RESEARCH IN HIGHER EDUCATION
Volume 6, pages 223-274

CONSISTENCY AND VARIABILITY AMONG COLLEGE STUDENTS IN RATING THEIR TEACHERS AND COURSES: A Review and Analysis

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

As indicated by the reliability of individual ratings, college students are only moderately consistent in rating their teachers and courses, although these modest interrater associations do produce substantial reliabilities for composite ratings when the ratings of at least 20 to 25 students in a class are averaged together. The patterning and correlates of variability of student ratings within classes are examined. Certain attributes and experiences of students are weakly related to their ratings, and inconsistently so, across studies; others are more strongly and consistently related. Various correlates of student ratings have also been found to interact as well as linearly combine with one another in their association with ratings. Moreover, certain kinds of "fit" between teachers and different students in their classes are related to ratings. Whether various correlates of within-class ratings are to be interpreted as biasing factors or as natural influences on social perception is analyzed in terms of whether students' ratings are objective descriptions or subjective, evaluative reactions.

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

For years now, the use of formal student assessment of college courses and instructors has been widespread in American universities. If anything, the past several years have seen an increase in the prevalence and popularity of student ratings as well as an increase in the frequency with which they are used in decisions concerning faculty status (Bejar, 1975). One concern of educators, administrators, and researchers is whether students are in enough agreement in their assessments of courses and teachers for their ratings to be taken seriously. To the degree that students within classes are not consistent, a concomitant concern arises—namely, whether there are differences among

students that may be producing systematic variability in ratings. These two concerns are generally discussed in terms of the reliability of student ratings (specifically, interrater reliability) and the student characteristics that may be associated with ratings. The research in both of these areas has, of course, received review and analysis (see, for example, Costin, Greenough, and Menges, 1971; Doyle, 1975; Dwyer, 1968; Flood Page, 1974; Kulik and McKeachie, 1975; Miller, 1974). Enough complexities and uncertainties remain, however, to merit a somewhat more extended consideration than is usually found.

CONSISTENCY AMONG STUDENTS IN THEIR RATINGS

Traditional Test Reliability Theory and its Application to Ratings

In terms of classical reliability theory, observed scores of individuals on psychological tests are conceived as comprising some "true" component, which is variously defined (see Lord and Novick, 1973, Chap. 2; Wiggins, 1973, Chap. 7), plus an "error" component. Reliability is then defined as the ratio of the variance in true scores to the variance in observed scores. Since the reliability coefficient is itself a coefficient of determination, it is interpreted directly (without being squared) as the proportion of variance in the obtained scores determined by variance in the true scores. (The correlation coefficient that has, in effect, been squared is that between the obtained and true scores, and is known as the index of reliability; see Anastasi, 1968, Chap. 4; Guilford, 1954, Chap. 13.)

Several approaches are taken in estimating a reliability—based on test-retest, parallel test, split-half, and internal-consistency procedures—essentially defining in different ways what is meant by error. Of particular significance to the present paper are split-half and internal-consistency procedures. As is well known, the split-half approach estimates reliability from a single testing by dividing a multi-item test into two presumably equivalent halves. The estimate of reliability is determined by correlating the scores of the two half-length tests and correcting for length by using the Spearman-Brown Prophecy Formula. The principle of fractionation of a test into halves has been broadened to divisions into smaller parts, even into single items (see especially Guilford, 1954, Chaps. 13 and 14). In the case of single items, the information sought concerns their equivalence for measurement purposes—in short, the internal consistency of the test. Internal consistency is thus interpreted as the degree to which items have something in common or the extent to which they measure the same trait. It is possible through the use of item statistics to estimate the reliability of a

single item (that is, the reliability per item); it is also possible to estimate the reliability of a composite score across items—for example, the reliability of a total score or a mean score (see especially Guilford, 1954, Chaps. 13 and 14).

As Crichton and Doyle (1975) point out, the traditional testing model deals with a persons (testees) × test items data matrix, with the items dimension collapsible into a total scores vector by virtue of the assumption that items are replicates, except for error. This approach has been adapted or generalized to the rating situation by substituting ratees for testees and raters for test items in the assumptions and deductions of the testing model. For example, it is assumed that raters (say, students) are replicates of one another, except for random error, in their "measuring" of the ratees' (say, teachers') attributes.

Analogous to the traditional testing model, either the reliability of individual ratings or the reliability of average ratings can be estimated (Ebel, 1951). The first reliability may be conceived as a type of average of the reliabilities of the individual raters, termed by Guilford (1954, p. 395) as either the "reliability of ratings for a single rater" or "the mean reliability for one rater." (For the present analysis, reference will be either to the "reliability of individual ratings" or to the "reliability of the single rater.") This reliability is an estimate of the proportion of variance among the observed individual ratings that can be "accounted for" by "true" variance in these ratings. The second reliability is that for the averages of the ratings (of each ratee) given by two or more raters, sometimes referred to as the reliability of composite ratings (see Tinsley and Weiss, 1975). This reliability is an estimate of the proportion of variance among the average ratings that can be "accounted for" by "true" variance in these ratings. Incidentally, as Ebel (1951) notes for the formulas presented in his analysis, "since the reliability of average ratings is determined completely by the reliability of the component ratings, it is always possible to determine the reliability of individual ratings, or of averages, no matter which value a formula gives initially" (p. 408).

Interrater Reliability of Student Raters

Several different procedures have been used by researchers to determine the degree of consistency among college students in rating their teachers and to estimate, thereby, the reliability of these ratings. Following is a brief survey of the results of these studies, categorized by type of procedure.

1. One way of determining the consistency among student raters is to calculate the product-moment correlation of their ratings (across

teachers). If there are more than two raters, the average intercorrelation among raters is used. In only one of the studies located for this review was an average interrater correlation actually calculated. Menne (1968) administered a multi-item rating scale to 15 randomly selected students from each of 76 classes at Iowa State University. Using a subroutine in a regular correlation program, the off-diagonal correlations between students' scores on this scale were arranged to give the average interrater correlation coefficient. This coefficient, itself the estimate of the interrater reliability of the single rater, was 0.34. Applying the Spearman-Brown Prophecy Formula, the reliability of the average ratings of the 76 groups of students (15 persons each) is 0.89. By the same formula, Menne estimated the reliability of the average class rating to be 0.91, 0.93, 0.94, and 0.965 for classes of size 20, 25, 30, and 50 students, respectively.

2. In studies by Guthrie (1945) and by Magoon, Bausell, and Price (see Bausell and Magoon, 1972a, 1972b, 1972d, Appendix A; Magoon and Bausell, unpublished; Magoon and Price, 1972; Price and Magoon, 1971), the reliability of individual ratings was estimated by correlating the ratings of randomly drawn pairs of raters for those students completing teacher rating forms in various courses. An interrater correlation was calculated for each of the rating items on the form. Results were typically in the 0.30s. Given an average class size of 30 students (which was roughly the median number of students in the classes in the studies by Magoon and his associates), the reliability of average ratings for the typical rating item would be in the 0.90s. A similar pairing procedure was used in an early study by Guthrie (1927), although in this case the data were based on students' rankings of the instructors they had had during the year in order of the quality of their teaching. The reliability of a single rank was 0.26; the reliability of an average of 16.5 rankings for a teacher (the average number of rankings per teacher) was 0.85.

3. A more often-used index of consistency among student raters (and estimate of the reliability of individual ratings) is the coefficient of intraclass correlation. Although this correlation may be viewed as giving "essentially an average intercorrelation" (Guilford, 1954, p. 395), the two correlations are not the same. Under specific conditions certain arithmetical identities do exist between these two types of correlation, and under some circumstances they do give either the same or closely similar results, but they are not logically identical measures (for details, see Haggard, 1958, Chap. 1; Winer, 1962, pp. 124–132; and Schuessler, 1971, Chaps. 6 and 8). The coefficient of intraclass correlation, as calculated and interpreted within an analysis of variance framework, is a measure of the relative homogeneity of the scores (say, ratings by stu-

dents of instructors or courses) within the classes of scores under consideration (say, those of different instructors or courses) in relation to the total variation among all the scores across the classes. The lower the estimate of reliability of the single student rater or individual student ratings (that is, the lower the intraclass correlation coefficient), the larger the variation in responses among the student raters within courses, or the lower the variation in average responses across courses (instructors), or both. Contrarily, the larger the intraclass coefficient, the more differentiation there is among courses or instructors relative to that among student raters within courses.

The following studies directly present reliabilities of individual ratings (for each single rating item or each multi-item scale taken separately) or give data from which such reliabilities can be calculated: Bendig (1953a); Centra (1972, 1973, 1974, 1975); Follman, Lavely, Silverman, and Merica (1974); Follman, Lucoff, Small, and Power (1974); Gillmore (1973, see Samples 1-3, 1975); Majer and Stayrook (1974); Sharon and Bartlett (1969); Veldman (1968). The estimates in these studies of the reliability of individual ratings (the coefficients of intraclass correlation) are primarily in the 0.10s and 0.20s; coefficients in the 0.30s or higher are infrequent. The average class size (that is, the average number of raters per course or teacher) in most of these studies was around 20 students or more, and most of the reliabilities of the average scores for various rating items or scales were therefore at least in the .70s and more often in the 0.80s and 0.90s.[1]

4. Several studies (Apt, 1966; Baker and Remmers, 1951, also see Remmers and Weisbrodt, 1964; Bendig, 1953a; Deshpande, Webb, and Marks, 1970; French-Lazovik, 1974; Snedeker, 1959; Voecks, 1962) report reliability of average ratings for instructors or courses, as determined by the generalized reliability formula developed by Horst (1949). (For a comparison of this formula with the one based on intraclass correlation, see Ebel, 1951.) Again the reliabilities of average ratings for various single items or multi-item scales are generally in the 0.80s and 0.90s. None of these studies report reliabilities of the single rater, but they can be calculated by using the Spearman-Brown Prophecy Formula (in reverse, as it were) when the average class size used is known (Ebel, 1951). Performing this calculation for these particular studies results in reliabilities of individual ratings that are typically in the 0.10s and 0.20s, although there are instances of both lower and higher reliabilities.

5. Estimates of reliabilities of average ratings have been gained by another method. Two mean scores for any particular rating item or scale are obtained for each class by (randomly) dividing each class into two subgroups of students (for each of which subgroups a mean score

on the rating item or scale of interest is calculated). Subgroups may be gotten by dividing each class into halves, although this is not the only way to do so; in some studies two samples are picked for each class, each of which has less than half of the students in class, but all of which have the same number of students. The two mean scores for each class are then correlated across classes. The resulting correlation is corrected by the Spearman-Brown Prophecy Formula, since the means are based on only half (or less) of the student raters in a class. Reliability estimates so generated have generally been in the 0.70s, 0.80s, and 0.90s (see Bausell and Magoon, 1972c, 1972d, Appendixes A and U; Gillmore, 1973, see Samples 5 and 6; Guthrie, 1949, 1954; Hoyt, 1969, 1973a, 1973b; Maslow and Zimmerman, 1956; Murray, 1975; Remmers and Weisbrodt, 1964, also see Baker and Remmers, 1951; Spencer, unpublished). In all these reports, except those by Gillmore (1973), Murray (1975), and Spencer (unpublished), there is sufficient information to use the Spearman-Brown Prophecy Formula as a "step-down" formula for obtaining a rough approximation of reliabilities of individual ratings. Across these studies, these reliabilities are usually within the range of 0.10 to 0.30.

6. Kane, Gillmore, and Crook (1977) and Gillmore, Kane, and Naccarato (1976) have suggested that the reliability of ratings of teachers and courses be approached through generalizability theory. This theory, as explicated by Cronbach, Gleser, Nanda, and Rajaratnam (1972), is a broadening of classical reliability theory to take advantage of information that can be provided by data collected with a more complex design. Unlike traditional reliability theory, which incorporates a univariate interpretation of error (in the case of either the traditional testing model or its application to ratings), generalizability theory allows for a multidimensional interpretation of error. Also, unlike the traditional reliability coefficient which is calculated for rating items or scales one at a time, a generalizability coefficient—the coefficient that "replaces" the traditional reliability coefficient—can be determined simultaneously across all items or scales of a rating form or any subset thereof. (For a particularly useful explanation and discussion of the application of generalizability theory to a split-plot design in which students are nested within classes—the typical situation when students rate their teachers—see Kane and Brennan, 1977.) Since generalizability theory is an extension of traditional reliability theory, it is not particularly surprising that the sizes of various generalizability coefficients reported by Kane et al. (1977) for students at the University of Illinois and by Gillmore et al. (1967) for students at the University of Washington are in the same ranges as those that have been reported in analyses using traditional reliability theory.

The Interpretation of Interrater Reliability Coefficients

Interpreting interrater reliability coefficients must be done with caution, especially when interest lies in the consistency among students' ratings of their teachers. At the outset, it should be pointed out that reliability coefficients of individual ratings indicate the degree of general or relative consistency among raters; they do not measure exact or absolute agreement. Although the terms interrater reliability and interrater agreement are often used interchangeably, they do not refer to the same thing (see Tinsley and Weiss, 1975). Interrater agreement is the extent to which different raters give exactly the same ratings for the rated subject. Perfect interrater agreement would be shown, for example, if raters assigned exactly the same values on a numerical scale when rating the same person. By contrast, interrater reliability represents the degree to which the ratings by different raters are proportional when expressed as deviations from their means. As such, interrater reliability is usually expressed in terms of correlation or analysis of variance indexes. Although several agreement indexes exist (see Frick and Semmel, 1974; Shapiro, 1974; Tinsley and Weiss, 1975) and have been used in various areas of research, no studies were found that used any of these agreement statistics in an analysis of college students' ratings of their teachers. Perhaps investigators in this area are more interested in determining the general or relative consistency among students in order to fulfill the "technical" requirement of reporting (hopefully high) interrater reliabilities than they are in measuring the exact agreement of raters as an interesting matter in its own right. At any rate, an agreement index would be a useful supplement to an interrater reliability (see Byrne, 1964).

"Error variance" in reliability formulas and theories refers to random error—that is, uncontrolled fluctuations or haphazard variation. These are errors reflecting momentary variations in the circumstances of measurement or the like that are unrelated to the measurement procedure itself. Consequently, "not every type of error, not every discrepancy from the value which an omniscient recording angel would register for the speciman in question, qualifies as a part of error variance" (Stanley, 1971, p. 360). Likewise, "true score" does not refer to the ontologically "real" or "correct" score (unless the somewhat unsatisfactory simple Platonic conception of true score is accepted, see Lord and Novick, 1968, Chap. 2; Wiggins, 1973, Chap. 7). Stanley (1971) puts the matter as follows:

> "As used, true score is not the ultimate fact in the book of the recording angel. Rather, it is *the score resulting from all systematic factors one chooses to aggregate*, including any systematic biasing factors that may pro-

> duce systematic incorrectness in the score. . . . The heart of any treatment of reliability involves recognition that true variance is *wanted* variance and that what is wanted will depend on the interpretation proposed by the investigator" (p. 361, emphasis in the original).

Given that the sample of student raters is a random sample from a population of comparable raters, and if certain other conditions or assumptions are met,[2] then one interpretation of the estimate of reliability of the average (or class) ratings is as follows: If the ratings of the same 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 (see Winer, 1962, pp. 129–132). The interpretation of the reliability of individual ratings follows along analogous lines. Random sampling of students is especially problematic, of course, since students largely self-select themselves into courses; even for each course considered separately, it is usually not possible to point to a specific population from which the students are a random sample. Thus the assumption of random sampling is relaxed—not without challenge, it may be noted (see Stanley, 1961, 1971)—by introducing the idea of an unspecified population of students "like those observed"; the estimate of interrater reliability is interpreted as the degree to which (observed) ratings would be expected to correlate with the ratings gained by another set of students "similar" to the ones who were used—that is, in effect, another set of students who might reasonably have taken the various courses and who in fact may take them in the future (see Cornfield and Tukey, 1956; Gillmore, 1973; Guthrie, 1945; Kane et al., 1977; Peters and Van Voorhis, 1940, Chap. 7).

It is important to realize that, although the reliabilities of *average* college student ratings tend to be high in the studies surveyed (generally in the 0.70s, 0.80s, and 0.90s), this does not mean that students within classes were highly consistent in their ratings. In fact, the magnitude of interrater consistency is empirically moderate, if not low (with indexes of various kinds generally being in the 0.10s, 0.20s, and 0.30s). Of course, these low-to-moderate interrater associations (and thereby low-to-moderate reliability estimates of *individual* ratings) "build up" to high reliabilities when the ratings of 20 to 25 or more students in a class are averaged together, and estimates of reliability are for these averages (directly or indirectly using the Spearman-Brown Prophecy Formula or its equivalent to calculate the reliability of average ratings). The rationale that is given for using averages is that taking the mean of individual observations tends to reduce errors; individual idiosyncracies and ignorances, as well as other (nonsystematic) errors of observation, tend to cancel out (see Cattell, 1957, Chap. 3; Hirschi and Selvin, 1967,

Chap. 12). In this respect, raters are considered as functioning very much as "items" do in conventional tests. Interrater reliability is related to the number of raters as given by the Spearman-Brown Prophecy Formula for test length; increasing the number of raters is viewed as a special type of lengthening (Wiggins, 1973, Chap. 7; Thorndike and Hagen, 1969, Chap. 6).

Whether within-class ratings are justifiably averaged—or, alternately put, whether these ratings are justifiably pooled to provide an estimate of variance due to error measurement—becomes the matter at issue. Ratings are justifiably averaged or pooled if they have been made independently; if so made, and assuming all else equal, "error" components will be independent and will tend to cancel out (Thorndike and Hagen, 1969, Chap. 13). For ratings to be independent, raters should reach their decisions individually rather than as a result of comparing ratings with one another, talking to one another about the ratee, or engaging in formal or informal group discussions and conferences (see Cattell, 1957, Chap. 3; Cochran, 1968; Helmstadter, 1964, Chap. 8; Hirschi and Selvin, 1967, Chap. 12; Horst, 1949; Thorndike and Hagen, 1969, Chap. 13). Otherwise, ratings may become dependent on the personality interaction among raters, the possible influence of the more persuasive or dominant raters, and other such factors that are involved in joint decisions.

In one sense, students do rate their teachers independently (or, at least, could do so). Presumably one or another kind of "conferencing" procedure is controlled (or could be) during the actual completion of teacher rating forms—by asking students not to compare ratings or to confer with one another, and monitoring their behavior to assure that they do not. In another sense, however, ratings by students are not altogether independent in typical classroom settings. To one extent or another, students in a class confer throughout the semester about their teacher and the course. They note each other's reaction to the teacher and course material, talk to one another about the teacher and the course, construct with one another the meaning and interpretations of the teacher's behaviors, mutually establish their own "hearsay" about the teacher, and the like. The more difficult and subtle problem, then, is not that of direct collaboration among raters at the time of rating, but rather of indirect contamination by what is sometimes referred to as the "local reputation" of the ratee (see Thorndike, 1949, Chap. 4), in this case a "reputation" specific to a particular classroom.[3]

This situation, in which indirectly collaborative assessments are permitted (if not facilitated), is unlike the conditions that generally prevail for trained individuals who code information from questionnaires or interviews, who score certain psychological tests, or who act as observ-

ers (and raters) in studies using systematic observational procedures. These persons may be trained together as a unit (although this is not inevitable, and depends on the investigator and the nature of the project). They may also work with one another in practice sessions when coding, scoring, or observing. But in the actual coding, scoring, or observing (and rating) sessions, not only do they make their decisions independently of each other but also the stimulus material, events, or persons are new to them. The teacher rating situation is different because the "object" to be rated (the teacher, the course, or both) is the very entity about which students have been, in part, mutually influencing one another's opinions and jointly forming their assessments; of course, they may be coming to certain independent conclusions in addition.

One consequence of this lack of independence is the possibility of a spuriously high estimate of reliability (Horst, 1949). If, to some degree, students have come to agree with one another more than they otherwise would, were they totally independent observers of their teachers or courses throughout the semester, then any measure of this consistency has entangled in it both jointly produced consensus and similarity of individual decisions independently made. This is not to say that consistency among students, whatever its magnitude, does not exist; such consistency, after all, is the empirical reality. Nor is it to say that consistency among students necessarily represents a stereotyped or false picture of the ratee (although this always remains a possibility). Similarity in rating that is based on interpersonal influence and indirect collaboration becomes a source of systematic variance in "true" scores across teachers or courses; it may or may not be a source of systematic error, which is really a question of validity rather than reliability (see Kerlinger, 1973, Chap. 26).

Rather, the point to be made is that the amount of observed consistency in ratings among students may be an "impure" base for estimating either the reliability of individual ratings or the reliability of average ratings. Thus interpretations of the observed consistencies and reliability estimates based on them should be made with caution. This is especially true if measures of consistency and reliability are to be compared across settings or conditions. For example, Guthrie (1927), in finding an increase in interrater reliability over the school year, writes that "whether the increased agreement found among the students at the end of a year is due to better acquaintance [with the teacher] or to exposure to student gossip is not determined" (p. 176). Likewise, there is more than one possible interpretation for the finding by Gillmore (1975) that interrater consistency, as measured by intraclass correlation, was somewhat greater for students in seminar-discussion classes than for

students in other kinds of classes, including those of approximately the same size as the seminar-discussion classes. It may be that seminar-discussion classes are settings in which students have a greater opportunity to observe the teacher and course more closely, and thus they rate with more knowledge (and consequently with more similarity). Or perhaps it is in exactly these kinds of courses that students get to know one another better, are more likely to discuss both class and nonclass matters with each other, and are more likely to influence one another in various areas (including the assessment of the teacher and course). Perhaps both (and other) factors are involved.

Good reasons exist for the use of average student ratings, both in terms of the increase in reliabilities that result and of the economies gained from data reduction (for purposes of research analyses as well as for administrative decisions). It must be remembered, however, that averaging ratings assumes that raters are replicates except for random error (see Crichton and Doyle, 1975; Magoon and Price, 1972; Remmers, Shock, and Kelly, 1927). As has been seen, the amount of consistency of student ratings within classes seems to be only moderate at best, even with the possibility that students have influenced one another in their cognitions and assessments of their teachers and courses. If the diversity of ratings within classes is indeed due to haphazard fluctations, then it makes sense to assume that raters are "replicates," although possibly not ones who are altogether independent of each other. However, within-class variability may be more than random error; there may be patterned differences in ratings linked to different student types or subgroups in classes. The more such differences exist, the less sound is the assumption that students are interchangeable, and the less easily interpretable are the averages of their ratings (which mask subcategory differences) as well as the reliabilities of these averages (see Crichton and Doyle, 1975). It is to the question of the patterning and correlates of the variability of student ratings within classes that the analysis now turns.

STUDENT TYPES AND CORRELATES OF STUDENT RATINGS

In an early study of student ratings, Wilson (1932) found that some teachers at the University of Washington had distinctly bimodal distributions of student responses on some of the rating items; more students checked the upper and lower extremes of the categories of responses than checked the central positions. Although Wilson notes that "an investigation of such cases showed ordinarily that two quite different types of students were in the class," the ways in which the groups were different are not given in the report.

Centra and Linn (1973), Singhal (unpublished), and Whitlock (1972) have also identified subgroups of students within classes who were distinguishable by their responses across rating items or scales. All three studies used obverse factor analysis to identify these subgroups. Such an analysis uses the subject correlation matrix, in contrast to the item or variable correlation matrix, in order to identify groups of individuals with similar pattern of responses across items or variables. Singhal speculates that the subgroups he found might result from students within the class having different value patterns and experiences, but he did not explore the nature of these values and experiences. Whitlock clustered students within classes according to factorial homogeneity, but the attempt of her study was to discover common clusters *across* teachers (and student attributes that might distinguish these common clusters) rather than to interpret clusters *within* classes and to find the characteristics of students by which these within-class clusters could be differentiated. Only Centra and Linn investigated whether certain characteristics of students would discriminate among the within-class subgroups of students identified in their research. For each of the three courses that they studied, a discriminant analysis was run using the subgroups within each class and five student characteristics (expected grade in the course, cumulative grade-point average, year in school, gender, and whether the course belonged to the student's major). In only one of these three courses were any of the discriminant functions statistically significant. Correlating the student characteristics with this function indicated that student year in school and grade expected in the course were the most highly associated, followed by cumulative grade-point average. Students in groups that were high on the function, compared to other students, tended to be freshmen and sophomores, to expect higher grades in the course, and to have higher cumulative grade-point averages.

Other researchers have searched for variation in students' background, experiences, values, attitudes, interests, and personality traits that might account for variation in ratings (although they have not used obverse factor analysis to identify different subgroups or types of students in a class with respect to patterns of responding across rating items). Stuit and Ebel (1952) directly asked 1,230 students at the University of Iowa whether certain things would seriously influence their judgment in rating their instructors. Approximately 46% of the students reported that liking or disliking the subject would influence their judgment, and 32% and 20% of them (respectively) felt that "personal like or dislike for the instructor" and their "standing in the course" would do so. Students at the University of Delaware were asked about vari-

ous aspects that might affect ratings of students in general (not necessarily their own ratings) (Purohit and Magoon, 1974). Over half of the approximately 250 students completing the questionnaire agreed that students expecting a higher grade in a course rate the professor and the course higher than those who expect a lower grade, and that male and female students do not rate the same rating item for the same instructor in the same way; and over half of the students disagreed that average rating for freshmen, sophomores, juniors, and seniors would not differ. A smaller proportion of these students (about one-third) felt that whether or not the course was elective or compulsory would have an effect on ratings.

Because students feel that certain aspects will influence their own or others' ratings does not mean, of course, that these aspects indeed do so when students actually rate their teachers and courses. Thus it is important to find out which experiences and attributes of students are, in fact, associated with their ratings. The results of the numerous studies in this area are briefly reviewed below. The emphasis of this review is on the strength of the associations that have been found as well as on the consistency of results across studies. The studies have been grouped by similarity of students' characteristics for which association with teacher ratings have been sought.

Unless otherwise specified, only studies dealing exclusively or primarily with undergraduate students at American and Canadian universities and colleges have been considered. Moreover, since interest lies in the correlates of the variability among individual students in the same classes, only studies in which the individual student is the unit of analysis are included and not studies in which the class or course itself is the exclusive unit of analysis. This procedure is important, for the two types of studies essentially ask and answer different questions (see Menzel, 1950). It is generally hazardous to draw inferences about the direction and strength of relationships at the group level of analysis (i.e., "ecological analysis") from the direction and strength of relationships at the individual level of analysis (i.e., "individual analysis"). At the level of individual analysis, two kinds of studies have been included for this review: those in which data or students have been pooled across classes or courses and those in which they have not. Including both has been done in order to gain sufficient studies from which to make generalizations and to draw conclusions, even though such pooling across classes potentially masks certain effects and mixes together within-class and between-class variation (see Doyle and Whitely, 1974; Feldman, 1976a; Linn, Centra, and Tucker, 1974; Sockloff, 1975; Whitely and Doyle, 1976).

Grades, Academic Facility, and Learning

From a previous review and analysis (Feldman, 1976a), the general conclusion was that students' expected and actual grades in a class were positively related to their ratings of the course and teacher. Considering only studies using correlational techniques of analysis, most of the correlations fall within a range encompassed by the mid 0.10s at one end and just below 0.30 at the other end, although there are a few studies in which the correlations are in the 0.30s, 0.40s, and higher. These correlation coefficients, assuming as they do the linearity of relationships, may somewhat underestimate the strength of the association between grades and ratings. In studies in which ratings by students in different grade categories were compared, the lowest ratings were usually (although not inevitably) given by students expecting or receiving a "D" in the course rather than by those expecting an "E" (or "F").

It was also concluded in this prior review that the grade-point average of students, as a general indicator of academic adeptness or facility, had little or no relationship with teacher or course ratings in various studies. By contrast, there is some evidence that the discrepancy between a student's grade-point average and his or her grade in a particular class is related to ratings, perhaps even a little more strongly so than the grade alone. Actual achievement of the student in the academic area covered in the case, as measured by "objective" and relatively standardized examinations and performance tests is also related to ratings at about the same strength (at best) as that found for the association between grades and ratings, whereas the student's perception of his or her own learning is associated much more strongly with course and teacher ratings.[4]

Interest and Motivation

Clearly related to teacher and course ratings are the students' interest in the subject matter of the course and their motivation to master it, especially when these dispositions are measured in direct ways. Thus there is a positive association—of substantial size in some studies—between items on a teacher rating form and students' report of their liking for or interest in the subject matter of the course, their wanting to take the course, and the effort they say they expend on the course, including the amount of out-of-class work they put into it (see Brooks, Tarvey, Kelley, Liberty, and Dickerson, 1971; Canter and Meisels, 1971; Christensen and Bourgeois, 1974; Doyle and Whitely, 1974; Granzin and Painter, 1973, 1975, 1976; Haslett, 1976; Miller, 1972, Ap-

pendix B; Pohlmann, 1972; Price and Magoon, 1971; Whitely and Doyle, 1976; Whitely, Doyle, and Hopkinson, 1973).[5] The degree to which the student's interest in the course and motivation to do well had been induced and maintained by the teacher, in contrast to these being dispositions brought by the student to the course, is generally not known from these studies (see Feldman, 1976a).

Indirect measures of students' interest and motivation in a course also are related to teacher and course ratings but not as strongly or consistently so across studies. Thus students taking a course as an elective tend to rate it and the teacher higher than do students taking it as a requirement; moreover, students taking a course in their major field or taking a course in an area or in a department in which they have already taken a number of other courses, tend to give higher ratings than do the other students in the class. It should be stressed that these particular relationships appear in some studies and not others, and even those that have occurred are generally very small in size. (A listing and somewhat more detailed review of relevant studies may be found in Feldman, 1976a; also see the following studies, all of which were located after that review: Kelley, 1972; Mallory, Huggins, and Steinberg, 1941; Miller, 1972, Appendix B; Perkins, 1971, Appendix D; Pohlmann, 1972; Pohlmann and Tuinen, 1972; Whitely and Doyle, 1976.)

General Impressions, First Impressions, and Preimpressions

There is some evidence that students' generalized impressions of instructors, instruction, and courses tend to be positively associated with their ratings of specific courses and teachers. Pohlmann (1972) reports correlations of approximately 0.20 in size between students' overall evaluation of instruction and classes at Southern Illinois University and their ratings of specific teachers and courses. In a study by Crowe (1974) at Purdue University, the direction of the association between these two sorts of ratings was also positive, but too small to be statistically significant ($r = 0.06$). Treffinger and Feldhusen (1970) found somewhat more substantial associations between students' general impressions of courses and instructors at Purdue University and their ratings of specific instructors and courses on various rating scales; in their study, the percent of variance in ratings that was "explained" by the variable of general impression ranged from about 4% ($r = 0.19$) to as high as 15% ($r = 0.39$). A possible interpretation of these results is that some students are predisposed to rate courses and teachers higher than other students, and they carry this disposition from course to course. Or perhaps certain students are just more generally impressed with the

teaching and courses they have had at their college and this general satisfaction "spills over" into the specific courses they are currently taking.

Of even greater strength than the association between general impression and ratings is that between initial impression of a specific instructor or course and the later ratings of either. Four studies were located in which undergraduates rated instructors or courses once at the very beginning of the semester and again near or at the end of the semester (Bejar and Doyle, 1975; Day, 1969; Kohlan, 1973; Widlak and Quereshi, 1972). Across these studies correlations between the two sets of ratings ranged from about 0.40 to 0.60 or so; therefore, between nearly one-fifth to over one-third of the variance in final ratings in these studies was accounted for by students' very early assessments.[6]

Not only initial impressions but also precourse impressions (and information) appear to be related to ratings. Miller (1972, Appendix B) found that students at Baldwin-Wallace College who had heard that their professors were good rated them higher than those who had heard otherwise. Bausell and Magoon (1972c; 1972d, Appendix U) found that on several rating items, both at the beginning of a course and its end, the students who had previously taken a course with the instructor rated him and the course more highly than did students who knew him only by reputation; the second group of students, in turn, rated the instructor and the course more highly than did students who reported knowing nothing about the instructor before taking the course. Similarly, Kohlan (1973) reports that students' degree of previous knowledge of the instructor and course was positively associated with the four ratings scales used in his study.

These findings presumably depend, in part, on a selection effect. Other things equal, students would be more likely to select courses taught by instructors with whom they had taken a course in the past (and liked), or about whom they have heard favorable reports, than to select courses with which they are unfamiliar or that are taught by instructors whom they do not know or about whom they have heard unfavorably. Having thus selected certain courses and teachers because of prior impressions and information, students might be more likely to rate them higher than would other students in the class. The most directly supportive evidence that this indeed is the case comes from a study by Mallory et al. (1941) of the reasons given by students (at a four-year liberal-arts college for women) for their choice of specific courses. It was found that students who chose particular courses because of their preference for the teacher were more likely to rank the course highly on various dimensions than were students who had other reasons for choosing their courses.

In addition to any direct effects these various impression factors have on ratings, it is possible that they also have indirect effects through their influence on students' motivation and interest. For example, perhaps variation in the general impressions that students have of their courses and teachers produces variation in the interest and motivation that students bring with them to specific courses, which in turn may create variability in their ratings. Likewise, a student's initial impression of the course and teacher might affect ratings through the influence it has on the student's interest in the course and motivation to do well in it. Similar reasoning would apply in proposing an indirect effect of precourse impressions on ratings—in conjunction with a selection effect. In this case, supportive evidence, albeit a little indirect, can be found in studies done at the University of Manitoba. Not only did different kinds of students at this university selectively register in different sections of a multisection course, but the ability and/or reputation of the particular section instructor was one of the primary reasons given by the students for their selection (Leventhal, Abrami, Perry, and Breen, 1975). Moreover, those students for whom this ability and/or reputation reason was of greater importance—and, thus, presumably those students who brought with them to the course greater interest and motivation—rated their teachers higher than did other students (Leventhal, Abrami, and Perry, 1976).

Year in School (College-Class Level)

A number of studies have compared the course or teacher ratings of students who differed in their year in college. Before reviewing the results of these studies, it should be noted that the exact college-class levels that are compared vary across the studies. In some of them, students in all four levels are compared (which, obviously, is only possible if all four level of students are taking a course). Other studies compare the ratings of fewer levels (for example, freshman versus sophomores). Some studies combine college-class levels before comparing ratings, creating additional variation due to differences in the composition of these combinations (for example, freshmen versus nonfreshmen; freshmen/sophomores versus juniors/seniors; seniors versus nonseniors).

Findings in this area are inconsistent. Some studies report essentially no relationship between student's class year and teacher or course ratings (Bausell and Magoon, 1972d, Appendix J; Delaney, 1976; Dick, 1967; Doyle, 1972; Elliott, 1950, also see Kapel, 1974; Remmers and Elliott, 1949; Office of Evaluation Services, 1972; Pohlmann, 1972;

Rayder, 1967, but see Rayder, 1968; Riley, Ryan, and Lifshitz, 1950; Sockloff and Deabler, 1971; Spencer, 1969; Walker, 1968; Whitely and Doyle, 1976). Other studies report a positive association (implicitly significant statistically, if not always explicitly so): The higher the class level of the student the higher the rating (Cooke, 1952; Downie, 1952; Doyle and Whitely, 1974; Frey, Leonard, and Beatty, 1975; Hillery and Yukl, 1971; Kohlan, 1973; Lovell and Haner, 1955; Lunney, 1974; Maas and Owen, 1973; Miller, 1972, Appendix B; Murray, unpublished; Perry and Baumann, 1973; Rosenshine, Cohen, and Furst, 1973). Still others find a negative relationship between class year and ratings (Bendig, 1952a; Centra and Linn, 1973; Christensen and Bourgeois, 1974; Cohen and Humphreys, unpublished; Crouch and Leathers, 1951; and Granzin and Painter, 1973).[7]

Within a few of these studies there are both positive and negative associations with class level, although the positive associations tend to outnumber the negative (see Hildebrand, Wilson, and Dienst, 1971, in conjunction with personal communication from Wilson; Carter, 1968; Bausell and Magoon, 1972d, Appendix I). Also, across these studies, statistically significant associations often are found for some of the rating items in a particular research but not for others, or for some of the subsamples within a study and not for others. Moreover, the strength of the significant relationships is generally very weak—with a few exceptions where the association between class level and ratings is rater substantial (see especially Centra and Linn, 1973; Frey et al., 1975). This lack of strength is most clearly seen in studies in which relationships are shown in the form of product-moment correlation coefficients,[8] wherein the proportion of explained variance in ratings generally ranges from 1% (and not always that) to 4% or so. Since correlational analysis assumes linearity of relationships, it is of interest that two studies report some evidence of a nonlinear relationship between class year and ratings. In both of them the lowest ratings were given by juniors, but in the first (Clark and Keller, 1954), the highest ratings were generally given by seniors whereas in the second (Nichols, 1967), the highest ratings were generally given by the freshmen.

Some of the inconsistencies across studies that have been noted in this section may be due, in part, to using year in school without taking into account the proportional distribution of the class level of students in a course. It may make a difference whether a certain class level (say, seniors) form a large majority of students in a class compared to those instances where they are only a minority. This is purely speculative, but may be worth further investigation.

Gender of Student

Many of the studies in which the ratings of male and female students in the classroom are compared find essentially no differences between the two groups (Bendig, 1953b; Caffrey, 1969; Christensen and Bourgeois, 1974; Cohen and Humphreys, unpublished; Colliver, 1972; Cooke, 1952; Corcoran, 1957; Delaney, 1976; Dick, 1967; Elmore and LaPointe, 1974; Granzin and Painter, 1973, 1975; Levinthal, 1974; Levinthal, Lansky, and Andrews, 1970; Lovell and Haner, 1955; Maas and Owen, 1973; Murray, unpublished; Null and Nicholson, 1972; Pohlmann, 1972; Riley et al., 1950; Singhal, 1968; Sockloff and Deabler, 1971; Spencer, 1969; Walker, 1968; Whitely and Doyle, 1976; Wilson and Doyle, 1976). Of the studies that do show statistically significant relationships between gender and ratings, the associations generally appear for only some of the rating items and not others, and are usually very small in size (Bausell and Magoon, 1972d, Appendix I; Bendig, 1952a; Centra and Linn, 1973; Crowe, 1974; Elliott, 1950, but see Remmers and Elliott, 1949; Elmore and LaPointe, 1975; Hildebrand, Wilson, and Dienst, 1971, and personal communication from Wilson; Kapel, 1974; Kelley, 1972; Kennedy, 1971, 1972; Kohlan, 1973; Office of Evaluation Services, 1972; Perkins, 1971; Perry and Baumann, 1973; Quereshi and Widlak, 1973, also see Widlak and Quereshi, 1972; Rayder, 1967, but see Rayder, 1968; Scott, Halpin, and Schnittjer, 1974; Touq, 1972; Touq and Feldhusen, 1973; Walter, 1971, also see Null and Walter, 1972).

Considering only studies where rating differences between males and females are statistically significant, most of them find that women students rate the teacher or course higher than do men. The results of three studies (Bausell and Magoon, 1972d, Appendix I; Bendig, 1952; and Kapel, 1974, in conjunction with personal communication from him), where just the reverse was true, are clear exceptions. A few studies report "mixed" results, in that women in the class(es) rate the teacher or course higher than do men on certain of the items whereas men rate the teacher or course higher on other of the items (usually only one or two of the items), with the remaining items in each of the studies showing no differences (see Doyle and Whitely, 1974; Haslett, 1976; Nichols, 1967; Rosenshine et al., 1973).[9]

Interactions between gender and other attributes of the student have been found to affect ratings, with some (but not complete) consistency across studies. Thus, statistically significant interaction effects between gender, on the one hand, and personality traits, attitudes and values, on the other, have been found—at least as the characteristics are

measured by the Edward Personal Preference Inventory (Rezler, 1965), the California Psychological Inventory (Carney, 1961), and the Allport-Vernon-Lindzey Study of Values (Walter, 1971; also see Null and Walter, 1972). However, Corcoran (1957) did not find an interaction effect of gender and dogmatism on course ratings. Bausell and Magoon (1972d, Appendix I) and Kohlan (1973) found evidence of an interaction effect on ratings between gender and year in college, although Bendig (1952a) did not. Quereshi and Widlak (1973; also see Widlak and Quereshi, 1972) found a statistically significant interaction effect on ratings between gender and class grades of students, as did Haslett (1976) between students' gender and their interest in and knowledgeability of the area of the course; and Scott, Halpin, and Schnittjer (1974) found the student's "academic achievement status" to be related to ratings for men but not for women. In other studies, however, interaction effects on ratings were not found between gender and anticipated grade in the class (Levinthal, 1974) or gender and overall grade-point average of the student (Kohlan, 1973).

Personality Traits, Attitudes, Values and Related Characteristics of Students

Researchers have searched for associations between teacher or course ratings and the personality traits, interests, preferences, opinions, attitudes, and values of students. With the exception of a study by Kovacs and Kapel (1976), in which need for achievement and feelings of personal control were related to certain rating items and scales, studies of one or two of these sorts of personality and related characteristics have not found them to be related to ratings of courses and teachers. Thus, although Freehill (1967) did find that students scoring high on authoritarianism were more critical with their college experience (including instructors in general), neither Maney (1959) nor Corcoran (1957) found degree of a student's authoritarianism to be related to the actual ratings of specific teachers and courses. Nor have these ratings been found to correlate with a student's score on the Bills' Inventory of Adjustment and Values (Riechmann, 1974), with one's score on Rotter's Internal-External Control Scale (Crowe, 1974), with personality type as measured by the Meyer-Biggs Type Indicator (Blank, 1970), with either the Achievement-via-Independence or Achievement-via-Conformance Scales of the California Psychological Inventory, with "achievement orientation" and "social orientation" as measured by particular clusters of scales of the California Psychological Inventory (Carney, 1961), with measures of creativity (Scott et al., 1974), or with a measure of the degree to which a person has an "abstract" rather

than a "concrete" personality structure (Tuckman and Orefice, 1973).

By contrast, studies that have used a number of dimensions and indicators of the personality, attitudes, or values of students (and usually a number of different rating items or scales) have found certain associations between the characteristics and ratings (Bausell and Magoon, 1972d, Appendix W; Grande and McCollester, unpublished; Kennedy, 1971, 1972; McKeachie, 1973; Phillips, 1960; Potter, 1969; Rezler, 1965; Walter, 1971, also see Null and Walter, 1972; Weinstein and Bramble, unpublished; Yonge and Sassenrath, 1968. Of course, when there are a relatively large number of comparisons within a study, some statistically significant results can be expected by chance alone. Calculating the percent of associations in each study that were statistically significant, for those studies where this could be done, reveals that some of the studies have proportions of statistically significant associations of less than 5% (Walter, 1971, also see Null and Walter, 1972; Potter, 1969); thus the significant associations that appeared in these studies were most likely chance occurrences. In the other studies (Grand and McCollester, unpublished; Kennedy, 1971, 1972; McKeachie, 1973; Phillips, 1960; Yonge and Sassenrath, 1968), the proportion of statistically significant results is more than 5%. Even so, it cannot definitely be said that the number of significant results in these studies is larger than would be expected by chance alone, since within each study the personality traits or related characteristics are most likely interrelated in various degrees; thus the associations of these variables with ratings are not independent from one another. Across these studies, the proportion of statistically significant results varies from just a little over 5% to a little under 25%. To take only one of these studies, consider that by Yonge and Sassenrath (1968), in which product-moment correlations were calculated between students' scores on each of 14 Omnibus Personality Inventory (OPI) scales and each of nine rating scales, for each of three teachers. This procedure produced 378 correlations, of which 56 (or approximately 15%) were statistically significant. These significant correlations were not unsubstantial in size, ranging as they did from 0.18 to 0.50, with a median correlation of 0.32 (one of the largest of the typical associations found in the various studies in this area).

It might be thought that characteristics of students that more closely reflect their learning styles and attitudes about the classroom would be related more consistently and strongly to their ratings than would less specifically relevant attitudes, values, and personality traits. It is true that Riechmann (1974) found certain types of students learning styles to be related (at moderately substantial levels) to multi-item scales as well as an overall teacher rating. However, the items on the inventory

measuring the various styles are so similar to teacher and class assessment items that positive correlations with actual teacher ratings forms are not surprising. Moreover, in the same study, the Jensen Inventory of Classroom Activity Preferences (designed to distinguish between students who most consistently prefer "student-centered" instructional methods and those who prefer "teacher-centered" methods) was not related to any of the teacher evaluation scales or overall ratings. Similarly, as part of a study by Corcoran (1957), the Preferred Instructor Characteristics Scale (a measure of relative preference for "cognitive" instruction versus "affective" instruction) was found not to be related to the course (lecture) rating scale of the study.

Various personality, attitudinal, and value characteristics of students do seem to interact with certain other of their experiences and attributes. Thus several studies testing for significant interactions between students' actual or expected grades in the course and personality or related characteristics did find them (Blass, 1974; Corcoran, 1957; Page and Roy, 1975; Walter, 1971, also see Null and Walter, 1972). Furthermore, Walter (1971, also see Null and Walter, 1972), Rezler (1965), and Carney (1961) all found statistically significant interactional effects on teacher ratings of gender and certain personality traits, attitudes or values (but see Corcoran, 1957, in which such effects were not found).

It is hard to generalize substantively across the studies reviewed in this section because of the variation in the personality and related characteristics measured and the wide variety of indicators used to measure them. Moreover, results are not always consistent across studies (or even within studies, for those studies presenting data separately for each teacher or course). Direction and content differences seem dependent on the nature of the rating items, the specific personality or related characteristics measured, differences in experiences and other attributes of the student, and the particulars of the courses and teachers.

Combinations of Student Attributes (R and R^2)

Of the studies located for this review, most of them deal separately with one or another of the experiences and attributes of students discussed so far, even when two or more of these variables have been included as part of the research. Some, but not many, of the studies have searched for interaction effects between certain of these variables, the results of which have been reviewed. Nor have many studies taken two or more of these variables together in order to explore the association between teacher ratings and the linear combinations of various of these student experiences and attributes. In studies that do so, the dependent

(or criterion) variable of teacher or course ratings is regressed on some set of student experiences and attributes (the predictor variables). Across studies by Crowe (1974), Doyle and Whitely (1974), Kelley (1972), Menard (1972), Pohlmann (1972), Rayder (1967, 1968), Scott et al. (1974); Sockloff and Deabler (1971), Treffinger and Feldhusen (1970), and Widlak and Quereshi (1972), the multiple correlation coefficients (R) range from a little above 0.10 to around 0.50 and even higher in a couple of instances—with the variance in teacher or course ratings explained by the set of student characteristics (R^2) thus ranging from between 2% or 3% to as much as 25% or even a little more—depending, among other things, on the population studied, the nature of the variables included in the predictor battery, and the particular rating item or scale serving as the criterion variable. Granzin and Painter (1973, 1975) and Riechmann (1974) report even higher R's and R^2's, but undue importance should not be attached to the size of the multiple correlations, since some of the predictors in the predictor set are themselves so similar to (or actually are) teacher or course rating items that rather large R's and R^2's would be expected.

The Student-Teacher Match

The match between students and teachers has also been studied as a possible source of variation in teacher ratings. For example, instances appear in the extant research in which some teachers at a school are rated more highly by the men than by the women in their classes, whereas other teachers at the same school are rated more highly by the female students than by the male students in their classes (see Bendig, 1953b; Potter, 1969). Although the factors that may be causing such differences are not known, the gender of the teacher (in comparison to that of the student) has been suggested as relevant and important. There are hints that under some circumstances similarity of teacher-student gender is associated with higher ratings, although the evidence from the few relevant studies is essentially inconclusive.

Both in a study by Ferber and Huber (1975) in which students rated teachers they had had in previous semesters, and in a study by Walker (1968) in which students rated teachers in whose classes they currently were, the highest ratings were received by female teachers rated by female students. In the first study, the lowest ratings were received by women teachers being rated by their male students, whereas, in the second study, the lowest ratings were received by men teachers being rated by their female students. In a study by Elmore and LaPointe (1975), only one of twenty rating items consistently showed a gender-of-student by gender-of-teacher interaction effect across the four sub-

analyses contained in the report: Women students rated women instructors higher than they did men instructors on "showed an interest in students," whereas men students rated men instructors higher than women instructors on this trait. Finally, in other studies, an interaction effect on ratings between the gender of the teacher and that of the student was generally not found (Bausell and Magoon, 1972d, Appendix R; Elmore and LaPointe, 1974; Levinthal, 1974; Wilson and Doyle, 1976). Wilson and Doyle (1976) suggest that, even though gender interactions may not be a particularly common occurrence, it is still possible that these effects may arise in certain specific kinds of situations—for example, a course on sex roles taught to a diversity of students from a strongly feminist (or antifeminist) point of view. If so, further research is needed in this area to establish the kinds of specific situations in which such interactions are to be expected.

Researchers have also explored the significance for teacher ratings of the match between students and teachers regarding their perceptions, values, attitudes, personality traits, and related characteristics. With the exception of Blank (1970) several investigators have found certain statistically significant associations between ratings and the *actual* similarity between students and teachers in these characteristics (Crowe, 1974; Good 1971; Good and Good, 1973; Levenson and LeUnes, 1974; Lewis, 1964; McDaniel, 1972; Purohit and Magoon, 1971; Taylor, 1968; also see the study by Menges, 1969, in which graduate students were the population studied). But these associations are found for some rating items and not for others, or are found in some classrooms and not others within certain of the studies; moreover, in most of the studies, the associations that have appeared are generally not very strong. Degree of *assumed* similarity (on the student's part) between the student's own characteristics and those of their teachers has also been found to be positively associated with teacher ratings (Day, 1969; Good, 1971; Good and Good, 1973; Fulcher and Anderson, 1974). Not surprisingly, in the one study where direct comparisons can be made (Good, 1971; also see Good and Good, 1973), assumed similarity is related to more items on a teacher rating form, and more strongly so, than is actual similarity.

There may be circumstances under which students might tend to rate the instructor more positively the more dissimilar they perceive him or her to be to themselves, if the divergence positively favors the instructor—that is, if the students regard the instructor as being superior in certain desirable traits. Davison (1973) and Riechmann (1974) asked students to rate themselves and their instructor on the trait adjectives in the Bills' Index of Adjustment and Values. In both studies, students who perceived the instructor as being superior to themselves

on the traits tended also to rate the teacher higher on the rating form. Grush, Clore, and Costin (1975) were able to predict successfully on just which personality traits such "positive dissimilarity," as they put it, would be associated with a three-item global rating of the instructor.

The similarity or dissimilarity between teacher and student attributes can be analyzed as part of the somewhat broader context of the "fit" or congruence between the student and the teacher (or course conditions). Three studies were found—all of which were essentially "field experiments"—that used this broader framework. In one of these studies (Domino, 1971) it was found that when the teaching style of the instructor was consonant with the students' achievement orientation (compared to when it was dissonant), students gave a higher overall rating to the course and rated the teacher as more effective. In the second of these studies (Parent, Forward, Canter, and Mohling, 1975), students in "mini-courses" in which the conditions of class discipline were congruent with the type of discipline they preferred were more satisfied with the "mini-course." Finally, Tuckman, and Orefice (1973) found some evidence that students tended to be the most positive about methods of instruction and instructional experiences that more closely "matched" their own personality structures and the least positive about those that did not.

The notion of "fit" or consonance can be extended to encompass the nature of instructors orientation to, and interaction with, different types of students in class. Elliott (1950) compared chemistry instructors who were relatively more effective with the higher-ability students than with the lower-ability students in their classes with chemistry instructors who proved to be more effective with the lower-ability students than with the higher-ability students in their classes. (Effectiveness was measured by the students' actual achievement in chemistry, controlled for their ability.) For the first group of instructors, the higher-ability students were more likely to rate the instructors higher on a variety of rating items than were the lower-ability students; for the second group of instructors, just the reverse was true. Elliott suggests that these differences may have been due to the two groups of teachers differing in the level at which their teaching was pitched. Consistent with this interpretation is the report by Wilson (1932) that some of the instructors in his study who had two different types of students in their classes apparently adapted their teaching methods to one or the other group, receiving, in turn, higher ratings from the group to which they so adapted. The findings of these two studies suggest, but do not directly document, that teachers interact in different ways with different kinds of students. Some documentation of such differences can be found, however, in a study by Mann and his associates (1970), an intensive

analysis of the interpersonal events and social interaction in four different classrooms over the course of a semester. There is also much evidence throughout the analysis that different types of students did vary in their perceptions and reactions to their teachers; moreover these differences were to some extent mirrored in the students' ratings of the teacher and the course (see especially Appendix A of the report).

RELIABILITY AND BIAS: THE OBJECTIVITY/SUBJECTIVITY OF STUDENTS' RATINGS

To this point, it has been shown that students are only moderately consistent in rating their teachers, and that the resultant variability in ratings within classes is associated with various student experiences and attributes. For some of these variables, associations are very weak as well as inconsistent across studies; for others, the associations are stronger and more consistent. Not only do these variables "explain" some of the variation in ratings when they are taken individually (in lesser or greater degree), they also do so when taken interactionally and in combination with one another. Furthermore, the match between teacher and student also is associated with teacher ratings under certain conditions. The question of whether these various associations indicate that students' ratings are biased is now examined. For this discussion, it is necessary first to explore the nature of these ratings in terms of their objectivity as opposed to their subjectivity.

Some Important Dimensions Involved in the Objectivity/Subjectivity of Ratings

At least one of the following three elements is stated or implied in discussions of the objectivity versus subjectivity of ratings (see, for example, Crichton and Doyle, 1975; Ghiselli and Ghiselli, 1972; Nunnally, 1967, Chap. 13; Sockloff, 1973; Wiggins, 1973, Chap. 6; Wilson, 1932):

Description-Evaluation. Presented in terms of the two ends of a continuum, ratings can be distinguished by whether students in essence are neutrally describing the attributes of teachers and courses or whether they are giving their evaluative reactions to them (see Coombs, 1964, pp. 334–341; Feldman, 1976b; Follman, 1975; Frey, 1974; Ghiselli and Ghiselli, 1972; Halstead, 1972; Levinthal et al., 1971; Menges, 1973; Rumery, Rhodes, and Johnson, 1975; Walter, 1971, pp. 16–17). With respect to the way in which rating items are framed, the distinction is exemplified by the following contrasting items: "To what degree was the course material organized by your instructor?" (to which the stu-

dent checks one of the following four alternatives: highly organized, somewhat organized, somewhat disorganized, highly disorganized) versus "How satisfied were you with your instructor's organization of the course?" (highly satisfied, somewhat satisfied, somewhat dissatisfied, highly dissatisfied).

Items on a teacher rating form may request a judgmental reaction without explicitly mentioning evaluation. Thus, if the student is asked whether the amount of assigned material for the course was "excessive," "just right," or "too little," the very categories available for response imply the student's judgmental reaction rather than neutral description of the amount of reading required in the course. Moreover, the mere use of the word "describe" or its equivalent does not automatically make the rating nonevaluative, especially if the item's content is global in nature. Thus the student may be asked to "describe" the degree of the teacher's overall effectiveness by marking an appropriate category ("highly effective," "somewhat effective," etc.), but it is unlikely that a "pure" description of the teacher's effectiveness will result. Rather, it is more likely that the student's overall evaluation and degree of approval of the teacher and course will be elicited.

It has been argued that ratings are rarely if ever pure descriptions, even under the best of conditions. Ghiselli and Ghiselli (1972) maintain that "ratings must be considered to be reports by the rater of his impressions or perceptions of the stimulus person, his opinions or judgments about him, and not reports of the objective, tangible properties of the stimulus person" (p. 270). Follman (1975) takes a somewhat less extreme position, arguing that ratings probably lie between description and evaluation. Even if it were true that the "pure" descriptive end of the dimension cannot be reached—and not all would agree—it nevertheless would still be true that rating items could vary in this regard, with some being closer to one or the other end of the continuum (depending, in part, on the content of an item and the way it is worded). Presumably rating items that are constructed to maximize neutral descriptions are more likely to elicit such descriptions than do items explicitly requesting students' evaluation and satisfaction. (Of course, evaluative judgments can still be inferred by administrators, researchers, other students, and teachers themselves from what are primarily students' descriptions of their teachers, but that is another matter. See Feldman, 1976b.)

Nonpersonal-personal stance. A second dimension—most clearly applicable to rating items involving teachers orientation to, and interaction with, students[10]—is the nature of the stance to be taken by the students in rating the teacher or course. The distinction here is between taking the stance of the group of students (in effect putting one-

self into the place of the other students, or the typical student, in the class) and taking a personal stance. This dimension is applicable to either the description or evaluation of the teacher and course.

In terms of description, the two extremes of this dimension are illustrated by the following two "stems" of rating items: "Describe the degree to which the instructor stimulated the interest of the class in the course material" versus "Describe the degree to which the instructor stimulated your interest in the course material." An ambiguous (in-between) case is one where it is not clear which of the two stances the student is to take: "Describe the degree to which the instructor stimulated interest in the course material." Although it is hard to know how often this sort of ambiguous item appears in teacher rating forms, it seems fair to say that it is not a particularly rare occurrence.

The two ends of this dimension for students' evaluations and satisfactions are similar to those for their descriptions—for example, "How satisfied was the class with the degree to which the teacher stimulated interest in the course material" versus "How satisfied were you with the degree to which the instructor stimulated your interest in the course material." Again, the ambiguous in-between case—"How satisfactory was the degree to which the instructor stimulated interest in the course material"—is probably not uncommon.

Amount of inference. A third dimension (cross-cutting the other two dimensions) is based on the amount of inference students must make—in either their description or evaluation, whatever the stance that is taken. Although there is not a one-to-one correspondence, the degree of inference is generally lower (a) the more visible to the student are the attributes of the teacher to be assessed and the more direct the information that the student has of these attributes, (b) the more a student is asked to consider behavioral attributes of the instructor rather than predispositional or attitudinal attributes, and (c) the more molecular (less molar) the behavior to be rated, or the more specific (less global) the attitudinal attribute of the teacher to be assessed (see Crichton and Doyle, 1975; Wiggins, 1973, Chap. 7).

Implications—Interrater Consistency (Reliability)

The more that evaluation is elicited by the rating item, the more that a personal stance is to be taken, and the greater is the degree of inference required on the student's part, the less "objective" is the rating and, presumably, the more random and systematic error that is created. The attempt to make ratings more objective—for example, by asking students to rate descriptively various delimited and visible behavioral attributes of teachers, requiring low inference and a nonper-

sonal stance from the student—is in part an attempt to bring about increased interrater reliabilities as well as other benefits (see Frey, 1974; Greenwood, Bridges, Ware, and McLean, 1973; Harari and Zedeck, 1974; McInnis, 1966; also see Thorndike and Hagen, 1969, Chap. 13, for a detailed discussion of the general problems involved in obtaining sound ratings and of the methods for improving the effectiveness of ratings, whether done by students in the classroom or by persons in other settings).

At least for student ratings of teachers, however, there is little direct and systematic evidence, one way or the other, that an increase in the objectivity of ratings brings about an increase in interrater reliability. The results of a study by Sharon and Bartlett (1969) do offer some support, although it is somewhat indirect. These investigators report a greater degree of interrater consistency (as indicated by intraclass correlation) for the one set of students in each class using graphic scales to rate teachers on 60 items than for the other set in these same classes who rated the teachers on exactly the same items grouped into 15 forced-choice tetrads. Since the forced-choice procedure pairs items that have been equated on degree of "favorability" or "social desirability" (Edwards, 1957), the forced-choice format may be assumed to lead to less subjective ratings, in the sense of involving less "evaluation," than do forms not controlling favorability or social desirability (Sharon, 1970). If so, then the interrater consistency in this study was greater for the less subjective ratings.[11]

Many of the techniques that are recommended for making student ratings of teachers more objective and more reliable are the same as, or similar to, those found in research using systematic observation procedures (Heyns and Lippitt, 1954; Medley and Mitzel, 1963; Weick, 1968). Of course, procedures used in teacher-rating settings cannot fully duplicate certain methods found in direct-observation studies—such as systematic sampling of the stimulus persons or events to be observed, observation (and consequent rating) that is immediate rather than retrospective, and extensive training of the observers and raters. Even so, techniques have been proposed that would modify conventional rating procedures in order to make them a little closer to the techniques typically used in direct-observation projects. For example, Guilford (1959, Chap. 7) and Thorndike and Hagen (1969, Chap. 13) discuss ways that raters in classroom or analogous settings can be trained, if only minimally. In this connection, Halstead (1972) documents actual differences in ratings of the same college teachers done by students who participated in training sessions lasting only 30 minutes compared to students who were not trained at all—differences interpretable as showing that the minimally trained students were more

objective in their ratings, in the sense of their being more descriptive and less evaluative in their ratings.

Implications—Bias in Student Ratings?

In response to the question of exactly which characteristics of students associated with teacher or course ratings should be considered as biasing elements in the ratings, one answer that has been given are those variables that are "irrelevant" or "extraneous" to the assessment of the teacher or the course (for example, see Murray, unpublished; Aleamoni, 1974). The criterion for irrelevance or extraneousness is usually put in terms of "non-teaching factors that affect student ratings" (Sheehan, 1975). For example, Crittenden and Norr (1975) distinguish between "factors theoretically related to teaching" and biasing factors. Similarly, Gillmore (1973) points out that nonbiased ratings are those "ratings given a course [that] are reflective of the content and teaching of that course, and not influenced greatly by noninstructional factors" (p. 22). Given a criterion of this sort, all the correlates of teacher ratings reviewed in the earlier section of the present analysis could be viewed as biasing elements—with the possible exceptions of any learning and motivational characteristics of students directly prompted by the teacher or course. Although these excepted correlates are not instructional factors per se, they are direct consequences of such factors, and therefore could be seen as legitimate influences on student ratings. Something of this reasoning seems to lie behind the argument that an association between grades and teacher ratings is not necessarily indicative of a bias in rating. The argument is that the association may be explainable in terms of differential student interest in the course and motivation to do well, which are a direct consequence of the teacher's attributes and actions, and which, in turn, tend to produce both better grades on the part of some students in the class and higher ratings of the teacher by the same students (Feldman, 1976a).

The matter is more complex really, being contingent on whether ratings are objective, subjective, or a mixture of both. If, on the one hand, ratings are meant or claimed to be objective, then ideally none of the attributes of the students nor any of their class experiences should be related to ratings. Any that do are biasing results. Clearly students' anticipated grades in the course, interest and motivation brought to the course, certain predispositions, and the like, should not be related to objective ratings of the teacher or course. Neither, for that matter, should motivation, interest, and learning induced by the teacher. Even if some students were more inspired by the teacher and learned more

from him or her (or thought they did), these experiences should not affect neutral descriptions and assessments of the teacher's degree of preparation and organization of the course, knowledge of the subject matter, or any of the specific areas in which students are asked to rate their teachers. Indeed, it is arguably the case that these class experiences also should not be related to overall ratings of the general "effectiveness" of the teacher, if these global ratings are meant to be objective assessments.[12]

If, on the other hand, ratings are the subjective assessment of the teacher—either theoretically or in practice—then teacher-inspired motivation and teacher-induced learning would be expected to be associated with students' evaluation of the teacher's overall "effectiveness" as well as some of the more specific areas of the teacher's performance. Considered more generally, other of the student's characteristics and experiences might also be expected to correlate with the student's evaluation of the teacher, given the general theory and research on the personal factors that affect individuals' perceptions and evaluation of the qualities and behaviors of others (Taguiri, 1969; Taguiri and Petrullo, 1958; Warr and Knapper, 1968).

Indeed, in recent years, certain analysts of teacher ratings have in effect taken these ratings to be subjective assessments—at least partially so—when they suggest that all of the variability of student ratings in a class is not necessarily "error" (see Bejar and Doyle, 1977; Crichton and Doyle, 1975; Magoon and Price, 1972; Majer and Stayrook, 1974) and that student characteristics that are associated with ratings need not be regarded as biasing elements (see Crittenden and Norr, 1973; Yonge and Sassenrath, 1968). Student ratings, it has been suggested, must be analyzed as the inevitable resultant of the characteristics of the student raters themselves, the characteristics of their teachers and courses, and the context in which the ratings are made (see Bejar and Doyle, 1977; Centra and Linn, 1973; Follman, 1975; Haslett, 1976; Kerlinger, 1963; Norr and Crittenden, 1975; Riechmann, 1974, Chap. 4; Tetenbaum, 1975; Yonge and Sassenrath, 1968).

Given that students, like other people, view and react to those around them through a screen of their own values, preferences and experiences, it has been proposed that students should not be seen as totally "impartial recorder[s] of events" in the first place, as Yonge and Sassenrath (1968) put it. Within this orientation, some degree of inconsistency among students in their evaluation of teachers is considered reasonable; such inconsistency reflects a genuine source of individual differences among students, under the assumption that a given teacher differentially appeals to different students in class. Differences

in certain of the attributes and experiences of students may indeed be a source of variation in their appreciation and evaluation of various aspects of the course and teacher, but they are seen as "legitimate inputs to the evaluation process" (Crittenden and Norr, 1973, p. 144) rather than sources of bias. Crichton and Doyle (1975) offer the following analysis:

> The psychometric literature . . . reveals a universal attitude of excluding all rater effects from true variance and therefore concluding that reliability means . . . relative absence of both random error and rater differences. The results of considering raters the source of at least some true variance . . . must be explored. . . . The traditional theory of reliability of ratings assumes that there exists a true value on a given trait for the ratee which every rater, if he is not biased or unmotivated or careless or unobservant, will give the ratee. This ignores the possibility that there may be a different "true" value for each student, for example, because the instructor satisfied his needs or desires with respect to the function named to a differing degree. This would imply the presence of ideographic true variance, true specific rating components of varying magnitude across raters. There should be inconsistencies among raters [even] if they rate without error" (p. 19, pp. 27–28).

These various suggestions do not imply that students rate without errors, random or systematic, but that it is analytically possible—and hopefully empirically so—to separate rater error from true rater variance. Researchers and practioners would still try to eliminate or reduce both random and systematic error by such procedures as making the rating items clear and easy to respond to, by using the most effective rating format, by asking about things the student has actually experienced in the classroom, by giving students the same set in using the rating scales, by trying to give students a uniform level of motivation to respond as well as they are able, and the like.

Nor should the various suggestions in this area be taken to mean that correlates of ratings are without consequence and therefore can be ignored. These correlates should still be taken into account when interpreting the ratings of teachers and courses. If different types of students are reacting somewhat differently to the teacher and course, then an average of their subjective ratings may be hard to interpret, since the average may not well represent any particular type of student. Thus the practice at the University of Kansas, as reported by Hoyt, Owens, and Grouling (1973), makes a good deal of sense. When results from the rating form used at the school are returned to individual teachers, average ratings on the items are given not only for the class as a whole but also separately for those students expecting "A's" and "B's" in the class (compared to those expecting lower grades) as well as for those students who attended class regularly (compared to

those who did not). Moreover, the manual that was developed to help faculty interpret the ratings includes norms for ratings in classes where the typical student's motivation to take the class was high and for classes where this motivation was low.

If the proportional distribution of different types of students varies across classes, then comparison of average evaluations across these classes is ambiguous since differences among these ratings may be due to differences in the proportion of various kinds of students in the class as well as to differences in the teachers and courses. Under such circumstances, it would help to "adjust" the evaluations for relevant student differences, so that class ratings could be more meaningfully compared with one another. Hoyt and Spangler (1976) have done just this in their study of the relationships between the research involvement of instructors and the student evaluation of their classes: Ratings of instructors were adjusted for students' initial desire to enroll in the course (by using this variable as the covariate in an analysis of covariance).

Assuming that in practice most student ratings are not altogether "objective" (just as they are not completely "subjective"), it would seem reasonable to search for ways "to separate the subjective component (depending to some degree on the rater) from the objective component (depending only on what the ratee does) in an individual rating," as Crichton and Doyle (1975, p. 21) put it. These authors suggest that being able to discriminate among the reactions of subgroups of different kinds of students would be a substantial beginning:

> ". . . perhaps the most realistic strategy to make composite [e.g., average] ratings—and the individual ratings which compose them—more useful would be to try to (a) minimize systematic and random error and (b) then find subgroups within which all total rater contribution (including the error component) approach a constant, or equivalently, in which the observed ratings approach equality. . . . Perhaps the groups will have distinguishable characteristics which will both be an aid in interpreting their ratings and lead to the development or discovery of an external instrument to identify kinds of raters to aid in the interpretation of ratings gathered in the future. Conversely, perhaps it will be possible to group raters according to some theory of how they will rate in a particular situation, and their ratings within subgroups will be more uniform than ratings within the total group" (p. 22).

It may be noted here, that if subgroups are to be formed, it would be important to compose them (if at all possible) of students who are similar with respect to the attitudes and behaviors they desire in a teacher. Doing so would help to assure that students within the subgroup are similar in the way in which they "translate" the degree of discrepancy

between what they prefer in an instructor and what they see the instructor as giving them into a particular rating—to help assure, in short, what Coombs (1964, Chap. 16) calls the "interpersonally comparable utility" of evaluations. This is important since averaging evaluative ratings assumes their interpersonally comparable utility, an assumption that Coombs notes is usually neither specified nor particularly warranted. (For some empirical work in this area, see Levinthal, 1974, and Levinthal et al., 1971.)

CONCLUDING COMMENTS

In many of the studies of college students' ratings of their teachers and courses, interest resides in the degree of consistency among students as a means of establishing the interrater reliability of these ratings. The present review and analysis has taken the reverse tack. Published information about interrater reliabilities has been used as the starting point for an analysis of the consistency among students in their ratings.

As indicated by the reliabilities of individual ratings or the single rater, consistency among students in their ratings is moderate, at best. In this regard, any one student's rating of his or her teacher or course is of limited usefulness. By contrast, under the assumption that students are "independent replicates" (except for random error), these modest interrater associations do produce substantial reliabilities when the ratings of 20 to 25 (or more) students in a class are averaged together, thus indicating that the average or composite ratings of teachers and courses are rather dependable measures. The estimates of average reliabilities are probably somewhat inflated, however, since there is a sense in which students in a class do not rate their teachers and courses with complete independence: The objects of the ratings (particular teachers or courses) are the very entities about which students, in part, mutually influence each other's assessments as the semester progresses. Although the empirically determined consistency in ratings among students within classes is sufficiently high for average or composite class ratings to be taken seriously, it should be kept in mind that this consistency may be due to indirect student collaboration and jointly produced consensus in addition to similar decisions that have been individually and independently reached. Because the amount of observed consistency among ratings may thus be an "impure" base for estimating the reliability of average ratings (or, for that matter, the reliability of individual ratings), interpretations of these consistencies and of the reliability estimates based on them must be made with caution.

Not only may students not be completely independent in their ratings, they may not be total "replicates" either. The possibility exists of patterned differences in ratings linked to differences in student types within classes. Students' ratings are based on what has been called "retrospective naturalistic observation," wherein the rater is called on to recollect earlier observations made in naturalistic settings (Wiggins, 1973, pp. 296–298). Given this circumstance, together with the fact that students are not trained as either observers or raters, it might well be expected that ratings done in typical classroom settings would be dependent to some extent on the characteristics and experiences of the student observers (Wiggens, 1973, Chap. 7). There is evidence that this is the case. Students are not altogether interchangeable as raters, and within-class variability seems not to be due to random error alone.

Although certain attributes and experiences of students are usually only weakly related to their ratings, and inconsistently so, across studies (the student's gender, grade-point average, year in school, and certain personality and related traits), other attributes and experiences are more strongly and more consistently related (the student's anticipated grade in the course, achievement in the content of the course, interest and motivation, general impression of teaching and courses at the school, and prior and initial impressions of the particular teachers to be rated). Whether strongly or weakly associated with ratings when considered separately, various correlates of student ratings interact as well as linearly combine with one another to "explain" variation in ratings, again in varying degrees of consistency and strength across studies. Moreover, certain kinds of "fit" (both perceived and actual) between teachers and different students in their classes are related to ratings. Despite the fairly large amount of research in the area, these generalizations about the consistency and size of relationships should be taken as suggestive rather than definitive, since the data on which they are based are not ideal. For example, although the studies of correlates that have been reviewed all use individual students as the unit of analysis, some of them pool students and data across classes; this procedure may mask useful information and does not give a "pure" indication of within-class relationships.

Whether the various correlates of within-class ratings are to be interpreted as "biasing" elements or as "natural" influences on social perception depends in large part on whether student ratings are claimed to be objective or subjective (descriptions of low inference from a nonpersonal stance compared to evaluative reactions of high inference from a personal stance). If these ratings are meant to be objective descriptions, then an argument can be made that any characteristics and experiences of students that relate to these ratings are biasing

factors since unwanted, systematic variance (error) has been introduced into the ratings. Particularly when the association between certain of these biasing factors and ratings become sizeable, as they can, steps should be taken to eliminate them if the claim for the objectivity of the ratings is to be warranted. Barring elimination, these factors need to be taken into account somehow (for example, by adjusting the ratings for their effects) for the ratings to be useful or meaningful.

If, on the contrary, the ratings are designed not so much to obtain objective descriptions of teachers and courses but to measure the subjective reactions of students to them, as important information in its own right, then some of the patterned variability in ratings represents so-called true variance and not systematic error. In this orientation, differences among the background, characteristics, and experiences of students are seen as legitimate or genuine sources of influences on their ratings. Even so, the more significant correlates should still not be ignored, since it is difficult to interpret and compare these evaluative ratings without knowing at least the approximate contribution to the ratings of the actual behavior and attitudes of the teacher(s), the characteristics and experiences of the students, and the properties of the context in which the ratings were made (Kulik and Kulik, 1974).

The attempt of the present analysis has been to raise certain issues concerning the consistency of student ratings of teachers and courses, and to clarify some of the problems involved in trying to resolve them. The analysis of consistency in ratings has hardly been exhausted thereby, for only interrater consistency has been considered. Other kinds of consistency—including consistency of students' ratings over time, intrastudent consistency across items of a rating form, consistency between ratings of teachers by their students and by other types of raters (for example, their colleagues), and consistency of ratings across different contexts and conditions—must also be carefully analyzed if student ratings are to be used appropriately and interpreted meaningfully.

FOOTNOTES

[1] Information based on analysis of variance is also given in Frey (1973, 1974, 1976) and Bendig (1952b, 1952c) but in a form from which estimates of reliability of individual and of average ratings cannot be determined.

[2] These others include the assumptions that the sample of ratees is a random sample from the population to which inferences are to be made, that error of measurement is uncorrelated with the true score, and that within-ratee variance may be pooled to provide an estimate of variance due to error of measurement (Winer, 1962, pp. 129–133). The necessity of meeting these and still other assumptions (as well as the practicality or possibility of doing so) in estimating reliability—in dealing either with traditional reliability estimates of multi-item psychological tests or with its application to multiratings of

ratees—is receiving debate in the psychometric literature (see, in particular, Cochran, 1968; Cronbach, Gleser, Nanda, and Rajaratnam, 1972; Loevinger, 1947, 1965; Nunnally, 1967, Chap. 6; Rozeboom, 1966; Stanley, 1961, 1971; and Tryon, 1957), only some of which is of concern in the present analysis.

[3] It is possible that "hearsay" about a teacher before students enter a class—that is, the more general "local reputation" of the teacher at the college—influences their ratings after they have been in the course. At a group level of analysis, little is systematically known about this (cf. McClelland, 1970; Perry, Niemi, and Jones, 1974). At the individual level of analysis, however, there is some evidence that a student's prior knowledge about, and impression of, an instructor or course is related to his or her ratings of that instructor or course (to be discussed in a later section of the present analysis).

[4] Citations to the extensive research on which the conclusions of this section have been based can be found in Feldman (1976a). The following additional reports, which either had not been located or were not yet available at the time of the earlier review, are generally supportive of one or another of the conclusions presented therein: Batista and Brandenburg (1975); Bausell and Magoon (1972, Appendix U); Blass (1974); Christensen and Bourgeois (1974); Crowe (1974); Delaney (1976); Endo and Della-Piana (1976); Fenker (1975); Frey (1976); Frey, Leonard and Beatty (1975); Gery (1972); Goldenbaum and Wheeler, as cited in Carter (1968); Granzin and Painter (1975, 1976); Hillery and Yukl (1971); Hocking (1976); Jernstedt (1976); Kelley (1972); Kline (1975); Kovacs and Kapel (1976); Lewis and Dahl (1972); Lewis and Orvis (1973); Miller (1972, Appendix B); Murdock (1969); Page and Roy (1975); Perkins (1971); Pohlmann (1972); Pratt and Pratt (1976); Riechmann (1974); Saunders (1972); Scott, Halpin, and Schnittjer (1974); Sherman and Winstead (1975); Sloane (1972); Weinrauch and Matejka (1973); Whitely and Doyle (1976).

[5] Similar findings usually appear when the group rather than the individual student is the unit of analysis; see Doyle and Whitely (1974); Elmore and Pohlmann (1976); Gillmore (1975); Gillmore and Naccarato (1975); Harry and Gouldner (1972); Hoyt, Owens and Groulin (1973); Jiobu and Pollis (1971); Pohlmann (1975, also see Pohlmann, 1973); Sorge and Kline (1973); Whitely and Doyle (1976).

[6] In Bejar and Doyle (1975) the data given are multiple correlation coefficients for each of ten rating items using all ten first-impression measures as predictors; in the other three studies, the data are zero-order correlation coefficients between the measures of first impressions and the corresponding rating items. In a study of graduate students in psychology courses, the beginning-of-course by end-of-course correlations were, in general, even larger than those reported in these four studies of undergraduates (Aleamoni, Yimer, and Mahan, 1972). As large or larger correlations are also reported for ecological analyses, in which correlations were calculated across classes between average instructor or course ratings at the beginning and at the end of the course (Bausell and Magoon, 1972c, also see Bausell and Magoon, 1972d, Appendix U; Oles, 1975).

[7] The following studies were found, which do not have enough information upon which to draw conclusions one way or another: Davis (1969); Perkins (1971, Appendix D); and Fenker (1975). Kohlan (1973) reports a statiscally significant interaction effect (on one of the four rating scales used in his study) between class year and grade-point average, but the nature of the interaction is not given.

[8] In general, studies using other than correlational analysis have not given enough information to determine the percent of explained variance.

[9] Gender was also a variable in the following studies, but there is not enough information in them to be able to include their results in the present review; Davis (1969); Doyle (1972); Menard (1972); Price and Magoon (1971).

[10] Reference is to such rating items as instructor's stimulation of students' interest, sensitivity to student reactions and progress, feedback to students, encouragement of discussions and openness to others' viewpoints, respect for students, friendliness, and availability and helpfulness—in contrast to items that deal with the teacher's organization, knowledgeability, intellectual expansiveness, enthusiasm for the subject matter and for teaching, elocutionary skills, use of supplementary materials, and the like (see Feldman, 1976b).

[11] Other research on student ratings of teachers and courses exists that has used forced-choice instruments, but comparison with other techniques has not been made in terms of

possible differences in interrater reliabilities (see Cobb, 1956; Cosgrove, 1959; Echandia, 1963; Leftwich and Remmers, 1962; Lovell and Haner, 1955; Patton and Meyer, 1955; Snedeker, 1959).

[12] The argument is different at the group level of analysis. Given certain controls, positive associations between average teacher or course ratings and the average performance of classes on purportedly objective or relatively standardized indicators of student achievement in the content of the course are generally expected (and usually found). A list of much of the relevant research is given in Feldman (1976a, p. 98), to which the following studies may be added: Capozza (1973); Centra (1977); Frey (1976); Frey, Leonard and Beatty (1975); Marsh, Fleiner, and Thomas (1975); Soper (1973); Turner and Thompson (1974); Whitely and Doyle (1976).

REFERENCES

Aleamoni, L. M. (1974). Typical faculty concerns about student evaluation of instruction. Presented at the Symposium on Methods of Improving University Teaching at the Technion, Israel Institute of Technology.

Aleamoni, L. M., Yimer, M., and Mahan, J. M. (1972). Teacher folklore and sensitivity of a course evaluation questionnaire. Psychological Reports 31: 607–614.

Anastasi, A. (1968). Psychological Testing (3rd ed.). New York: Macmillan.

Apt, M. H. (1966). A measurement of college instructor behavior. Ph.D. dissertation, University of Pittsburgh.

Baker, P. C., and Remmers, H. H. (1951). Progress in research on personnel evaluation. Journal of Teacher Education 2: 143–146.

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

Bausell, R. B., and Magoon, J. (1972a). Expected grade in a course, grade point average, and student ratings of the course and the instructor. Educational and Psychological Measurement 32: 1013–1023.

Bausell, R. B., and Magoon, J. (1972b). Instructional methods and college student ratings of courses and instructors. Journal of Experimental Education 40: 29–33.

Bausell, R. B., and Magoon, J. (1972c). The persistence of first impressions in course and instructor evaluations. Presented at the Annual Meeting of the American Educational Research Association.

Bausell, R. B., and Magoon, J. (1972d). The Validation of Student Ratings of Instruction: An Institutional Research Model. Newark, DE: College of Education, University of Delaware.

Bejar, I. I. (1975). A survey of selected administrative practices supporting student evaluation. Research in Higher Education 3, 77–86.

Bejar, I. I., and Doyle, K. O., Jr. (1975). Student Ratings of Instruction: Expectations, First Impressions, and Evaluations. Minneapolis, MN: Measurement Services Center, University of Minnesota.

Bejar, I. I., and Doyle, K. O., Jr. (1977). The effect of prior expectations on the structure and elevation of student ratings of teaching behavior. Journal of Educational Measurement, in press.

Bendig, A. W. (1952a). A preliminary study of the effect of academic level, sex, and course variables on student rating of psychology instructors. Journal of Psychology 34: 21–26.

Bendig, A. W. (1952b). A statistical report on a revision of the Miami Instructor Rating Sheet. Journal of Educational Psychology 43: 423–429.

Bendig, A. W. (1952c). The use of student-rating scales in the evaluation of instructors in introductory psychology. Journal of Educational Psychology 43: 167–175.

Bendig, A. W. (1953a). Comparison of psychology instructors and national norms on the Purdue Rating Scale. Journal of Educational Psychology 44: 435–439.

Bendig, A. W. (1953b). Student achievement in introductory psychology and student ratings of the competence and empathy of their instructors. Journal of Psychology 36: 427–433.

Blank, L. F. (1970). Student-faculty psychological types and student instructional ratings. Oshkosh, WI: Wisconsin State University, Oshkosh, 1970. ERIC Document Reproduction Service No. ED 040 422.

Blass, T. (1974). Measurement of objectivity-subjectivity: Effects of tolerance for imbalance and grades on evaluations of teachers. Psychological Reports 34: 1199–1213.

Brooks, T. E., Tarver, D. A., Kelley, H. P., Liberty, P. G., Jr., and Dickerson, A. D. (1971). Dimensions underlying student ratings of courses and instructors at the University of Texas at Austin: Instructor Evaluation Form 2. Research Bulletin RB-71-4. Austin, TX: Measurement and Evaluation Center, University of Texas at Austin.

Byrne, D. (1964). Assessing personality variables and their alteration. In P. Worchel and D. Byrne (Eds.), Personality Change. New York: Wiley.

Caffrey, B. (1969). Lack of bias in student evaluations of teachers. Proceedings of the 77th Annual Convention of the American Psychological Association 4: 641–642.

Canter, F. M., and Meisels, M. (1971). Cognitive dissonance and course evaluation. Improving College and University Teaching 19: 111–113.

Capozza, D. R. (1973). Student evaluations, grades and learning in economics. Western Economic Journal 11: 127.

Carney, R. E. (1961). An analysis of university student behaviors with measures of ability, attitude, performance and personality. Ph.D. dissertation, University of Michigan.

Carter, R. E. (1968). The effect of student characteristics on three student evaluations of university instruction. Ph.D. dissertation, Indiana University.

Cattell, R. B. (1957). Personality and Motivation Structure and Measurement. Yonkers-on-Hudson, NY: World Book.

Centra, J. A. (1972). Two studies on 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 Report No. 2. Princeton, NJ: Educational Testing Service.

Centra, J. A. (1973). The Student Instructional Report: Comparisons with alumni ratings; item reliabilities; the factor structures. SIR Report No. 3. Princeton, NJ: Educational Testing Service.

Centra J. A. (1974). College teaching: Who should evaluate it? Findings 1 (No. 1): 5–8.

Centra, J. A. (1975). Colleagues as raters of classroom instruction. Journal of Higher Education 46: 327–337.

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

Centra, J. A., and Linn, R. L. (1973). Student points of view in ratings of college instruction. Research Bulletin RB-73-60. Princeton, NJ: Educational Testing Service.

Christensen, L. B., and Bourgeois, A. E. (1974). Student ratings of instructional effectiveness. Presented at the Annual Meeting of the American Psychological Association.

Clark, K. E., and Keller, R. J. (1954). 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, MN: University of Minnesota Press.

Cobb, E. B. (1956). Construction of a forced-choice university instructor rating scale. Ph.D. dissertation, University of Tennessee.

Cochran, W. G. (1968). Errors of measurement in statistics. Technometrics 10: 637–666.

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

Colliver, J. A. (1972). 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.

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

Coombs, C. H. (1964). A Theory of Data. New York: Wiley.

Corcoran, M. E. (1957). The role of personal attitudes in student evaluation of an introductory education course. Ph.D. dissertation, University of Minnesota.

Cosgrove, D. J. (1959). Diagnostic rating of teacher performance. Journal of Educational Psychology 50: 200–204.

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

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

Crichton, L. I., and Doyle, K. O., Jr. (1975). Reliability of ratings. Minneapolis, Minn.: Measurement Services Center, University of Minnesota.

Crittenden, K. S., and Norr, J. L. (1973). Student values and teacher evaluation: A problem in person perception. Sociometry 36: 143–151.

Crittenden, K. S., and Norr, J. L. (1975). Some remarks on "Student Ratings": The validity problem. American Educational Research Journal 12: 429–433.

Cronbach, L. J., Gleser, G. C., Nanda, H., and Rajaratnam, N. (1972). The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles. New York: Wiley.

Crouch, H. B., and Leathers, C. M. (1951). The validity of student opinions in evaluating a program of college biology. Science Education 35: 73–76.

Crowe, M. H. (1974). Selected student characteristics and their relationship to course ratings. Ph.D. dissertation, Purdue University.

Davis, R. H. (1969). Student Instructional Rating System (SIRS): Technical bulletin. East Lansing, MI: Michigan State University.

Davison, D. C. (1973). Perception of instructor in relation to self and evaluation of instructor's performance. Perceptual and Motor Skills 36: 533–534.

Day, C. R. (1969). Assumed similarity to others: Some determinants and consequences. Ph.D. dissertation, Ohio State University.

Delaney, E. L. (1976). The relationships of student ratings of instruction to student, instructor and course characteristics. Presented at Annual Meeting of the American Educational Research Association.

Deshpande, A. S., Webb, S. C., and Marks, E. (1970). Student perceptions of engineering instructor behaviors and their relationships to the evaluation of instructors and courses. American Educational Research Journal 7: 289–305.

Dick, W. (1967). Course Attitude Questionnaire: Its development, uses and research results. Report No. 67-1 (revision of No. 106, revised by D. Stickell). University Park, PA: Office of Examination Services, University Division of Instructional Services, Pennsylvania State University.

Domino, G. (1971). Interactive effects of achievement orientation and teaching style on academic achievement. Journal of Educational Psychology 62: 427–431.

Downie, N. M. (1952). Student evaluation of faculty. Journal of Higher Education 23: 495–496, 503.

Doyle, K. O., Jr. (1972). Construction and evaluation of scales for rating college instructors. Ph.D. dissertation, University of Minnesota.

Doyle, K. O., Jr. (1975). Student Evaluation of Instruction. Lexington, MA: Heath.

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

Dwyer, F. (1968). A review of characteristics and relationships of selected criteria for evaluating teacher effectiveness. University Park, PA: University Division of Instructional Services, Pennsylvania State University.

Ebel, R. L. (1951). Estimation of the reliability of ratings. Psychometrika 16: 407–424.

Echandia, P. P. (1963). A methodological study and factor analytic validation of forced-choice performance ratings of college accounting instructors. Ph.D. dissertation, New York University.

Edwards, A. L. (1957). The Social Desirability Variability in Personality Assessment and Research. New York: Dryden.

Elliott, D. N. (1950). Characteristics and relationships of various criteria of college and university teaching. Purdue University Studies in Higher Education 70: 5–61.

Elmore, P. B., and LaPointe, K. A. (1974). Effects of teacher sex and student sex on the evaluation of college instructors. Journal of Educational Psychology 66: 386–389.

Elmore, P. B., and LaPointe, K. A. (1975). Effect of teacher sex, student sex, and teacher warmth on the evaluation of college instructors. Journal of Educational Psychology 67: 368–374.

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

Endo, G. T., and Della-Piana, G. (1976). A validation study of course evaluation ratings. Improving College and University Teaching 24: 84–86.

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

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

Fenker, R. M. (1975). The evaluation of university faculty and administrators: A case study. Journal of Higher Education 46: 665–686.

Ferber, M. A., and Huber, J. A. (1975). Sex of student and instructor: A study of student bias. American Journal of Sociology 80: 949–963.

Flood Page, C. (1974). Student Evaluation of Teaching: The American Experience. London: Society for Research into Higher Education.

Follman, J. (1975). Student ratings of faculty teaching effectiveness: Rater or ratee characteristics. Research in Higher Education 3: 155–167.

Follman, J., Lavely, C., Silverman, S., and Merica, J. (1974). Student raters' referents in rating college teaching effectiveness. Journal of Psychology 86: 247–249.

Follman, J., Lucoff, M., Small, L., and Power, F. (1974). Kinds of keys of student ratings of faculty teaching effectiveness. Research in Higher Education 2: 173–179.

Freehill, M. F. (1967). Authoritarian bias and evaluation of college experiences. Improving College and University Teaching 15: 18–19.

French-Lazovik, G. (1974). Predictability of students' evaluations of college teachers from component ratings. Journal of Educational Psychology 66: 373–385.

Frey, P. W. (1973). Student ratings of teaching: Validity of several rating factors. Science 182: 83–85.

Frey, P. W. (1974). The ongoing debate: Student evaluation of teaching. Change February: 47–48, 64.

Frey, P. W. (1976). Validity of student instructional ratings as a function of their timing. Journal of Higher Education 47: 327–336.

Frey, P. W., Leonard, D. W., and Beatty, W. W. (1975). Student ratings of instruction: Validation research. American Educational Research Journal 12: 435–444.

Frick, T., and Semmel, M. (1974). Observational records: Observer agreement and reliabilities. Bloomington, IN: Center for Innovation in Teaching the Handicapped, School of Education, University of Indiana.

Fulcher, D. G., and Anderson, W. T., Jr. (1974). Interpersonal dissimilarity and teaching effectiveness: A relational analysis. Journal of Educational Research 68: 19–25.

Gery, F. W. (1972). Does mathematics matter? In A. L. Welsh (Ed.), Research

Papers in Economic Education. New York: Joint Council on Economic Education.

Ghiselli, E. E., and Ghiselli, W. B. (1972). Ratings—*Kundgabe* or *Beschreibung*. Journal of Psychology 80: 263–271.

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

Gillmore, G. M. (1975). 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 503. Seattle, WA: Educational Assessment Center, University of Washington.

Gillmore, G. M., Kane, M. T., and Naccarato, R. W. (1976). The generalizability of student instructional ratings: General theory and application to the Washington Instructional Assessment System. EAC Report 74-16. Seattle, WA: Educational Assessment Center, University of Washington.

Gillmore, G. M., and Naccarato, R. W. (1975). The effect of factors outside the instructor's control on student ratings of instruction. EAC Report 283A. Seattle, WA: Educational Assessment Center, University of Washington.

Good, K. C. (1971). Similarity of student and instructor attitudes and student's attitudes toward instructors. Ph.D. dissertation, Purdue University.

Good, K. C., and Good, L. (1973). Assumed attitude similarity and instructor evaluation. Journal of Social Psychology 91: 285–290.

Grande, P. P., and McCollester, C. W. Psychological correlates of students' evaluation of teaching. Unpublished.

Granzin, K. L., and Painter, J. J. (1973). A new explanation for students' course evaluation tendencies. American Educational Research Journal 10: 115–124.

Granzin, K. L., and Painter, J. J. (1975). A multivariate analysis of factors underlying student evaluations of college instructors. California Journal of Educational Research 26: 96–106.

Granzin, K. L., and Painter, J. J. (1976). A second look at cognitive dissonance and course evaluation. Improving College and University Teaching 24: 113–115.

Greenwood, G. E., Bridges, C. M., Jr., Ware, W. B., and McLean, J. E. (1973). Student Evaluation of College Teaching Behaviors instrument: A factor analysis. Journal of Higher Education 44: 596–604.

Grush, J. E., Clore, G. L., and Constin, F. (1975). Dissimilarity and attraction: When difference makes a difference. Journal of Personality and Social Psychology 32: 783–789.

Guilford, J. P. (1954). Psychometric Methods (2nd ed.). New York: McGraw-Hill.

Guthrie, E. R. (1927). Measuring student opinion of teachers. School and Society 25: 175–176.

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

Guthrie, E. R. (1949). The evaluation of teaching. Educational Record 30: 109–115.

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

Haggard, E. A. (1958). Intraclass Correlation and the Analysis of Variance. New York: Dryden.

Halstead, J. S. (1972). Students' ratings of college classroom verbal interaction as related to ratings of instructor teaching effectiveness. Ph.D. dissertation, Purdue University.

Harari, O., and Zedeck, S. (1974). Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology 58: 261–265.

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

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

Helmstadter, G. C. (1964). Principles of Psychological Measurement. New York: Appleton-Century-Crofts.

Heyns, R. W., and Lippitt, R. (1954). Systematic observational techniques. In G. Lindzey (Ed.), Handbook of Social Psychology, Vol. I. Reading, MA: Addison-Wesley.

Hildebrand, M., Wilson, R. C., and Dienst, E. R. (1971). Evaluating University Teaching. Berkeley, CA: Center for Research and Development in Higher Education, University of California at Berkeley.

Hillery, J. M., and Yukl, G. A. (1971). Convergent and discriminant validation of student ratings of college instructors. Presented at the Annual Meeting of the Midwestern Psychological Association.

Hirschi, R., and Selvin, H. C. (1967). Delinquency Research: An Appraisal of Analytic Methods. New York: Free Press.

Hocking, J. M. (1976). College students' evaluations of faculty are directly related to course interest and grade expectation. College Student Journal 10: 312–316.

Horst, P. (1949). A generalized expression for the reliability of measures. Psychometrika 14: 21–31.

Hoyt, D. P. (1969). Instructional effectiveness. II. Identifying effective classroom procedures. Report No. 7. Manhatten, KS: Office of Educational Research, Kansas State University.

Hoyt, D. P. (1973a). Identifying effective educational procedures. Improving College and University Teaching 21: 73–76.

Hoyt, D. P. (1973b). Measurement of instructional effectiveness. Research in Higher Education 1: 367–378.

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

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

Jernstedt, G. C. (1976). The relative effectiveness of individualized and traditional instruction methods. Journal of Educational Research 69: 211–220.

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

Kane, M. T., and Brennan, R. L. (1977). The generalizability of class means. Review of Educational Research, 47: 267–292.

Kane, M. T., Gillmore, G. M., and Crooks, T. J. (1977). Student evaluations of teaching: The generalizability of class means. Journal of Educational Measurement, in press.

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

Kelley, A. C. (1972). Uses and abuses of course evaluations and measures of educational output. Journal of Economic Education 4: 13–18.

Kennedy, W. R. (1971). The relationship of selected student characteristics to components of teacher/course evaluations among freshmen English students at Kent State University. Ph.D. dissertation, Kent State University.

Kennedy, W. R. (1972). The relationship of selected student characteristics to components of teacher/course evaluations among freshman English students at Kent State University. Presented at the Annual Meeting of the American Educational Research Association.

Kerlinger, F. N. (1963). Educational attitudes and perceptions of teachers: Suggestions for teacher-effectiveness research. School Review 71: 1–11.

Kerlinger, F. N. (1973). Foundations of Behavioral Research (2nd ed.). New York: Holt, Rinehart and Winston.

Kline, C. R., Jr. (1975). Students rate profs in accord with grade expectations. Phi Delta Kappan 57: 54.

Kohlan, Richard G. (1973). A comparison of faculty evaluations early and late in the course. Journal of Higher Education 44: 587–595.

Kovacs, R., and Kapel, D. E. (1976). Personality correlates of faculty and course evaluations. Research in Higher Education 5: 335–344.

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

Kulik, J. A., and McKeachie, W. J. (1975). The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), Review of Research in Education, Vol. 3. Itasca, IL: F. E. Peacock.

Leftwich, W. H., and Remmers, H. H. (1962). A comparison of graphic and forced-choice ratings of teaching performance at the college and university level. Purdue University Studies in Higher Education, No. 92, 3–31.

Levenson, H., and LeUnes, A. (1974). Student evaluation of an instructor: Effects of attitude similarity. Psychological Reports 34: 1074.

Leventhal, L., Abrami, P. C., and Perry, R. P. (1976). Do teacher rating forms reveal as much about students as about teachers? Journal of Educational Psychology 68: 441–445.

Leventhal, L. Abrami, P. C., Perry, R. P., and Breen, L. J. (1975). Section selection in multi-section courses: Implications for the validation and use of teacher rating forms. Educational and Psychological Measurement 35: 885–895.

Levinthal, C. F. (1974). An analysis of the teacher evaluation process. Final Report, U.S. Department of Health, Education, and Welfare, National Institutes of Education, Project No. 2B089. Hempstead, NY: Hofstra University.

Levinthal, C. F., Lansky, L. M., and Andrews, O. E. (1971). Student evaluations of teacher behaviors as estimations of real-ideal discrepancies: A critique of teacher rating methods. Journal of Educational Psychology 62: 104–109.

Lewis, E. C. (1964). An investigation of student-teacher interaction as a determiner of effective teaching. Journal of Educational Research 57: 360–363.

Lewis, D. R., and Dahl, T. (1972). Factors influencing performance in the principles course revisited. In A. L. Welsh (Ed.), Research Papers in Economic Education. New York: Joint Council on Economic Education.

Lewis, D. R., and Orvis, C. C. (1973). A training system for graduate student instructors of introductory economics at the University of Minnesota. Journal of Economic Education 5: 38–46.

Linn, R. L., Centra, J. A., and Tucker, L. R. (1974). Between, within, and total group factor analyses of student ratings of instruction. Research Bulletin RB-74-39. Princeton, NJ: Educational Testing Service.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs 61 (4, Whole No. 285).

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review 72: 143–155.

Lord, F. M., and Novick, M. R. (1968). Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley.

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

Lunney, G. H. (1974). 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.

Maas, J. B., and Owen, T. R. (1973). Cornell Inventory for Student Appraisal of Teaching and Courses: Manual of instructions. Ithaca, NY: Center for Improvement of Undergraduate Education, Cornell University.

Magoon, A. J., and Bausell, R. B. The pass fail option and course and instructor ratings: A discriminant analysis. Unpublished.

Magoon, A. J., and Price, J. R. (1972). Rating dimensions of course and instructor characteristics: The eye of the beholder. Presented at the American Educational Research Association.

Majer, K., and Stayrook, N. (1974). Reliability of college classroom course evaluations. Presented at the annual meeting of the National Council on Measurement in Education.

Mallory, E. B., Huggins, M., and Steinberg, B. (1941). Journal of Educational Psychology, 32: 13–22.

Maney, A. C. (1959). The authoritarianism dimension in student evaluations of faculty. Journal of Educational Sociology 32: 226–231.

Mann, R. D., Arnold, S. M., Binder, J. L., Cytrynbaum, S., Newman, B. M., Ringwald, B. E., Ringwald, J. W., and Rosenwein, R. (1970). The College Classroom: Conflict, Change, and Learning. New York: Wiley.

Marsh, H. W., Fleiner, H., and Thomas, C. S. (1975). Validity and usefulness of student evaluations of instructional quality. Journal of Educational Psychology 67: 833–839.

Maslow, A. H., and Zimmerman, W. (1956). College teaching ability, scholarly activity and personality. Journal of Educational Psychology 47: 185–189.

McClelland, J. N. (1970). The effect of student evaluations of college instruction upon subsequent evaluations. California Journal of Educational Research 21: 88–95.

McDaniel, E. D. (1972). Student preferences and evaluation of faculty. Presented at the Annual Meeting of the American Psychological Association.

McInnis, T. (1966). Some methodological considerations and a report of some research findings concerning course and/or teacher evaluations by students. Research Report No. 231. Urbana-Champaign, IL: Measurement and Research Division, Office of Instructional Resources, University of Illinois.

McKeachie, W. J. (1973). Correlates of student ratings. In A. L. Sockloff (Ed.), Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia, PA: Measurement and Research Center, Temple University.

Medley, D. M., and Mitzel, H. E. (1963). Measuring classroom behavior by systematic observation. In N. L. Gage (Ed.), Handbook of Research on Teaching. Chicago: Rand McNally.

Menard, T. L. (1972). An analysis of the relationship between teacher effectiveness and teacher appearance. Ph.D. dissertation, University of Northern Colorado.

Menges, R. J. (1969). Student-instructor cognitive compatibility in the large lecture class. Journal of Personality 37: 444–458.

Menges, R. J. (1973). The new reporters: Students rate instruction. New Directions for Higher Education 1: 59–75.

Menne, J. (1968). Students' evaluation of instructors. Presented at the Annual Meeting of the National Council on Measurement in Education.

Menzel, H. (1950). Communication on Robinson's "Ecological Correlations and the Behavior of Individuals." American Sociological Review 15: 674.

Miller, R. I. (1972). Evaluating Faculty Performance. San Francisco: Jossey-Bass.

Miller, R. I. (1974). Developing Programs for Faculty Evaluation. San Francisco: Jossey-Bass.

Murdock, R. P. (1969). The effect of student ratings of their instructor on the student's achievement and ratings. Office of Education, U.S. Department of Health, Education, and Welfare Project No. 9-H-014. Salt Lake City: University of Utah.

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

Murray, H. G. (1975). Predicting student ratings of college teaching from peer ratings of personality traits. Teaching of Psychology 2: 66–69.

Nichols, M. G. (1967). A study of the influences of selected variables involved in student evaluations of teacher effectiveness. Ph.D. dissertation, University of South Dakota.

Norr, J. L., and Crittenden, K. S. (1975). Evaluating college teaching as leadership. Higher Education 4: 335–350.

Null, E. J., and Nicholson, E. W. (1972). Personal variables of students and their perception of university instructors. College Student Journal 6: 6–9.

Null, E. J., and Walter, J. E. (1972). Values of students and their ratings of a university professor. College Student Journal 6: 46–51.

Nunnally, J. C. (1967). Psychometric Theory. New York: McGraw-Hill.

Office of Evaluation Services. (1972). Student Instructional Rating System responses and student characteristics. SIRS Research Report No. 4. East Lansing, MI: Michigan State University.

Oles, H. J. (1975). Stability of student evaluation of instructors and their courses with implications for validity. Educational and Psychological Measurement 35: 437–445.

Page, M. M., and Roy, R. E. (1975). Internal-external control and independence of judgment in course evaluations among college students. Personality and Social Psychology Bulletin 1: 509–512.

Parent, J., Forward, J., Canter, R., and Mohling, J. (1975). Interactive effects of teaching strategy and personal locus of control on student performance and satisfaction. Journal of Educational Psychology 67: 764–769.

Patton, H. M., and Meyer, P. R. (1955). A forced choice rating form for college teachers. Journal of Educational Psychology 46: 499–503.

Perkins, E. R. (1971). Relationships among empathy, genuineness, nonpossessive warmth, and college teacher effectiveness and selected characteristics. Ph.D. dissertation, University of Kentucky.

Perry, R. P., Niemi, R. R., and Jones, K. (1974). Effect of prior teaching evaluations and lecture presentation on ratings of teaching performance. Journal of Educational Psychology 66: 851–856.

Perry, R. R., and Baumann, R. R. (1973). 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, PA: Measurement and Research Center, Temple University.

Peters, C. C., and Van Voorhis, W. R. (1940). Statistical Procedures and Their Mathematical Bases. New York: McGraw-Hill.

Phillips, B. N. (1960). Authoritarian, hostile, and anxious students' ratings of an instructor. California Journal of Educational Research 11: 19–23.

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

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

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

Pohlmann, J., and Tuinen, M. V. (1972). Norms for required and elective course level for IIQ subscales. Technical Report 11.1-72. Carbondale, IL: Counseling and Testing Center, Southern Illinois University, Carbondale.

Potter, N. R. (1969). The relationships of selected student characteristics to teacher ratings. Ph.D. dissertation, Colorado State College.

Pratt, M., and Pratt, T. A. E. C. (1976). A study of student-teacher grading interaction process. Improving College and University Teaching 24: 73–81.

Price, J. A., and Magoon, A. J. (1971). Predictors of college student ratings of instructors. Presented at the Annual Meeting of the American Psychological Association.

Purohit, A., and Magoon, A. J. (1971). The validity of student-run course evaluations. Presented at the Annual Meeting of the American Educational Research Association.

Purohit, A., and Magoon, A. J. (1974). Congruence in attitude of instructors and students towards course evaluation. College Student Journal 8: 29–36.

Quereshi, M. Y., and Widlak, F. W. (1973). Students' perception of a college teacher as a function of their sex and achievement level. Journal of Experimental Education 41: 53–57.

Rayder, N. F. (1967). College student ratings of instructors. Ph.D. dissertation, Colorado State College.

Rayder, N. F. (1968). College student ratings of instructors. Journal of Experimental Education 37: 76–81.

Remmers, H. H., and Elliott, D. N. (1949). The Indiana College and University Staff-Evaluation Program. School and Society 70: 168–171.

Remmers, H. H., Shock, N. W., and Kelly, E. L. (1927). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue Rating Scale. Journal of Educational Psychology 18: 187–195.

Remmers, H. H., and Weisbrodt, J. A. (1964). Manual of Instructions for Purdue Rating Scale of Instruction. Purdue, IN: Purdue Research Foundation.

Rezler, A. G. (1965). The influence of needs upon the student's perception of his instructor. Journal of Educational Research 58: 282–286.

Riechmann, S. W. (1974). The relationship between student classroom-related variables and students' evaluations of faculty. Ph.D. dissertation, University of Cincinnati.

Riley, J. W., Jr., Ryan, B. F., and Lifshitz, M. (1950). The Student Looks at His Teacher: An Inquiry into the Implications of Student Ratings at the College Level. New Brunswick, NJ: Rutgers University Press.

Rosenshine, B., Cohen, A., and Furst, N. (1973). Correlates of student preference ratings. Journal of College Student Personnel 14: 269–272.

Rozeboom, W. W. (1966). Foundations of the Theory of Prediction. Homewood, IL: Dorsey.

Rumery, R. E., Rhodes, D. M., and Johnson, H. C., Jr. (1975). The role of student reports in the evaluation of teaching in higher education. Higher Education Bulletin 3: 93–99.

Saunders, P. (1972). Student learning and instructor ratings: The Carnegie-Mellon experience in introductory economics. In A. L. Welsh (Ed.), Research Papers in Economic Education. New York: Joint Council on Economic Education.

Schuessler, K. (1971). Analyzing Social Data: A Statistical Orientation. Boston: Houghton Mifflin.

Scott, O., Halpin, G., and Schnittjer, C. (1974). Student characteristics associated with student perceptions of college instruction. Presented at the Annual Meeting of the National Council on Measurement in Education.

Seldin, P. (1975). How Colleges Evaluate Professors: Current Policies and Practices in Evaluating Classroom Teaching Performance in Liberal Arts Colleges. Croton-on-Hudson, NY: Blythe-Pennington.

Shapiro, P. (1974). After data collection: Coding—an educational research tool. SRIS Quarterly 7: 16–23.

Sharon, A. T. (1970). Eliminating bias from student ratings of college instructors. Journal of Applied Psychology 54: 278–281.

Sharon, A. T., and Bartlett, C. J. (1969). Effect of instructional conditions in producing leniency on two types of rating scales. Personnel Psychology 22: 251–263.

Sheehan, D. S. (1975). On the invalidity of student ratings for administrative personnel decisions. Journal of Higher Education 46: 687–700.

Sherman, T. M., and Winstead, J. C. (1975). A formative approach to student evaluation instruction. Educational Technology 15: 34–39.

Singhal, S. Inter-group differences on Course Evaluation Questionnaire. Research Report No. 262. Urbana-Champaign, IL: Measurement and Research Division, Office of Instructional Resources, University of Illinois.

Singhal, S. (1968). Illinois Course Evaluation Questionnaire items by rank of instructor, sex of instructor and sex of the student. Research Report No. 282. Urbana-Champaign, IL: Measurement and Research Division, Office of Instructional Resources, University of Illinois.

Sloane, P. E. (1972). The relationship of performance to instruction and student attitudes. In A. L. Welsh (Ed.), Research Papers in Economic Education. New York: Joint Council on Economic Education.

Snedeker, J. H. (1959). The construction of a forced-choice rating scale for college instruction. Ph.D. dissertation, Indiana University.

Sockloff, A. L. (1973). Instruments for student evaluation of faculty: Ideal and actual. In A. L. Sockloff (Ed.), Proceedings of the First Invitational Conference on Faculty Effectiveness as Evaluated by Students. Philadelphia, PA: Measurement and Research Center, Temple University.

Sockloff, A. L. (1975). Behavior of the product-moment correlation when two heterogeneous subgroups are pooled. Educational and Psychological Measurement 35: 267–276.

Sockloff, A. L., and Deabler, V. T. (1971). The construction of the Faculty and Course Evaluation Instrument. Research Report 71-2. Philadelphia, PA: Testing Bureau, Temple University.

Soper, J. C. (1973). Soft research on a hard subject: Student evaluations reconsidered. Journal of Economic Education 5: 22–26.

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

Spencer, R. E. Judge consistency of Course Evaluation Questionnaire ratings. Research Report No. 211. Urbana-Champaign, IL: Office of Instructional Research, Measurement and Research Division, University of Illinois.

Spencer, R. E. (1969). A history of the development of the Illinois Course Evaluation Questionnaire. Research Report No. 306. Urbana-Champaign, IL: Measurement and Research Division, Office of Instructional Resources, University of Illinois.

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

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

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

Tagiuri, R. (1969). Person perception. In G. Lindzey and E. Aronson (Eds.), The Handbook of Social Psychology (2nd ed.), Vol. 3. Reading, MA: Addison-Wesley.

Tagiuri, R., and Petrullo, L. (Eds.). (1958). Person Perception and Interpersonal Behavior. Stanford, CA: Stanford University Press.

Taylor, R. E. (1968). An investigation of the relationship between psychological types in the college classroom and the student perception of the teacher and preferred teaching practices. Ph.D. dissertation, University of Maryland.

Tetenbaum, T. J. (1975). The role of student needs and teacher orientations in student ratings of teachers. American Educational Research Journal 12: 417–429.

Thorndike, R. L. (1949). Personnel Selection: Test and Measurement Techniques. New York: Wiley.

Thorndike, R. L., and Hagen, E. (1969). Measurement and Evaluation in Psychology and Education (3rd ed.). New York: Wiley.

Tinsley, H. E., and Weiss, D. J. (1975). Interrater reliability and agreement of subjective judgments. Journal of Counseling Psychology 22: 358–376.

Touq, M. (1972). The relationship between student participation in classroom discussion and student ratings of instructors at the college level. Ph.D. dissertation, Purdue University.

Touq, M. S., and Feldhusen, J. F. (1973). The relationship between student ratings of instructors and their participation in classroom discussion. Presented at the Annual Meeting of the National Council on Measurement in Education.

Treffinger, D. J., and Feldhusen, J. F. (1970). Predicting students' ratings of instruction. Proceedings of the 78th Annual Convention of the American Psychological Association 5: 621–622.

Tryon, R. C. (1957). Reliability and behavior domain validity: Reformulation and historical critique. Psychological Bulletin 54: 229–249.

Tuckman, B. W., and Orefice, D. S. (1973). Personality structure, instructional outcomes, and instructional preferences, Interchange 4: 43–48.

Turner, R. L., and Thompson, R. P. (1974). Relationships between college student ratings of instructors and residual learning. Presented at the Annual Meeting of the American Educational Research Association.

Veldman, D. J. (1968). Student evaluation of College of Education courses, fall semester, 1968. Unpublished.

Voeks, V. W. (1962). Publication and teaching effectiveness. Journal of Higher Education 33: 212–218.

Walker, B. D. (1968). An investigation of selected variables relative to the manner in which a population of junior college students evaluate their teachers. Ph.D. dissertation, University of Houston.

Walter, J. E. (1971). Relationships between selected values of students and their perception of a university instructor. Ph.D. dissertation, Purdue University.

Warr, P. B. and Knapper, C. (1968). The Perception of People and Events. New York: Wiley.

Weick, K. E. (1968). Systematic observational methods In G. Lindzey and E. Aronson (Eds.), The Handbook of Social Psychology (2nd ed.), Vol. 2. Reading, MA: Addison-Wesley.

Weinrauch, J. D., and Matejka, J. K. (1973). Are student ratings of business communication teachers honest feedback? Journal of Business Communication 11: 31–37.

Weinstein, P., and Bramble, W. J. "Student press": Student course ratings as a function of student variables. Unpublished.

Whitely, S. E., and Doyle, K. O., Jr. (1976). The validity and generalizability of student ratings from between-class and within-class data. Minneapolis, MN: Measurement Services Center, University of Minnesota.

Whitely, S. E., Doyle, K. O., Jr., and Hopkinson, K. (1973). Student ratings and criteria for effective teaching. Report 731 F. Minneapolis, MN: Measurement Services Center, University of Minnesota.

Whitlock, L. G. (1972). The dimensions of observer perceptions of teacher performance. Ph.D. dissertation, University of Tennessee.

Widlak, F. W., and Quereshi, M. Y. (1972). Student characteristics and instructor ratings: A person-perception approach. Presented at the Annual Meeting of the American Psychological Association.

Wiggins, J. S. (1973). Personality and Prediction: Principles of Personality Assessment. Reading, MA: Addison-Wesley.

Wilson, D., and Doyle, K. O., Jr. (1976). Student ratings of instruction: Student and instructor sex interactions. Journal of Higher Education 47: 465–470.

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

Winer, B. J. (1962). Statistical Principles in Experimental Design. New York: McGraw-Hill.

Yonge, G. D., and Sassenrath, J. M. (1968). Student personality correlates of teacher ratings. Journal of Educational Psychology 59: 44–52.