhaps, the interaction of these effects with level. None of the year by initial rank interactions depended on course level. The quadratic component of year did not interact with initial rank for any of the 11 SEEQ scores, but the linear effect of the year interacted with initial rank for 6 of 11 SEEQ scores (see Table 5). Inspection of the correlations between SEEQ scores and year for each initial rank (Table 5), however, demonstrates that these effects are very small. There is, however, a consistent pattern in which correlations are most negative (or least positive) for assistant professors and least negative (or most positive) for full professors.

These results suggest that the effect of year may vary slightly depending on initial rank, but the results are not consistent with the trend suggested by Feldman (1983). The lack of nonlinearity was consistent across the three groups. Furthermore, Feldman's speculations suggested that ratings would increase for the least experienced teachers and decrease for the most experienced teachers, whereas our weak trends were in the opposite direction.

Separate Analyses of Responses For Each Instructor

In an alternative approach to this problem of the stability of mean ratings, we evaluated the linear and nonlinear effects separately for each of the 195 instructors. Separate analyses were conducted for the 19 to 61 classes taught by each instructor (mean = 30.9) and were summarized using the techniques of meta-analysis (e.g., Glass, McGaw, & Smith, 1981; Hedges & Olkin, 1985). Means and standard deviations are presented for the simple correlations, and for the linear and quadratic components of year in Table 6. Normalizing transformations made little difference because the coefficients are consistently close to zero and approximately normally distributed, and so only the untransformed values are presented. Interpretations of these mean statistics are essentially the same as those in earlier analyses (Table 3) and so are not repeated. This similarity across different analytic techniques having somewhat different assumptions, however, contributes to confidence in the earlier interpretations.

A new aspect considered here is tests of the homogeneity of effect sizes. Whereas the mean effect sizes are consistently close to zero, the homogeneity tests indicate the variability of the effect sizes is significantly larger than would be expected by chance alone. For the linear effects, for example, this means that there were subgroups of instructors whose ratings declined significantly over time and others who increased significantly over time. Similarly, for quadratic effects there were subgroups with significant U-shaped trends and others with inverted U-shaped trends. Inspection of these

Table 5

Correlations Between Students' Evaluations and Year for Instructors with an Initial Rank of Assistant, Associate, or Full Professor During the First 3 Years

	Classes taught by:			
	Assist Prof. (n = 767 classes)	Assoc Prof. (n = 934 classes)	Full Prof. (n = 1154 classes)	Significance of differences
Factor scores				
Learning/Value	−.010	−.065	.028	ns
Enthusiasm	−.074	−.085	.044	*
Organization	−.119	−.056	−.002	*
Group intereact	−.061	−.043	−.051	ns
Indiv. rapport	−.105	−.027	−.057	ns
Breadth	−.070	−.047	−.017	ns
Exams	−.070	−.081	.031	ns
Assignments	−.074	−.031	.009	*
Workload	.03	.025	.039	*
Overall ratings				
Course	−.090	−.098	.017	*
Instructor	−.117	−.113	.021	*

Note. Multiple regression was used to predict factor scores and overall ratings from year (linear and quadratic components), course level (graduate or undergraduate), and initial rank and all possible interactions. Significance of differences refers to a significant linear year-by-initial-rank interaction.
*$p < .05$.

tests (Table 6) indicates that the nonhomogeneity of effects is substantially larger for the linear effects than the nonlinear effects. Also, whereas the homogeneity tests are reasonably similar across the different SEEQ scores, the lack of homogeneity is substantially larger for the Workload/Difficulty score.

As noted elsewhere (e.g., Alexander, Scozzaro, & Borodkin, 1989; Hedges & Olkin, 1985; Hunter, Schmidt, & Jackson, 1982; also see Marsh, Balla, & McDonald, 1988) this test of homogeneity is so powerful that small deviations will produce significant effects when N is large. Hedges and Olkin (1985), for example, frequently use this test for evaluating homogeneity within 15 or fewer studies compared to our omnibus test across all 195 instructors. One approach to evaluating the practical significance considered here was to count the number of instructors in which the observed effect sizes differ significantly from the mean of observed effect sizes across all instructors. If the results are homogeneous, then one would expect 5% of the samples would be significant. Across all 11 SEEQ scores, these tests were significant ($p < .05$) for approximately 16% (linear effects) and 10% (nonlinear effects) of the instructors. This difference between expected and observed homogeneity does not appear to be large, but may be sufficient to warrant further research. For example, earlier analyses indicated that the linear effect of time varied somewhat depending on the initial rank of the instructor. Nevertheless, we interpret these results to indicate that the lack of linear and nonlinear effects of time generalizes reasonably well across instructors.

Discussion

SETEs of the same teachers over a 13-year

Table 6

Linear and Nonlinear Changes in Separate Analyses of 195 Instructors for the 1976–1988 Period and a χ^2 Test of the Homogeneity of Effects Across Instructors (Hom χ^2)

	r			Linear beta			Quadratic beta		
	M	*SD*	Hom χ^2	*M*	*SD*	Hom χ^2	*M*	*SD*	Hom χ^2
Factor scores									
Learning/Value	.011	.293	462.5**	.008	.282	475.1**	−.054	.239	302.6**
Enthusiasm	−.030	.276	449.2**	−.036	.299	488.9**	−.028	.263	329.4**
Organization	−.068	.278	502.0**	−.087	.296	471.1**	−.012	.267	350.8**
Group interact	−.008	.249	354.9**	−.008	.273	366.4**	−.039	.261	313.3**
Indiv. rapport	−.025	.273	434.9**	−.044	.302	435.9**	.031	.254	310.7**
Breadth	.019	.258	385.3**	.015	.278	395.6**	−.012	.241	304.0**
Exams	−.036	.256	374.2**	−.043	.274	368.4**	−.004	.257	302.5**
Assignments	.017	.252	381.5**	−.004	.266	368.5**	−.019	.264	313.9**
Workload	−.014	.302	603.9**	−.002	.323	590.1**	−.011	.258	329.1**
Overall ratings									
Course	−.028	.261	417.8**	−.036	.273	398.6**	−.029	.242	293.1**
Instructor	−.054	.275	452.1**	−.064	.293	446.5**	−.022	.260	333.1**

Note. Year (1976–1988) was related to instructor ratings separately for each of the 195 instructors. The number of courses for each instructor varied from 19 to 61 (mean = 30.9). Presented are mean and standard deviations of the simple correlation (*r*), beta weights for the linear component, and beta weights for the quadratic component. Chi-square tests for the homogeneity of effect sizes (Hom χ^2) were computed using *r* to *z* transformations and weighting for sample size (Hedges & Olkin, 1985; Alexander, Scozzaro, & Borodkin, 1989).
**$p < .01$.

period of time are remarkably stable. The mean ratings for our cohort of 195 teachers showed almost no systematic changes over this period. We interpret this finding to indicate that teaching effectiveness as perceived by students is stable. Supplemental analyses suggested that the standards that students use apparently did not change over this period as ratings of all teachers—those in our longitudinal sample and those who were not—were also stable. The very weak (inverse) effects that we did observe were primarily linear. Nonlinear effects were not observed for either the total sample, or subsamples of teachers with little, intermediate, or substantial amounts of teaching experience at the start of the 13-year longitudinal study. Extremely powerful tests indicated that the effect sizes—particularly the linear effects—were not entirely consistent over instructors. For example, ratings of instructors who had an initial rank of assistant professor declined somewhat more over the 13-year period considered here than did those who were full professors (Table 5) at the start of the study. From a practical perspective, however, the lack of homogeneity did not appear to be large and the overall conclusion that mean ratings are stable over time appears to generalize reasonably well across instructors.

The results of the present investigation are important because they demonstrate that teaching effectiveness is stable over time and because they counter findings of other research that has found either inverted U-shaped relations or weak negative relations between SETEs and teaching experience. It is important, however, to note that most of this previous research is based on cross-sectional designs instead of a longitudinal design like the one considered here. In cross-sectional designs the ratings of teachers with many years of experience are used to infer how less experienced teachers will be evaluated many years in the future, whereas the ratings of inexperienced teachers are used to infer how experienced teachers were (or would have been) evaluated many years in the past;

both these inferences must be made cautiously. Particularly during the last few decades, many assistant professors do not receive tenure and leave the university. For this reason alone, it is problematic to make comparisons between less experienced teachers (including those who will not get tenure) and more experienced teachers who were tenured. This issue is further complicated by the role that SETEs, in combination with research performance and other criteria, play in the decision as to who is granted tenure. Our longitudinal study, because it is based on such a large and diverse sample of teachers considered continuously over such a long period of time provides a stronger basis of inference about the mean stability of SETEs and is apparently unique in this area of research.

References

Alexander, R. A., Scozzaro, M. J., & Borodkin, L. J. (1989). Statistical and empirical examination of the chi-square test for homogeneity of correlations in meta-analysis. *Psychological Bulletin*, **106**, 329–331.

Barnes J. (1985). Experience and student achievement/teacher effectiveness. In T. Husen & T. N. Postlethwaite (Eds.), *International encyclopedia of education: Research and studies* (pp. 5125–5128). Oxford: Pergamon Press.

Bausell, R. B., & Bausell, C. R. (1979). Student ratings and various instructional variables from a within-instructor perspective. *Research in Higher Education*, **11**, 167–177.

Braskamp, L. A., Brandenburg, D. C., & Orgy, J. C. (1985). *Evaluating teaching effectiveness: A practical guide*. Beverly Hills, CA: Sage.

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

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. *Research in Higher Education*, **18**, 3–124.

Feldman, K. A. (1984). Class size and students' evaluations of college teacher and course: A closer look. *Research in Higher Education*, **21**, 45–116.

Feldman, K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics: A review and synthesis. *Research in Higher Education*, **24**, 139–213.

Feldman, K. A. (1987). Research productivity and scholarly accomplishment: A review and exploration. *Research in Higher Education*, **26**, 227–298.

Feldman, K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities. *Research in Higher Education*, **28**, 291–344.

Feldman, K. A. (1989a). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education*, **30**, 137–194.

Feldman, K. A. (1989b). Association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education*, **30**, 583–645.

Gilmore, G. M., Kane, M. T., & Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. *Journal of Educational Measurements*, **15**, 1–13.

Glass, G. V., McGaw, B., & Smith, M. L. (1981). *Meta-analysis in social research*. Beverly Hills, CA: Sage.

Goldstein, H. (1979). *The design and analysis of longitudinal studies: Their role in the measurement of change*. London: Academic Press.

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. New York: Academic Press.

Hunter, J. E., Schmidt, F. L., & Jackson, G. B. (1982). *Meta-analysis: Cumulating research findings across studies*. Beverly Hills, CA: Sage.

Kulik, J. A., & Kulik, C. (1974). Student ratings of instruction. *Teaching of Psychology*, **1**, 51–57.

Marsh, H. W. (1981). The use of path analysis to estimate teacher and course effects in student ratings of instructional effectiveness. *Applied Psychological Measurement*, **6**, 47–60.

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

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

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

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

Marsh, H. W., Balla, J. R., & McDonald, R. P. (1988). Goodness-of-fit indices in confirmatory factor analysis: The effect of sample size. *Psychological Bulletin*, **102**, 391–410.

Marsh, H. W., & Dunkin, M. J. (in press). Students' evaluations of university teaching: A multidimensional perspective. In J. Smart (Ed.), *Higher education: Handbook of theory and research*. New York: Agathon.

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.

Overall, J. U., & Marsh, H. W. (1980). Students' evaluations of instruction: A longitudinal study of their stability. *Journal of Educational Psychology*, **72**, 321–325.

Pedhazur, E. J. (1982). *Multiple regression in behavioral research* (2nd Ed.). New York: Holt, Rinehart and Winston.

Plewis, I. (1985). *Analyzing change: Measurement and explanation using longitudinal data*. New York: Wiley.

Rogosa, D. R. (1979). Causal models in longitudinal research: Rationale, formulation, and interpretation. In J. R. Nesselroade & P. M. Baltes (Eds.), *Longitudinal research in the study of behavior and development* (pp. 263–302). New York: Academic Press.

Rogosa, D. R., Floden, R., & Willett, J. B. (1984). Assessing the stability of teacher behavior. *Journal of Educational Psychology*, **76**, 1000–1027.

Rogosa, D. R., & Willett, J. B. (1985). Satisfying a simplex structure is simpler than it should be. *Journal of Educational Statistics*, **10**, 99–107.

Ryans, D. G. (1960). Prediction of teacher effectiveness. In C. W. Harris (Ed.), *Encyclopedia of educational research* (pp. 1486–1491). New York: Macmillan.

Willett, J. B. (1988). Questions and answers in the measurement of change. In E. Z. Rothkopf (Ed.), *Review of research in education* (Vol. 15). Washington, DC: American Education Research Association.

Received 4 April 1991 □