The Use of Path Analysis to Estimate Teacher and Course Effects in Student Ratings of Instructional Effectiveness

Herbert W. Marsh
University of Sydney

A path analytic technique is described for obtaining separate estimates of variables that are normally confounded. The particular problem involves ascertaining the relative contributions of the instructor and the course being taught in determining student ratings of teaching effectiveness. A series of six different path analytic models were used to estimate teacher and course effects, along with a variety of other parameters. The best model indicated that the effect of the teacher is about five times as large as the effect of the course and that the difference is even larger for components such as Overall Instructor and Instructor Enthusiasm. In contrast to the student rating items, background/demographic variables such as class size, students' prior subject interest, and reason for taking a course were largely a function of the course rather than the instructor.

An important problem in applied psychological research is determination of the separate effects of two naturally confounded variables. The classic experimental approach (Campbell & Stanley, 1967; Cook & Campbell, 1979) is to randomly assign subjects to different treatment conditions, thus imposing an artificial independence between the different effects. However, this approach may not be possible or even desirable in some conditions, since many variables are not amenable to random assignment. Furthermore, the basic nature of the variable may not be the same if random assignment is imposed, thus threatening the external validity of the study and the generalizability of the findings to the original setting from which the study was generated. Such a situation exists in the study of students' evaluations of teaching effectiveness and in the attempt to unconfound the effects of the teacher and the course. If the effect of the particular course being taught has substantial impact on student ratings, particularly relative to the size of the teacher effect, then the practice of comparing ratings of different instructors for tenure/promotion decisions is dubious. In any one class setting, the course being taught and the instructor teaching it are completely confounded. Furthermore, random assignment is not appropriate; there is little interest in knowing how well a historian would do at teaching an advanced calculus course, or even how well a personality theorist would fare in a physiological psychology course. The purpose of this study was to develop a set of models that unconfounded these effects.

Students' evaluations are designed to measure instructional effectiveness. Researchers, using a construct validation approach, have sought to demonstrate that the ratings are positively related to variables indicative of effective teaching and unrelated to variables that are not. Student ratings have been validated against such criteria

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 6, No. 1, Winter 1982, pp. 47–59
© Copyright 1982 Applied Psychological Measurement Inc.
0146-6216/82/010047-13$1.65

as the retrospective ratings of former students (Centra, 1974; Marsh, 1977; Overall & Marsh, 1980), the affective course consequences (Marsh & Overall, 1980), and the self-ratings of the faculty being evaluated (Marsh, Overall, & Kesler, 1979; Marsh, in press–a). However, the most common criterion has been student learning measured by an objective examination. Researchers have considered different sections of the same multisection course that are taught by different teachers and have found that those sections that perform best on the standardized examinations are also evaluated most favorably (Centra, 1977; Frey, 1973; Frey, Leonard, & Beatty, 1975; Marsh, Fleiner, & Thomas, 1975; Marsh & Overall, 1980; Sullivan & Skanes, 1974). In spite of this display of the validity of student ratings across a variety of criteria, validity coefficients are generally in the range of .35 to .80, indicating that there is considerable variance left to be explained by other variables.

The demonstration that student ratings are valid does not preclude the possibility that they are also biased by factors unrelated to teaching excellence, and faculty generally believe this to be the case (Marsh, in press–a). However, literature reviews have typically concluded that a variety of potential biases in student ratings apparently have little impact (Costin, Greenough, & Menges, 1971; Hildebrand, Wilson, & Dienst, 1973; Marsh, 1980a, 1980b; McKeachie, 1973, 1979; Remmers, 1963).

Arguing that background and course characteristics beyond the instructor's control do not influence student ratings involves trying to prove the null hypothesis of no effect. Difficult under ideal situations, this problem is particularly troublesome in an area where most data are correlational, making causal inferences problematic (but see March, 1980a); and the definition of what constitutes a potential bias is fuzzy. An alternative approach is to try to determine independent effects of the course and the teacher, demonstrating that the teacher effect is large relative to the course effect. If the influence of the course is large and the ratings depend more upon what course is being taught than upon who is teaching it, then the use of student ratings as a measure of the *instructor's* effectiveness is questionable.

Student ratings of any one teacher, collected in one class setting, confound the effect of teacher and course. Ideally, each of a large set of randomly selected courses would be taught by a randomly selected group of teachers and evaluated with a well-developed multifactor rating form. This would allow the partition of variance into independent effects due to the teacher, the course, and their interaction, and the determination of how these effects vary across different components of effective teaching. In fact, instructors generally teach only a few different courses, any given course is only taught by a few different teachers, and the teaching assignments are definitely not random.

Researchers have considered alternative approaches to this problem (Bausell, Schwartz, & Purohit, 1975; Gillmore, Kane, & Naccarato, 1978; Kulik & Kulik, 1974; Marsh & Overall, 1981). Each of these approaches considers the consistency of one instructor's ratings across different offerings of the same course or across offerings of different courses, or the similarity in ratings of the same course taught by different instructors. In two of these studies (Bausell et al., 1975; Kulik & Kulik, 1974) the investigators correlated ratings from pairs of courses in which (1) both the instructor and the course were the same, (2) the instructor was the same but the course was different, and (3) the course was the same but the instructor was different. In both studies, correlations between ratings of two different courses taught by the same teacher tended to be substantially higher than when two different teachers taught the same course. This argues that the teacher effect is larger than the course effect. There were some items—particularly those related to Workload/Difficulty, course relevance, and rapport—in which the course effect was as large or larger than the instructor effect.

Kulik and Kulik (1974) extended this analysis by developing an explicit path-analytic model to estimate the course and teacher effects. The simplest such model consists of only two parameters—a teacher effect (T) and a course effect (C). Parameters of this model can be estimated from the following equations:

1. $r_1 = T + C =$ correlation between ratings of same teacher in two different offerings of the same course.
2. $r_2 = T =$ correlations between ratings of same teacher in two different offerings of different courses.
3. $r_3 = C =$ correlations between ratings of different teachers in the same course.

Since there are two parameters and three equations, this model can be tested by comparing the sum of r_2 and r_3 with r_1. This simple model assumes that a number of possible sources of covariance do not exist. For example, two courses taught by the same teacher are assumed to be unrelated with respect to conduciveness to good ratings (course covariance effect), as are the different teachers assigned to teach the same courses (teacher covariance effect). Furthermore, it assumes that no instructor performs systematically better or worse in a particular course (uniqueness covariance effect) and that the effectiveness of a teacher is unrelated to the course effect (teacher-course covariance effect, e.g., teachers who are particularly skillful at leading small group discussions are not more likely to teach courses that are more conducive to small group discussion). Each of these assumptions represents a covariance term that could be added to the model as an additional parameter. However, since there are only three equations, there can be no more than three unknowns (i.e., the teacher effect, the course effect, and one of the covariance terms).

Kulik and Kulik (1974) considered only one model that contained parameters for the teacher, the course, and the teacher-course covariance. Their analysis suggested that the teacher effect was substantially larger than the course effect for ratings of *skill* and *structure*, that the two effects were approximately equal for ratings of *rapport*, and that the course effect was larger for ratings of *difficulty*. The covariation term was positive for rapport ratings, suggesting that teachers who are best at establishing rapport are more likely to teach courses conducive to establishing rapport. However, the covariation term for the skill factor was negative, which led them to the unlikely conclusion that "the most skillful teachers are assigned to (or choose to teach in) courses in which it is hardest to shine." However, Kulik and Kulik stressed that their data was based upon a small sample, that the data were limited, that their calculations provided only rough approximations, that many assumptions were untested, and that more elaborate models were possible and probably more plausible. Correlations between pairs of ratings of the same course taught by the same instructor, for example, were based upon only 30 data points, allowing a large sampling fluctuation. In one instance (out of four comparisons) the correlation between ratings of the same instructor teaching the *same* course was less than the correlation between ratings of the same instructor teaching *different* courses. This led to the conclusion that at least one of these correlations must be inaccurate. Additionally, their sampling procedure further confused the issue in that the instructors and courses used in the different comparisons were not the same and were chosen according to different criteria.

The purpose of the present study was to develop alternative models to estimate the effects of the teacher, the course, and various sources of covariation in determining student ratings. Ratings from 1,364 classes were selected from a data base containing evaluations of more than 8,000 classes, eliminating many of the problems of a small sample size experienced in previous studies. Four different comparisons were made, thus allowing the estimation of four parameters and the fitting of more elaborate models. Data were all based upon the same evaluation instrument, which has a well developed factor structure, thus allowing a determination of how the effects varied over different components of effective teaching. Finally, comparisons were also made for different background/demographic characteristics that were expected to be more a function of the course than the teacher (e.g., class size).

Method

During the period between 1976 and 1980, students' evaluations were collected in 8,277 classes, representing 35 different academic departments at the University of Southern California. The different academic units included Social Sciences, Humanities, Business, Engineering, Education, Gerontology, and the Institute of Systems and Safety Management. Although an academic unit's participation in this particular program was voluntary, the University required that each unit systematically collect some form of students' evaluations and would not consider tenure/promotion recommendations that did not contain such documentation. Consequently, most of the academic units that did participate in the program required that all their faculty be evaluated in each course they taught. The student evaluation forms were distributed to faculty shortly before the end of each semester, administered by a student in the class, and taken to a central office where they were processed. The program, Students' Evaluations of Educational Quality (SEEQ), the instrument upon which it is based, and the research that led to its development are described by Marsh (in press-b).

The evaluation instrument consisted of 35 evaluation items and several additional background/demographic items. Except for the determination of reliability that considered the responses of individual students within each class, all analyses in this study were based upon class-average responses. Ratings on SEEQ are summarized by 11 evaluations scores—9 evaluation factor scores and 2 overall summary items. The 9 factors had been previously identified by factor analysis of a sample of undergraduate social science classes (Marsh, 1978, in press-b). The same 9 evaluation factors also resulted from a factor analysis of faculty self-evaluations of their own teaching effectiveness on the SEEQ form (Marsh, in press-a; Marsh & Cooper, 1981). A factor analysis of the ratings from the 1,364 courses considered in this study was performed to test the replicability of the factors in a larger, more diverse sample. Factor scores, based upon this analysis, were produced by the SPSS program (Nie, Hull, Jenkins, Steinbrenner, & Bent, 1975) and were used in subsequent analyses.

Several criteria were used for the selection of courses to be considered in this study. Courses taught by teaching assistants and courses that were evaluated by fewer than 12 students were eliminated from consideration. Courses were then arranged into sets such that each set contained two evaluations of the same instructor teaching the same course on two different occasions (A and B), evaluations of a different course taught by the same instructor (C), and ratings of the same course as in A and B that was taught by a different instructor (D). Once an instructor had appeared in position A, he/she was not considered for that position again but could appear in position D for other sets. Through this selection process, a total of 341 sets of courses were selected, each set consisting of four courses that fit the configuration described above. From this data, four correlations were obtained for each of the evaluation scores and for several background/demographic items. These were

r_1: the correlation between A and B courses —ratings of the same instructor teaching the same course on two different occasions;

r_2: the average of the correlation between A and C courses and B and C courses—ratings of the same instructor teaching two different courses;

r_3: the average of the correlations between A and D courses and B and D courses—ratings of the same course taught by two different instructors;

r_4: the correlation between C and D courses —ratings of two different instructors (who sometimes teach the same courses) teaching two different courses (which are sometimes taught by the same instructor).

Analysis was conducted in two different stages. In the first stage the ratings from all 1,364 courses were factor analyzed and the results of this factor analysis were used to compute factor scores. An Anova model was also used to determine the reliability (intraclass correlations; see Marsh & Overall, 1979; Winer, 1971) of the class-average ratings for each of the evaluation factors. In the second phase, each of the evaluation factors and their respective reliabilities were used to fit a variety of models that estimated teacher and course effects under different sets of assumptions.

Results

Factor Analysis and Reliability

The results of the present investigation (see Table 1) clearly identify each of the nine factors the instrument was designed to measure. Each item loaded highest on the factor it is designed to measure; each of these loadings was at least .37 and most exceeded .60. All other loadings were less than .30, and were generally less than .20. These results are quite similar to previous factor analyses of both student ratings and instructor self-evaluations of their own teaching effectiveness (Marsh, in press-a; Marsh & Cooper, 1981). The findings offer further support for the contentions that student ratings are multidimensional and that the SEEQ instrument measures distinct components of teaching effectiveness.

The reliability of student ratings had been estimated with internal consistency measures (Marsh, in press-a), intraclass correlation coefficients (Marsh & Overall, 1979), and stability over time (Marsh & Overall, 1979; Overall & Marsh, 1980). However, Marsh (in press-b) and Gillmore et al. (1978) have both argued that some form of the intraclass correlation is the most appropriate measure. According to this conceptualization, ratings are reliable if there is relative agreement among students within the same class (within class variance) and significant differentiation in the average ratings of different classes (between class variance). A conceptually similar approach would be to take random halves of the students in each class, to correlate the mean ratings of these split halves, and to correct the resulting correlation with the Spearman-Brown equation. For purposes of this analysis, the ratings of the items defining a given factor were averaged separately for each of the 39,580 students in the 1,364 classes. Mean ratings for the entire set of ratings were substituted for missing values. A one-way Anova was then conducted in which each of the 1,364 courses was a level, and responses from students within each of these classes were used to determine the MS_{within} effect. The results of this analysis were then used to estimate the reliability (Winer, 1971) of each of the evaluation scores (see Table 2). Reliability coefficients for all the evaluation factors and the two overall summary ratings exceeded .9, indicating that the student ratings were quite reliable.

Correlations Between Different Sets of Courses

Values for each of the four sets of correlations described earlier are presented in Table 2. Correlations between ratings of the same instructor teaching the same course were consistently near .7 for all the evaluation scores, but these values were much lower than the reliabilities, which all exceeded .9. This suggests that there is a substantial portion of the variance (over 20%) in student ratings that is reliable but is unique to a particular offering of a given course. Correlations between ratings of the same instructor teaching two different courses—one possible estimate of the teacher effect—averaged .52, ranging from values near .4 for ratings of Workload/Difficulty and Assignments to values over .6 for Instructor Enthusiasm and the Overall Instructor rating. Correlations between ratings of different instructors teaching the same course were substantially lower, averaging .14, and provided one estimate of the course effect. This third set of correlations was the most variable, having near zero values for many factors but

Table 1
Factor Analysis of Students' Evaluations
of Teaching Effectiveness (N=1364 Classes)

	Factor Pattern Coefficients								
Evaluation Items (paraphrased)	I	II	III	IV	V	VI	VII	VIII	IX
1. Learning/Value									
Course Challenging/Stimulating	.41	.23	.11	.03	-.01	.16	.15	.14	.28
Learned Something Valuable	.61	.07	.06	.02	.00	.14	.11	.19	.13
Increased Subject Interest	.65	.10	.02	.05	.04	.18	.08	.13	.02
Learned/Understood Subject Matter	.54	.00	.18	.15	.02	.05	.12	.16	-.23
Overall Course Rating	.37	.26	.19	.06	.02	.11	.16	.18	.08
II. Enthusiasm									
Enthusiastic About Teaching	.13	.52	.19	.09	.22	.12	-.01	.06	.02
Dynamic & Energetic	.11	.67	.15	.08	.12	.10	-.01	.07	.03
Enhanced Presentations with Humour	.13	.62	.07	.12	.10	.11	-.04	.12	-.06
Teaching Style Held Your Interest	.18	.57	.21	.14	.03	.06	.05	.10	.01
Overall Instructor Rating	.15	.40	.30	.10	.12	.10	.11	.11	.05
III. Organization									
Instructor Explanations Clear	.21	.13	.49	.16	.02	.07	.14	.11	-.07
Course Materials Prepared & Clear	.10	.04	.66	.03	.02	.07	.23	.08	.00
Objectives Stated & Pursued	.18	.01	.47	.03	.00	.05	.30	.20	.04
Lectures Facilitated Note Taking	.02	.09	.59	-.12	.02	.16	.25	-.03	.01
IV. Group Interaction									
Encouraged Class Discussions	.09	.11	-.01	.80	.04	.04	.07	.06	.00
Students Shared Ideas/Knowledge	.11	.04	-.02	.83	.08	.09	.03	.05	-.04
Encouraged Questions & Answers	.07	.12	.22	.57	.12	.10	.13	.08	.02
Encouraged Expression of Ideas	.06	.08	.07	.71	.17	.16	.07	.07	.00
V. Individual Rapport									
Friendly Towards Students	.07	.14	.06	.20	.65	.05	.01	.10	-.04
Welcomed Seeking Help/Advice	.05	.02	.14	.10	.81	.05	.02	.07	.01
Interested In Individual Students	.10	.09	.06	.18	.68	.07	.06	.10	.02
Accessible to Individual Students	.01	.07	.18	.02	.61	.14	.14	.07	.05
VI. Breadth of Coverage									
Contrasted Implications	.07	.07	.13	.05	.04	.69	.08	.09	.04
Gave Background of Ideas/Concepts	.13	.06	.16	.02	.07	.67	.04	.08	.00
Gave Different Points of View	.05	.05	.13	.14	.10	.67	.03	.12	-.04
Discussed Current Developments	.24	.10	.08	.11	.01	.55	.02	.05	.01
VII. Examinations/Grading									
Examination Feedback Valuable	.01	.13	.00	.02	.16	.10	.66	.13	-.05
Eval.Methods Fair/Appropriate	.03	.13	-.09	.05	.22	.11	.69	.16	-.03
Tests Emphasized Course Content	.07	.15	.03	.00	.15	.07	.64	.15	.01
VIII. Assignments									
Reading/Text Valuable	.04	-.02	.04	.00	-.01	.06	.00	.89	.08
Added to Course Understanding	.09	.02	.07	.04	.04	.00	.07	.81	.08
IX. Workload/Difficulty									
Course Difficulty (Easy-Hard)	-.07	.08	.09	-.10	-.04	.12	.07	.06	.83
Course Workload (Light-Heavy)	.12	-.08	.02	.05	.03	-.02	-.04	.07	.90
Course Pace (Too Slow-Too Fast)	-.10	.17	.05	-.11	-.05	.07	.05	.06	.70
Hours/week Outside of Class	.12	-.06	-.08	.09	-.09	-.08	.00	.08	.82

Note: Factor loadings in boxes are the loadings designed to measure each factor. The analysis was performed with the commercially available SPSS routine (See Nie, et al., 1975). Correlations among the oblique factors varied between .03 and .49 (Median r = .30).

values of .3 and higher for others. This estimate of the course effect was particularly high for the Workload/Difficulty factor, the only factor where the course effect approached the size of the teacher effect.

The values for each of the four sets of correlations are also presented for several Background/Demographic variables that were suspected of being largely a function of the particular course (see Table 3). The Workload/Difficulty factor is included with these to emphasize that it might more appropriately be considered a background variable than an evaluation of effective teaching. Correlations between ratings for the same course taught by the same instructor were again near .7, but for these variables the course effect was much larger than the teacher effect. Apparently, the particular course rather

Table 2
Correlations Among Different Sets of Classes for Each of the Evaluation Scores

Evaluation Scores	Reliability (intraclass correlation)	r_1 Same Teacher Same Course	r_2 Same Teacher Different Course	r_3 Different Teacher Same Course	r_4 Different Teacher Different Courses
Learning/Value	.935	.696	.563	.232	.069
Enthusiasm	.956	.734	.613	.011	.028
Organization/Clarity	.937	.676	.540	-.023	-.063
Group Interaction	.939	.699	.540	.291	.224
Individual Rapport	.929	.726	.542	.180	.146
Breadth of Coverage	.922	.727	.481	.117	.067
Examinations/Grading	.920	.633	.512	.066	-.004
Assignments	.900	.681	.428	.332	.112
Workload/Difficulty	.933	.773	.400	.392	.215
Overall Course	.918	.712	.591	-.011	-.065
Overall Instructor	.934	.719	.607	-.051	-.059
Mean Coefficient	.929	.707	.523	.140	.061

Table 3
Correlations Among Different Sets of Courses for Background/Demographic Items

Background/ Demographic Items	r_1 Same Teacher Same Course	r_2 Same Teacher Different Courses	r_3 Different Teachers Same Course	r_4 Different Teachers Different Courses
Prior Subject Interest	.635	.312	.563	.209
Reason for Taking Course (Percent Indicating General Interest)	.770	.448	.671	.383
Class Average Expected Grade	.709	.405	.483	.356
Workload/Difficulty	.773	.400	.392	.215
Course Enrolment	.846	.312	.593	.058
Percent Attendance on Day Evaluations Administered	.406	.164	.214	.045
Mean Coefficient	.690	.340	.491	.211

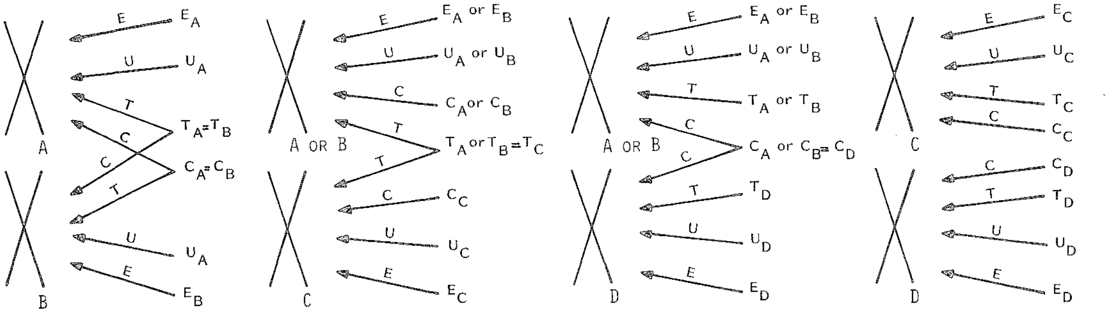
than the instructor who happens to be teaching it is the primary determinant of such variables as class size, students' prior subject interest, and reason for taking the class. Intuitively, these findings make sense and provide support for the logic behind the analyses conducted.

Models of the Teacher and Course Effects

A somewhat more elaborate model of each of the correlations is depicted in Table 4 and Figure 1. The most simple model, which is illustrated in Figure 1 and whose solution is given under Model I, assumes that each rating consists of an error term (E), a reliable but unique component (U), a teacher effect (T), and a course effect (C). The error is one minus the reliability shown in Table 2, and the uniqueness is the difference between the reliability and the correlation of ratings for the same instructor teaching the same course (r_1). The correlation between any pair of courses is a function of the number of shared effects. In the simple model, having no covariance terms, there are four equations to estimate two parameters—the teacher and course effects. This allows two tests of the model. The first (see Table 4 and Figure 1) is that r_1 (same teacher-same course correlation) will equal the sum of r_2 and r_3. The second test is that r_4 (different teacher-different course correlation) will be zero. Inspection of Table 5 (Model I) indicates that this model did reasonably well for the average values across all the evaluation scores but that it did not fare nearly so well when each of the evaluation scores was considered separately. Deviations (errors in the predictions) for the first test were .1 or more for six scores, whereas two of the deviations for the second test exceeded .2. These failures suggest that a more elaborate model is necessary.

Four possible covariance terms that were discussed earlier were considered. For example, Model II includes a course covariation term and suggests that there is a relationship (with respect to conduciveness to high ratings) between two courses that are taught by the same instructor. The diagrams reflecting this model would differ from those in Figure 1 only in that double-headed curved arrows would be drawn between "C_A or C_B" and "C_C" for r_2, and "C_C" and "C_D" for r_4. Only these two relationships (r_2 and r_4) would be affected by this covariance term. The equations for the four correlations are presented under Model II. In this case, there are four equations to estimate three unknowns and one degree of freedom still remains to test the model.

Figure 1
Path Diagrams for Correlations $r_1 - r_4$ as Hypothesized in Model I (see Table 4) with Each Course Evaluation ($X_A - X_D$) a Function of Teacher Effect (T), Course Effect (C), Reliable Uniqueness (U), and Error (E)



r_1
Correlations Between Ratings of the Same Teacher in the Same Course.

r_2
Correlations Between Ratings of the Same Teacher in Different Courses.

r_3
Correlations Between Ratings of Different Teachers in the Same Course.

r_4
Correlations Between Ratings of Different Teachers in Different Courses.

Table 4
Six Alternative Models For Determining Teacher (T), Course (C), and Various Covariance (COV) Effects

Model I (no covariances)	Model II (Course Covariance - COV_c)	Model III (Teacher Covariance - COV_T)	Model IV (Teacher-Course Covariance - COV_{TC})	Model V (Uniqueness Covariance - COV_U)	Model VI (Uniqueness, Teacher & Course Covariances)
$r_1 = T + C$	$r_1 = T + C$	$r_1 = T + C$	$r_1 = T + C + COV_{TC}$	$r_1 = T + C + COV_U$	$r_1 = T + C + COV_U$
$r_2 = T$	$r_2 = T + COV_C$	$r_2 = T$	$r_2 = T + COV_{TC}$	$r_2 = T$	$r_2 = T + COV_C$
$r_3 = C$	r_3	r_3	r_3	r_3	r_3
$r_4 = 0.0$	r_4	r_4	r_4	r_4	r_4
Definition of Effects	Definition of Effects	Definition of Effects	Definition of Effects	Definition of Effects	Definition of Effects
$T = r_2$	$T = r_1 - r_3$	$T = r_2$	$T = r_1 - r_3$	$T = r_2$	$T = r_2 - (r_4)/2$
$C = r_3$	$C = r_3$	$C = r_1 - r_2$	$C = r_1 - r_2$	$C = r_3$	$C = r_3 - (r_4)/2$
	$COV_C = r_2 + r_3 - r_1$	$COV_T = r_2 + r_3 - r_1$	$COV_{TC} - r_2 + r_3 - r_1$	$COV_U = r_1 - r_2 - r_3$	$COV_C = COV_T = (r_4)/2$
Tests	Tests	Tests	Tests	Tests	$COV_U = r_1 + r_4 - r_2 - r_3$
1) $r_1 - r_2 + r_3 = 0$ 2) $r_4 = 0$	$r_1 + r_4 - r_2 - r_3 = 0$	$r_1 + r_4 - r_2 - r_3 = 0$	$r_1 + r_4 - r_2 - r_3 = 0$	$r_4 = 0$	(No tests available)

r_1 = correlation between ratings of the same teacher in the same course

r_2 = correlation between ratings of the same teacher in two different courses

r_3 = correlation between ratings of different teachers in the same course

r_4 = correlation between ratings of different teachers in different courses

Note: Student ratings are a function of the Teacher Effect (T), Course Effect (C), reliable Uniqueness (U), and Error (E). The correlation between ratings in different courses is a function of the number of shared influences, including possible covariance effects. A path analysis diagram of Model I is shown in Figure 1.

The test in this case is that the sum of r_1 and r_4 will equal the sum of r_2 and r_3. The estimates generated by this model are shown in Table 5.

This model fared better than the simple model in that none of the deviations exceeded .2 and only four were greater than .1. However, there is still ample room for improvement. It is also disconcerting to note that the estimated covariance term was negative for 7 of the 11 scores, implying that two courses taught by the same instructor are negatively related in terms of conduciveness to Enthusiasm, Organization, Breadth of Coverage, and even Overall Instructor effectiveness. Logically this does not make sense and casts suspicion on the model. It is interesting to note that the same model applied to the data presented by Bausell (Bausell et al., 1975) also produced negative covariance terms in 27 of 36 comparisons that they considered.

Two of the remaining models (Models III and IV) are very similar to the one just discussed. In fact, although each differed in the way the teacher and course effects were defined, each involved the same test and resulted in the same set of deviation scores. Each of these models also resulted in an unlikely preponderance of negative covariance terms. Model III, involving a teacher covariance term, suggested that two teachers who teach the same course will be negatively related in terms of such characteristics as Enthusiasm, Organization, Breadth of Coverage, and Overall Instructor effectiveness. Model IV suggested that teachers who were particularly well organized tended to teach courses in which it was difficult to be organized and that particularly enthusiastic teachers will teach courses that were not conducive to being enthusiastic.

Model V had quite a different rationale from those already discussed. It assumed that the uniqueness components were correlated when the same instructor taught the same course. In this case (see Table 5) the covariance terms tended to be positive. This implies that when an instructor did reliably better (or worse) than would be expected in one offering of a particular course, the same thing would tend to occur when he/she taught the same course again. The majority of these covariance terms were positive, providing a picture that is more intuitively appealing than those already discussed. However, this model still did not fit the data well, predicting that r_4 will always be zero. Consequently, this model must be rejected in favor of one that contains a covariance term that can explain the nonzero correlations for r_4.

Since each of the models involving a single covariation term still has one remaining degree of freedom, more elaborate models could be postulated that contain four parameter estimates. In fact, however, models involving any two of the teacher, course, and teacher-course covariance terms contain an additional dependency that makes their solution impossible. Only models involving a combination of the uniqueness covariance term and one of the three other covariance terms can be solved. An additional possibility is demonstrated in Model VI, which contains five parameters but adds the additional constraint that the teacher and course covariance terms must be equal. Consequently, the number of unknowns is effectively reduced to four and a solution is possible.

Inspection of the estimated parameters in Model VI provides results that are more intuitively appealing than any considered so far. The teacher and course covariance terms tended to be small and were predominantly positive. The uniqueness covariance term is generally larger and is also positive in most cases. The course effect averaged about .10, is predominantly positive, and is substantially larger for those effects that seem to have the most to do with the course (e.g., Workload/Difficulty, Assignments, and Group Interaction). The teacher effect is by far the largest effect, about five times as large as the course effect. These differences are even larger in scores such as the Overall Instructor rating and the Instructor Enthusiasm factor where the course effect is minimal. It might also be argued that the uniqueness term should also be in-

Table 5
Parameter Estimates for Six Models Used to Determine
Teacher, Course, and Various Covariance Effects

	Evaluation Scores											
Models & Effects	Learn	Enths	Organ	Group	Indiv	Brdth	Exams	Assgn	Work	Crse	Instr	Mean
Model I												
Teacher (Tch)	.563	.613	.540	.540	.542	.481	.512	.428	.400	.591	.607	.523
Course (crse)	.232	.011	-.023	.291	.180	.117	.066	.332	.392	-.011	-.051	.140
Deviations (Test 1)	.099	.110	.159	.132	.004	.129	.055	.079	.019	.132	.163	.098
Deviations (Test 2)	.069	.028	.063	.224	.146	.067	.004	.112	.215	.065	.059	.096
Model II												
Teacher	.464	.723	.699	.408	.546	.610	.567	.349	.381	.723	.770	.567
Course	.232	.011	-.023	.291	.180	.117	.066	.332	.392	-.011	-.051	.140
Crse Covar	.099	-.110	-.159	.132	-.004	-.129	-.055	.079	.019	-.132	-.163	-.038
Deviations	.030	.094	.092	.092	.150	.196	.051	.112	.196	.067	.104	.108
Model III												
Teacher	.563	.613	.540	.540	.542	.481	.512	.428	.400	.591	.607	.529
Course	.133	.121	.136	.159	.184	.246	.121	.253	.373	.121	.112	.178
Tch-Covar	.099	-.110	-.159	.132	-.004	-.129	-.055	.079	.019	-.132	-.163	-.038
Deviations	.030	.094	.092	.092	.150	.196	.051	.112	.196	.067	.104	.108
Model IV												
Teacher	.464	.723	.699	.408	.546	.610	.567	.349	.381	.723	.770	.567
Course	.133	.121	.136	.159	.184	.246	.121	.253	.373	.121	.112	.178
Tch-Crs Covar	.099	-.110	-.159	.132	-.004	-.129	-.055	.179	.019	-.132	-.163	-.038
Deviations	.030	.094	.092	.092	.150	.196	.051	.112	.196	.067	.104	.108
Model V												
Teacher	.563	.613	.540	.540	.542	.481	.512	.428	.400	.591	.607	.529
Course	.232	.011	-.023	.291	.180	.117	.066	.332	.392	-.011	-.051	.140
Unique Covar	-.099	.110	.159	-.132	.004	.129	.055	-.079	.019	.132	.163	.038
Deviations	.069	.028	.063	.224	.146	.067	.004	.112	.215	.065	.059	.096
Model VI												
Teacher	.528	.599	.572	.428	.467	.447	.514	.372	.292	.623	.637	.498
Course	.198	-.033	.009	.179	.105	.083	.068	.276	.284	.021	-.021	.103
Unique Covar	-.030	.138	.095	.092	.154	.197	.051	.033	.197	.068	.103	.106
Tch Covar	.034	.014	-.032	.112	.075	.034	-.002	.056	.108	-.032	-.030	.031
Crs Covar	.034	.014	-.032	.112	.075	.034	-.002	.056	.108	-.032	-.030	.031

Note: Abbreviated labels for the evaluation scores correspond to those shown in Table 2. The models, parameter estimates and tests are presented in Table 4 and Figure 1. Deviations represent the absolute difference between the predicted and obtained value for each of the tests.

cluded as part of the teacher effect, as it is a teacher effect that is specific to a particular course.

The previous discussion has emphasized the differences between each of the various models. However, the estimated teacher and course effects were reasonably consistent across all the different models. The size of the mean teacher effect varied between .50 and .57, while the mean estimates of the course effect varied between .10 and .18. The consistency of these results, generated under a wide variety of different assumptions, adds further strength to the conclusion about the relative size of the teacher and course effects. The pattern of findings for the different evaluation components was also quite consistent across the different models; teacher effects were larger and course effects were smaller for Overall Instructor and Instructor Enthusiasm ratings, while the opposite was the case for Workload/Difficulty, Assignments, and Group Interaction ratings.

Discussion

The purpose of this study was to describe an analytic approach and to develop specific models designed to provide estimates of the effects of variables that are normally confounded. Specifically, separate estimates of the effect of the particular instructor and the particular course being taught were generated for students' evaluations of teaching effectiveness. In the initial analysis, it was demonstrated that the effect of the Teacher predominated, to varying extents, in each of the components of the students' evaluations but that the course was more important in determining such characteristics as class size, students' prior subject interest, and reason for taking the course.

A series of path-analytic models was then developed, each involving a teacher and course effect as well as a variety of other effects. The final such model, having the greatest intuitive appeal as well as fitting the data, suggested that the different courses taught by the same teacher tend to be similar, as do different teachers who teach the same courses. The model also implies that instructors who are uniquely effective (or ineffective) in any one particular offering of a given course will tend to perform similarly in another offering of the same course. Finally, this model shows that the teacher is the most important component, that his/her effect is about five times as large as the effect of the course, and that this difference is even larger for components such as the Overall Instructor rating and Instructor Enthusiasm. This final conclusion was further strengthened in that the estimates of the teacher and course effects were reasonably consistent over a wide set of different assumptions incorporated into the various models.

References

Bausell, R. B., Schwartz, S., & Purohit, A. An examination of the degree to which various student rating parameters replicate across time. *Journal of Educational Measurement,* 1975, *12,* 273-280.

Campbell, D. T., & Stanley, J. C. *Experimental and quasi-experimental designs for research.* Chicago: Rand-McNally, 1966.

Centra, J. A. The relationship between student and alumni ratings of teachers. *Educational and Psychological Measurement,* 1974, *34,* 321-326.

Centra, J. A. Student ratings of instruction and their relationship to student learning. *American Educational Research Journal,* 1977, *14,* 17-24.

Cook, T. D., & Campbell, D. T. *Quasi-experimentation: Design and analysis issues for field settings.* Chicago: Rand-McNally, 1979.

Costin, F., Greenough, W. T., & Menges, R. J. Student ratings of college teaching: Reliability, validity, and usefulness. *Review of Educational Research,* 1971, *41,* 511-535.

Frey, P. W. Student ratings of teaching: Validity of several rating factors. *Science,* 1973, *182,* 83-85.

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

Gillmore, G. M., Kane, M. T., & Naccarato, R. W. The generalizability of student ratings of instruction: Estimating the teacher and course components. *Journal of Educational Measurement,* 1978, *15,* 1-15.

Hildebrand, M., Wilson, R. C., & Dienst, E. R. *Evaluating university teaching.* Berkeley: University of California, Center for Research and Development in Higher Education, 1971.

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

Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. *American Educational Research Journal,* 1977, *14,* 441-447.

Marsh, H. W. *Students' evaluations of instructional effectiveness: Relationship to student, course, and instructor characteristics.* Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978. (ERIC Document Reproduction No. ED 155 217)

Marsh, H. W. The influence of student, course, and instructor characteristics on evaluations of university teaching. *American Educational Research Journal,* 1980, *17,* 219-237. (a)

Marsh, H. W. Research on students' evaluations of teaching effectiveness. *Instructional Evaluation,* 1980, *4,* 5-13. (b)

Marsh, H. W. Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. *Journal of Educational Psychology,* in press. (a)

Marsh, H. W. SEEQ: A reliable, valid and useful instrument for collecting students' evaluations of university teaching. *British Journal of Educational Psychology,* in press. (b)

Marsh, H. W., & Cooper, T. L. Prior subject interest, students' evaluations, and instructional effectiveness. *Multivariate Behavioral Research,* 1981, *16,* 82-104.

Marsh, H. W., Fleiner, H., & Thomas, C. S. Validity and usefulness of student evaluations of instructional quality. *Journal of Educational Psychology,* 1975, *67,* 833-839.

Marsh, H. W., & Overall, J. U. Long-term stability of students' evaluations: A note on Feldman's "Consistency and variability among college students in rating their teachers and courses." *Research in Higher Education,* 1979, *10,* 139-147.

Marsh, H. W., & Overall, J. U. Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. *Journal of Educational Psychology,* 1980, *70,* 468-475.

Marsh, H. W., & Overall, J. U. Relative influence of course level, course type, and instructor on students' evaluations of instruction. *American Educational Research Journal,* 1981, *18,* 103-112.

Marsh, H. W., Overall, J. U., & Kesler, S. P. Validity of student evaluations of instructional effectiveness: A comparison of faculty self-evaluations and evaluations by their students. *Journal of Educational Psychology,* 1979, *71,* 149-160.

McKeachie, W. J. Correlates of student ratings. In A. L. Sockloff (Ed.), *Proceedings: The first invitational conference on faculty effectiveness as evaluated by students.* Philadelphia: Temple University, Measurement and Research Center, 1973.

McKeachie, W. J. Student ratings of faculty: A reprise. *Academe,* 1979, *65,* 384-397.

Nie, N. H., Hull, C. H., Jenkins, J. G., Steinbrenner, K., & Bent, D. H. *Statistical package for the social sciences.* New York: McGraw-Hill, 1975.

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

Remmers, H. H. Ratings methods in research on teaching. In N. Gage (Ed.), *Handbook on teaching.* Chicago: Rand-McNally, 1963.

Sullivan, A. M., & Skanes, G. R. Validity of student evaluation of teaching and the characteristics of successful instructors. *Journal of Educational Psychology,* 1974, *66,* 584-590.

Winer, B. J. *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill, 1971.

Acknowledgments

Grateful acknowledgments are extended to John Schutz, Joseph Kertes, and Robert Linnell, administrators at the University of Southern California who have supported the student evaluation program. Thanks are also extended to J. U. Overall and Dennis Hocevar, who made helpful suggestions on earlier drafts of this paper.

Author's Address

Send requests for reprints or further information to Herbert W. Marsh, Department of Education, The University of Sydney, Sydney, NSW 2006, Australia.