RESEARCH IN HIGHER EDUCATION
Volume 9, pages 199-242

COURSE CHARACTERISTICS AND COLLEGE STUDENTS' RATINGS OF THEIR TEACHERS: What We Know and What We Don't

Kenneth A. Feldman, *State University of New York at Stony Brook*

From showing in a general way that there is "room" for course context to influence class (average) ratings of instruction, this review proceeds to a search for specific course characteristics that are associated with these ratings. Extant research has centered around five such characteristics: class size, course level, the "electivity" of the course, the particular subject matter of the course, and the time of day that the course is held. Although statistically significant zero-order relationships do not appear in every piece of research located for review, such relationships are more likely to be found than not for the first four of these characteristics. The associations may not be particularly strong, but rather clear-cut patterns do emerge. Of the studies reporting an association between size of class and class ratings, most find it to be inverse, although several studies show a curvilinear (U-shaped) relationship. Teacher (and course) ratings tend to be somewhat higher for upper division courses and elective courses. Compared to other instructors, those teaching humanities, fine arts, and languages tend to receive somewhat higher ratings. The possible reasons for these relationships are many and complex. A precise understanding of the contribution of course characteristics to the ratings of teachers (and the courses themselves) is hampered by two circumstances. Studies in which relevant variables are controlled are far fewer in number than are the studies in which only the zero-order relationships between course characteristics and ratings are considered. More importantly, existing multivariate studies tend to underplay or ignore the exact place of course characteristics in a causal network of variables.

Key words: evaluation of college teachers; course evaluation; student ratings; course characteristics; bias in ratings

An earlier analysis (Feldman, 1977) raised certain issues concerning the degree of consistency among college students in rating their teachers and courses, and attempted to clarify some of the problems involved in trying to resolve these issues. Analysis of consistency in stu-

Address reprint requests to Kenneth A. Feldman, Dept. of Sociology, State University of New York at Stony Brook, Stony Brook, NY 11794.

dent ratings was hardly exhausted thereby, for the primary focus of the earlier analysis was only on interrater consistency within classrooms. As was pointed out, other kinds of consistency must also be analyzed if student ratings are to be used appropriately and interpreted meaningfully. These include consistency of students' ratings over time, intra-student consistency across items of a rating form, consistency between ratings of teachers by their students and by other types of raters (for example, teachers' colleagues), and consistency of ratings across different contexts and conditions. The present review and analysis focuses on one aspect of these matters—namely, the consistency of ratings across different course contexts. The specific concern is with the way in which course characteristics are related to (and may affect) students' ratings.

As in the other analyses of this series (Feldman, 1976a, 1976b, 1977), the purview of the present paper is restricted to studies of students and teachers at colleges and universities in the United States and Canada. The major concern is again with the undergraduate population at these schools, although studies that have samples consisting of graduate students as well as undergraduates have been included when there was no way to separate out the data for the undergraduates alone. Unlike the earlier efforts, the primary unit of analysis in the present review is the classroom as a whole and not the individual student. Thus so-called ecological correlations are the point of interest rather than individual correlations (see Robinson, 1950; Goodman, 1959; Hammond, 1973). The search is for the characteristics of course settings that are associated with the *average* or *aggregate* ratings of the students in the courses. These will be called class ratings.

COMPARING RATINGS OF INSTRUCTORS IN THE SAME AND DIFFERENT COURSES: RELIABILITY AND BEYOND

If class (average) ratings are not reliable measures, then variation in ratings across course settings (even for the same teacher) would not be particularly surprising or informative, since the variability could be due merely to unreliability of measurement. Thus as background to the study of the possible effects of course conditions and contexts, it is useful to know something about the reliability of class ratings. In the earlier analysis of interrater consistency (Feldman, 1977), it was seen that the several different procedures that have been used by researchers to determine the degree of consistency among students in rating their teachers have produced similar results. Although within-class consistency among students in their ratings is moderate at best (as indicated by the size of the estimated reliabilities of individual ratings or

the single rater), this modest within-class agreement is sufficient to produce substantial estimates of reliability—at least in the .70's, and more often in the .80's and .90's—when the ratings of at least 20 to 25 students in the same classroom are averaged together. Moreover, the larger the typical number of students in the classes at a school, the higher the estimated reliabilities of the class ratings that can be expected.

Under the assumption that the student raters constitute a random sample from a population of comparable raters, and if certain other conditions or assumptions are met (Winer, 1962, pp. 129–132), then one interpretation of the estimate of the reliability of class (average) ratings is as follows: if the ratings of the same set of teachers were to be repeated with another random sample of student raters, the correlation between the mean ratings obtained from the two sets of data on the same teachers would be approximately equal to the reliability estimate. The assumption of random sampling of students is problematic, of course, since students in part self-select themselves into courses rather than randomly distributing themselves among courses; moreover, even for each course considered separately, it is usually not possible to point to a specific population from which the students in the course are a truly random sample. Thus the assumption of random sampling is often relaxed by introducing the idea of an unspecified population of students "like those observed." The estimate of reliability is then interpreted as the degree to which the observed average ratings of the classes under consideration would be expected to correlate with the class ratings gained by another set of students "similar" to the ones who were used—in effect, that is, another set of students who might reasonably have taken the various courses, and who in fact may do so in the future (Cornfield and Tukey, 1956; Gillmore, 1973; Guthrie, 1945; Kane, Gillmore, and Crooks, 1976; Peters and Van Voorhis, 1940, chap. 7; but see Stanley, 1961, 1971).

Given this latter interpretation, it is of particular interest to observe that when the two sets of class ratings of teachers who have taught the same course (or the same section of a course) to two separate classes of students are actually correlated, the correlations are generally in the .60's and .70's (Bausell and Magoon, 1972, Appendix M; Bausell, Schwartz, and Purohit, 1975; Derry, 1977; Gillmore, 1973; Gillmore, Kane, and Naccarato, 1977; Hogan, 1973; Pohlmann, 1973; Schwab, 1976; Seiler, Weybright, and Stang, 1977; Spencer, 1969b; and Wilson, 1932).[1] It would seem, then, that consistency or agreement between classes of students who are rating under the same-instructor/same-course circumstance, while substantial, is far from total. In any case, the typical size of the correlation is lower than the aforementioned reli-

ability estimates based on consistency among individual students (or, in some cases, subsets of students) within classrooms. This discrepancy may be due to the following factors: that instructors may not teach the same course in exactly the same way to two different classes, and may in effect display different pedagogical skills and characteristics; that the different sets of students involved may be somewhat different in their characteristics (which may differentially influence their reactions even if the teachers themselves taught their paired courses identically); or that the conditions and circumstances under which the same course is taught may vary; or some combination of these reasons may apply. Moreover, the estimates of reliability based on the degree of interrater consistency among students within their classes may themselves be artifically high (see Feldman, 1977) which could in part be responsible for the discrepancy. Indeed, it could be argued that the somewhat lower correlations found in the same-instructor/same-course condition are the better estimates of reliability.

An obvious question at this point, and the one of greater importance to the present analysis, is what happens to ratings when the instructor is teaching a different course rather than the same one, and when the same course is being taught by a different instructor. Table 1 presents information from the few studies in which same-instructor/same-course correlations are compared with each of the two other relevant sets of correlations—those between ratings of the same instructor teaching two different courses and (when available) those between ratings of two different instructors teaching the same course. The correlation between ratings is highest when the same instructor is teaching the same course and is lower when the same instructor is teaching different courses; for the various samples of the studies in Table 1, the range of correlations of the first set of ratings is .62 to .80, whereas the range of correlations of the second set of ratings is .29 to .54. Correlations between ratings are even smaller when the same course is being taught by different instructors (the third set of ratings); across the studies, correlations range from .04 to .20. It might be added here that, although the generalizability coefficients presented by Gillmore et al. (1977) for equivalent comparison groups (in a study not included in Table 1) are not strictly comparable to correlation coefficients, and although the procedures of their study differ from those used in the studies presented in Table 1, these coefficients do fall in the same ranges as those in the table.

The difference between the first and second sets of correlations (same-instructor/same-course correlation vs. same-instructor/different-course correlation) provides a rough index of the importance of the course context to the ratings while the difference between the first and

TABLE 1. Summary of Results of Studies Showing Correlations of Instructor Ratings for Same and Different Instructors Teaching the Same and Different Courses

Reference	Data	Selected rating item	Same instructor, same course	Same instructor, different course	Different instructor, same course
Bausell, Schwartz, and Purohit (1975)	*Sample 1.* University-wide rating of instructors/courses, Fall, 1969, and Spring, 1970	Overall instructor evaluation	$r = .75$ ($N = 41$)	$r = .46$ ($N = 125$)	$r = .20$ ($N = 73$)
	Sample 2. Rating of instructors/courses, College of Business and Economics, Fall semesters, 1972 and 1973	Instructor's rating compared to others	$r = .80$ ($N = 39$)	$r = .30$ ($N = 62$)	$r = .04$ ($N = 67$)
	Sample 3. Rating of instructors/courses, College of Arts and Science, Spring and Fall, 1973	"Instructor recommendation"	$r = .78$ ($N = 37$)	$r = .29$ ($N = 74$)	$r = .08$ ($N = 90$)
Hogan (1973)	Voluntary university-wide rating of instructors/courses, University of Wisconsin-Green Bay, Fall, 1971, Spring, 1972	5-item global rating scale (overall, summative judgment about the course and instructor)	$r = .67$ ($N = 30$)	$r = .39$ ($N = 45$)	$r = .16$ ($N = 39$)

TABLE 1 *(Continued)*

Reference	Data	Selected rating item	Same instructor, same course	Same instructor, different course	Different instructor, same course
Seiler, Weybright, and Stang (1977)	University-wide rating of instructors, C.U.N.Y., Queens College:				
	Data Set No. 1. Ratings of instructors are from the same semester (either Fall, 1972, or Fall, 1973). Average correlations across different subsamples are shown here.	Overall instructor rating	$\bar{r}$ = .69 (N = 188, 187)	$\bar{r}$ = .54 (N = 195, 133, 180, 129)	—
	Data Set No. 2. Ratings of instructors are from different semesters (Fall, 1972, and Fall, 1973). Average correlations across different subsamples are shown here.	Overall instructor rating	$\bar{r}$ = .62 (N = 183, 123, 117, 99, 126)	$\bar{r}$ = .45 (N = 134, 74, 133, 67)	—

third sets of correlations (same-instructor/same-course correlation vs. different-instructor/same-course correlation) provides a rough index of the importance of the teacher (see Kulik and Kulik, 1974, and Gillmore et al., 1977, for attempts to get a more precise estimate of teacher effects and course effects as well as the teacher-course covariance and interaction effects). As might be expected, the characteristics of the teacher clearly contribute more to the ratings than do those of the course setting.[2] Nevertheless, course setting does seem to make a definite contribution to rating differences, even when the teacher remains the same.

COURSE CHARACTERISTICS AND RATINGS

Essentially shown so far is that there is "room" for the course context (and factors associated with it) to influence class (average) ratings. Doubtlessly, each course has its own unique contribution to ratings. In addition, however, there may be certain ways in which courses generally differ among themselves that are associated with, and possibly affect, these ratings. It is to a review of what these characteristics may be that the analysis now turns. Research in this area has centered around the characteristics of class size, course level, the "electivity" of the course, the time of day that the course is held, and the particular subject matter of the course.

Class Size

College professors have been known to complain that it is difficult to get high ratings for themselves and for the courses they teach when their class enrollments are large; and at least one study (Scott, 1977) has found that instructors who felt that one of their classes was too large for them to present the material of the course in an adequate way did in fact receive somewhat lower ratings compared to their fellow instructors who did not feel this way. This is an important piece of information, for, assuming that the instructors in question did not merely use large class size as an "excuse" for anticipated low ratings, it can be taken as evidence that instructors' feelings about class size may affect their teaching performance and their ratings. (A complete picture, of course, would include consideration of instructors who prefer to teach large classes and may be more effective in them than in small courses; cf. Rohrer, 1957.) Scott's finding alone, however, does not tell us whether actual class size is indeed related to ratings of the teacher (and the course itself). For this, the many empirical studies in which the existence of this particular relationship is investigated must be re-

viewed. Because different types of procedures and statistical analyses have been used in these studies, their results will be summarized separately for each type.

Establishing the Existence and Nature of the Relationship between Class Size and Ratings: Zero-Order Associations. Nearly 30 studies were located in which data on the relationship between the size of class enrollment and class ratings are presented in the form of a product-moment correlation between the two variables. About one third of these studies find essentially no relationship between size and ratings (Colliver, 1972; Delaney and Kojaku, 1977; Gillmore and Naccarato, 1975; Hanke, 1970; Heilman and Armentrout, 1936; Hillery and Yukl, 1971; Jiobu and Pollis, 1971; Murray, n.d.; Overall, Marsh, and Kesler, 1977; and Spencer, 1969a). The rest (roughly two thirds) of these correlational analyses find indications of a negative relationship—the smaller the size of the class, the higher the ratings (Aleamoni and Thomas, 1977; Bausell and Magoon, 1972, Appendix P; Brandenburg, Slinde, and Batista, 1977; Brown, 1976; Centra and Creech, 1976; Cashin and Slawson, 1977a, Table 5; Delaney, 1976; Elmore and Pohlmann, 1976; Gillmore, 1975b; Hidebrand, Wilson, and Dienst, 1971; Linsky and Straus, 1972, 1975; Lunney, 1974; McDaniel and Feldhusen, 1970; Marsh, 1976a, also see Marsh, 1976b, and Overall et al., 1977, Study No. 2; Marsh, 1978; Marsh, Overall, and Thomas, 1976; Perlman, 1973; Pohlmann, 1975; Scott, 1975; and Van Horn, 1968). Across the studies, correlations generally fall in the range of –.10 to just under –.30, although there are some instances of both smaller and larger negative correlations (the larger correlations generally being for rating items about the teacher's encouragement of questions from the students, general class discussion of course material, and the like). Thus the variable of size in these studies usually explains somewhere between 1% and 7% or 8% of the variance in class ratings, depending on the particular rating item and the population under study.

Other studies report results in terms of rating differences between classes that have been divided into the two categories of "large" and "small." Actually, "larger" and "smaller" are the better designations, since the so-called large category in some of these studies is not exclusively restricted to classes that are large in any absolute sense. Of these studies, Cooke (1952), Cornwell (1974), Downie (1952), and Miller (1972, Appendix B) all report that ratings were somewhat lower for larger classes than for smaller ones. However, Solomon (1966) found no differences between larger and smaller night classes—although, in this case, larger classes were anything over nine students! Moreover, in an early study comparing students at Purdue University in classes of over and under 50 students, Remmers (1929a) found that the larger

classes, if anything, had slightly more positive ratings than did the smaller classes on some of the items of the Purdue Rating Scale for Instructors. Likewise, Villano (1975) found that larger compared to smaller science and math classes rated higher on one of the three factor scales in his study, the other two scales showing no differences; but note that any class of 21 students or over was designated as large in this study.

Both the use of product-moment correlations and the comparison of two categories of size in effect assume that the relationship between size and ratings is linear. If, however, this assumption is incorrect, the true degree of the association between the variables is underestimated by these procedures. If the actual relationship between size and ratings is essentially curvilinear, product-moment correlations and two-category comparisons would be expected to show only relatively weak correlations or none at all. The possibility that the actual relationship in some of the studies reviewed to this point may indeed have been non-linear is raised by the results of studies by Delaney and Kojaku (1977), Marsh (1976b) (also see Overall et al., 1977), Pohlmann (1975), and Wood, Linsky, and Straus (1974). In these studies the investigators used a polynomial trend analysis to see whether a second-degree (or parabolic) curve fitted the data better than did a straight line (see Blalock, 1974, pp. 459–462, for a brief explanation of polynomial regression). In all four analyses, it did. This result is most clearly detailed in Delaney and Kojaku (1977) and in Marsh (1976b), where, although a simple (negative) linear trend was able to describe to some degree the relationship between size and ratings, there was a statistically significant increment in rating variance accounted for by the quadratic component. All four studies found a negative curvilinear relationship between size and student rating—that is, a U-shaped curve—whereby both relatively smaller and relatively larger classes tended to receive higher ratings than did the medium-sized classes.

Another set of studies helps in determining the exact form of the relationship between size and ratings, albeit in a less precise way. These studies compare the ratings made by classes that have been divided into three or more categories of size. Of these studies, four found that size had little or no relationship to class ratings (Aleamoni and Graham, 1974; Grant, 1971; Riley, Ryan, and Lifshitz, 1950; and Weerts and Whitney, 1975a). The remainder of them did find a relationship between the two variables, but its exact nature varied. Supporting the results of the majority of the simple correlational studies, some of the studies using three or more categories of size show a generally (if not always perfectly) linear decline in ratings as courses get larger (Bausell and Magoon, 1972, Appendix P; Cashin and Slawson,

1977a, Table 4; Crittenden, Norr, and LeBailly, 1975; Francis, 1976; Kohlan, 1973; and Perry and Baumann, 1973). In the rest of the studies, the relationship is curvilinear. Congruent with the studies using polynomial trend analysis, the results of studies by Centra and Creech (1976), Clark and Keller (1954), and Gage (1961) suggest a U-shaped relationship; and this particular curvilinear relationship was also the predominant pattern (although not the only one) across various rating items in a study by Haslett (1976). Four other studies, however, found that the highest rated classes tended to be the medium sized classes, thus producing an inverted U-shaped curve (Kapel, 1974; Lovell and Haner, 1955; Nichols, 1967; and Starrack, 1934).

Considering all studies reviewed to this point,[3] it can be seen that it is much more likely for a researcher to find an association between class size and ratings than not to find one. When a relationship is found, it is most likely to be an inverse association (the larger the class size, the lower the class rating). Also, the possibility clearly exists of a U-shaped curvilinear relationship between class size and ratings (with the lowest ratings going to medium sized classes). Although this particular pattern has not been found as often as the inverse relationship, the extent of its existence may be somewhat underestimated by the analytic and statistical procedures generally used by the studies in the area. An inverted U-shaped relationship between size and ratings is not a particularly likely possibility (although this pattern has occurred in a few studies), while a positive association between the two variables is not likely at all.

Exploring the Meaning of the Association between Class Size and Class Ratings: Multivariate Analysis. As has been seen, certain patterns of findings predominate in studies relating class size to class ratings, although results are far from uniform. Some of the variation in the findings may reflect differences in research procedures and data analysis. Thus inconsistency in results across studies may be due in part to differences in the degree to which the sample of classes are representative of the classes at the school, the statistical techniques used to establish relationships, the number of categories of class size (in studies where differently sized classes are classified into a smaller number of categories), the particular cutting points used to generate these categories, and the like. However, it seems unlikely that such procedural differences could completely account for the differences across the studies. In short, there may exist actual differences among colleges that affect the way in which size is related to ratings. If so, research is needed to establish more clearly the conditions under which one rather than another relationship between class size and ratings can be expected.

Because of the many empirical instances of a "simple" inverse size-

rating relationship, the alternate finding in several studies that evaluation of the teacher (and the course) did not continue to become less favorable beyond a certain point as class enrollment increased but at some point actually became increasingly favorable as courses got larger is especially intriguing. Assuming that ratings are valid indicators of teaching effectiveness, various reasons can be suggested for this particular pattern (see, especially, Centra and Creech, 1976; Linsky and Straus, 1972; Overall et al., 1977; and Wood et al., 1974). It may be that some colleges, or departments within colleges, make available increased resources for particularly large courses and select instructors to teach these courses on the basis of their expressed interest and demonstrated success in teaching large courses. Or perhaps instructors in large courses at some universities (or under certain conditions within universities) feel an increased challenge in teaching classes of such size, thus carefully tailoring their teaching methods to the size of the course and increasing their own preparation for the course. It is even possible in some cases that highly rated instructors in part "cause" the size of the class, in that instructors with prominent reputations for teaching effectiveness may draw large number of students to their courses.

Since teachers of different general teaching ability and effectiveness may be differentially located in variously sized classes, it is particularly important to know how the same teacher gets rated in settings of varying sizes. Yet there are almost no studies with this information. Research by Bausell and Vinograd (1977) is an important exception. These investigators identified 254 pairs of courses at the University of Delaware, each pair of which was taught by the same instructor. One of the instructor's two courses was randomly selected as the "criterion" course for the study, with the instructor's rating in the course considered to be the criterion rating. The instructor's other course was then designated as the "predictor" course, with the instructor's rating in this course taken as one of the predictors of the criterion rating. Other predictors in the study (course size, course level, instructional method used, and whether or not the course was required) were defined in terms of the relationship of characteristics of the predictor course to its paired (criterion) course. For example, since the enrollment of the predictor course was either larger than, smaller than, or equal to the enrollment of the same instructor's criterion course, a ratio was computed between the two enrollments (subtracting the enrollment of the criterion course from that of the predictor course and dividing by the larger enrollment). As would be expected, the strongest predictor of an instructor's class rating in the criterion course was the rating in the instructor's other course, accounting as it did for 31% of the crite-

rion variance ($r = .56$). Of importance here is that even after controlling for instructor's rating on this predictor course, the ratio of the enrollment differences between the instructor's two courses was still related (inversely) to the criterion ratings, predicting an additional 10% of the criterion variance.

It could be argued that the rating differences found by Bausell and Vinograd were not necessarily due to differences in size alone, since pairing courses by teacher does not automatically pair courses that are identical (that is, courses with the same course number or course title). Therefore, from their sample of 254 paired courses these investigators also identified those course pairs that differed with respect to class size (in this case, the larger's enrollment exceeded the smaller's by at least 50%) but whose course numbers were identical.[4] Of the 26 course pairs thus determined, instructors received higher ratings in the smaller of the two classes in 19 cases. This same tendency was clearly found in a study by Holland (1954) and to some degree in an early study by Remmers, Hadley, and Long (1932), each of these studies also having the matched-teacher, matched-course, varying-size design but with a much smaller number of cases.

No other studies were found that controlled for instructor (yet alone for instructor and the course itself) when analyzing the relationship between size and ratings. However, Wood et al. (1974) did control for variables indicative of the teaching experience of the instructors in their study. The original relationship they found between size and ratings—in this case, a U-shaped curve—generally held when controls were introduced for the instructor's rank, highest degree, and the number of years since receiving his or her highest degree. Likewise, McDaniel and Feldhusen (1970) found not only that there was a zero-order inverse relationship between course size and the rating of the teacher but also that course size was still a significant predictor of teacher rating when various indicators of faculty scholarly production and service were controlled in a multiple regression analysis.

It is important to emphasize at this point that size of a course does not necessarily vary independently from other characteristics of the course. For example, smaller courses at many colleges are probably more likely to be upper division courses than are larger courses. If this or some other covariate of size is itself related to teacher or course ratings it is important to control for the particular covarying course characteristic in the analysis of the relationship between size and ratings. Put a little too simply, this procedure of controlling helps the analyst to determine whether size as such is directly contributing to variation in ratings or whether size is associated with ratings indirectly, or even coincidentally, due to its association with some other course feature that

may be determining rating differences. As it turns out, current evidence suggests that the relationship between size and ratings tends to hold (if not always consistently so across all rating items and conditions) when two other course characteristics, course level and/or the subject matter of the course, are controlled or in some way taken into account (Lunney, 1974; Marsh, 1976b; Perlman, 1973; and Perry and Baumann, 1973).

Not only might different sized classes vary in other of their course features and in the characteristics of their teachers, they might also vary in terms of the aggregate characteristics of their students—such as the proportion of each sex in the course, mean grade-point average of the students in the class, students' knowledgeability in the subject matter of the course, students' general values and attitudes, and so forth. Only one study (Cashin and Slawson, 1977a, 1977b) was located that controlled for one or more of such student characteristics alone, although a number of studies (to be reviewed shortly) have included student characteristics among a battery of control variables, the others being teacher and/or course characteristics. Pooling data gathered from a large number of colleges and universities, Cashin and Slawson found a clear, inverse association between size of class and various rating items (twenty items about specific aspects of the teacher's pedagogical procedures, ten items about the students' perceived progress in specific areas, and two global items). As part of their study, these investigators divided the thousands of classes available for analysis into five ordered categories of a variable they call either "students' level of motivation" or "class motivation," as determined by the average response of students in a class to a question asking how much the student (near the end of the semester) agreed with the statement that "I had a strong desire to take this course." In a series of tables, data on the relationship between size of the class and each of the rating items is given for each of the five levels of class motivation. By surveying the results of these tables, a relatively clear-cut pattern can be seen, although the investigators do not point it out: for rating items where the zero-order inverse relationship between size and the rating is relatively strong, the inverse relationship tends to remain within each of the five levels of motivation but at a generally weaker degree of strength; for items where the overall inverse relationship is weak to begin with, the relationship tends to disappear altogether within the levels.

Although the pattern of results is rather clear-cut, its meaning is not. The conceptual-empirical referent of "students' level of motivation" or "class motivation" is ambiguous, as is its exact location in a causal network of variables. On the one hand, the question put to students, with its phrase "had a strong desire to take this course," may be meas-

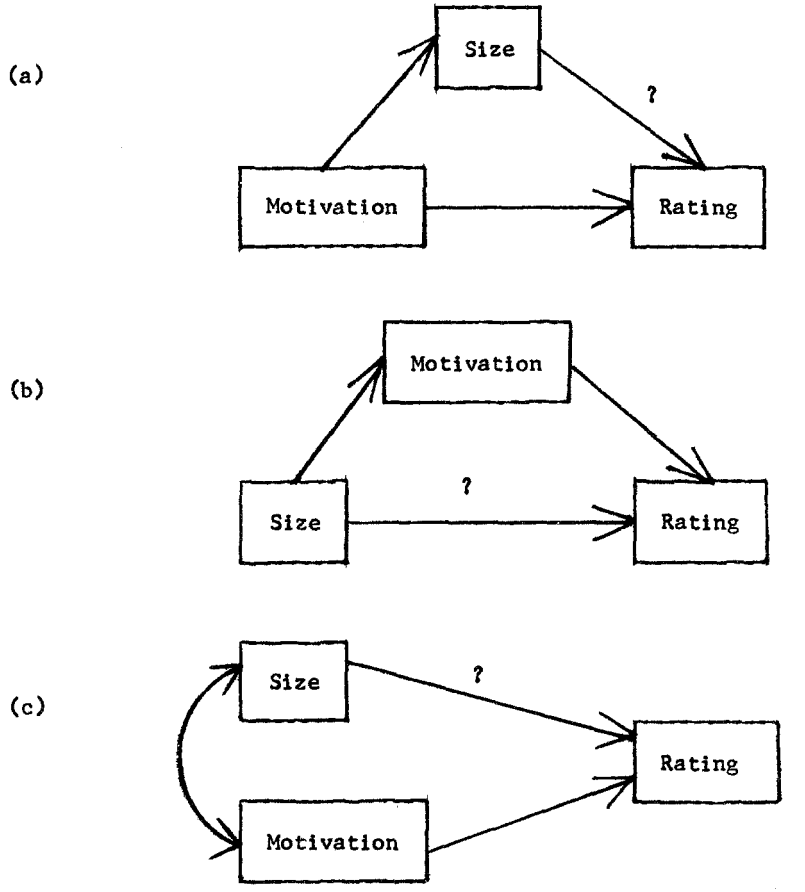

FIGURE 1. Three possible causal models of the relationships among the variables of class motivation, class size, and class rating of instructor. (The question mark indicates the relationship "at risk" under statistical controls.)

uring a pre-course characteristic of students; students may have felt that they were being asked about their desire to take the course even before they enrolled in it. Supposing this so, it might be argued that the weakening or disappearance of the zero-order relationship should lead to discarding it (in part or totally) as causally "spurious." To the degree that the pre-course motivation of students determines the kinds of courses they take (and, as a consequence, the size of the course) and influences the ratings given to the teacher, to that degree the relationship between size and ratings is "explained" by this causally prior variable of students' level of motivation (see Figure 1a). The problem with this reasoning in this particular instance, however, is that the rela-

tionship produced spuriously should be positive rather than negative (the empirical case), since it would be expected that level of motivation as a pre-course variable would be associated positively with the size of the course, and the data themselves show level of motivation to be associated positively with ratings.

On the other hand, the question may be taken by students as asking them about their general reactions or motivations during the course, reactions or motivations themselves affected by the size of the course. If so, it could be argued that the original relationship is mediated by this intervening variable of class motivation; that is, the size of the course affects class motivation, which in turn affects class ratings (see Figure 1b). Of course, the original relationship may neither be "explained" (the first case) nor "interpreted" (the second case) by students' level of motivation.[5] Another possibility is that, for whatever reason, class motivation and size are seen to be so inextricably bound up with one another that the investigator is unable (or unwilling) to assign a particular causal direction between the two (see Figure 1c). Controlling for either variable would weaken the effect of the other on the dependent variable. As still another possibility, it is even conceivable that students may in part react to a question (near the end of a semester) about their desire to take the course as though they were being asked for their overall assessment of the course experience. In short, to some degree the students in the study may have responded to the question as though it, too, were a rating item. If so, controlling for this variable would obviously weaken the original relationship between size and ratings, for no other reason than that one has, in effect, partially controlled on the dependent variable itself.

The previous few paragraphs have been concerned with the significance (to the relationship between class size and ratings) of teacher characteristics alone, or course characteristics alone, or student characteristics alone. A number of studies have controlled or taken into account some combination of characteristics from two or all three of these sets of characteristics. The most general, although somewhat oversimplifying, statement that can be made about the results of these studies is that an association between size and ratings is still found after taking into account one or another combination of these various types of characteristics (Brandenburg et al., 1977; Brown, 1976; Centra and Creech, 1976; Crittenden et al., 1975; Delaney, 1976; Delaney and Kojaku, 1977; Grant, 1971; Haslett, 1976; Kohlan, 1973; Villano, 1975; and Villano, Rosenstock, and Estes, 1974; but see Aleamoni and Graham, 1974, and Marsh, 1978).[6] It should be noted that this statement is not always true for all rating items used in a study. For example, Delaney (1976) reports a statistically significant beta coefficient for class size in a multiple regression analysis when the dependent variable

was the class rating on a scale assessing the teacher-student relationship, but a statistically insignificant beta coefficient when the dependent variable was the class rating on a scale assessing aspects of the organization and objectives of the course.

To say that an association between class size and ratings is still found under one or another combination of controls is not to say that certain covarying aspects of differently sized courses may not, in part, be producing the rating differences. Variations in these associated characteristics most likely do explain some of the variation in the ratings although the degree to which they do cannot be clearly or precisely determined from these studies (just as the degree to which the initial association between size and ratings has been reduced in strength by the introduction of controls cannot be determined), due to the nature of either the data analysis of the study or the information reported in it.

Nor should the proffered generalization be taken as saying that the relationship between size and ratings is unlikely to differ in strength (and occasionally even in direction) within different categories or levels of a control variable, or that there is no interaction between size and other characteristics. To the contrary, specification and interaction effects have been reported. Thus Crittenden et al. (1975) and Lunney (1974) found that the inverse relationship between class size and ratings in the overall sample of each of their studies is considerably weaker in strength (or even reversed in direction) for upper-level courses and for natural science courses (but see Perlman, 1973, and Perry and Baumann, 1973). Likewise, in studies using one or another type of analysis of variance procedure, interaction effects (for certain rating items) have been found between size and such characteristics as students' year in school (Kohlan, 1973), teaching load of instructor (Centra and Creech, 1976), course level (Aleamoni and Graham, 1974, Villano, 1975, Villano et al., 1974), and rank of instructor (Aleamoni and Graham, 1974, Villano, 1975, Villano et al., 1974; but see Grant, 1971, where this is not the case).

In the present review, one factor has been singled out for separate analysis and discussion in terms of its potential importance to the association between class size and ratings, and that is the typical grade expected by (or actually given to) students in their classes. This variable of grades is a difficult one to classify in terms of the type of characteristic it represents, for variation among classes on this variable may be considered as variation in a student characteristic (say, differential achievement in course work by students in different classes), variation in a course characteristic (say, differential difficulty of different courses), variation in a characteristic of teachers (say, differential grading "generosity" or differential grading practices of different teachers),

or variation in some combination or amalgamation of these sorts of characteristics. At any rate, because expected grade (or actual grade, as its proxy) is known to be associated at low to moderate strength with ratings, at the individual level of analysis as well as the group level (see Feldman, 1976a, 1977), and because it is possible that differences in this variable covary with differences in size of class, it is important to see whether the association between size and ratings can somehow be accounted for in terms of a grades factor.

Indeed, in a study of instructor ratings dealing only with class size and grades, Batista and Brandenburg (1975) found that although expected grade was significantly related to total score on a rating form both when taken by itself as well as when class size had already been taken into account, the reverse was not true. That is, although class size was found to be significantly related to ratings when taken by itself, this was not the case when expected grade was taken into account. However, this particular finding is generally not supported in some other studies that also have a control for grades (usually the average grade expected by students in each class). These studies, all but one of which (Crittenden et al., 1975) use multiple regression techniques to analyze the data, not only control for this grades factor but also for one or more teacher, student, and/or course characteristics. In general, although the grade variable clearly tends to be the strongest predictor of ratings in these studies, size is still a predictor (that is, still has a statistically significant beta coefficient) even when controlling for grades and teacher, course, and/or student characteristics (Brandenburg et al., 1977; Brown, 1976; Crittenden et al., 1975; Delaney, 1976; and Delaney and Kojaku, 1977; but see Cornwell, 1974; Jiobu and Pollis, 1971; Marsh, 1978; and Nichols and Soper, 1972).[7]

The possibility remains—one not really much explored in the extant research—that controlling for a grades factor can be expected at least to weaken the zero-order association between class size and ratings (if not eliminate it as a causally relevant variable at certain colleges or under certain conditions). If future research in this particular area does show grades to be important in this way, one interpretation is that the influence of the size of a class on instructor ratings is partly indirect (if not totally so under specified conditions) through its influence on the expected grade of students in the class. Another interpretation, although not a very plausible one, would be that the relationship between class size and ratings is to one or another degree "spurious," as a consequence of these two variables being causally dependent on the prior variable of expected grade. It may well be true that students are more likely to enroll in certain courses because they expect their grades to be relatively high in them, which tendency, other things equal, would build up the enrollment in these courses, producing a positive associa-

tion between expected grade and course size. Assuming that students on the average have no reason to change their minds about their expected grade once they are in the course, and if expected grade is also positively associated with ratings, a spuriously produced positive association between size and ratings would be expected. But, as has been shown, when an association between size ratings is found, it is almost never a positive one.

Course Level

This section reviews the results of studies that relate variation in students' ratings of teachers (and courses) to variation in the level of the course, either rating differences between upper division and lower division courses or for some finer-grained comparison (freshman-level courses, sophomore-level courses, etc.). Also included in this summary are the few studies that have related average year in school of the students in the class to their ratings. The rationale for this inclusion is that average year in school of students in a class can be considered a rough indicator of course level and that results for the studies using this variable are about the same as for the set of studies considering course level directly.

Many of the studies that were located show that the higher the course level the higher the average student rating of the teacher (and the course itself): Aleamoni (1972a); Aleamoni and Thomas (1977); Brown (1976); Cashin and Slawson (1977a); Clark and Keller (1954); Elmore and Pohlmann (1976); Francis (1976); Gage (1961); Gillmore (1975b); Heilman and Armentrout (1936); Linsky and Straus (1972, 1975); Marsh (1976a, 1978); Pohlmann (1975; also see Pohlmann, 1973); Pritchard (1972); Spencer (1966, 1969a); and Stuit and Ebel (1952). (Data in Remmers, 1929a, Haslett, 1976, and Hildebrand et al., 1971, also show this positive association, although it is not as clear-cut and is less consistent across the rating items used in each of the studies.) It should be noted that although the positive association between course level and ratings is clear and relatively consistent across various rating items in almost all of these studies, it does tend to be quite weak in strength.

A positive association between the two variables under consideration is not universally found. There are studies in which course level (or average class year of the students in the class) is essentially unrelated to class ratings (Bausell and Magoon, 1972, Appendix N; Delaney, 1976; Doyle and Whitely, 1974, also see Whitely, Doyle, and Hopkinson, 1973; Grant, 1971; Jiobu and Pollis, 1971; and Weerts and Whitney, 1975a). Moreover, results in studies by Cope, McMillin, and Richardson (1972), Office of Evaluation (1971), Marsh et al. (1976), and

Villano (1975) are basically inconsistent across rating items, with indications in some of these studies that the higher the course level the lower the ratings on certain items (also see Cohen and Humphreys, n.d.).[8]

When controlling for one or another set of instructor, course, and student characteristics, associations between course level and class ratings are still found in Brandenburg et al. (1977) and Lunney (1974); they are not found in Bausell and Vinograd (1977), Brown (1976), Cashin and Slawson (1977a, Table 5), Marsh (1978), Mirus (1973), Nichols and Soper (1972), and Shingles (1977), although in some of these studies it is not clear whether there was a zero-order association between course level and ratings to begin with. In Haslett (1976) results are mixed when control variables are introduced. There are hints in these various studies that when a relationship is found between course level and ratings it may have less (if at all) to do with differences in course level in any direct way and more to do with differences that may accompany course-level differences, such as differences in the size of course enrollment (Bausell and Vinograd, 1977, Cashin and Slawson, 1977a, Table 5, and Haslett, 1966), in grades given to and expected by students in the class (Brandenburg et al., 1977, Brown, 1976, Marsh, 1978, Mirus, 1973, Nichols and Soper, 1972), in the degree of "electivity" of the course as well as the students' academic motivation and knowledgeability of and general interest in the subject matter of the course (Brandenburg et al., 1977, Cashin and Slawson, 1977a, Table 5, Haslett, 1976, and Marsh, 1978), and in instructor characteristics (Bausell and Vinograd, 1977).[9]

The fact that an association between course level and ratings does not appear under statistical controls does not necessarily mean that the level of the course makes no contribution whatsoever to teacher ratings, nor does it automatically show that the relationship between course level and ratings is "spurious." It could be that the course level is indirectly contributing to ratings through its association with certain of the aforementioned variables introduced into the analysis, some of which may be intervening variables. Unfortunately, it is not known whether or not this is the case, since the place of course level in a causal network of factors contributing to variation in class ratings has not in general been considered in the studies.

Elective Courses versus Required Courses

There is some evidence that the teachers of nonrequired or elective courses as well as the courses themselves receive somewhat higher ratings than do their counterparts (required courses and their teachers) (Downie, 1952; Evans, 1960; Gage, 1961; Lovell and Haner, 1955; Mur-

ray, n.d.; but see the nonsupporting evidence in Heilman and Armentrout, 1936, and Nichols, 1967). Likewise, the relationship between the percentage of students taking a course as an elective (that is the average "electivity" of the class for students in it) and the ratings of the teacher and the course is generally positive and of small to moderate strength (Aleamoni and Thomas, 1977; Brandenburg et al., 1977; Landis and Pirro, 1977; Marsh, 1976a; Marsh, 1978; Marsh et al., 1976; Pohlmann, 1972, 1975; Weerts and Whitney, 1975a; but see Elmore and Pohlmann, 1976).[10]

Congruent with the results of these two sets of studies are studies showing a small positive relationship (correlations primarily in the .10's and .20's) between class ratings and the average *intrinsic* interest of students in the course, as indicated, for example, by students' reported interest in the subject matter before enrolling in the course (see Gillmore, 1975b; Gillmore and Naccarato, 1975; Marsh, 1978; Marsh et al., 1976; and Murray, n.d.).[11] Whether this sort of interest in fact somehow accounts for the positive association between elective courses and class ratings cannot be determined from these particular studies.

More generally, whether rating differences between required and elective courses are directly due to the "requiredness" or the "electivity" of the course as such rather than to one or another set of teacher, course, and/or student characteristics covarying with required–elective courses has received only a little study, and results in the few existing pieces of research are inconsistent. Brandenburg et al. (1977) did find that the proportion of students in a class taking the course as an elective was positively related to average ratings even when controlling for the average expected grade of students in the class, class size, the level of the course, and the gender and rank of the instructor. Controlling for similar variables, Marsh (1978) found that differing proportions of students in a class with different reasons for taking the course (requirement for the major, elective in the major, general interest only, etc.) was still a predictor of various rating factors, although not a strong one. (Stronger predictors were prior interest of the students in the subject matter, their expected grade, and their perceptions of the workload and difficulty of the course.) Results, however, are mixed in Shingles (1977); and Mirus (1973) found that whether or not a course was required was unrelated to overall ratings of faculty in a business school when controlling on a number of relevant variables. Moreover, the results of the carefully designed study by Bausell and Vinograd (1977), as described in some detail in the present analysis in the section on class size, suggest that differences in ratings may be directly accounted for by differences among teachers themselves in conjunction with differences in the size of courses rather than by the "electivity" of the

course (or the course level and the mode of instruction used in the course, the other variables in this study).

When the Class Meets

It might be thought that students' general preference for some class times rather than others might "spill over" into their ratings of the courses and the instructors themselves. Little support for this notion exists. No differences in ratings of courses and teachers among classes grouped by their meeting time were found in studies by Christensen and Bourgeois (1974), Cornwell (1974), Gillmore and Naccarato (1975), Lunney (1974), Mirus (1973), Murray (n.d.), and Overturf and Price (1966). By contrast, Aleamoni (1972b), Clark and Keller (1954), Nichols (1967), and Nichols and Soper (1972) have reported (usually) slight differences among ratings, but the pattern of results are not consistent across these four studies.

Academic Field and Subject Matter

What connection, if any, exists between the subject matter being taught and the rating of the instructor who is doing the teaching? As an initial step in answering this question, the results of studies comparing student evaluations of instructors of different subject matter are summarized in Table 2. Each of the studies on which this table is based has information about the class rating of instructors in at least three different academic departments or divisions, from which a rank ordering of these academic fields can be derived with respect to the ratings of the teachers in them (see note to Table 2).[12] Because the number of academic fields varies from study to study, a direct comparison of these ranks would be problematic. For instructors of a particular academic field to rank in third place, say, when fifteen fields are involved is hardly equivalent to ranking third when only four fields are involved. Therefore, in order to increase comparability among the studies, the ranks in each study have been standardized by dividing each of them by the number of academic departments or divisions in the study; the smaller the resulting fraction the higher the standardized ranking of the department or division with respect to favorability of instructors' ratings.[13]

Table 2 shows the frequency with which the standardized rankings of various academic fields fall into the highest third of the rankings (i.e., standardized ranks from .00 to .33), the medium third (.32–.67), and the lowest third (.68–1.00). As can be seen from this table, the academic

TABLE 2. Summary of the Rankings in Eleven Studies of Academic Areas with Respect to Class Ratings of Instructors[a]

Academic area as given in the original reference	High rank	Medium rank	Low rank
Classics (8)	/		
Philosophy (6)(8)	//		
Drama (2)	/		
Art (2)(8)	//		
Fine and Industrial Arts (7)			/
Music (2)(6)(7)	//		/
Humanities (1)(3)(4)(5)(9)(10)	/////	/	
European Languages (8)	/		
Other [than European] Languages (8)	/		
French (2)(6)	/	/	
Spanish (2)(6)	/	/	
German (2)(6)	/	/	
English (2)(6)(8)		//	/
Language and Linguistics (8)		/	
Literature and Languages (7)		/	
English, Humanities, and Languages (11)		/	
History (2)(8)	//		
Anthropology (6)(8)	/	/	
Political Science (6)(8)		//	
Sociology (2)(6)(8)		//	/
Psychology (2)(6)(8)		/	//
Economics (6)(8)			//
Social Science(s) (1)(3)(4)(5)(10)(11)		////	//
Social Studies (7)(9)		/	/
Education (2)(3)(7)(8)(10)	//	//	/
Education and Applied Areas (4)	/		
Secretarial Studies (2)	/		
Business (Administration) (2)(3)(8)(10)(11)		///	//
Home Economics (2)(3)		//	
Health and Physical Education (Women) (7)		/	
Health and Physical Education (Men) (7)			/
Physical Education and Psychology (11)			/
Agriculture (2)			/
"Professional" (Agriculture and Natural Resources, Architecture and Environmental Design, Communications, Computer and Information Sciences, Health Professions) (3)	/		

TABLE 2 ***(Continued)***

Academic area as given in the original reference	High rank	Medium rank	Low rank
Botany (8)	/		
Genetics (8)	/		
Physiology (8)		/	
Zoology (8)		/	
Biology and Microbiology (8)		/	
Biological Sciences (4)			/
Astronomy (8)	/		
Geology (6)(8)	/		/
Geography (2)(6)(8)	/		//
Meteorology (8)			/
Mathematics (2)(3)(6)(8)	/	/	//
Chemistry (2)(6)(8)		/	//
Biochemistry (8)			/
Engineering (2)(3)(8)		/	//
Mechanical Engineering (6)			/
Physics (6)(8)			//
Physical Science (8)			/
Natural Science(s) (1)(5)(8)	/	/	/
Science(s) (3)(7)		/	/
Science and Mathematics (9)(10)(11)	/		//
Mathematics, Physical Sciences, Engineering (4)			/

[a]This table shows the frequency with which the standardized rank (rank divided by the number of categories of academic fields in the study) of each category of academic fields falls into high, medium, and low thirds (see text). The categories of academic fields are exactly those used in the studies cited. The academic fields have been combined only when this was done in the original study itself.

Numbers in parentheses refer to the study or studies for which the results in the row obtain. Following are the code numbers of these studies and information about the particular class rating in the study that was ranked and standardized for the present analysis:

(1) Apt (1966): Mean ratings of instructors at the University of Pittsburgh (on an overall rating item), as divided into three academic divisions (see Table 15, p. 85), have been ranked from 1 to 3.

(2) Bausell and Magoon (1972): Grand mean ratings, of two sets of mean ratings for different academic years, of instructors at the University of Delaware (on an item of overall evaluation of instructor), as divided into 19 academic departments (see Table 28, p. 160) have been ranked from 1 to 19.

(3) Cashin and Slawson (1977b): Mean ratings of instructors at about 120 colleges and universities (on an item asking degree to which student "would like to take another course from this instructor"), as divided into nine academic fields (see Table H, Item No. 37), have been ranked from 1 to 9.

(4) Centra (1972): Mean ratings of teachers at five colleges (on an overall) rating item on teaching effectiveness), as divided into five subject areas (see Appendix H, p. 75), have been ranked from 1 to 5.

TABLE 2 *(Continued)*

(5) Centra and Creech (1976): Mean ratings of teachers at a number of colleges and universities (on an overall rating item on teacher effectiveness), as divided into three subject areas (see p. 40), have been ranked from 1 to 3.
(6) Cope, McMillin, and Richardson (1972): Ratings of teachers at the University of Washington on an eight-item Scale for Student Assessment of Teaching, as divided into 19 departments; for the present analysis the two sets of departmental rankings (based on two ways of getting average teacher rating, see Table 11, p. 23) have been averaged to produce a new ranking from 1 to 19.
(7) Heilman and Armentrout (1936): Mean ratings of teachers at Colorado State College of Education (on the ten-item Purdue Rating Scale for Instructors), as divided into 9 divisions (see Table 3, p. 208), have been ranked from 1 to 10 (the division of Health and Physical Education was subdivided into men and women).
(8) Linsky and Straus (1972): Ratings of teachers at 16 colleges and universities (on ratings of classroom performance derived from published course critiques, and expressed as average z scores), as divided into 31 academic disciplines (see Table 1, p. 13) have been ranked from 1 to 31.
(9) Lunney (1974): Mean ratings of instructors at Centre College of Kentucky (across 10 items of a ratings form), as divided into three academic divisions (see Table 4), have been ranked from 1 to 3.
(10) Pohlmann (1976): Mean ratings of teachers at Southern Illinois University, Carbondale (across 21 items of a rating form), converted to standardized T scores, as divided into 5 disciplines (see Table 1, p. 341), have been ranked from 1 to 5.
(11) Walker (1968): Mean ratings of teachers at Lee Junior College (on the Purdue Rating Scale of Instruction), as divided into five subject matter areas (see Table 4, p. 35, and Table 22, p. 73) have been ranked from 1 to 5.

areas encompassed by English, humanities, arts, and languages fall mostly in the high and medium ranks with respect to class ratings of teachers (with certain specific fields falling predominantly in the high third alone). The social sciences tend to be in the medium or low third of rankings; this is especially so for political science, sociology, psychology and economics. With the exception of certain subareas of the biological sciences (which are in the higher two thirds of the rankings), the other fields of science, as well as mathematics and engineering, are also usually in the lower two thirds of the rankings, although, in this case, more frequently in the lower than the medium third.

It must be remembered that Table 2 gives information about the placement of academic fields relative to one another; it does not show that teachers in certain fields are more likely than others to receive absolutely high (or low) ratings. Moreover, quite apart from the fact that the generalizations that have been drawn from the table are based on only eleven studies, caution must be exercised in arriving at conclusions about the existence and nature of the connection between academic fields and ratings of their instructors. Even if the general results of these studies were to be duplicated in other studies, this fact alone would tell us nothing about whether differences in subject matter as such are directly affecting students' ratings of their teachers. Nor would we learn about the significance to ratings of differences in the

attributes of instructors who teach different subject matter, differences among course characteristics that may vary with subject matter, and differences in the kinds of students who prefer (and thus enroll in) courses in one academic area rather than another. Differences in ratings of teachers in different academic departments and divisions could merely be signalling the effects of say, differences in the size of the course, the "requiredness" of the course, student motivation to take the course, students' expected grades in the course, or the gender, rank, teaching experience or teaching load of the instructor. This seems not to be true, however. At least in the studies that have controlled or in some way taken into account one or another set of these particular characteristics, differences among the ratings of teachers in varying academic fields still generally appear (Bausell and Magoon, 1972, Appendix O; Centra, 1972; Centra and Creech, 1976; Delaney, 1976; Hanke, 1970; Hoyt and Spangler, 1976; Linsky and Straus, 1973; and Lunney, 1974).

Still, the possibility that course, teacher, and/or student characteristics are important to the association between academic fields and ratings should not be ruled out too quickly. An association does not inevitably appear between field and each and every rating item in the studies that have controlled for one or more variables. More importantly, some of the associations that have been found under (statistically) controlled conditions are very weak in strength. Indeed, the weakness of association may indicate that the size of the partial relationship is generally smaller than the original zero-order association, although unfortunately this information cannot generally be gained from the research reports one way or the other. Furthermore, the variables that have been introduced into the analysis in these studies hardly exhaust the possibilities. Many of the studies have only controlled on one or two variables. Concomitantly, most of the specific variables that have been mentioned have been controlled in only one or two studies. For example, the average grade expected by the student in the course, which would seem to be an important variable to consider, has been controlled only in the study by Delaney (1976). None of the studies controlled for the proportion of men or women in the class. This may be an important factor to be brought into the analysis, for not only do the fields that were found in the present analysis to have somewhat higher teacher ratings tend to be the same ones in which women are proportionately overrepresented (see, for example, Davis, 1965), but also there is some evidence that proportion of women in a class may have a small, positive relationship with teacher ratings at certain universities (Aleamoni and Thomas, 1977; Elmore and Pohlmann, 1976; and Pohlmann, 1973, 1975; but see Delaney, 1976, and Jiobu and Pollis, 1971).

So far the analysis of this section has been of research on the ratings of teachers in widely varying disciplines. Since in most modern universities and colleges it is highly unusual for the same instructor to be teaching courses in disparate academic areas (say a course in introductory chemistry and one in introductory sociology), it is of especial importance to know if the content or subject area of a course *within* the more delimited area of a particular department or division is associated with teacher ratings. Here only a few studies were located—three for the divisions (or schools) of business administration, one for sociology departments, and one for chemistry departments. Linsky and Straus (1973), pooling data across 16 schools, found that three types of sociology courses (theory, statistics, and social psychology) generally received higher ratings than did other types (introductory sociology, methods, social problems, and others). Cornwell (1974) reports a mixture of statistically insignificant and statistically significant-but-weak differences among various kinds of chemistry courses and subject areas with respect to ratings of the instructors. Schwab (1976) found moderate differences when comparing instructor ratings for different subject areas of business courses (see especially Table 24, p. 60, of his report). Controlling for the average grade given by the instructor, Bassin (1974) found that instructors teaching qualitatively oriented business courses compared to those teaching quantitatively oriented courses received higher ratings on the quality of their lectures but not on such matters as their consideration for students or the quality of their exams. Finally, Mirus (1973) found no differences in overall teacher rating between instructors of quantitatively oriented and qualitatively oriented business courses, at least when controlling for such factors as whether or not the course was required, proportion of enrolled students attending class, level of the course, the time the course met, and the average grade expected by the students in the class.[14]

Throughout this section, as in other sections, the importance of simultaneously controlling on relevant variables has been emphasized. To be kept in mind is that the purpose of controlling is not merely to see if the zero-order association between ratings and academic field or subject matter (or any other course characteristic) weakens or disappears, but to find out how the introduction of particular variables elaborates the original relationship. The quest is to understand more fully the nature of the connection between variables and the *pattern* of influences at work. As a simple example, one important question is whether a particular control variable "explains" away the original association or whether it "interprets" the relationship by providing an intervening mechanism. Questions such as this one (not to mention more complex ones) have yet to be systematically addressed in the research and anal-

ysis of the possible effects of subject area on teacher ratings.

DISCUSSION AND SUGGESTIONS

Considering the substantial number of studies in which course characteristics are variables, surprisingly little is known in any depth about the exact role that such characteristics play with respect to the ratings given by a class of students to its teacher (and to the course itself). Lack of base-line information is not the problem. Studies are sufficiently plentiful and informative to point to some generalizations about the zero-order relationships between certain course characteristics and ratings. Although statistically significant relationships have not inevitably appeared in every piece of research located for this review, at least for four of the five characteristics under consideration such relationships are more likely to be found than not. The associations may not be particularly strong, but rather clear-cut patterns do emerge. Thus, as the clearly primary pattern across studies, the larger the size of the class the lower the rating given to the teacher as well as to the course itself; a secondary pattern is also evident, for several studies found a U-shaped relationship between class size and rating. Teacher (and course) ratings tend to be somewhat higher for upper division courses and for elective courses than for lower division courses and required courses. In terms of the broad divisions of academic endeavors, teachers of courses in the humanities, fine arts and languages tend to receive somewhat higher ratings than do teachers of social science or of physical science, mathematics and engineering.

Because any particular course characteristic may covary with other course characteristics, as well as with characteristics of the teacher and of the students in the class as a whole, these other variables must be considered. Studies in which one or more such (control) variables have been introduced into the analysis of the relationship between the course characteristic of interest and class rating do indeed exist, although they are far fewer in number than are studies considering only the zero-order association between the two variables. One tentative generalization to be made from these studies is that, at least until further research shows otherwise, the association between ratings and the size of the course, and that between ratings and the course's subject matter, appear to be more resistant to "disappearing" when control variables are introduced into the analysis than are the associations between ratings and the other course characteristics under consideration in the present analysis. The original associations between course characteristics and ratings may well weaken, rather than disappear, when controls are introduced, although it is not always clear in the extant

studies whether they do, and, if so, by what degree. Furthermore, the number and nature of the control variables differ from study to study, making it difficult to ascertain the exact contribution of course characteristics to ratings.

Whether or not particular course characteristics are related to one another, they may still statistically combine and statistically interact in explaining variation in class ratings of teachers. Only a little work has been accomplished in detailing which course characteristics in fact do so, and to what effect, so not much in the way of generalizations can be made. When course characteristics are known to be associated with each other, additional considerations enter, the full details of which also await future research. It may turn out that some of the associated course characteristics are best viewed as what are called "correlated causes" of teacher ratings (see Figure 1c, substituting a course characteristic such as class level for the variable labeled motivation). By contrast, some of the course characteristics may themselves be causally dependent on still other course characteristics (in which case Figure 1b would apply, again with appropriate substitution of a course characteristic for the motivational variable), so that the effect on teacher ratings is by means of a causal chain.

As either student or teacher characteristics are introduced into the analysis, more attention needs to be paid to the nature of these variables than has been the case. For instance, it is important to distinguish between the following two kinds of student attributes: (1) those that are brought to particular types of courses (and in some cases may be responsible in part for why the courses were chosen in the first place) but are themselves mostly, if not totally, uninfluenced by the course or its teacher (for example, the gender, motivation for college study, intelligence, and overall grade-point average of the student); and (2) those that are themselves influenced by the teacher and nature of the course (say, motivation to achieve in the particular class). Likewise, the following question can be asked about variables representing teacher characteristics: Are they pre-existing characteristics brought to the course and perhaps even responsible for the teacher picking or being assigned to the course in the first place (for example, teachers of higher rank in the school may be more likely to teach higher level courses or courses of smaller size) or have they themselves been influenced by the nature of the course (for example, the size of the course may influence the teacher's disposition and practices concerning grades).

In sum, a precise understanding of the contribution of course characteristics to the ratings of teachers is hampered by the fact that the multivariate studies that have been done in the area tend to underplay or

ignore the nature of the relationships among the control variables themselves as well as the way in which each of them is related to the particular course characteristic under consideration and to the ratings. As techniques of analysis neither partial association nor regression analysis alone is explicitly informative about these matters, although certain assumptions about them will have been made, if only implicitly, in order for calculations to have proceeded. Indeed, if the underlying anatomy that has been assumed for the system of variables under consideration—explicitly or implicitly, knowingly or otherwise—is ambiguous or incorrect, the results of the analysis will be either basically uninterpretable or misleading (cf. Duncan, 1970, and Hirschi and Selvin, 1967, chap. 9). It is exactly at this point that the usefulness of applying a technique such as path analysis to the area under consideration becomes evident. One of the virtues of this method is that, in order to apply it, researchers must make explicit the theoretical framework within which they operate; this is so because an explicit commitment to a particular (causal) structure of the variables must be made before calculations can be done.[15]

This matter may be approached from a slightly different direction. A definite tendency can be noted in existing multivariate studies to consider a given course characteristic as unimportant and consequently ignorable if its association with class ratings disappears or greatly weakens when controlling for teacher, student, and (other) course characteristics. (This same tendency is present when either partial associations or the beta coefficients of multiple regression analyses are found to be very small and/or statistically insignificant, but when the original zero-order associations are not given.) Although often only vaguely articulated, the conclusion usually drawn in such cases seems to be that the relationship between ratings and the course characteristic is "merely" due to some other factor(s), as though an instance of a spuriously generated association had been automatically discovered. As noted more than once in the present analysis, however, another possibility is that the variables that have been introduced into the analysis are intervening variables. It is not that the particular course characteristic is unimportant to ratings, but that it can now be seen as indirectly important through its direct influence on the newly "discovered" mediating characteristics. As an example, it may be that a multiple regression analysis shows the typical grade expected by students in a class to have a relatively large beta coefficient, while the size of the class or its subject matter (or some other course characteristic) has a relatively small one. Even though the course characteristic thus has only a small direct effect on ratings, this does not necessarily make it unimportant to ratings. It might be that students' expectations about

their grades (which influence their ratings) are themselves largely determined in a direct way by the particular characteristic of the course (and, for that matter, in an indirect way by its influence on teachers' grading practices). The size of the course, say, or its subject matter, may be partially responsible for the average expected grade of the students in the class, which in turn influences the ratings given.[16] Moreover, still other reasons exist for the weakening of original zero-order relationships and consequently small betas, only some of which imply that a particular variable is unimportant (see Blalock, 1964, 1968).

Put in more general terms, the issue is really one of how best to determine the relative "importance" of particular variables (for instance, course characteristics) for explaining a given phenomenon (for instance, class ratings of the teacher). Although certain correlation and regression techniques are commonly employed, their use can be questioned under certain conditions. In particular, the use of beta coefficients to indicate the relative importance of a variable has been noted to be ambiguous and even misleading when the various "predictor" variables in the multiple regression analysis are themselves interrelated and when their location in a causal network is unexplored (see, for example, Duncan, 1970, Kerlinger and Pedhazur, 1973, chap. 11, and Lewis-Beck, 1974). Again, the use of path analysis can be suggested. Its advantage for assessing the relative importance of variables lies in its provision of a decomposition technique for clarifying the form and strength of relationships. Provided certain necessary conditions are met, a zero-order correlation (say that between a course characteristic and ratings) can be decomposed into its component "effects," consisting of one or more of the following, a value for each of which can be calculated: (1) the direct effect of one variable on the other; (2) the indirect effect of the one on the other (operating through intervening variables); (3) an unanalyzed effect due to the association between exogenous variables, i.e., that part of the correlation due to unanalyzed or predetermined associations (correlated causes); and (4) analyzed or unanalyzed prior effects, i.e., that part of the correlation due to joint dependence on either common or correlated causes (spuriousness) (see especially Lewis-Beck, 1974; also see Alwin and Hauser, 1975, Duncan, 1971, Addenda, and Finney, 1972).

An important issue remains to be discussed, and that is the question of whether or not course characteristics directly "bias" the ratings made by students of their teachers, if, by this, is meant that certain course features directly and "inappropriately" influence students' judgments about or evaluations of teachers. Statistically significant zero-order associations between course characteristics and ratings give

little, if any, information in this respect, for it could be argued that they indicate only that teachers who in actuality are differentially effective are assigned to or select different kinds of courses. Adding a variety of control variables is not necessarily of help, regardless of the particular kind of multivariate analysis used (including path analysis as well as partial correlation, multiple regression analysis, and the like). Suppose that, whatever the analytic procedure to control the influence of relevant variables, a "direct" effect on ratings of one or another course characteristic is still found, either exclusively so or, more likely, in addition to indirect effects through intervening variables and any other component "effects" (including spuriousness). (From present, albeit incomplete, evidence, this seems most likely the case for size of the course and for its general subject matter, but perhaps may also occur in certain circumstances for the level of the course and its "requiredness.") It is still uncertain whether the fact that rating differences are found after relevant control variables have been taken into account implies that a particular course characteristic has directly biased ratings, for this is not the only explanation left for these differences (even ruling out the possibility that important control variables were mistakenly left out of the analysis).

Consider the stringently controlled situation in which each instructor of a set of instructors teaches exactly the same subject matter to exactly similar classes of students, the only predetermined difference between the two courses for each instructor being class size. If each teacher receives a lower rating in the larger course, is this because students are reacting negatively to the size of the class which then spills over into (hence biasing) their evaluation of the teacher or because the teacher is actually less effective in the larger class than in the smaller one, which is then correctly mirrored in students' ratings? The same question arises for other course characteristics, although for some of them less stringent control conditions may apply. For instance, it would be hard to imagine finding conditions (or even setting them up) whereby the same instructor, across a set of instructors, was teaching courses in different academic disciplines to similar sets of students. In this instance, where instructors, students and course characteristics vary, interest focuses on differences in student ratings of teachers that remain after statistically controlling for differences in the characteristics of the teachers (such as teaching experience, rank, gender, and the like) as well as pertinent student and course differences. Even so, there are still at least two interpretations for any rating differences that are found. Either the course characteristic directly biased students' reactions to the teacher or it created certain conditions that influenced the teachers' behaviors and effectiveness (consequently reflected in the

ratings). Both things may be true, of course; that is, both the students' evaluation of the teacher and the quality of teaching may be directly affected by one or another course characteristic.

It is important to emphasize, then, that whether a course characteristic directly "biases" the class rating of a teacher cannot be answered from knowledge of the correlations between the two variables alone, nor from partial associations between them (no matter how many of the variables typically controlled for are included). It is necessary in addition to measure directly students' feelings about and reactions to the features of the courses themselves, as an initial step in determining how these feelings and reactions may be related to their ratings. Moreover, in assessing teachers' behaviors and effectiveness, in order to find out how they are affected by course characteristics or course conditions, the procedure used to measure them must be independent from the procedure used to gather student evaluations (for example, by using descriptions and ratings of instructors made by outside observers, by assessing students' performance on independently constructed and administered examinations of course material, and the like). In this regard, the degree to which course characteristics bias a class' ratings cannot be divorced from the questions of what it is that students ratings measure and how validly they do so.[17]

It should be obvious by this point that a review of existing research on the relationship between course characteristics and ratings of college teachers raises many more questions than it answers. The association between one or another course characteristic is generally small to (at best) moderate in size, although these characteristics may combine to explain somewhat larger proportions of variance in the ratings of teachers (and courses). When an association between some characteristic of courses and the ratings of their teachers (or the courses themselves) is found, the possible reasons for the relationship are many and complex. It does seem highly unlikely that the "advantage," however slight, of teaching certain types of courses can be explained away as somehow merely spurious. The possible direct and indirect effects of course characteristics can thus only be ignored at some risk of losing a certain degree of comparability of ratings across instructors. In order to take variation in course conditions into account, procedures to adjust faculty ratings (Shingles, 1977) or to establish norms for appropriate comparison groups (Hoyt, Owens, and Grouling, 1973, and Instructor and Course Evaluation System, n.d.) have been suggested. If procedures such as these are not practical, less systematic methods might be used to some avail. For example, consideration of relevant course characteristics could be achieved informally by only comparing the ratings of instructors who are teaching under somewhat similar course

conditions. Thus, one would try to compare ratings of instructors who are teaching roughly similar sized courses, or broadly similar subject matter, or courses of about the same general degree of "requiredness" for students. (At certain schools, it might also be important to consider separately the ratings for instructors of upper-division and lower-division courses.) Such procedures and methods should serve to increase the usefulness of these ratings to students, teachers, and administrators alike.

ACKNOWLEDGMENT

I would like to thank Herbert W. Marsh for his careful reading of a draft of this paper.

REFERENCES

Aleamoni, L. M. The Illinois Course Evaluation Questionnaire: Manual of interpretation (rev.). Research Rep. No. 331. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1972. (a)

Aleamoni, L. M. A review of recent reliability and validity studies on the Illinois Course Evaluation Questionnaire (CEQ). Research Memo. No. 127 (May). Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1972. (b)

Aleamoni, L. M., and Graham, M. H. The relationship between CEQ ratings and instructor's rank, class size, and course level. *Journal of Educational Measurement,* 1974, *11,* 189–202.

Aleamoni, L. M., and Thomas, G. S. Is the instructor's rating of the class related to the class' rating of the instructor? Research Rep. No. 1. Tucson, Ariz.: Office of Instructional Research and Development, University of Arizona, 1977.

Alwin, D. F., and Hauser, R. M. The decomposition of effects in path analysis. *American Sociological Review,* 1975, *40,* 37–47.

Anderson, J. G., and Evans, F. G. Causal models in educational research: Recursive models. *American Educational Research Journal,* 1973, *11,* 29–39.

Apt, M. H. A measurement of college instructor behavior. Unpublished doctoral dissertation, University of Pittsburgh, 1966.

Bassin, W. M. A note on the biases in students' evaluations of instructors. *Journal of Experimental Education,* 1974, *43,* 16–17.

Batista, E., and Brandenburg, D. C. Expected grades, class size, and student ratings of instructors. Research Rep. No. 357. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1975.

Bausell, R. B., and Magoon, J. The validation of student ratings of instruction: An institutional research model. Newark, Del.: College of Education, University of Delaware, 1972.

Bausell, R. B., Schwartz, S., and Purohit, A. An examination of the conditions

under which various student rating parameters replicate across time. *Journal of Educational Measurement,* 1975, *12,* 273–280.

Bausell, R. B., and Vinograd, C. J. Student ratings and various instructional variables from a within instructor perspective. Unpublished manuscript, 1977.

Bejar, I., and Doyle, K. O., Jr. Relationship of curriculum area and course format with student ratings of instruction. *American Educational Research Journal,* in press.

Blalock H. M., Jr. *Causal inferences in nonexperimental research.* Chapel Hill, N. C.: University of North Carolina Press, 1964.

Blalock, H. M., Jr. Theory building and causal inferences. In H. M. Blalock, Jr., and A. B. Blalock (Eds.), *Methodology in Social Research.* New York: McGraw-Hill, 1968.

Blalock, H. M., Jr. *Social statistics* (2nd ed.). New York: McGraw-Hill, 1974.

Brandenburg, D. C., Slinde, J. A., and Batista, E. E. Student ratings of instruction: Validity and normative interpretations. *Research in Higher Education,* 1977, *7,* 67–78.

Brown, D. L. Faculty ratings and student grades: A university-wide multiple regression analysis. *Journal of Educational Psychology,* 1976, *68,* 573–578.

Carney, R. E., and McKeachie, W. J. Personality, sex, subject matter and student ratings. *Psychological Record,* 1966, *16,* 137–144.

Cashin, W. E., and Slawson, H. M. IDEA technical report No. 2: Description of data base, 1976–77. Manhattan, Kan.: Center for Faculty Evaluation and Development in Higher Education, 1977. (a)

Cashin, W. E., and Slawson, H. M. IDEA technical report No. 3: Description of data base, 1977–78. Manhattan, Kan.: Center for Faculty Evaluation and Development in Higher Education, 1977. (b)

Centra, J. A. Two studies on the utility of student ratings for improving teaching: I. The effectiveness of student feedback in modifying college instruction. II. Self-ratings of college teachers: A comparison with student ratings. SIR Rep. No. 2, Princeton, N. J.: Educational Testing Service, 1972.

Centra, J. A., and Creech, F. R. The relationship between student, teacher, and course characteristics and student ratings of teacher effectiveness. PR-76-1. Princeton, N. J.: Educational Testing Service, 1976.

Chermersh, R. Students' rating of their faculty—primary impression or dynamic process? *Sociology of Education,* 1977, *50,* 290–299.

Christensen, L. B., and Bourgeois, A. E. Student ratings of instructional effectiveness. Paper read at the annual meeting of the American Psychological Association, 1974.

Clark, K. E., and Keller, R. J. Student ratings of college teaching. In R. E. Eckert and R. J. Keller (Eds.), *A university looks at its program: The report of the University of Minnesota Bureau of Institutional Research, 1942–1952.* Minneapolis, Minn.: University of Minnesota Press, 1954.

Cohen, J., and Humphreys, L. G. Report on the student evaluation of undergraduate courses, Department of Psychology, University of Illinois, n.d. (Mimeographed)

Cole, S. *The sociological method* (2nd ed.). Chicago: Rand McNally, 1976.

Colliver, J. A. A report on student evaluation of faculty teaching performance

at Sangamon State University. Technical Paper No. 1. Springfield, Ill.: Division of Academic Affairs, Office of the Vice President, Sangamon State University, 1972.

Cook, V., Gillmore, G., Hodgson, T. F., and Tomandl, D. The training and effectiveness of graduate student teaching assistants. Prepared for the Committee on Evaluation and Improvement of Teaching by the Subcommittee on the Effectiveness of Graduate Student Teaching Assistants. Seattle, Wash.: University of Washington, 1975.

Cooke, L. S. An analysis of certain factors which affect student attitudes toward a basic college course, effective living. Unpublished doctoral dissertation, Michigan State College, 1952.

Cope, R. C., McMillin, J. G., and Richardson, J. M. A study of the relationship between quality instruction as perceived by students and research productivity in academic departments. Final Report, Project No. 1-J-010, Grant No. OEC-X-72-0021, U.S. Department of Health, Education and Welfare, Office of Education, 1972.

Cornfield, J., and Tukey, J. W. Average values of mean squares in factorials. *Annals of Mathematical Statistics,* 1956, *27,* 907–949.

Cornwell, C. D. Statistical treatment of data from student teaching evaluation questionnaires. *Journal of Chemical Education,* 1974, *51,* 155–160.

Crittenden, K. S., Norr, J. L., and LeBailly, R. K. Size of university classes and student evaluations of teaching. *Journal of Higher Education,* 1975, *46,* 461–470.

Davis, J. A. *Undergraduate career decisions: Correlates of occupational choice.* Chicago: Aldine, 1965.

Delaney, E. L. The relationships of student ratings of instruction to student, instructor and course characteristics. Paper read at the annual meeting of the American Educational Research Association, 1976.

Delaney, E. L., Jr., and Kojaku, L. K. The influence of teaching experience and instructional development activities on student ratings of instruction obtained by beginning university professors. Paper read at the annual meeting of the American Educational Research Association, 1977.

Derry, J. O. Strengths and vulnerabilities of the CAFETERIA model. Paper read at the annual meeting of the American Educational Research Association, 1977.

Downie, N. M. Student evaluation of faculty. *Journal of Higher Education,* 1952, *23,* 495–496; 503.

Doyle, K. O., Jr., and Whitely, S. E. Student ratings as criteria for effective teaching. *American Educational Research Journal,* 1974, *11,* 259–274.

Duncan, O. D. Partials, partitions, and paths. In E. F. Borgatta and G. W. Bohrnstedt (Eds.), *Sociological Methodology 1970.* San Francisco: Jossey-Bass, 1970.

Duncan, O. D. *Path analysis: Sociological examples.* In H. M. Blalock, Jr. (Ed.), *Causal models in the social sciences.* Chicago: Aldine-Atherton, 1971.

Elmore, P. B., and Pohlmann, J. T. Effect of teacher, student, and class characteristics on the evaluation of college instructors. Technical Rep. 2.1-76. Carbondale, Ill.: Student Affairs Research and Evaluation Center, Southern Illinois University, 1976.

Evans, E. D. Student activism and teaching effectiveness: Survival of the fittest? *Journal of College Student Personnel,* 1969, *10,* 102–108.

Feldman, K. A. Measuring college environments: Some uses of path analysis. *American Educational Research Journal,* 1971, *8,* 51–70. (a)

Feldman, K. A. Using the work of others: Some observations on reviewing and integrating. *Sociology of Education,* 1971, *44,* 86–102. (b)

Feldman, K. A. Grades and college students' evaluations of their courses and teachers. *Research in Higher Education,* 1976, *4,* 69–111. (a)

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

Feldman, K. A. Consistency and variability among college students in rating their teachers and courses: A review and analysis. *Research in Higher Education,* 1977, *6,* 223–274.

Feldman, K. A. Using the work of others: Some observations on reviewing, integrating, and consolidating findings. In R. B. Smith, B. Anderson, and P. Manning (Eds.), *Handbook of Social Science Research Methods.* New York: Irvington, 1979, in press.

Feldman, K. A., and Newcomb, T. M. *The impact of college on students.* San Francisco: Jossey-Bass, 1969.

Finney, J. M. Indirect effects in path analysis. *Sociological Methods and Research,* 1972, *1,* 175–186.

Francis, J. B. Faculty ratings of course evaluation items. *Research in Higher Education,* 1976, *4,* 23–40.

Gage, N. L. The appraisal of college teaching: An analysis of ends and means. *Journal of Higher Education,* 1961, *32,* 17–22.

Gillmore, G. M. Estimates of reliability coefficients for items and subscales of the Illinois Course Evaluation Questionnaire, Research Rep. No. 341. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1973.

Gillmore, G. M. The relationship between graduating senior nominations of valuable and non-valuable courses and end-of-course student ratings. EAC Report 271b. Seattle, Wash.: Educational Assessment Center, University of Washington, 1975. (a)

Gillmore, G. M. Statistical analysis of the data from the first year of use of the Student Rating Forms of the University of Washington Instructional Assessment System. EAC Report 76-9. Seattle, Wash.: Educational Assessment Center, University of Washington, 1975. (b)

Gillmore, G. M., Kane, M. T., and Naccarato, R. W. The teacher and the course as units of analysis in the generalizability of student ratings of instruction. EAC Report 77-9. Seattle, Wash.: Educational Assessment Center, University of Washington, 1977.

Gillmore, G. M., and Naccarato, R. W. The effect of factors outside the instructor's control on student ratings of instruction. Seattle, Wash.: Educational Assessment Center, University of Washington, 1975.

Goldsmid, C. A., Gruber, J. E., and Wilson, E. K. Perceived attributes of superior teachers (PAST): An inquiry into the giving of teacher awards. *American Educational Research Journal,* 1977, *14,* 423–440.

Goodman, L. A. Some alternatives to ecological correlation. *American Journal of Sociology,* 1959, *64,* 610–625.

Grant, C. W. Faculty allocation of effort and student course evaluations. *Journal of Educational Research,* 1971, *64,* 405–410.

Guthrie, E. R. Evaluation of faculty service. *American Association of University Professors Bulletin,* 1945, *31,* 255–262.

Guthrie, E. R. The evaluation of teaching: A progress report. Seattle, Wash.: University of Washington, 1954.

Hammond, J. L. Two sources of error in ecological correlations. *American Sociological Review,* 1973, *38,* 764–777.

Hanke, J. E. Teacher and student perceptions as predictors of college teacher effectiveness. Unpublished doctoral dissertation, University of Northern Colorado, 1970.

Hanke, J. E., and Houston, S. R. Teacher and student perceptions as predictors of college teacher effectiveness. *College Student Journal,* 1972, *6,* 45–46.

Harry, J., and Goldner, N. S. The null relationship between teaching and research. *Sociology of Education,* 1972, *45,* 47–60.

Haslett, B. J. Student knowledgeability, student sex, class size, and class level: Their interactions and influences on student ratings of instruction. *Research in Higher Education,* 1976, *5,* 39–65.

Heilman, J. D., and Armentrout, W. D. The rating of college teachers on ten traits by their students. *Journal of Educational Psychology,* 1936, *27,* 197–216.

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

Hill, W. R. Student rating of teachers. *Engineering Education,* 1969, *60,* 107–108.

Hillery, J. M., and Yukl, G. A. Convergent and discriminant validation of student ratings of college instructors. Paper read at the annual meeting of the Midwestern Psychological Association, 1971.

Hirschi, T., and Selvin, H. C. *Delinquency research: An appraisal of analytic methods.* New York: Free Press, 1967.

Hogan, T. P. Similarity of student ratings across instructors, courses, and time. *Research in Higher Education,* 1973, *1,* 149–154.

Holland, J. B. The image of the instructor as it is related to class size. *Journal of Experimental Education,* 1954, *23,* 171–177.

Hoyt, D. P., and Cashin, W. E. IDEA technical report No. 1: Development of the IDEA system. Manhattan, Kan.: Center for Faculty Evaluation and Development, 1977.

Hoyt, D. P., Owens, R. E., and Grouling, T. Interpreting "Student Feedback on Instruction and Courses": A manual for using student feedback to improve instruction. Manhattan, Kan.: Office of Educational Resources, Kansas State University, 1973.

Hoyt, D. P., and Spangler, R. K. Faculty research involvement and instructional outcomes. *Research in Higher Education,* 1976, *4,* 113–122.

Instructor and Course Evaluation System. ICES: Its rational and description. Newsletter Number 2. Urbana-Champaign, Ill.: Office of Instructional Resources, Measurement and Research Divisions, University of Illinois at Urbana-Champaign, n.d.

Jiobu, R. M., and Pollis, C. A. Student evaluations of courses and instructors. *American Sociologist,* 1971, *6,* 317–321.

Kane, M. T., Gillmore, G. M., and Crooks, T. J. Student evaluations of teaching: The generalizability of class means. *Journal of Educational Measurement,* 1976, *13,* 171–183.

Kapel, D. E. Assessment of a conceptually based instructor evaluation form. *Research in Higher Education,* 1974, *2,* 1–24.

Kerlinger, F. N., and Pedhazur, E. J. *Multiple regression in behavioral research.* New York: Holt, Rinehart and Winston, 1973.

King, A. P. The self-concept and self-actualization of university faculty in relation to student perceptions of effective teaching. Unpublished doctoral dissertation, Utah State University, 1971.

Kohlan, R. G. A comparison of faculty evaluations early and late in the course. *Journal of Higher Education,* 1973, *44,* 587–595.

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

Land, K. C. Principles of path analysis. In E. F. Borgatta (Ed.), *Sociology Methodology 1969.* San Francisco: Jossey-Bass, 1969.

Landis, L. M., and Pirro, E. B. Required/elective student differences in course evaluations. *Teaching Political Science,* 1977, *4,* 405–422.

Lasher, H., and Vogt, K. Student evaluation: Myths and realities. *Improving College and University Teaching,* 1974, *22,* 267–269.

Lewis-Beck, M. S. Determining the importance of an an independent variable: A path analytic solution. *Social Science Research,* 1974, *3,* 95–107.

Linsky, A. S., and Straus, M. Dimensions of academic competence: The relationship of classroom and research performance of college faculty. Final Report, Project No. 0-A-045, Grant No. OEG-70-000045-0014(509), Office of Education, U.S. Department of Health, Education and Welfare, 1972.

Linsky, A. S., and Straus, M. A. Student evaluations of teaching: A comparison of sociology with other disciplines. *Teaching Sociology,* 1973, *1,* 103–118.

Linsky, A. S., and Straus, M. Student evaluations, research productivity, and eminence of college faculty. *Journal of Higher Education,* 1975, *46,* 89–102.

Lovell, G. D., and Haner, C. F. Forced-choice applied to college faculty rating. *Educational and Psychological Measurement,* 1955, *15,* 291–304.

Lunney, G. H. Attitudes of senior students from a small liberal arts college concerning faculty and course evaluation: Some possible explanations of evaluation results. Research Report No. 32. Danville, Ky.: Office of Institutional Research, Centre College of Kentucky, 1974.

Marsh, H. W. The relationship between background variables and students' evaluations of instructional quality. OIS 76-9. Los Angeles, Calif.: Office of Institutional Studies, University of Southern California, 1976. (a)

Marsh, H. W. The relationship between students' evaluations of instruction and course enrollments. OIS 76-15. Los Angeles, Calif.: Office of Institutional Studies, University of Southern California, 1976. (b)

Marsh, H. W. The validity of students' evaluations: Classroom evaluations of instructors independently nominated as best or 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 read at the annual meeting of the American Educational Research Association, 1978.

Marsh, H. W., Overall, J. U., and Thomas, C. S. The relationship between student evaluations of instruction and expected grade. Paper read at the annual meeting of the American Educational Research Association, 1976.

McDaniel, E. D., and Feldhusen, J. F. Relationships between faculty ratings and indexes of service and scholarship. *Proceedings of the 78th Annual Convention of the American Psychological Association,* 1970, *5,* 619–620.

Miller, R. I. *Evaluating faculty performance.* San Francisco: Jossey-Bass, 1972.

Mirus, R. Some implications of student evaluation of teachers. *Journal of Economic Education,* 1973, *5,* 35–37.

Murray, H. G. The reliability and validity of student ratings of faculty teaching ability. Unpublished manuscript, n.d.

Nichols, A., and Soper, J. Economic man in the classroom. *Journal of Political Economy,* 1972, *80,* 1069–1073.

Nichols, M. G. A study of the influences of selected variables involved in student evaluations of teacher effectiveness. Unpublished doctoral dissertation, University of South Dakota, 1967.

Office of Evaluation Services. Student Instructional Rating System: Analysis of responses for Winter term 1970. SIRS Research Rep. No. 1. East Lansing, Mich.: Michigan State University, 1971.

Overall, J. U., Marsh, H. W., and Kesler, S. P. Class size and students' ratings of instruction: A clarification of relationship. Paper read at the annual meeting of the American Educational Research Association, 1977.

Overturf, C. L., Jr., and Price, E. C. Student rating of faculty at St. Johns River Junior College. Administrative Team Report. Palatka, Fla.: St. Johns River Junior College, 1966.

Perlman, D. Class size and students' ratings of university courses. Paper read at the annual meeting of the Canadian Psychological Association, 1973.

Perry, R. R., and Baumann, R. R. Criteria for the evaluation of college teaching: Their reliability and validity at the University of Toledo. In A. L. Sockloff (Ed.), Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia, Penn.: Measurement and Research Center, Temple University, 1973.

Peters, C. C., and Van Voorhiss, W. R. *Statistical procedures and their mathematical bases.* New York: McGraw-Hill, 1940.

Pohlmann, J. T. Summary of research on the relationship between student characteristics and student evaluations of instruction at Southern Illinois University, Carbondale. Technical Rep. 1.1-72. Carbondale, Ill.: Counseling and Testing Center, Southern Illinois University, Carbondale, 1972.

Pohlmann, J. T. Evaluating instructional effectiveness with the Instructional Improvement Questionnaire. Technical Rep. 5.1-73. Carbondale, Ill.: Counseling and Testing Center, Southern Illinois University, Carbondale, 1973.

Pohlmann, J. T. A multivariate analysis of selected class characteristics and student ratings of instruction. *Multivariate Behavioral Research,* 1975, *10,* 81–92.

Pohlmann, J. T. A description of effective college teaching in five disciplines as measured by student ratings. *Research in Higher Education,* 1976, *4,* 335–346.

Pritchard, W. M. Student evaluation of college physics teaching. *Journal of Research in Science Teaching,* 1972, *9,* 383–384.

Remmers, H. H. The college professor as the student sees him. *Bulletin of Purdue University,* 1929, *29* (6, Purdue University Studies in Higher Education No. 11). (a)

Remmers, H. H. Departmental differences in the quality of instruction as seen by students. *School and Society,* 1929, *30,* 332–334. (b)

Remmers, H. H., Hadley, L., and Long, J. K. Learning, effort, and attitudes as affected by class size in beginning college engineering mathematics. *Bulletin of Purdue University,* 1932, *32* (9, Purdue University Studies in Higher Education No. 19).

Riley, J. W., Jr., Ryan, B. F., and Lifshitz, M. *The student looks at his teacher: An inquiry into the implications of student ratings at the college level.* New Brunswick, N.J.: Rutgers University Press, 1950.

Robinson, W. S. Ecological correlations and the behavior of individuals. *American Sociological Review,* 1950, *15,* 351–357.

Rohrer, J. H. Large and small sections in college classes. *Journal of Higher Education,* 1957, *28,* 257–279.

Schwab, D. P. Manual for the Course Evaluation Instrument. Madison, Wis.: Graduate School of Business and Industrial Relations Research Institute, University of Wisconsin-Madison, 1976.

Scott, C. S. Correlates of student ratings of professorial performance: Instructor defined extenuating circumstances, class size, and faculty member's professional experience and willingness to publish results. Paper read at the annual meeting of the American Educational Research Association, 1975.

Scott, C. S. Student ratings and instructor-defined extenuating circumstances. *Journal of Educational Psychology,* 1977, *69,* 744–747.

Seiler, L. H., Weybright, L. D., and Stang, D. J. How useful are published evaluation ratings to students selecting courses and instructors? Unpublished manuscript, 1977.

Shingles, R. D. Faculty ratings: Procedures for interpreting student evaluations. *American Educational Research Journal,* 1977, *14,* 459–470.

Solomon, D. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. *American Educational Research Journal,* 1966, *3,* 35–47.

Sorge, D. H., and Kline, C. E. Verbal behavior of college instructors and attendant effect upon student attitudes and achievement. *College Student Journal,* 1973, *7,* 24–29.

Spencer, R. E. Course Evaluation Questionnaire: Results by course level. Research Rep. No. 213. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1966.

Spencer, R. E. Some dimensions of the Illinois Course Evaluation Questionnaire. Research Rep. No. 303. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1969. (a)

Spencer, R. E. A study of the inter-judge reliability of the Illinois Course Evaluation Questionnaire. Research Rep. No. 305. Urbana-Champaign, Ill.: Measurement and Research Division, Office of Instructional Resources, University of Illinois, 1969. (b)

Stanley, J. C. Analysis of unreplicated three-way classifications, with applications to rater bias and trait independence. *Psychometrika,* 1961, *26,* 205–219.

Stanley, J. C. Reliability. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd ed.). Washington, D.C.: American Council on Education, 1971.

Starrack, J. A. Student rating of instruction. *Journal of Higher Education,* 1934, *5,* 88–90.

Stuit, D. B., and Ebel, R. L. Instructor rating at a large state university. *College and University,* 1952, *27,* 247–254.

Trent, C., and Johnson, J. F. The influence of students' values and educational attitudes on their evaluation of faculty. *Research in Higher Education,* 1977, *7,* 117–125.

Van Horn, C. An analysis of the 1968 course and instructor evaluation report. Institutional Research Bulletin No. 2-68. West Lafayette, Ind.: Measurement and Research Center, Purdue University, 1968.

Villano, M. W. The relationship of certain course characteristics to student ratings of science and mathematics teaching at four-year and two-year colleges. Paper read at the annual meeting of the American Educational Research Association, 1975.

Villano, M. W., Rosenstock, E. H., and Estes, C. A decade with a student course evaluation form at a major university. Paper read at the annual meeting of the American Educational Research Association, 1974.

Walker, B. D. An investigation of selected variables relative to the manner in which a population of junior college student evaluate their teachers. Unpublished doctoral dissertation, University of Houston, 1968.

Weerts, R. R., and Whitney, D. R. The effect of student, course, and instructor characteristics on types of items used in student evaluation of instruction. Paper read at the annual meeting of the National Council on Measurement in Education, 1975. (a)

Weerts, R. R., and Whitney, D. R. Student Perceptions of Teaching (SPOT): V. Relationships between averaged student responses and selected course characteristics. Research Rep. No. 78. Iowa City, Iowa: Evaluation and Examination Service, University of Iowa, 1975. (b)

Whitely, S. E., Doyle, K. O., Jr., and Hopkinson, K. Student ratings and criteria for effective teaching. Rep. No. 731F. Minneapolis, Minn.: Measurement Services Center, University of Minnesota, 1973.

Wilson, W. P. Students rating teachers *Journal of Higher Education,* 1932, *3,* 75–82.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

Witheiler, P., and Yuker, H. E. Course evaluations at Hofstra University, 1969. Rep. No. 90. Hempstead, N.Y.: Center for the Study of Higher Education, Hofstra University, 1970.

Wood, K., Linsky, A., and Straus, M. A. Class size and student evaluations of faculty. *Journal of Higher Education,* 1974, *45,* 524–534.

Wood, P. H. The description and evaluation of a college department's faculty rating system. Paper read at the annual meeting of the American Educational Research Association, 1977.

FOOTNOTES

1. From these studies, incidentally, it seems to make little difference in the size of correlations whether the paired courses were taught by the teacher in the same semester or in different semesters. Another set of studies was considered separately for the present analysis (Bausell and Vinograd, 1977; Colliver, 1972; Heilman and Armentrout, 1935; Shingles, 1977; Van Horn, 1968; and Wood, 1977). In these studies either (1) only some but not all of the teachers in the sample were teaching exactly the same course to two different classes of students or (2) it is not clear whether, for each instructor, the paired class ratings were for exactly the same course. The correlations between the paired ratings for each teacher are still high in these studies, although perhaps typically a little lower in size, for correlations generally ranged from the .50's to the .70's.

2. This predominant contribution of the instructor is clear for rating items that focus on the evaluation of the instructor (as do those items selected for presentation in Table 1). Not surprisingly, however, course setting has a greater influence than does the instructor when rating items pertain to certain aspects of the course itself, such as the difficulty of the course or its relevance to the student (see Hogan, 1973, and Bausell, Schwartz, and Purohit, 1975).

3. In addition to the research that has already been cited, it should be noted that the following studies also report there being a relationship between size and ratings, but they were not included in the review because the exact nature of the relationship between the two variables cannot be clearly determined from the information given in the studies: Batista and Brandenburg (1975); Hoyt and Cashin (1977); Hoyt, Owens, and Grouling (1973); Lasher and Vogt (1974); Overall, Marsh, and Kesler (1977, Study No. 2); Rohrer (1957); and Villano, Rosenstock, and Estes (1974). Also, it is not exactly clear in Weerts and Whitney (1975b) and Hill (1969) whether size has a weak association with ratings or none at all. Not generally included in the present review are studies of teachers or courses who have been selected as the "best" through one or another nominating procedure (Gillmore, 1975a; Goldsmid, Gruber, and Wilson, 1977; Guthrie, 1954; and Marsh, 1977). It may be noted here only that, in these studies, size of class does not distinguish outstanding teachers (courses) from other teachers (courses), with the one exception of the study by Goldsmid and his associates, where the direction of the association is not given.

4. Because the investigators write that "course numbers were identical," this would obviously seem to mean that exactly the same course (but one given to different classes of students) is involved. It should be noted, however, that the impression is also given in their report that only general course level (freshman level, sophomore level, and so on) was controlled.

5. Readers unfamiliar with the logic of "elaborating" relationships through "explanation" or "interpretation" are referred to expositions in Cole (1976, chap. 2) and Hirschi and Selvin (1967, chaps. 5 and 6).

6. As would be expected, size is not related to ratings in studies by Hanke (1970) and Jiobu and Pollis (1971) when other relevant variables are controlled, since the initial zero-order association between size and ratings was not statistically significant in these studies. Size is also not related to ratings, or results are mixed, when relevant variables are controlled in studies by Cornwell (1974), Nichols and Soper (1972), and Shingles (1977); however, it is not known from these studies whether the controlling procedure "washed away" the zero-order relationships or whether there were no such relationships in the first place, for the zero-order associations between size and ratings are not given. The beta coefficient for class enrollment is unexpectedly positive and statistically significant in the regression analysis reported in Mirus (1973), although the zero-order correlation is not given.

7. Whether the fact of statistically insignificant beta coefficients (for the predictor variable of class size) in the studies by Cornwell (1974), Jiobu and Pollis (1971), and Nichols and Soper (1972), can be taken as furnishing contrary evidence is not clear (for reasons given in footnote 6). The clearest exception to the general finding is shown in the study by Marsh (1978), which thus offers the most support to Batista and Brandenburg's analysis. The data presented in Marsh's report show that the weak, inverse relationship between class size and rating in effect did not hold when controlling for a set of other variables, the most important of which turned out to be average expected grade as well as average interest in the subject matter prior to entering the course and average (reported) workload and difficulty of the course.

8. Aleamoni and Graham (1974), Brandenburg, Slinde, and Batista (1977), Lasher and Vogt (1974), and Villano, Rosenstock, and Estes (1974) all found statistically significant relationships between course level and ratings, but the exact nature of the relationship cannot be clearly or easily determined from the information given in their reports. Also, because of the form in which information is reported in the following studies, it is not clear whether or not zero-order associations between course level and ratings were found: Cook et al. (1975); Mirus (1973); Nichols and Soper (1972); Shingles (1977); Weerts and Whitney (1975b); and Witheiler and Yuker (1970).

9. For evidence of interactive effects (or lack of them) between course level and other course characteristics as well as instructor characteristics, see Aleamoni and Graham (1974), Cook et al. (1975), Grant (1971), Villano (1975), and Villano, Rosenstock, and Estes (1974).

10. Lasher and Vogt (1974) report that "required course offering" is related to evaluation of teachers, but the direction of results is not given. Weerts and Whitney (1975b) only report correlations over .30.

11. As an aside, it may be noted that higher correlations (than that between ratings and the intrinsic interest in the subject matter of the course presumably brought to it by the students) would be expected if the variable being related to ratings is the degree of stimulation by the teacher (and by the course) of students' interest in the subject matter of the course. Indeed, Harry and Goldner (1972) report a correlation of .82 between the percent of students reporting increased interest in the course matter over the semester (due presumably to the teacher as well as the course experience in general) and the overall rating of the instructor. Likewise, Cashin and Slawson (1977a, Table 5) report a correlation of .79 between the average degree to which students in a class agree with the statement that "as a result of taking this course, I have more positive feelings toward this field of study" and the average degree of agreement with the statement that they "would like to take another course from this instructor." The fact that some students change their interest in the course's subject matter, due to the efforts of the instructor and their experiences in the class, may also account for the high correlations (primarily in the .60's and .70's), found by Jiobu and Pollis (1971) and Sorge and Kline (1973) between "interest in" or "attitude toward" (respectively) the subject matter of the course and the average student rating. For, without the specification to students that reference is to their stable interest or attitude brought to the course, the correlations probably include a component due to teacher-induced (and course-induced) changes in these interests and attitudes. A similar explanation might be offered for the correlations in the .40's (in a study by Hoyt and Cashin, 1977) between the average degree to which students (near the end of the course) agreed that they "had a strong desire to take this course" and their ratings of various pedagogical procedures of the instructor, since it is not clear whether the question about their desire is referring to their reactions and academic motivation before they entered the course or during the course (or both).

12. The following reports also have information on academic area and ratings, although it could not be included in Table 2, primarily because of the restricted range or unusual combination of academic areas studied or because the academic fields could not be ranked in terms of level of instructor ratings (due to insufficiencies in the data or to the form in which the data are presented): Bejar and Doyle (in press); Clark and Keller (1954); Delaney (1976); Hanke (1970; also see Hanke and Houston, 1972); Hoyt and Spangler (1976); King (1971); Lasher and Vogt (1974); Marsh (1976a, 1976b); Remmers (1929b); Riley, Ryan, and Lifshitz (1950); Solomon (1966); and Villano (1975).

13. This same procedure has been used to compare the typical personality and attitudinal attributes of students majoring in different academic areas (see Feldman and Newcomb, 1969, chap. 6). Discussion of some of the advantages as well as the problems in this procedure can be found in Feldman (1971b, 1979) and Feldman and Newcomb (1969, Appendix G).

14. It is even possible that the nature of the topics within a course may affect class ratings. At least in a study by Carney and McKeachie (1966), it was found not only that students in introductory psychology courses generally preferred life-oriented to science-oriented topics but also that the students rated the entire course (including lectures, discussions, and tests) higher when life-oriented topics were being taught than when science-oriented topics were being taught. Whether the overall class rating of the teachers varied accordingly is not known from the report.

15. Two studies were found in which path analysis was used to study factors related to students' ratings of their teachers (Chermesh, 1977, and Trent and Johnson, 1977); in neither of them, however, are course characteristics part of the path model. For relatively simple expository material on path analysis, see Anderson and Evans (1973), Feldman (1971a), Kerlinger and Pedhazur (1973, chap. 11), and Land (1969).

16. Although the matter is put rather simply here, the exact way in which expected grade intervenes between a course characteristic and a class rating (assuming it does so at all) may turn out to be more complicated, due to the ways in which expected grade is probably linked to such other factors as students' innate interest in the subject matter of the course, the difficulty of the course in terms of content and workload, and students' actual achievement in the class. For some hints at the complexities that may be involved, see Marsh (1978).

17. The discussion here has been of the possibility of course characteristics directly biasing the class ratings of teachers. It is also possible that course characteristics may indirectly bias these ratings through their effect on intermediary variables. An example of such a variable would be the expected grade of students in the class, since the grade typically expected by students in a class could be influenced by some characteristic of the course and might in turn bias ratings. Of course, it must be shown, rather than assumed, that expected grade does indeed bias ratings, in that students tend to "unjustly" reward (or punish) teachers by raising (or lowering) their ratings according to the grades they anticipate. Otherwise, a relationship between (average) expected grade and class rating could just as well indicate that differences in the achievement of different classes (and, consequently, differences in the grades typically expected) are produced by differences in the actual effectiveness of teachers who are consequently and deservedly rated differentially by their classes (cf. Feldman, 1976a, and Marsh, 1978).