EFFECTIVENESS OF STUDENT-RATING FEEDBACK FOR IMPROVING COLLEGE INSTRUCTION: A Meta-Analysis of Findings

Peter A. Cohen

This article applied meta-analytic methodology to integrate findings from 22 comparisons of the effectiveness of student-rating feedback at the college level. On the average, feedback had a modest but significant effect on improving instruction. Instructors receiving mid-semester feedback averaged .16 of a rating point higher on end-of-semester overall ratings than did instructors receiving no mid-semester feedback. This corresponded to a gain of over one-third of a standard-deviation unit, or a percentile gain of 15 points. The effects of student-rating feedback were accentuated when augmentation or consultation accompanied the ratings. Other study features, such as the length of time available to implement changes and the use of normative data, did not produce different effect sizes.

Over the last decade, there has been a dramatic increase in the use of student ratings of instruction in colleges and universities across the country. Generally, student ratings may serve three functions: (1) aiding administrative evaluations of teaching effectiveness for decisions concerning pay increases, promotion, and tenure; (2) providing feedback to teachers for the purpose of improving instruction; and (3) helping students select courses and instructors. The present paper focuses on the use of student ratings for improving instruction.

With the prevalent use of students as a data source for information on teaching, the body of literature concerning student ratings has become voluminous. The concensus seems to be that student ratings are reliable,

Peter A. Cohen, Office of Instructional Services and Educational Research, Dartmouth College. The research reported here was conducted while the author was at the Center for Research on Learning and Teaching, The University of Michigan.

Research in Higher Education Vol. 13, No. 4, 1980

0361-0365/80/080321-21$01.50

valid, and not influenced to an undue extent by extraneous factors (Costin, Greenough, & Menges, 1971; Kulik & McKeachie, 1975; McKeachie, 1979; Seibert, 1979; Stumpf, Freedman, & Aguanno, 1979). While there is some reservation concerning the use of student-rating data as sufficient criteria for administrative decisions of teaching effectiveness (e.g., Cohen & McKeachie, in press; Menges, 1979), there is little controversy over their use for purposes of improving instruction.

Efforts to improve instruction may be classified into two categories: (1) instructor development, the improvement of general teaching abilities in the instructor over time, and (2) within-class improvement, improvement in instructional effectiveness evidenced over the course of a semester.

Instructor development is evidenced by long-range outcomes. For example, the instructor should show improvement in course design skills (e.g., clarity and appropriateness of course goals and objectives, measuring student attainment of objectives, etc.). Besides developing general teaching abilities, the instructor should also learn to incorporate diagnostic feedback and improvement strategies into his or her conception of the teaching-learning process. In addition, the individual's commitment to and enjoyment of teaching as a career role should be strong for those instructors who have strived to improve their teaching over a period of time.

Within-class improvement involves increased effective interaction among teacher, instructional method(s), and students throughout a semester. While the instructor can improve general teaching abilities over the course of his or her career, many teaching dimensions are relevant only in the context of a particular class of students. A large component of instructional improvement involves adapting, adjusting, and changing instruction to meet the combined needs of the students and the instructor in a particular classroom context.

In terms of providing feedback, the use of student ratings seems most appropriate for within-class improvement efforts. Cohen and Herr (1979a) note three potential advantages of using within-class improvement strategies. First, students may get increasingly better instruction as the semester progresses. At the very least, efforts to improve the course will be perceived positively by students; for example, they will feel that they have some stake in the rating process. Second, the instructor becomes actively involved in a faculty development process; not only does teaching improve but also the resistance to instructional evaluation dissipates, since these strategies are seen as facilitative of improvement rather than as judgmental. Third, the intellectual and interpersonal satisfactions derived from teaching-related experiences are likely to become more salient when the instructor can overcome the frustrations associated with successfully

executing the tasks of teaching. Implementation of within-class improvement strategies should increase the instructor's sense of competence. Both McKeachie (1979) and Bess (1977) stress the importance of enhancing the instructor's intrinsic motivation for teaching.

While there seems to be substantial support for using student ratings as a source of feedback, it is not clear how effective such information can be in terms of improving instruction. A number of reviewers have noted mixed findings—some studies reveal a significant influence of student-rating feedback, while other studies show no such effects. Two recent articles have been published dealing specifically with the effects of student evaluative feedback on instructional improvement. However, neither review draws firm conclusions from the literature in this area. Rotem and Glasman (1979) report that "feedback from student ratings (as was elicited and presented to teachers in the studies reviewed) does not seem to be effective for the purpose of improving performance of university teachers" (p. 507). Using a somewhat different set of studies, Abrami, Leventhal, and Perry (1979) state, "there seems to be enough evidence to conclude that feedback from student ratings leads some instructors to improve their subsequent student ratings. However, the effect is not reliable judging from the inconsistency of findings across studies" (p. 361).

According to other reviewers in this area (Kulik & Kulik, 1974; Kulik and McKeachie, 1975; McKeachie, 1979), there seem to be four possible explanations for the failure of teachers to improve following student-rating feedback. First, the feedback should provide new information. Some studies assessing both student and self-ratings have found more improvement for instructors whose self-ratings were discrepant from their student ratings (e.g., Centra, 1973; Pambookian, 1974). Second, it may be difficult to implement changes within the short time span of a semester. Centra's (1973) study indicates that teaching improvement may not be realized until subsequent semesters. Third, normative data may be needed to help instructors determine where their strengths and weaknesses lie. Finally, instructors may not know *how* to modify their teaching once they receive student rating feedback. Consultants may play an important role here, by helping to interpret feedback and suggesting change strategies.

Characteristics of the feedback implementation, which vary from study to study, may contribute to the lack of consistent findings across studies. Therefore, drawing overall conclusions from the reviews in this area proves difficult. These reviewers did not use objective, statistical methods to find the characteristics that distinguished between studies reporting substantial effects and studies reporting negligible results. In his presidential address to the American Educational Research Association, Glass (1976) proposed a method for handling the difficulties posed by the diver-

sity of findings in the social sciences. This approach, which he called "meta-analysis," involved the statistical analysis of a collection of results from individual studies for the purpose of integrating findings.

This paper applies Glass's meta-analytic methodology to research on the effectiveness of student-rating feedback for improving instruction at the college level. The analysis focuses on three questions. First, how effective is student-rating feedback in the typical comparative study? Second, is student-rating feedback especially effective for certain dimensions of instruction? Third, under which conditions does student-rating feedback appear to be most effective? It is intended that by addressing these questions from a meta-analytic framework, more precise conclusions concerning the effects of student ratings on instructional improvement can be reached.

METHODS

This section describes the methods that were used to locate studies, to code study features, and to quantify outcomes. The methods were similar to those used in previous meta-analyses of instructional technologies in higher education by Kulik and his colleagues (Kulik, Cohen, & Ebeling, in press; Kulik, Kulik, & Cohen, 1979a, 1979b).

Locating Studies

The first step in the meta-analysis was to collect a number of studies that compared the effects of student-rating feedback versus no feedback on instructional improvement. Primary sources for these studies were major reviews (Abrami et al., 1979; Rotem & Glasman, 1979) and four data-based indices: *Psychological Abstracts, Comprehensive Dissertation Abstracts, Cumulative Index to Journals in Education* and *Research in Education.* Secondary sources for reports on student-rating feedback were bibliographies contained in articles located through these reviews and indices.

To be included in the analysis, an article had to meet four basic criteria. First, it had to describe a study that was conducted in an actual college class. Articles on the effectiveness of student-rating feedback in secondary schools were not included (e.g., Tuckman and Oliver, 1968). Second, the study had to compare teaching improvement outcomes in two groups of instructors, those receiving student-rating feedback and those not receiving feedback. Studies without a control group (e.g., Pambookian, 1974; Vogt & Lasher, 1973) or studies using students rather than instructors as the unit of analysis (e.g., Marsh, Fleiner, & Thomas, 1975) were not included. Third, the instructors being compared had to hold the major

responsibility for teaching the class. Studies using graduate student teaching assistants (TAs) were included only if this condition was met. Fourth, the study had to be free from major methodological flaws, e.g., uncontrolled differences in initial levels of teaching ability between groups. Only comparisons that used random assignment of instructors to groups or that made statistical adjustments for differing initial teaching abilities were included. In fact, many of the studies used in the following analysis met both of these conditions.

Additional guidelines helped insure that the set of comparisons used in the analysis was as complete and representative as possible. When several papers reported the same comparison, the single, most complete report was used. When there were two distinct implementations of feedback in the same study (such as receiving consultation with the ratings or receiving ratings only), a separate comparison was obtained for each implementation. When two or more groups received the same type of feedback within a study, a single comparison was obtained.

Using these criteria, a total of 17 articles were located containing data that could be used in the meta-analysis. The 17 articles reported on 22 separate comparisons of student-rating feedback versus no feedback. The 17 studies are listed in Table 1.

Description of Studies

The 17 studies located for this analysis took place in a variety of settings and measured outcomes in various ways. The next step in the meta-analysis was to describe the relevant study characteristics and outcomes.

Study Characteristics. To characterize the studies more precisely, nine variables were defined. Two of these described methodological features: method of subject assignment and control for prefeedback ratings. Three other variables described ecological conditions under which feedback and no feedback groups were compared. These conditions included the duration of time that occurred between the feedback implementation and the outcome measure, the type of institution at which the comparison took place, and the experience of the instructors. Another three variables described the nature of the feedback. These variables specified whether or not the student-rating feedback was accompanied by consultation and/or improvement strategies, whether a self-rating discrepancy analysis was utilized, and whether or not normative data were included in the feedback given instructors. The final variable described whether or not the study was published. The coding categories for each of these variables and the number of comparisons in each category are listed in Table 2.

Study Outcomes. The next step in the meta-analysis was to express the

TABLE 1. Major Characteristics of Student-Rating Feedback Studies.

Study	Design features	Nature of feedback	Outcome measures
Aleamoni (1978)	Control group consisted of instructors who volunteered for consultation but could not schedule a meeting with the consultant due to time constraints. End-of-semester rating results obtained for the following semester.	15–20 min personal consultation: descriptive information, normative data, problem identification, change strategies.	Ratings of instructor: skill.
Braunstein, Klein, & Pachla (1973)	Random assignment of instructors to feedback and no feedback groups. Data obtained on faculty expectation regarding student ratings.	Descriptive item feedback, normative data.	Ratings of instructor.
Carter (1974)	Random assignment of teaching assistants to feedback and no feedback groups. Data obtained on TA perception concerning credibility of students as a feedback source.	Descriptive item feedback, normative data, student written comments.	Ratings of instructor: skill, rapport.
Centra (1973)	Random assignment of instructors to feedback and no feedback groups. End-of-semester rating results also obtained for the following semester.	Descriptive item feedback, normative data.	Ratings of instructor: skill, feedback.

Cohen & Herr (1979b)	Random assignment of teaching assistants to three groups: no feedback, descriptive feedback (feedback only), and descriptive feedback with self-programmed improvement guide (interactive). Self-rating analysis. Covariance analysis.	Feedback only group: descriptive item feedback. Interactive group: descriptive item feedback, student-self-rating discrepancy profile, self-programmed improvement booklet.	Ratings of instructor: skill, rapport, structure, interaction.
Erickson & Erickson (1979)	Random assignment of instructors to no feedback and consultation groups. Self-rating analysis.	Consultation procedure included feedback on mid-semester ratings, interview, classroom observation, and videotape.	Ratings of instructor: skill, rapport structure, interaction.
Howard (1977)	Random assignment of instructors to no feedback and faculty development groups. Covariance analysis.	Groups of eight instructors engaged in group sessions/dyads: discussions of student rating feedback, presentations on instructional issues and teaching methods, use of classroom observation and videotape.	Ratings of instructor: skill, rapport, structure.
Hoyt & Howard (1978)	Instructors divided into groups post hoc based on degree of contact with faculty development office (much, some, none). Covariance analysis.	Consultation on interpreting student ratings, assessing student performance, instructional development matters.	Ratings of instructor: skill, rapport, difficulty. Student progress ratings. Attitude toward subject ratings.

TABLE 1 *(Continued)*

Study	Design features	Nature of feedback	Outcome measures
McKeachie & Lin (1975)	Random assignment of teaching assistants to no feedback, printed feedback, and personal feedback groups.	Printed feedback group: descriptive item feedback, normative data. Personal feedback group: printed feedback plus consultation, including analysis of discrepancies between actual student ratings, expected student ratings, self-ratings, and ideal ratings. Discussion of change strategies.	Ratings of instructor: skill, rapport, structure, difficulty, interaction, feedback. Attitude toward subject ratings. Student achievement.
Miller (1971)	Random assignment of teaching assistants to feedback and no feedback groups. Data obtained on TA attitudes toward value of student ratings. Covariance analysis.	Descriptive item feedback.	Ratings of instructor.
Murphy & Appel (1978)	Random assignment of instructors to augmented feedback, feedback only, and no feedback groups. Data obtained on instructors' performance standards.	Feedback only: descriptive item feedback. Augmented feedback: descriptive item feedback/performance standards discrepancy analysis, remedial alternatives.	Ratings of instructor.

Overall & Marsh (1979)	Random assignment of teaching assistants to no feedback and consultation groups. Covariance analysis.	Consultation on interpreting student ratings (descriptive item and factor feedback, normative data). Discussion of strategies for improving instruction.	Ratings of instructor: skill, rapport, structure, difficulty, interaction. Student progress ratings. Attitude toward suject ratings. Student achievement.
Rotem (1975)	Random assignment of instructors to feedback and no feedback groups. "Actual-ideal" discrepancies obtained for both students and instructors.	Descriptive item feedback for student actual-ideal ratings.	Ratings of instructor: skill, rapport, structure, difficulty, interaction, feedback.
Scott (1976)	Random assignment of instructors to feedback and no feedback groups.	Descriptive item feedback.	Ratings of instructor.
Smith (1977)	Random assignment of team-teaching instructors to feedback and no feedback groups.	Descriptive item feedback.	Ratings of instructor.
Thomas (1969)	Random assignment of instructors to feedback and no feedback groups. "Difference score" for real-ideal teacher ratings obtained.	Descriptive item feedback (profile) for "real" and "ideal" student ratings.	Ratings of instructor.
Weerts (1978)	Random assignment of teaching assistants to no feedback, printed feedback, and verbal feedback groups.	Printed feedback: descriptive item feedback, normative data. Verbal feedback and meeting with supervisor concerning ratings. Discussion of improvement strategies.	Ratings of instructor: skill, rapport, interaction, feedback. Student progress ratings.

TABLE 2. Categories for Describing Studies and Number of Studies in Each Category.

Coding categories	No. of comparisons
Methodological features	
Random assignment of comparison groups	
1. No	2
2. Yes	20
Statistical control for pre-feedback ratings	
1. No	12
2. Yes	10
Ecological conditions	
Duration	
1. One semester	19
2. More than one semester	3
University setting	
1. Comprehensive, liberal arts or community college	7
2. Doctorate-granting institution	15
Instructor experience	
1. Teaching assistant	9
2. Faculty member	13
Nature of the feedback	
Type of use	
1. Ratings only	13
2. Ratings and augmented feedback/consulting	9
Use of self-rating discrepancies	
1. No	14
2. Yes	8
Use of normative data	
1. No	12
2. Yes	10
Publication feature	
Source of study	
1. Unpublished	12
2. Published	10

outcomes of each comparison in quantitative terms. Outcomes used in the studies were of four major types: student ratings of the instructor, student ratings of their own learning, student attitudes toward the subject matter, and student achievement.

Student ratings of the instructor were categorized on seven dimensions. Initially, data were collected for an overall teaching effectiveness dimension. These data came from either a single rating item concerning overall teaching effectiveness or an average of all items or dimensions on a rating form in a particular study. Because instructional improvement may be more likely to occur for some aspects of teaching as opposed to others, rating data were also collected on six additional dimensions of teaching. Four of these dimensions were identified by Kulik and McKeachie (1975) as "common" factors in their review of factor analytic studies of student ratings. These four dimensions are skill, rapport, structure, and difficulty. Two other dimensions, interaction and feedback to students, were interpreted by Isaacson et al. (1964), and were used in a number of the studies collected for the present meta-analysis. All ratings for these six dimensions and for the overall teaching effectiveness dimension were converted to a five-point scale, where 5 represented the highest rating (i.e., high skill, high difficulty level, etc.) and 1 represented the lowest possible rating.

The other outcomes were student ratings of their learning, student attitude toward the subject matter, and student achievement. Student ratings of their progress and their attitude toward the subject matter were converted to the same five-point scale that was used for the instructor ratings. Student achievement was measured by performance on common final examinations.

Cohen's (1977) "pure" measure of effect size was calculated as the basic index of effect for these major outcomes. Cohen's *d*, defined as the difference between the means for two groups divided by the standard deviation common to the two populations, gives the size of effect when group averages are compared. One advantage of using *d* is that it not only provides a common metric, but also it can be calculated from different sources of data.

In a previous meta-analysis of research on audio-tutorial instruction, Kulik et al. (1979b) reported that different effect size measures agreed remarkably well when applied to the same data set. Because the correlations were so high, they were able to write regression equations for "plugging" missing data on specific effect size measures. In the present analyses, the two indices of effect were (1) the difference between ratings for feedback and no feedback groups on a five-point scale, and (2) Cohen's *d*. The correlation between these two measures of effect was .98 for the total rating measuring overall teaching effectiveness. Therefore, if a study did not report mean student ratings, but did report data from which Cohen's *d* could be calculated, Cohen's *d* was used to predict the difference on a five-point rating scale separating feedback and no feedback groups. Likewise, student rating differences were used to predict Cohen's *d*. In this

way, both measures of effect were obtained for each comparison in the analysis.

RESULTS

This section reports the results of two different analyses. The first analysis examined the overall size and significance of effects of student-rating feedback. The second analysis was conducted to determine whether reported effects of student-rating feedback were different for different types of studies and under different conditions.

Overall Effects

In the first set of analyses, simple descriptive statistics were used to compare the results of student-rating feedback versus no feedback. Results were compiled on four outcome criteria: (1) student ratings of the instructor; (2) student ratings of their progress; (c) student attitude toward the subject matter area; and (4) student achievement. Table 3 presents the means, average effect sizes, and the results of a statistical analysis in which each comparison was treated as a single case for these outcome measures.

Student Ratings of the Instructor. The feedback group received higher end-of-term "total" ratings than the no feedback group in 20 of the 22 comparisons. A total of 10 of the 22 comparisons reported statistically significant differences between feedback and no feedback groups, and in each case, the comparison favored the feedback group.

Continuous measures of effect size permit a more exact description of the influence of student-rating feedback on end-of-term ratings. As shown in Table 3, the average total rating was 3.86 for the feedback groups; the average total rating was 3.70 for the no feedback groups. Total end-of-term ratings for feedback and no feedback groups differed, therefore, by .16 of a rating point (on a five-point scale); the standard deviation of this difference was .19. It is statistically very unlikely ($p < .001$) that a difference of this size would be found if there were no overall difference in effectiveness of receiving or not receiving midterm student-rating feedback.

The average Cohen's d for the 22 comparisons was .38. Thus the effect of student-rating feedback in a typical comparison was to raise end-of-term ratings by more than one-third of a standard-deviation unit. This implies that a typical instructor receiving student-rating feedback was performing at the 65th percentile (as measured by end-of-term "total" rating), whereas the typical control group instructor performed at the 50th percentile. Cohen (1977) described effects of this magnitude as modest. This is to be contrasted with small ($d = .2$) and large ($d = .8$) effects.

TABLE 3. Mean Ratings and Effect Sizes for Major Outcome Measures.

Outcome	Number of comparisons	Means			Average effect size
		Feedback	Control		
Total Rating	22	3.86	3.70	4.21**	.38
Skill	14	3.96	3.80	4.34**	.47
Rapport	11	4.13	4.02	2.06	.25
Structure	8	3.65	3.46	2.05	.29
Difficulty	5	2.45	2.40	1.46	.09
Interaction	9	4.01	3.95	1.33	.09
Feedback to students	7	3.65	3.49	2.79*	.40
Student progress	4	3.87	3.77	3.05	.30
Attitude toward subject	4	3.30	3.15	4.41*	.42
Student achievement	4				.19

*$p < .05$.
**$p < .001$.

Table 3 also presents results of the analyses concerning data collected on the six dimensions of student ratings of instruction. For two of these dimensions, skill and feedback to students, midterm student-rating feedback had a significant impact on end-of-term ratings. Average effect sizes for these dimensions were .47 and .40, respectively. For the dimensions of rapport, structure, difficulty, and interaction, however, midterm feedback did not lead to significant increases in end-of-term ratings.

Other Outcome Measures. Student progress ratings measured students' perception of their learning or progress in a course. Three studies (four comparisons) reported data for this outcome. All four comparisons favored the feedback group; however, only one comparison was statistically significant. Continuous measures of effect size showed that students whose instructors received midterm feedback did not rate their own learning significantly higher than students whose instructors did not receive feedback. Cohen's *d* for this measure averaged .30, a small to modest effect.

Another three studies (four comparisons) presented data on student attitudes toward the subject area. For all four comparisons, differences in student attitudes toward the subject favored the feedback group. In two of these comparisons, the difference was statistically significant. For these four comparisons, continuous measures of effect size showed that students whose instructors received midterm feedback were significantly more positive in their attitudes toward the subject matter than students whose instructors did not receive feedback. This effect corresponded to a Cohen's *d* of .42.

Student achievement data were available from three studies (four comparisons). Not enough information was presented in these studies to convert achievement data to a common metric (i.e., difference in examination score between the two groups, expressed as a percentage). Three of the four comparisons showed achievement to be greater for students whose instructors received feedback, while one comparison favored students whose instructors did not receive feedback. In none of the comparisons was the difference statistically significant. For this set of comparisons, Cohen's *d* averaged .19, a small effect.

Study Characteristics and Study Outcomes

The purpose of the second set of analyses was to determine whether the comparisons that reported large effects differed systematically from those reporting small effects. Table 4 presents the correlations between total rating effect sizes and study characteristics. The table shows that only one variable, augmentation of feedback, was significantly related to effect size. Figure 1 presents the distribution of total rating effect sizes for augmented

TABLE 4. Correlations of Study Characteristics with Total Rating Effect Size.

Study characteristic	Correlation with effect size
Random assignment of comparison groups	−.06
Statistical control for prefeedback ratings	−.07
Duration	.09
University setting	.09
Instructor experience	.03
Use of augmentation of feedback	.64*
Use of self-rating discrepancies	.26
Use of normative data	.11
Source of study	.08

*$p < .01$.

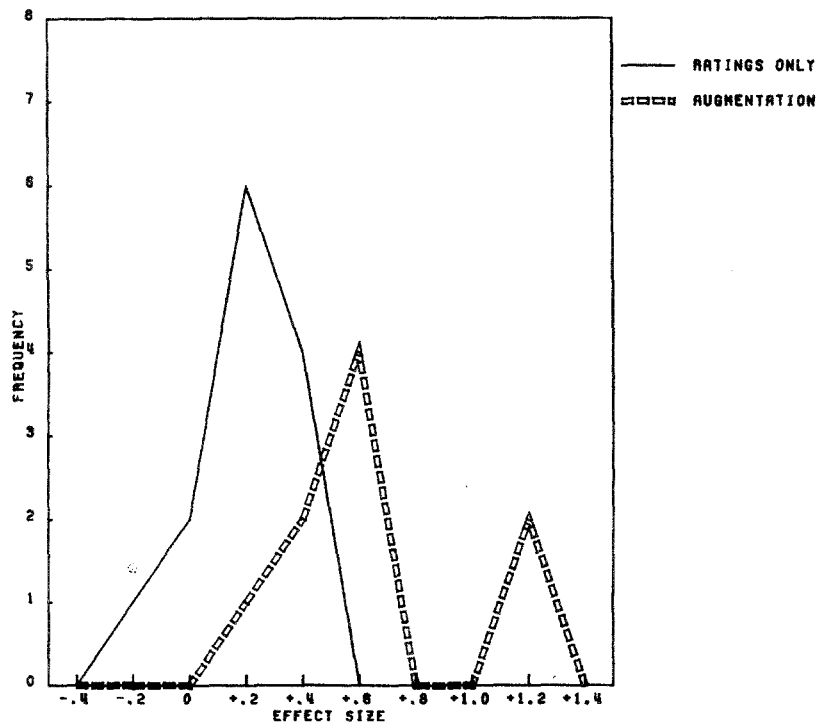

FIGURE 1. Distribution of total rating effect sizes for augmented and ratings only studies.

TABLE 5. Comparison of Rating Effect Sizes for Augmented and Ratings Only Groups.

Rating dimension	Average effect size	
	Augmented	Rating only
Total rating	.64	.20
Skill	.59	.30
Rapport	.44	.03
Structure	.39	.11
Difficulty	.00	.23
Interaction	.21	−.06
Feedback	.50	.36

feedback comparisons and ratings only comparisons. The distributions are clearly different.

For total rating, Cohen's *d* averaged .64 for comparisons that used some sort of augmentation or consultation in conjunction with student rating feedback. Effects were smaller when ratings only feedback was used. Cohen's *d* averaged only .20 for these comparisons. Thus a typical instructor receiving augmented feedback was performing (at the end of the semester) at the 74th percentile. This can be compared with the typical instructor receiving only student ratings (58th percentile) and the typical instructor receiving no mid-semester feedback (50th percentile). Effect size comparisons for the other six rating dimensions are shown in Table 5.

Other study features were less highly related to effect size. None of the correlations between effect size and the remaining variables could be considered significantly different from zero. To investigate the possibility that a combination of study features might predict effect sizes more accurately than a single predictor, a stepwise multiple regression analysis was conducted. Results of this analysis were clear-cut. Once augmentation of feedback was taken into account, none of the variables was significantly related to effect size.

DISCUSSION

This meta-analysis showed that for the most part student-rating feedback has made a modest but significant contribution to the improvement of college teaching. These findings do not totally support the conclusions of recent reviewers in this area. For example, Abrami et al. (1979) and Rotem and Glasman (1979) have suggested that although student-rating feedback

may be useful for some instructors, it has had little overall impact on increasing instructional effectiveness. The use of meta-analytic techniques, however, makes it possible to reach more exact conclusions about the effects of student-rating feedback.

The first set of analyses reported on the overall size and significance of effects of student-rating feedback. In the typical implementation, feedback raised instructors' end-of-term overall rating by .16 of a rating point, or over one-third of a standard deviation. In addition to this overall effect, feedback led to increased end-of-term ratings for two of the six teaching dimensions, skill and feedback to students. The skill dimension is undoubtedly highly related to the "total" rating. Kulik and Kulik (1974) state that this dimension is the overriding quality to which students respond when rating instructors. For the present collection of studies, feedback on the average raised instructors' skill ratings by almost a half standard deviation, a moderate effect. The feedback to students dimension measures the instructor's concern with the quality of students' work. Typical items for this dimension are "The instructor tells students when they have done a particularly good job," and "The instructor checks to see that we have learned well before we go on to new material." For studies reporting feedback to students data, Cohen's *d* averaged .40.

It is interesting to note that feedback does not lead to significant increases in ratings for all dimensions of teaching. Ratings on some dimensions, for example, difficulty and interaction, may be more influenced by course setting characteristics (e.g., subject matter differences, class size) and therefore may not be a true reflection of an instructor's effectiveness. It has also been suggested (Cohen & Herr, 1979a; Sherman & Winstead, 1975) that feedback must be of a specific nature to be useful. Ratings on a dimension such as feedback to students provides the instructor with specific information, giving more direction to possible instructional changes.

Data for the other outcome measures were not reported as fully by the studies used in this meta-analysis. Only four comparisons were available for student progress ratings, attitude toward the subject and student achievement. Students whose instructors received midterm rating feedback did not learn more than students whose instructors did not receive feedback. This is evidenced by both students' self-report of learning and their scores on achievement measures. However, there is some indication that students whose instructors received feedback did rate the subject-matter area higher than did students whose instructors did not receive feedback. Perhaps a noticeable improvement effort on the part of the instructor generates more student enthusiasm for the subject matter. These results, though, are at best suggestive because of the small number of available comparisons.

The second set of analyses reported on the relationship of study features and study outcomes. In general, there was little relationship between methodological features and study outcomes. Almost all studies randomly assigned instructors to feedback groups. Studies statistically controlling for pre-feedback ratings produced the same results as studies without this control. Nor did settings influence findings in any substantial way. Findings were similar in comparisons lasting one semester and more than one semester. Findings were the same for different types of schools and for instructors with differing amounts of experience. Concerning the nature of the feedback, neither comparisons using self-rating/student rating discrepancies nor those using normative data produced differing results.

The only variable that correlated significantly with total rating outcome was use of augmentation of feedback. Comparatively large effect sizes emerged from studies using augmented feedback or consultation in conjunction with student-rating feedback. Studies using only student-rating feedback produced much smaller effects. These results clearly suggest that instructors need more than just student-rating feedback to markedly improve their instruction.

The studies by Erickson and Erickson (1979) and McKeachie and Lin (1975) are important because these investigators reported especially strong effects of consultation in conjunction with student-rating feedback. In each of these implementations, the consultants were "experts" in the area of college teaching. As in many of the other studies using consultation, improvement goals and change strategies were discussed.

In at least two respects, the meta-analysis does not confirm the impressions of other reviewers concerning the relationship of student-rating feedback to instructional improvement. First, the present analysis shows that the length of time available to implement changes may not be a critical factor in determining whether or not instructional improvement takes place. The findings from the three studies that measured overall teaching effectiveness in subsequent semesters did not differ from studies assessing change from mid- to end-of-semester. Second, the use of normative data does not seem to enhance instructional improvement. Comparisons with one's colleagues perhaps belong with a more evaluative use of student ratings. For formative purposes, personal norms may be more appropriate.

What implications do these findings have for using student-rating feedback to facilitate within-class improvement? It is evident that when instructors are left to their own resources, ratings provide little help. Augmented feedback, or more specifically expert consultation, seems to be the key element for making student-rating data useable for improvement purposes. It is also clear that instructional change cannot be accomplished on all teaching dimensions. Therefore, instructors should only request

midterm student feedback for aspects of teaching that they are able to modify. Finally, rating items and dimensions that supply specific information should be preferred to global ratings. With these considerations, student ratings are a valuable data source for improving instruction at the college level.

ACKNOWLEDGMENTS

The author thanks James Kulik and Wilbert J. McKeachie, who contributed to the development of the research described in this article.

REFERENCES

Abrami, P. C., Leventhal, L., & Perry, R. P. Can feedback from student ratings help to improve college teaching? *Proceedings of the 5th International Conference on Improving University Teaching*. London: 1979.

Aleamoni, L. M. The usefulness of student evaluations in improving college teaching. *Instructional Science*, 1978, *7*, 95-105.

Bess, J. L. The motivation to teach. *Journal of Higher Education*, 1977, *48*, 243-258.

Braunstein, D. N., Klein, G. A., & Pachla, M. Feedback expectancy and shifts in student ratings of college faculty. *Journal of Applied Psychology*, 1973, *58*, 254-258.

Carter, K. R. The effect of student feedback in modifying teaching performance. Unpublished doctoral dissertation, University of Georgia, 1974.

Centra, J. A. Effectiveness of student feedback in modifying college instruction. *Journal of Educational Psychology*, 1973, *65*, 395-401.

Cohen, J. *Statistical power analysis for the behavioral sciences* (Rev. ed.). New York: Academic, 1977.

Cohen, P. A., & Herr, G. A procedure for diagnostic instructional feedback: The Formative Assessment of College Teaching (FACT) model. *Educational Technology*, 1979, *19*, 18-23. (a)

Cohen, P. A., & Herr, G. *Improving instruction in the college classroom: does mid-semester rating feedback make a difference?* Ann Arbor: The University of Michigan, Center for Research on Learning and Teaching, 1979(b).

Cohen, P. A., & McKeachie, W. J. *The role of colleagues in the evaluation of college teaching. Improving Collegeand University Teaching*, in press.

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

Erickson, G. R., & Erickson, B. L. Improving college teaching: An evaluation of a teaching consultation procedure. *Journal of Higher Education*, 1979, *50*, 670-683.

Glass, G. V. Primary, secondary, and meta-analysis of research. *Educational Researcher*, 1976, *5*, 3-8.

Howard, G. S. A program to improve instruction: A promising area for psychologists. *Professional Psychology,* 1977, *8,* 316-327.

Hoyt, D. P., & Howard, G. S. The evaluation of faculty development programs. *Research in Higher Education,* 1978, *8,* 25-38.

Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Y. G., Hofeller, M., Baerwaldt, J. W., & Zinn, K. L. Dimensions of student evaluations of teaching. *Journal of Educational Psychology,* 1964, *55,* 344-351.

Kulik, J. A., Cohen, P. A., & Ebeling, B. J. Effectiveness of programmed instruction in higher education: A meta-analysis of findings. *Educational Evaluation and Policy Analysis,* in press.

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

Kulik, J. A., Kulik C-L. C., & Cohen, P. A. A meta-analysis of outcome studies of Keller's personalized system of instruction. *American Psychologist,* 1979, *34,* 307-318. (a)

Kulik, J. A., Kulik, C-L. C., & Cohen, P. A. Research on audiotutorial instruction: a meta-analysis of comparative studies. *Research in Higher Education,* 1979, *11,* 321-341. (b)

Kulik, J. A., & McKeachie, W. J. The evaluation of teachers in higher education. In F. N. Kerlinger (Ed.), *Review of research in education* (Vol. 3). Itaska, Ill.: F. E. Peacock Publishers, 1975.

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

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

McKeachie, W. J., & Lin, Y. G. *Use of student ratings in evaluation of college teaching.* Ann Arbor: The University of Michigan, Department of Psychology, 1975.

Menges, R. J. Evaluating teaching effectiveness: What is the proper role for students? *Liberal Education,* 1979, *65,* 356-370.

Miller, M. T. Instructor attitudes toward, and their use of, student ratings of teachers. *Journal of Educational Psychology,* 1971, *62,* 235-239.

Murphy, J. B., & Appel, V. H. The effect of mid-semester student feedback on instructional change and improvement. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.

Overall, J. U., & Marsh, H. W. Midterm feedback from students: Its relationship to instructional improvement and students' cognitive and affective outcomes. *Journal of Educational Psychology,* 1979, *71,* 856-865.

Pambookian, H.S. The initial level of student evaluation of instruction as a source of influence on instructor change after feedback. *Journal of Educational Psychology,* 1974, *66,* 52-56.

Rotem, A. *The effects of feedback from students to university professors: an experimental study.* Unpublished doctoral dissertation, University of California, Santa Barbara, 1975.

Rotem, A., & Glasman, N. S. On the effectiveness of students' evaluative feedback to university instructors. *Review of Educational Research,* 1979, *49,* 497-511.

Scott, N. H. *A pilot study testing effects of feedback on teaching behavior as measured by student ratings at New River Community College, Dublin, Virginia.* Unpublished doctoral dissertation, Virginia Polytechnic Institute and State University, 1976.

Seibert, W. F. Student evaluations of instruction. In S. C. Ericksen (Ed.), *Support for teaching at major universities.* Ann Arbor, MI: The University of Michigan, Center for Research on Learning and Teaching, 1979.

Sherman, T. M., & Winstead, J. C. A formative approach to student evaluation of instruction. *Educational Technology,* 1975, *15,* 34-39.

Smith, D. L. *The relationship of feedback to professors of the results of student ratings of their teaching effectiveness at mid-semester to their end-of-semester ratings.* Unpublished doctoral dissertation, University of Southern California, 1977.

Stumpf, S. A., Freedman, R. D., & Aguanno, J. C. A path analysis of factors often found to be related to student ratings of teaching effectiveness. *Research in Higher Education,* 1979, *11,* 111-123.

Thomas, H. B., Jr. *Changing teacher behavior: an experiment in feedback from students to teachers.* Unpublished doctoral dissertation, Purdue University, 1969.

Tuckman, B. W., & Oliver, W. F. Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology,* 1968, *59,* 297-301.

Vogt, K. E., & Lasher, H. *Does student evaluation stimulate improved teaching?* Bowling Green, Ohio: Bowling Green State University, 1973 (ERIC Document Reproduction Service No. ED 078 748).

Weerts, R. R. The use of feedback from student ratings for improving college teaching. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March 1978.