Journal of Educational Psychology
1973, Vol. 65, No. 3, 395-401

EFFECTIVENESS OF STUDENT FEEDBACK IN MODIFYING COLLEGE INSTRUCTION[1]

JOHN A. CENTRA[2]

Educational Testing Service, Princeton, New Jersey

An experimental study was conducted at five colleges to investigate the extent to which college teachers modify their instructional practices after receiving student feedback. Variables included teaching experience, sex, and self-ratings of the instructor, as well as course subject area. On the basis of equilibrium theory, a major hypothesis of this study was that student ratings would produce changes in teachers who had rated themselves more favorably than their students had rated them. Results of a regression analysis generally supported this hypothesis. A second conclusion of the study was that additional time (i.e., more than a half of a semester), along with comparative data to help the individual teacher interpret his feedback, also helped produce modest changes in teachers' instructional practices.

Formal student evaluations of courses and teaching have been receiving a good deal of attention and use at many colleges and universities. The results of these evaluations most often are seen only by instructors and are intended to help improve their teaching. Underlying this intended use is the assumption that the instructors will use the information to alter and improve their teaching. It is an assumption open to question.

Presumably, instructors value student opinion enough to change their teaching behavior when it is evaluated less favorably than the instructors might desire or expect. The theoretical justification for this belief, as developed by Gage, Runkel, and Chatterjee (1963) and by Daw and Gage (1967), may be found in equilibrium theory. Accordingly, when student feedback creates a condition of imbalance (Heider, 1958), asymmetry (Newcomb, 1959), or dissonance (Festinger, 1957) in an instructor, one might expect the instructor to change in the direction desired by students in order to restore a condition of "equilibrium." Following a suitable lapse of time, such changes should be reflected in a second description of teacher behavior.

There is some evidence that student feedback does indeed have a positive effect on teaching performance, although the evidence is far from conclusive, particularly at the college level. Tuckman and Oliver (1968), using 286 teachers of vocational subjects in high school and technical institutes, found that instructors who received student feedback showed greater "gains" in student ratings, as measured by changes in students' ratings after a 12-week interval, than instructors who received no feedback. (Actually all the change scores were negative, with positive changes or gains being simply less of a negative score.) Changes in ratings of teaching were also reported by Bryan (1963), using teachers of academic subjects at the secondary level, and by Gage et al. (1963), who experimented with sixth-grade teachers.

The results at the college level, however, have thus far been less positive. Miller (1971) reported that end-of-semester student ratings for teaching assistants who had received midsemester feedback did not differ from end-of-semester ratings for teaching assistants who did not receive the feedback. But because of the small and limited

[1] This study was supported by a grant from the Exxon Education Foundation.

[2] Requests for reprints should be sent to John A. Centra, Educational Testing Service, Princeton, New Jersey 08540.

nature of the sample (36 teaching assistants assigned to discussion sections in three courses), the results of the Miller study are very tentative.

The preceding studies neglected to include a number of relevant variables that might be expected to be related to changes in teaching. None of the studies investigated the instructor's awareness of his own teaching practices as indicated by self-ratings. On the basis of equilibrium theory, one could hypothesize that the greater the gap between student ratings and faculty self-ratings, the greater the likelihood that there would be change in instruction, since large differences would create the greatest amount of imbalance or dissonance in instructors. None of the preceding studies, furthermore, looked at possible variations in changes across subject areas nor did they investigate the sex of the instructor as still another variable. Finally, the number of years of teaching experience is a particularly critical variable, which was included in only one of the preceding studies. In that study (Tuckman & Oliver, 1968), the expectation that less experienced teachers would be more likely to change was not supported by the results.

The primary purpose of the present study was to investigate in some depth the effects of student feedback on teaching at the college level. Included as variables in the study were the instructor's sex, teaching experience, and self-ratings, as well as the subject area of the course. These instructor or course characteristics were examined to determine their relationship to instructional changes. The study was carried out at several types of postsecondary institutions in order to investigate also the possibility that changes occur at some colleges but not at others.

Method

Sample of Institutions

Five colleges, which did not have a formal program of student ratings of instruction, participated in the study. At these institutions the faculty generally would not be familiar with the way students viewed their instruction, and the formal feedback might therefore result in changes in teaching practices.

Procedure

All teaching faculty from four of the five institutions were asked to participate in the first phase of the study. At the fifth institution all but 30 members of the faculty were invited to participate in the study. Those 30, chosen at random, were subsequently asked to participate in the study in a second follow-up at the end of the second semester.

In only one of the four colleges was the faculty told the full details of the study and, in particular, that student feedback would be purposely withheld from some of them. At the other four colleges the faculty was told that the project was "investigating what students are able to evaluate in the classroom and how useful this information might be to the individual instructor."

Faculty members were assured that only they would see their individual rating reports. This assurance undoubtedly contributed to the excellent cooperation from the faculties; in fact, between 70%–90% of those at each institution participated in one or more phases of the study.

Teachers within each department of each institution were randomly assigned to one of three groups.

1. The feedback group administered a rating form at midsemester and received a summary of results (feedback) within a week; this group also received comparative data based on responses in 75 classes at a sixth institution, which had tried out several of the items during the previous year.

2. The no-feedback group had student ratings collected, but these were withheld at midsemester.

3. The posttest group used the rating form only at the end of the semester in order to determine whether the midsemester ratings had a sensitizing effect on student raters or teachers.

Each teacher was asked to use the questionnaire in *one* class of his choice. End-of-semester as well as midsemester ratings were collected for both the feedback and no-feedback groups. Both midsemester and end-of-semester ratings were collected during the fall semester of 1971.

In Table 1 the number of teachers participating in the study at midsemester and at the end of the semester is listed by college and by group. As might be expected, some of the teachers who used the form at midsemester did not remain in the study for the critical end-of-semester administration. The question then is whether those who dropped out of the study after using the form at midsemester biased the final sample. Were the dropouts, for example, generally the more poorly rated teachers? To examine this question, comparisons were made between three sets of scores: teachers who dropped out versus those who stayed in from Group 1, the same comparisons for Group 2, and teachers who stayed in from both Group 1 and Group 2. Thus, in–drop comparisons within the two groups and in–in comparisons across the groups were made for the 23 items used at midsemester. Out of 69 tests of significance (3×23), differences were statistically significant ($p < .05$)

for four of the items. Since differences for that many items could be expected to occur on the basis of chance ($p < .05$), it is safe to conclude that the teachers who dropped out of the study were very much like those who continued (at least in ratings given by students) and, equally important, that the feedback and no-feedback groups were very similar in their student ratings at midsemester.

Instruments

Student ratings or descriptions of instruction were measured at midsemester by a 23-item student instructional report. Included were items that faculty members in an earlier study had identified as providing information they would like from students (Centra, 1972). In particular, items that reflected instructional procedures or behavior that teachers presumably could change were used in the study. Among the areas included were course objectives, instructor preparation and organization, student–faculty interaction, student effort, and course difficulty and scope.

The end-of-semester student instructional report contained the same 23 items slightly rearranged. In addition, it contained several additional items eliciting overall rather than specific ratings; since most of the teachers would be administering the items for the second time in the same course, it was hoped that the additional items would encourage its repeated use.

Item responses for 19 of the items were on a 4-point agree–disagree scale. The remaining four items employed a 4- or 5-point scale with varying responses. Analysis of variance reliability estimates for each item were generally above .70 for 20 or more students in a class. Each instructor received at the appropriate time a summary report that included the mean and standard deviation for each item and the percentage of students in the class who gave each response.

Instructors in the feedback and no-feedback groups also completed an instructor's form at midsemester, which elicited their self-ratings on 21 of the 23 student instructional report items. Instructors also indicated the number of years they had been teaching (1 or 2, 3–6, 7 or more), and the subject field of the course being rated. Courses were grouped into four subject areas for subsequent analyses: natural sciences, social sciences, humanities, and education and applied subjects.

Results and Discussion

In the first three analyses, end-of-semester item scores were compared among the feedback (treatment) group, the no-feedback (control) group, and the posttest group. Because of the large number of dependent variables (items), the multivariate analyses of variance were done in two stages. First, the 15 items that were thought to have the best chance of reflecting instructional changes were analyzed. Then, using those 15 items as covariates, the remaining items were analyzed. Because the student instructional report items are not independent, using the 15 items as covariates served to minimize their effect on the succeeding analysis. The second group of items consisted of the remaining 8 repeated items plus, for two of the analyses, 2 items from the end-of-semester form dealing with the overall effectiveness of the instructor and the overall value of the course to students.

TABLE 1

Number of Teachers Participating in the Study at Midsemester and at the End of the Semester, by College and Group

College	Group				
	Feedback		No feedback		Posttest
	Mid-semester	End of semester	Mid-semester	End of semester	End of semester only
1	26	22	25	22	25
2	50	35	55	45	34
3	33	21	32	23	19
4	49	42	52	48	45
5	19	17	24	21	17
Total	177	137	188	159	140

The results of the first multivariate analysis of variance indicated that there were no differences among the three groups of instructors, nor were there differences in any of the interactions of subject area, years of teaching, and groups. There were, however, differences in the end-of-semester ratings given to teachers in various subject areas (for the first 15 items, $F = 2.40$, $df = 45/1118$, $p < .001$; for the remaining 10 items, using the first 15 as covariates, $F = 2.41$, $df = 30/1072$, $p < .001$) and, to a lesser extent, for teachers with varying numbers of years in teaching ($F = 1.91$, $df = 20/730$, $p < .01$).

In light of the differences in ratings related to the subject area of the course, the second multivariate analysis of variance included that variable once again, along with the sex of the instructor. Group differ-

ences were again insignificant, as were the two-way and three-way interactions of groups, subject area, and sex. Once again, subject area was highly significant (for the first 15 items, $F = 4.67$, $df = 45/1153$, $p < .001$; for the remaining 10 items, using the first 15 as covariates, $F = 2.68$, $df = 30/1134$, $p < .001$), as was sex of the instructor (for the first 15 items, $F = 3.15$, $df = 15/388$, $p < .001$; for the remaining 10 items, using the first 15 as covariates, $F = 2.27$, $df = 10/386$, $p < .01$).

The fourth variable investigated for its possible interaction with treatment effects was the college. Feedback did not result in significant instructional changes at any of the five colleges, although the faculty ratings across the colleges did differ significantly (for the first 15 items, $F = 3.63$, $df = 60/1587$, $p < .001$; for the remaining 10 items, using the first 15 as covariates, $F = 3.49$, $df = 32/1469$, $p < .001$).

To summarize the findings to this point, end-of-semester ratings of instructors who were given midsemester feedback did not differ from either the no-feedback or the posttest groups. Moreover, teacher ratings for the three groups did not differ when subject area, sex of instructor, college, or amount of teaching experience were taken into account.

But a major hypothesis of this study was that changes in instruction would be related to instructor self-ratings. Specifically, the expectation was that student feedback would lead to improved instruction for those teachers who had rated themselves much better than their students had rated them. The relationship, moreover, was predicted to be linear: the greater the discrepancy, the greater the likelihood of improvement. To test this hypothesis, the following regression equation was employed with the feedback and no-feedback groups:

$$R_2 = a_1 + b_1 R_1 + c (I - R_1) \quad ,$$

where R_2 is the predicted second-semester rating, R_1 is the midsemester rating, I is the teacher self-rating (thus $I - R_1$ is the difference between the instructor self-rating and the midsemester rating), and the as, bs, and cs are the regression weights. If the hypothesis is supported, there should be a significant difference between the regression weights for $I - R_1$ (i.e., c) for the feedback and no-feedback groups, with c for the feedback group being positive and greater.

For these analyses, instructors in both groups were divided into those who rated themselves more favorably and those who rated themselves less favorably than their students rated them on each item. Of particular interest were teachers who rated themselves more favorably, since the prediction was that student feedback would effect changes only for those teachers. For most of the items, about 60%–65% of the sample had rated themselves more favorably; also, for this group the size of the discrepancy between self-ratings and student ratings was much greater than for the group that rated themselves less favorably.

Results of the regression analyses for 17 of the 19 agree–disagree items appear in Table 2. (Instructors did not respond to Items 10 and 20 because they were not appropriate as self-rating items; also, the first 4 items were not scored appropriately for this analysis.) Listed are regression weight c and t-test results of the difference in c between the feedback and no-feedback groups. These results are presented for instructors who had rated themselves less favorably (left side of the table) and more favorably (right side) than their students rated them. Results for instructors who rated themselves less favorably indicate fairly random differences in c, and on only one item did the feedback and no-feedback groups differ.

But differences in c for teachers who rated themselves more favorably were significant ($p < .05$) for 5 of the 17 items, as indicated in the last column. Equally important is the fact that for 13 of the 17 items, the direction of the differences also supports the hypothesis. That is, for those items the cs for the feedback group (Group 1 in the table) are higher than those for the no-feedback group (Group 2). Thus the major hypothesis of this study—that student feedback would effect changes in teachers who had rated themselves more favorably than their students had rated them

—was generally supported by the regression analysis.

Changes in Instruction Following a Longer Time Period

Thus far, the results indicate that only those teachers who rated themselves much better than did their students changed after receiving midsemester feedback. It may be, however, that more teachers would change in due time. Perhaps teachers need more time to think about and develop new practices, and perhaps they find changes easier to make with the start of a new course.

To investigate this possibility, additional data were collected at the end of the spring semester at one of the five colleges. The particular college was one at which 30 teachers had been randomly selected to use the student instructional report rating form at a later time rather than during the fall semester. In addition to this group of 30, teachers in the fall feedback and posttest groups at this same college were asked (*a*) if, during the spring semester, they were teaching the same course in which they used the student instructional report form during the fall, and (*b*) if they would be willing to administer the form in that course at the end of the spring semester. (They were surveyed about one month prior to the end of the semester.) Eight teachers from the feedback group and 13 from the posttest group responded affirmatively to both questions and did administer the form once again at the end of the spring semester. Although a larger sample would have been desirable, the teacher ratings were mean scores and thus more reliable than individual scores. In sum, the 8 teachers in the feedback group were using the form for the third time, while the posttest group, which had administered the form only at the end of the fall semester, were using it for the second time. Of the 30 instructors asked to use the student instructional report form for the first time, 21 were able to do so.

While the multivariate analysis of variance of the fall data did not reveal significant differences between the feedback and comparison groups, there were 8 items for which the univariate F values for one or more of the analyses had approached significance ($.05 < p < .20$). On the basis of this prior finding and because they would appear to be most sensitive to change, these 8

TABLE 2

SUMMARY OF RESULTS OF REGRESSION ANALYSES

Item[a]	Group[b]	Instructors who rated themselves *less* favorably[c]		Instructors who rated themselves *more* favorably[d]	
		Regression weight c	t difference[e]	Regression weight c	t difference[e]
5	1	−.077	−.407	.189	1.23
	2	−.009		.045	
6	1	.120	−.0356	.026	−1.24
	2	.129		.167	
7	1	−.234	−1.97*	.310	1.11
	2	.137		.168	
8	1	.079	.248	.094	.738
	2	.049		−.033	
9	1	.007	−.182	.010	−.232
	2	.026		.033	
11	1	.202	.443	.142	2.466*
	2	.083		−.077	
12	1	—	—	.167	−.368
	2			.217	
13	1	.097	.451	−.043	.484
	2	−.001		−.075	
14	1	.117	−.411	.015	1.44
	2	.176		−.070	
15	1	.228	−.067	.235	2.196*
	2	.254		−.150	
16	1	.077	−.363	−.120	−.826
	2	.130		.049	
17	1	−.118	−.625	.104	.883
	2	−.002		.003	
18	1	.079	.607	.172	2.062*
	2	.015		−.031	
19	1	−.042	−1.337	.130	3.932*
	2	.132		−.155	
21	1	−.005	−1.174	.111	2.618*
	2	.087		−.093	
22	1	−.035	.831	.124	1.21
	2	−.240		.004	
23	1	−.098	−1.225	.190	1.45
	2	.061		.002	

Note. Results of the regression analyses are for the formula $R_2 = a_1 + b_1R_1 + c\ (I - R_1)$.

[a] Item numbers refer to the midsemester form. Instructors did not respond to Items 10 and 20. For Item 12, all instructor responses for one of the groups were identical, thus c could not be computed.

[b] 1 = feedback group; 2 = no-feedback group.

[c] Less favorably was defined as $(I - R_1) > 0$.

[d] More favorably was defined as $(I - R_1) < 0$.

[e] Test of the difference in regression weights c for $(I - R_1)$ in Groups 1 and 2.

* $p < .05$.

TABLE 3
SUMMARY OF UNIVARIATE F TEST RESULTS AND MEANS FOR THE SPRING DATA ON EIGHT ITEMS

Item	p	Group M[a]		
		Fall feedback (n = 8)	Fall posttest (n = 13)	Spring administration only (n = 21)
3. Instructor used class time well	<.07	1.45	1.80	1.85
5. Instructor knew when students didn't understand material	<.91	2.04	2.05	2.00
9. Instructor made helpful comments on papers or exams	<.07	1.79	2.27	2.06
12. Instructor was well prepared for each class	<.03	1.21	1.53	1.56
14. Instructor summarized or emphasized major points in lectures or discussions	<.01	1.40	1.75	1.71
16. Scope of the course was too limited	<.61	3.31	3.25	3.22
19. Instructor was open to other viewpoints	<.05	1.81	1.98	1.71
20. Instructor has accomplished his objectives for the course	<.05	1.51	1.82	1.74

[a] Lower score indicates agreement, except for Item 16.

items rather than the entire set of 23 were selected for further analysis with the spring groups.

The multivariate analysis of variance results clearly indicated that the groups differed ($F = 2.18$, $df = 16/264$, $p < .015$). An inspection of the univariate F values and the means for each group, presented in Table 3, further indicated that for most of the items the feedback groups received more favorable scores than either of the other two groups. (With the exception of Item 16, lower scores are more favorable.)

These results suggest then that student feedback did effect some changes in instruction over time, in that teachers who had received feedback twice during the previous semester did receive better ratings than instructors who had received feedback once or not at all. Before jumping to that conclusion, however, some alternative explanations need to be considered. Perhaps teachers in the feedback group who chose to readminister the items again in the spring were better to begin with. To investigate that possibility, differences in scores on each of the eight selected items at the *end of the fall* were tested for three pairs of groups: (*a*) the 8 teachers from the feedback group versus 35 from the same group who did not participate in the spring, (*b*) the 13 teachers from the posttest group versus 32 from the same group who did not participate in the spring, and (*c*) the 8 teachers in the feedback group versus the 13 in the posttest group.

Multivariate analysis of variance tests for the three pairs of comparisons did not yield significant differences. Differences at the end of the spring term, consequently, were less likely to be due to prior differences or self-selection. It would seem safe to conclude that student feedback did effect changes in the feedback group, and that these changes were reflected in certain spring ratings.

But what about the fall posttest group? Because their spring ratings were very similar to the group that used the form for the first time, it would appear that the single feedback had little effect in changing instruction. There are several possible explanations for this lack of change. The posttest group, unlike the feedback group, had not received any comparative or "normative" information to help them interpret their scores. It may be, therefore, that the lack of interpretive information did not enable those instructors to understand fully their ratings (particularly since student ratings are typically skewed in a positive direction). Consequently, they may not have thought they needed to change. Or it may

be that again not enough time had lapsed for changes to be made (one semester vs. a semester and a half for the feedback group). Or finally, perhaps at least two sets of student ratings are needed before many teachers see a pattern of weaknesses that they might improve.

REFERENCES

BRYAN, R. C. *Reactions to teachers by students, parents and administrators.* (U.S. Office of Education, Cooperative Research Project 668) Kalamazoo: Western Michigan University, 1963.

CENTRA, J. A. *The student instructional report: Its development and uses.* (SIR Rep. No. 1) Princeton, N. J.: Educational Testing Service, 1972.

DAW, R. W., & GAGE, N. L. Effect of feedback from teachers to principals. *Journal of Educational Psychology,* 1967, **58,** 181–188.

FESTINGER, L. *A theory of cognitive dissonance.* Evanston, Ill.: Row, Peterson, 1957.

GAGE, N. L., RUNKEL, P. J., & CHATTERJEE, B. B. Changing teacher behavior through feedback from pupils: An application of equilibrium theory. In W. W. Charters & N. L. Gage (Eds.), *Readings in the social psychology of education.* Boston: Allyn & Bacon, 1963.

HEIDER, F. *The psychology of interpersonal relationships.* New York: Wiley, 1958.

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

NEWCOMB, T. M. Individual systems of orientation. In S. Koch (Ed.), *Psychology: A study of a science.* Vol. 3. New York: McGraw-Hill, 1959.

TUCKMAN, B. W., & OLIVER, W. F. Effectiveness of feedback to teachers as a function of source. *Journal of Educational Psychology,* 1968, **59,** 297–301.

(Received November 9, 1972)

Manuscripts Accepted for Publication in the

Journal of Educational Psychology

Attention and Reading Achievement in First-Grade Boys and Girls. S. Jay Samuels (Reading Research Project, Center for Research in Human Learning, University of Minnesota, Minneapolis, Minnesota 55455) and James E. Turnure.

Formal Discipline Revisited: Affective Assessment and Nonspecific Transfer. Joseph F. Rychlak (Department of Psychological Sciences, Purdue University, West Lafayette, Indiana 47907), Ngyuen Duc Tuan, and William E. Schneider.

Personality Traits Associated with Effective Teaching in Rural and Urban Secondary Schools. Kenneth D. Mattsson (Center for Curriculum and Learning Strategies, Mankato State College, Mankato, Minnesota 56001).

Construct Validity of Test Items Measuring Acquisition of Information from Line Graphs. Jay R. Price (College of Education, University of Delaware, Newark, Delaware 19711), Victor R. Martuza, and James H. Crouse.

Learning by Observing versus Learning by Doing. Douglas K. Chalmers (School of Social Sciences, University of California, Irvine, California 92664) and Milton E. Rosenbaum.

Effect of Type of Objective, Level of Test Questions, and the Judged Importance of Tested Materials upon Posttest Performance. Orpha K. Duell (College of Education, Wichita State University, Wichita, Kansas 67208).

Effects of Feedback, Learner Control, and Cognitive Abilities on State Anxiety and Performance in a Computer-Assisted-Instruction Task. Joe B. Hansen (Education Service Center, Region XIII, 6504 Tracor Lane, Austin, Texas 78721).

Effects of Imagery on Learning Incidental Material in the Classroom. Frank Goldberg (716 Montgomery Street, Brooklyn, New York 11213).

Training Imagery Production in Young Children through Motor Involvement. William H. Varley, Joel R. Levin (Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706), Roger A. Severson, and Peter Wolff.

A Test of the Theory of Fluid and Crystallized Intelligence in Middle- and Low-Socioeconomic-Status Children: A Cross-Lagged Panel Analysis. Frank L. Schmidt and William D. Crano (Department of Psychology, Michigan State University, East Lansing, Michigan 48823).

Individual Differences in Learning from Pictures and Words: The Development and Application of an Instrument. Joel R. Levin (Wisconsin Research and Development Center for Cognitive Learning, University of Wisconsin, 1025 West Johnson Street, Madison, Wisconsin 53706), Patricia Divine-Hawkins, and Stephen W. Kerst.

Social-Emotional, Cognitive, and Demographic Determinants of Poor School Achievement: Implications for a Strategy of Intervention. Martin Kohn (William Alanson White Institute, 20 West 74th Street, New York, New York 10023) and Bernice L. Rosman.