ED 227 739 HE 015 976

AUTHOR Levinson, Judith L.; Menges, Robert J.
TITLE Improving College Teaching: A Critical Review of Research. Occasional Paper Number Thirteen.
INSTITUTION Northwestern Univ., Evanston, Ill. Center for the Teaching Professions.
PUB DATE Dec 79
NOTE 138p.; Appendix A will not reproduce clearly due to small print.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Information Analyses (070)

EDRS PRICE MF01/PC06 Plus Postage.
DESCRIPTORS *College Instruction; *Faculty Development; Faculty Evaluation; Higher Education; *Instructional Improvement; Intervention; Microteaching; Professional Training; Program Evaluation; Protocol Materials; *Research Methodology; Research Reports; Student Evaluation of Teacher Performance; Teacher Attitudes; *Teacher Effectiveness; Teaching Skills

ABSTRACT

Research on instructional improvement interventions for college teachers is reviewed and implications for practice are briefly considered. Attention is directed to both the methodological soundness of the program evaluations and the findings of the studies. Interventions to assist faculty to change their teaching attitudes, roles, or activities are considered. The following methods of assessing program impacts are identified: measures of the professors' attitudes, observations of their classroom behavior, student ratings, and measures of students' learning. Systematic research studies on the following types of interventions are analyzed in detail: grants to support faculty projects, workshops and seminars, practice with feedback (microteaching and minicourses), feedback from ratings by students, and concept-based training (protocols). The evaluative research is evaluated and charted according to the following categories: author/date; purpose; components of design (including description of participants, duration, and instrumentation); stated results; threats to validity; strengths; weaknesses; and confidence rating. In addition, each study is coded according to Campbell and Stanley's notation system, and specific threats to each type of validity are listed. Approximately 115 references are appended. (SW)

IMPROVING COLLEGE TEACHING:

A CRITICAL REVIEW OF RESEARCH

Judith L. Levinson and Robert J. Menges

Occasional Paper Number Thirteen

December, 1979

Northwestern University
Evanston, Illinois 60201

IMPROVING COLLEGE TEACHING:

A CRITICAL REVIEW OF RESEARCH

Judith L. Levinson and Robert J. Menges

December, 1979

Occasional Paper Number Thirteen
The Center for the Teaching Professions
Northwestern University
Evanston, IL
60201

Preface

This paper attempts to review and synthesize research on interventions designed to improve college teaching. Our review is more successful than our synthesis, since this relatively small body of research shrinks even further after critical analysis is applied.

Nevertheless, some implications for practice emerge. We hope that our critique and particularly our use of "confidence ratings" for studies will assist investigators to produce better designed studies. We hope that the next generation of teaching improvement efforts will be informed by these findings and evaluated more effectively than many of the studies we describe.

This paper is a working document in two senses. First, we ask readers to suggest to us pertinent studies which we may have missed. Second, we solicit comments on our interpretations and findings which will improve subsequent discussion of these issues.

J. L.
R. J. M.

Table of Contents

IMPROVING COLLEGE TEACHING: A CRITICAL REVIEW OF RESEARCH

For more than a decade, the movement for faculty development and instructional improvement has been generating projects and programs, research reports, conferences, and professional meetings. Agencies on many campuses support activities which promise to benefit faculty and in turn to enrich the education of students. This paper describes attempts to assist faculty to improve their teaching and critically reviews research evaluating the impact of such efforts.

We have two purposes for this review. First, we wish to assess the methodological soundness of these studies and to make suggestions for their improvement. Second, we wish to derive implications for practice, i.e., what guidance does this research provide for those who plan and administer instructional improvement programs?

Interventions with Faculty

Interventions to improve instruction take a variety of forms and have a variety of purposes. Their users seek to modify institutional climate, to restructure the curriculum, to clarify attitudes about teaching and learning, to increase knowledge of alternative instructional strategies, to introduce technologically sophisticated teaching techniques, to increase the clarity of lectures, to improve the quality of examinations, and so on. Because faculty members are the agents of instruction, each of these activities ultimately requires that teachers change what they do. Our concern is with interventions designed to promote such faculty change.

In this paper, we examine studies which evaluate programs to assist faculty as teachers to change their attitudes, roles, or activities. We are not concerned here with evaluations of particular instructional techniques, unless there is also an attempt to change faculty behavior and to monitor the success of that attempt. For example, we are not concerned with the large literature on the effects of the Personalized System of Instruction (for a comprehensive review see Kulik, Kulik, and Cohen, 1979), but we are concerned with evaluations of attempts to assist professors to become more proficient users of that approach.

Much of the research we have consulted merely documents program activities and assesses participants' satisfaction; but some of it assesses the relative impact of approaches and is thus potentially more useful in program design. While we draw upon descriptive research for illustrative purposes, the studies to which we give critical attention are those which individually or in combination can inform the choice of alternative interventions for teaching improvement. Whether studies are experimental or quasi-experimental in design and whether they use qualitative or quantitative methods is less important than that they be systematically executed and completely reported.

Impact of these interventions may be assessed using data of several types supplied both by students and professors. We have identified five types of evaluation data and the likely data sources for each. The categories begin with the participating professor's opinions about the activity and extend to changes in what their students learn.

a. Teacher attitude, assessed by self-report
b. Teacher knowledge, inferred from test or by observer
c. Teacher skill, recorded by observer or reported by student
d. Student attitude, self reported
e. Student skill, inferred from test or recorded by observer

The most powerful evidence for an intervention is its impact upon students (categories d and e), and the weakest evidence consists of the self-reported opinions of participating faculty members. Yet much of the research we reviewed fails to go beyond data collected on the spot from participants (categories a and b). Although we cite in our discussion some descriptive research with data in categories a and b, most studies in the appendix are those with data of types c, d, and e. In short, in order to be summarized in Appendix A, a study includes data other than opinions and attitudes of participants gathered during the intervention itself.

In the sections below, we first describe our procedures for the literature search and the criteria for our critical review. Several evaluations at the institutional or interinstitutional level are then discussed, studies which are primarily descriptive. Next, more systematic research on five types of interventions is analyzed in some detail. These types are the following: grants to support faculty projects, workshops and seminars, practice with feedback (microteaching and minicourses), feedback from ratings by students, and concept-based training (protocols). The final section of the paper presents some implications for researchers and for practitioners.

Procedures for the Literature Search

A systematic search was carried out for relevant instructional improvement research with faculty in postsecondary education. Procedures developed with precollege teachers, such as microteaching, are also discussed when they hold promise for higher education. The search was conducted through abstract indices, texts, and bibliographies, as well as major educational and psychological journals. Program officers at public and private funding agencies were contacted. Pertinent conference papers were also reviewed. In all, more than 100 studies were evaluated for inclusion in this review. The papers finally selected for critical attention are summarized in Appendix A.

Secondary sources, including review articles, were consulted when the body of original research on a topic was very large or if the original study could not be obtained. Of course, reliance on secondary sources does not permit an evaluation of quality, and where such is the case, it is duly noted in the discussion. We are confident that the studies included for final review are representative of the research from the mid-sixties to the present.

Evaluating the Quality of Studies

The following categories are used for the summary of studies in Appendix A: (a) author/date, (b) purpose, (c) components of design (including design code, description of participants, duration, and instrumentation), (d) stated results, (e) threats to validity, (f) strengths, (g) weaknesses, and (h) confidence rating. Several of these categories deserve further elaboration.

Design code. The design of each study is coded according to the notation system used by Campbell and Stanley (1963). In their system, 0 denotes a point in time at which data are collected (observation). An intervention or treatment is denoted by X. An X in parentheses, (X), represents an alternate intervention or treatment unrelated to the major experimental questions under study. Its usual purpose is to control for the time and attention received by members of the experimental group. If research participants are randomly assigned to groups, the designation R is used. A horizontal broken line between groups indicates that they were not randomly formed. A vertical broken line means that data gathered before the intervention came from different persons than data gathered after the intervention.

For example, consider a study in which students made ratings of their instructors' teaching at midterm and end-of-term. Instructors in the randomly-formed experimental group received their midterm ratings but other instructors did not. End-of-term ratings for the two groups were compared in order to assess the effect of feedback from student ratings.

This example is coded as follows:

```
R   0   X   0
R   0       0
```

From this design code, the major features of the study are immediately apparent. It can readily be seen that there are two groups, differing in that only one received the intervention, X. The R indicates that participants were randomly assigned to groups. Data are gathered, 0, from both groups before and after the intervention.

Even quite complex designs are easily comprehended by this notation system.

Threats to validity. Validity refers to the extent to which the propositions which express conclusions of a study approximate truth. Cook and Campbell (1979) discuss four types of validity, each of which asks particular questions about the components of an investigation.

1. Are the independent and dependent variables statistically related? This question tests the statistical conclusion validity of a study.

2. Is the demonstrated statistical relationship between independent and dependent variables a causal relationship? This question, which requires that we rule out noncausal reasons for the statistical relationship, tests the internal validity of a study.

3. Is the demonstrated statistical and causal relationship generalizable to more abstract constructs? This question requires that the operations used to gather data are adequate representations of the constructs under investigation; it tests the construct validity of the study.

4. Does the relationship among constructs generalize to other persons, settings, and times? This question moves outside the operations and the logic of the study itself to test its external validity.

For illustration, consider how these types of validity apply to the investigation described above in which professors in one group receive ratings feedback from students at midterm. The investigator's stated purpose is to determine the impact on teaching effectiveness of information from students about teaching performance. Statistical conclusion validity requires, among other things, that measures have adequate reliability and that statistical tests have adequate power. Internal validity requires that an observed end-of-term difference in ratings between groups be due to feedback rather than to some other variable such as a differential dropout rate in the two groups. Construct validity requires that the procedures used when professors receive "feedback" and the questions asked of students regarding "teaching effectiveness" are adequate representations of those constructs. External validity requires recognition that conclusions may generalize only to persons, places, and times like those of the study itself.

A number of specific threats to each type of validity are listed in Appendix B. Using this list, we have reviewed each study in order to determine the appropriateness of design, plausible alternative explanations to claimed results, and the degree of confidence that may be placed in the results. Threats pertinent to this review are mentioned as each study is outlined in Appendix A.

This approach to validity flows from the quantitative tradition of social science research, and most of the research we discuss has placed itself in that tradition. We argue that the last three types of validity are appropriate for analyzing qualitative research. Whether data are qualitative or quantitative, the threats associated with internal validity, construct validity, and external validity must be confronted by all investigators who wish to make causal inferences.

As an example of careful qualitative analysis of a teaching improvement project we cite the American Sociological Association's Project on Teaching Undergraduate Sociology. Even though that project's evaluation dealt primarily with national task groups rather than with interventions at the local level, it is notable for its methodological stance. Project evaluators were concerned with what they discern as problems imposed by the objective/quantitative tradition. They argue convincingly that, if it is to be useful, an evaluation should violate at least four rules of this tradition: the rule of objectivity, the rule of measurable outcomes, the rule of nonreactivity, and the rule of the scientific report. As participants and as observers these evaluators compiled field notes as a basis for portraying and analyzing events and for attempting to explain why events occurred as they did. An evaluation from the quantitative tradition, they expect, would have proved impossible in light of the project's very broadly stated purposes or would have resulted in data of little importance.

Qualitative methodology does not exempt investigators from an obligation to rule out threats to validity. The issue is not lost on the evaluators of the American Sociological Association project.

> It is regrettable that, without experimental evidence, it is not possible to attribute causation to the program as the agent of change. Since most authorities on program evaluation agree that experimentation is difficult, if not impossible, under program conditions, little is lost by abandoning the effort.
>
> Without the support of experimental logic, our efforts to attribute causation must rest on plausible explanations which our data fail to contradict and appear to support. If we can rule out alternative explanations, so much the better. However, the evaluation cannot provide conclusive evidence that the world or any part of it is different as a result of the program. (Deutscher and Gold, 1979, p. 135)

We are not as willing to advocate the exclusive use of qualitative methodology in program evaluation as Deutscher and Gold seem to be. They propose that because experimentation is difficult with respect to some program evaluation, little is lost by abandoning such efforts and using qualitative methodology only. We believe that the approaches jointly contribute toward ruling out alternate explanations and thus allow us more closely to approach causal attributions. The statistical conclusion validity that quantitative methodology can provide is important, even if it is available for only a few of the many experimental questions of a program evaluation. Along with qualitative information, it can provide a more complete picture of cause and effect. Campbell (1974) discusses the qualitative-quantitative methodological conflict and elaborates the relationship between the two:

> ...I have sought to remind my quantitative colleagues that in the successful laboratory sciences, quantification both builds upon and is, cross-validated by, the scientist's pervasive qualitative knowledge. The conditions of mass-produced quantitative social science in program evaluation are such that much of this qualitative base is apt to be lost. If we are to be truly scientific, we must reestablish this qualitative grounding of the quantitative in action research. (Campbell, 1974, p. 30)

Strengths and weaknesses. In evaluating the quality of each study, major strengths and weaknesses have been delineated. Many of them are directly related to the validity threats listed for a particular study. For example, "low statistical power" may have been noted as a statistical conclusion validity threat because a study used a small sample which may have contributed to nonsignificant results. Hence, "small N" would be listed under the weakness category. Strengths of a study might be the use of randomization or a thorough discussion of its limitations. Only the most pertinent strengths and weaknesses are noted in the table; indeed, for some studies, none have been specified.

Confidence rating. A rating of high, fair, or low has been assigned each study to suggest how much confidence should be placed in its results. It is difficult to set criteria by which all studies can be evaluated. Some factors are more important than others, depending in large part on the specific circumstances of each study. Thus, the ratings are tentative, meant only to suggest the general level of quality of research on a particular topic.

With respect to design, randomized studies using two or more groups have been regarded with greater confidence than studies using one group in a pretest-posttest design. Limited generalizability of findings is discussed as an external validity threat (e.g., selection by treatment biases), but generally we give more weight to internal than external validity. In assigning the final rating, however, all threats to validity have been considered.

Confidence in the findings of a study (our confidence rating) should be differentiated from a judgment of the study's ultimate importance. A high quality study may deal with a problem of little consequence. Likewise, a flawed study may merit attention because it is one of the very few attempts to deal with a problem of significance.

Interinstitutional Projects and Campus Agencies

Our literature search led us to reports of instructional improvement projects that involve groups of institutions or that evaluate the full range of activities of a campus agency. In most cases these reports consist of little more than program descriptions, sometimes bolstered by comments from participants. To illustrate evaluations at this level, we present three examples. One project brought together several institutions for a cooperative venture in faculty development (PIRIT). The second developed a special publication to convey information about teaching improvements (Change Magazine Reports on Teaching). The third assessed a campus-wide faculty development program on a particular campus (Memphis State University).

The Project on Institutional Renewal through the Improvement of Teaching (PIRIT) spent three years fostering collaborative activities which reached sixteen colleges and universities. On these campuses, teaching improvement programs of varied forms were begun. In some cases, program became embodied in a center. In other cases, existing instructional activities were redesigned to provide new roles and experiences for students and faculty. An issue of New Directions in Higher Education is devoted to a description of the project and includes a report of its evaluation (Gaff and Morstain, 1978).

Evaluation was based on a questionnaire distributed at the close of the project to all faculty at fourteen of the PIRIT schools. Case studies by team members from participating institutions have also been prepared. Those who returned completed questionnaires were judged to be representative of all faculty in age, field, and rank. Respondents who participated in project activities (479) were compared with those who had not (442) and were judged to be similar in age, field, rank, and profile of interests and activities, including self-assessed teaching effectiveness. This implies that faculty reached by the project were representative of faculty in general. Since the groups had not differed on these items when the survey was given at the start of the project, it also suggests that project participation had no impact on the particular characteristics measured by these items.

When asked specific questions about benefits derived from the project, faculty gave very positive responses to such items as "contact with interesting people from other parts of the institution," "increased motivation or stimulation for teaching excellence," and "personal growth or renewal." Lower but still positive benefit was indicated for "better relationships with colleagues," "skill in using new instructional techniques," and "better relationships with students." Rating overall benefit, 33 percent said that they would recommend project activities to a friend or colleague, and 61 percent indicated that they were using new techniques or approaches as a result of their participation. Most regarded these changes as important. In general, those who reported greatest involvement in the project also reported greatest benefit. Most nonparticipants, too, were knowledgeable about the project and positively disposed toward it.

We must be cautious about the self-report data employed in this evaluation, but the project does seem to have generated considerable satisfaction and knowledge among participating faculty. Impacts on faculty skills and on students are unknown, and we can make no inferences about the differential impact of strategies or about their cost effectiveness. Therefore, the findings are not sufficient for confidently deriving principles which would be useful in designing future teaching improvement projects for faculty.

The National Teaching Project of Change Magazine was a three-year effort which produced six magazine-sized reports as its major products. Each report dealt with three disciplines, profiling up to 30 professors and describing their teaching practices. The project is relevant to this review because its declared purpose was to make an impact upon college teaching through mass distribution of a publication "which celebrates successful teaching improvements."

An evaluation of four of the reports appeared in the final publication in that series (Francis, 1978). The evaluation employed a variety of methods ("heuristic evaluation") and assessed the impact of the reports on a variety of audiences including the magazine itself, the disciplinary associations who had selected teachers to be written about, the professors whose work was featured, and readers of the reports. Magazine staff members were pleased by the response to the series--about 50,000 copies of each of these first four reports were distributed--but they were disappointed that this response did not also increase sales of regular subscriptions. Little effect, at least of significant duration, could be documented for disciplinary associations. Case studies of professors who were profiled revealed some positive and some not-so-positive effects.

Regarding the larger audience, a questionnaire survey of readers of the reports revealed general satisfaction. Seventy-six percent said that they found ideas about teaching in the reports, 25 percent planned to incorporate those ideas, and 16 percent said that they were actually using the ideas. Twenty-eight percent indicated their own teaching had improved as a result of the reports and, of those, about three out of four were able to describe the improvement.

While a project of this nature and scope is unprecedented, only an equivocal judgment of its impact can be made from these data. It is particularly unfortunate that since there was only one "treatment," the evaluation can ask only one question, namely, to what extent did this strategy work? If the project had systematically varied media and dissemination techniques, their relative impact and cost could have been assessed. We could then ask which strategy was useful with whom for what purposes and at what cost, and use the findings for subsequent decision making and research.

On individual campuses, staff or committees charged with teaching improvement are typically expected to report on their activities. According to recent surveys, these reports are likely to include little evaluative data. McMillan (1975) found 16 of the 35 faculty development agencies which he surveyed had attempted evaluation but only four of them went beyond faculty reactions. According to Centra's (1978) survey of 756 institutions, fewer than one-fifth had attempted evaluation, most using unsophisticated designs. A survey of institutions in Ohio (Brown and Inglis, 1978) documented evaluation at 14 percent of the four-year colleges and universities and at just over half of the two-year institutions. That survey also revealed that institutions with teams participating in a series of statewide conferences on instructional development were no more likely than nonparticipants to conduct evaluations of their teaching improvement efforts.

Most of the campus-level evaluations we have seen are limited to user reports of satisfaction and dissatisfaction. Despite these limitations, such surveys can provide some useful information. For example, Mayo's (1979) survey listed each of the objectives of the center at Memphis State University. He asked users and nonusers how important each objective is, how well it is being achieved, what changes if any should be made in it, and how the center's performance could be improved regarding this objective. Such responses can guide staff who want to know the ostensible preferences of their faculty. One provocative finding of this survey was that, in general, both users and nonusers rated more highly those center services which are chosen freely by faculty, such as production of audiovisual materials, than those activities which are initiated by the center and require changes in faculty behavior, such as workshops on new teaching techniques.

From time to time reports are prepared for the purpose of reviewing a center's performance and making decisions about its future. A study of such documents, if they could be obtained, would no doubt be interesting. We suspect, however, that they might not tell us much about effective program features generalizable to other institutions, since they serve a purpose which is fundamentally political. We do not review such documents here.

In conclusion, it appears that few campus-wide and interinstitutional programs for instructional improvement are evaluated with the care necessary to permit conclusions which usefully inform program design.

Grants to Support Faculty Projects

Many faculty development agencies, particularly those established with external funding, award small grants competitively to faculty who propose projects for increasing their teaching effectiveness. Grants may purchase needed material or provide personnel such as proctors, tutors, and clerical staff. Consultation with instructional development professionals may also be supported. In more generously funded programs, released time is given or summer salary is paid. Centra's (1978) survey of faculty development practices found that 58 percent of the 756 responding institutions (two- and four-year colleges and universities) said they had a program of summer grants for projects to improve instruction or courses.

Because of their visibility, grant programs help to create a positive image for professional development centers. The awards also lend credibility to instructional ideas originating in the faculty. Programs vary in size and in purpose. Davis (1979) points out other distinctions among programs including whether or not funds are distributed from a centralized source, whether funds reach many faculty (breadth) or few faculty (depth), whether the object is institutional change of individual recognition, whether proposals are evaluated by administrators or by teaching faculty, and what criteria govern awards.

Research on grant programs is needed to answer a number of questions. What changes in instruction do project grants produce? Do these changes persist? What is the impact of the changes on students? How are such programs best organized with regard to size and duration of grant as well as characteristics of the project or person funded? How do project benefits compare with project costs? Unfortunately, most grant programs are documented only at a descriptive level. Reports for internal circulation or for funding agencies may tell only what awards were made and for what purpose. In addition, the recipient may prepare an account of how the grant was used. Since givers and receivers of awards are obviously self-interested, data should also be gathered from objective observers, from comparison groups of faculty, and from students intended to benefit from the program.

Most reported evaluations of granting programs find that participants are satisfied. For example, 70 percent of Centra's respondents whose institutions had summer grant programs said they felt the program was effective or very effective.

At the national level, one large scale grant program seeking to affect college instruction is the Institutional Grant Program of the National Endowment for the Humanities. From 1971 to 1977 approximately $44 million were awarded for development of grants and approximately $8 million for pilot grants. Impact on teaching and learning was one of the evaluation criteria; "the need to break the molds of custom in teaching and learning" was identified as most important among the goals of the program. Results of evaluation of a sample of grants according to that criterion was not encouraging: 50 percent of the developmental grants and 32 percent of the pilot grants were judged successful in this regard, 32 percent of the developmental grants and 40 percent of the pilot grants were judged partially successful, and 18 percent of the developmental grants and 30 percent of the pilot grants were judged unsuccessful (Curtis, 1978). Even these estimates may be inflated, since judgments were made by site visitors who had no data from students.

A granting program was part of the American Sociological Association's Project on Teaching Undergraduate Sociology. Members were invited to submit proposals for creating nine experimental programs on any aspect of undergraduate education in sociology. During the first year, two proposals were recommended for funding by the Association to the Fund for the Improvement of Post-secondary Education (FIPSE) which was supporting the Association's project. Only one proposal was funded by FIPSE. In the second year, five proposals were recommended by the project and none was approved by FIPSE. The proposal solicitation program was discontinued in the third year since its results did not justify the required resources.

In a preliminary draft of their evaluation of this program, Deutscher and Beattie (1978) offer several reasons for its apparent failure. First, the project had devised procedures for selecting proposals for funding, but FIPSE insisted that allocations be governed by its routine review procedures. Second, in order for FIPSE to screen proposals, the interval during which the Association could solicit and review them was quite short. Third, association members who submitted proposals were not as skilled in preparing proposals as expected. Fourth, the Association's review group was inexperienced in the review task. Fifth, the Association had insufficient resources to assist those submitting proposals during the revisions and resubmission process.

Given the innovative nature of this project, it is understandable that many of these problems were not anticipated. Deutscher and Beattie caution against interpreting this enterprise as a failure, although it assuredly did not meet its original objective. They point out that the project director's decision to discontinue the proposal competition after two years made resources available to other project activities which had a greater likelihood of success. The decision was, therefore, an appropriate adaptive response. As a qualitative study, this evaluation is richly suggestive of difficulties which may also afflict campus-level granting programs.

At the state level, an instructional minigrant program has been administered out of the Office of the Chancellor of the California State University and Colleges System. Since 1974 between $200,000 and $300,000 have been awarded annually. An evaluation reviewing four years of the program gathered data from grant recipients, deans, department chairs, local campus coordinators and local faculty senators (almost 1400 persons). Final project reports (N=560) were also examined.

The resulting report reveals a great deal about the program's public relations value. It documents that funded projects were in fact instructional in nature and concludes that the overall response has been favorable. It also concludes that "local campuses have not developed a formally structured and reportable mechanism for evaluating the projects they funded" (Bogdanoff, 1979, p. 3). Since the study relies only on "collective professional opinion," no data are available on how the grant proposals were implemented or what effects they may have had on students.

Like other studies reviewed so far, this study is severely limited because its data come solely from grant recipients and those working intimately with them. No alternative treatments are evaluated, no attempt is made to verify independently how grants were implemented, and impact on students is not studied.

A campus-level program has functioned for several years at Michigan State University. Grant-making activities of the Educational Development Program were evaluated in a survey of persons who received grants for undergraduate classroom projects from 1970 through 1975. Grant recipients were found to be representative of all faculty in age, rank, college affiliation, and self-perception. A factor analysis of questionnaire responses suggested that these instructional innovators were of three types: the reward seeker, the information seeker, and the dissatisfied maverick. Nearly all recipients reported that they were pleased with the results of their work under a grant. The innovations developed were reported still to be in use in 81 percent of the departments and by 74 percent of the developers.

All grantees had been asked to submit evaluations of their projects, and these reports served as the data base for preliminary assessment of the granting program. Of the 98 projects (1970-1974) examined in the study, no report had been received for 33, reports containing no evaluation were received for 14, and reports including evaluation were received for the remaining 51. Most evaluations were impressionistic and only 10 "could be considered of high quality" (Davis, Abedor, & Witt, 1976, pp. 97-98).

This study, therefore, identifies two kinds of information useful in evaluating such programs, namely innovator characteristics and grant recipients' reports.

At the University of Michigan, Kozma (1978) assessed a program to increase the use of instructional technology by faculty. The project awarded several faculty fellowships for released time, seminars, and technical assistance. Incidence of use of teaching innovations based on instructional technology was assessed before and after the fellowship period. Data were collected from several groups: faculty fellows (N=10), chairpersons of university departments (N=13), unsuccessful applicants for fellowships (N=8), holders of instructional development grants (a support program of smaller grants, N=25), and a randomly selected faculty comparison group (N=137). Since a given amount of funds can support considerably more instructional development grants than fellowships, comparing these groups provides a test of breadth versus depth in a granting program. Kozma found that both groups reported significantly increased use of innovations at the second survey compared with the first. Fellowship applicants also increased their use of innovations (but not significantly), and chairpersons and general faculty did not. These data are limited to self-reports which might be biased to please the investigator, but they do suggest that both programs had positive effects. There may also have been a predisposition among unsuccessful applicants to the program toward adopting innovations. In addition, the study presents evidence for diffusion of knowledge about these innovations. Fellows kept records of contacts with colleagues during which their projects were discussed. When later contacted independently by staff, these colleagues verified the conversations but indicated that they themselves had not adopted the innovations.

Research like this last study begins to document the impact of grant programs with varying features. It appears from this study that less expensive programs can be effective, but, as Kozma says, that finding may be a product of interaction with the characteristics of these participants. Since fellows were more innovative at the time of the first survey than were instructional development grant recipients, the more intensive (and expensive) program may have been appropriate for them, while a less demanding program was appropriate for professors just beginning to consider innovations in their teaching. Future studies, building on this one, should develop more reliable and sophisticated measures to assess instructional impacts.

In conclusion, we can say little about variables which would permit more intelligent design of granting programs. Such programs have attractive face validity, since persons completing a grant-supported project are likely to have gained new knowledge and skills. Nevertheless, impact on students is uncertain and remains to be studied in relation to specific features of particular programs.

Workshops and Seminars

Perhaps the most frequent but least carefully evaluated instructional improvement activities are workshops and seminars. These are occasions when faculty and prospective faculty gather to discuss or otherwise explore some topic related to teaching and learning. The gathering may be an informal conversation over brown bag lunches, a presentation by an off-campus consultant, a highly structured weeklong summer workshop, or any number of variations. It may or may not involve students and may or may not carry academic credit or financial remuneration. Attendance is sometimes voluntary and sometimes coerced.

The goals of these gatherings also vary. Purposes include helping faculty to get acquainted with one another, stimulating examination of attitudes about teaching and learning, generating interpersonal support for teaching improvement activities, increasing knowledge about research on teaching, developing a shared vocabulary for talking about teaching, mastering specific skills for course development or for communicating subject matter or for assessing student learning, and so on.

A number of courses to train graduate teaching assistants have been systematically evaluated. Activities for experienced faculty, on the other hand, are typically evaluated rather informally by questionnaires distributed at the close of the event or soon thereafter. Participants are likely to be asked how they felt about the activity and what they learned from it. These comments, at least as described in reports and published articles, are usually positive, but permit no conclusions about impacts which persist beyond the event itself.

In the discussion below, we deal with two types of workshops and seminars. The first type aims at changes in attitude and affect and the second type is oriented toward changes in skill.

Changes in Attitude

Research findings in social psychology suggest that exposure to diverse points of view facilitates attitude change. The likelihood that such changes will persist is greatest when persons are confronted by opposing views, that is when they grapple with dilemmas which they perceive as relevant, which generate emotional involvement, and for which possible solutions can be identified (Cook and Flay, 1978). We suspect that discussions meeting these conditions are what many faculty have in mind when they refer to a "good" discussion.

If one is to understand opposing views, however, mere exposure to them may not be sufficient. Tjosvold and Johnson (1977) had college students discuss their views about a moral dilemma. Before discussing the dilemma with another student (who was actually a confederate of the experimenter) some were led to believe the other's views were the same as their own (no controversy condition) and some were led to believe that the other's views disagreed with their own (controversy condition). Those not exposed to controversy were more confident that they understood the other person's view than those who were exposed to controversy, but a direct measure of understanding revealed greater knowledge for those in the controversy than the no-controversy condition.

These findings imply that a faculty group which merely discusses opposing views may report inaccurately high levels of satisfaction, while those dealing with incompatible views actually represented in the group will learn more, at least about those views. Educational research in general supports the value of controversy in classrooms for promoting curiosity, problem solving, and intellectual growth (Johnson & Johnson, 1979).

A number of devices are useful for stimulating controversy in groups. Surveys may be used to introduce findings which violate group members' expectations. Case studies may pose dilemmas. Role playing may promote identification with positions other than one's own. Discussions stimulated by videotapes of college classes may also satisfy these conditions. One series of such tapes, produced at Northwestern University, is used at workshops which aim at attitude examination and change. College Classroom Vignettes are discussion stimulus videotapes showing classroom incidents and interviews with professors and students. Because the taped segments are brief and are presented out of context, they elicit a variety of reactions from viewers. Not all of these reactions are compatible, and subsequent discussions must confront opposing views. Controversy is heightened if a later segment of the tape provides information, such as student comments, which contradicts a viewer's reaction to the first part of the tape. Or one may have to reconcile a negative reaction to an event on the tape with the fact that such events are typical of one's own teaching.

In some vignette discussions alternative models of "good" teaching emerge. For example, Brock (1976) notes that viewers may contend a particular action of a taped teacher is "bad" because (presumably) everyone knows that it is bad. Others, however, may judge the action according to its effects. For example, a teacher's interruption of a student is seen as bad only if it stifles subsequent class discussion. Thus the viewers must deal with contradictions between what might be called the "consensus model" and the "effects model" of good teaching.

Participants in vignette sessions report that the discussions expose them to a variety of views. Content analysis of vignette discussions document that controversies do occur. That is, discussion moves from general concerns and dependence on the leader to free expression of disagreements and relative independence from the leader (Menges, 1979).

There is no systematic evidence to show that vignette discussions have an impact beyond the sessions themselves, although anecdotes indicate that some faculty are subsequently motivated to try new teaching methods or to have their own classes videotaped. One activity which could build upon these video-stimulated discussions is the small, peer-led group. Blumenthal (1978) describes one such group in which members shared tapes of their own classes. Group activity may become a source of significant interpersonal support for continued attention to teaching improvement. Blumenthal points out similarities between such sessions and encounter or consciousness-raising groups. Support groups of this kind are potentially powerful vehicles for attitude and affective change. Still another approach involves teams of faculty members who visit one another's classrooms and share their reactions. Sweeney and Grasha (1979) describe a large-scale program of this kind which was positively evaluated by participants.

Faculty workshops may also aim at more complex affective characteristics. Goldman (1978) evaluated the College Center of the Finger Lakes (CCFL) Faculty Development Program to determine its impact on personal development, which was one stated objective of the program. Participants' level of self-actualization represented personal development and was measured by Shostrom's Personal Orientation Inventory (POI). A pretest-posttest design with matched controls was used. The six-day CCFL Basic Instructional Workshop consisted of instruction in diagnosis of teaching and learning styles, instructional methods and techniques, selection of teaching strategies, new instructional media and resources, and personal values and life plans as they affect instruction. Activities included discussion, role playing, skills training, and a series of micro-colleges. Included in the study were 12 college professors who participated in the workshop and 10 professors matched on age and academic division who were not involved in the workshop. Significant increases for six of 12 sub-scales (inner directedness, self-actualizing values, existentiality, feeling reactivity, acceptance of aggression and capacity for intimate contact) were noted for the experimental group while no significant changes occurred for the control group. Goldman points out that there were no significant differences between groups at the pretest. He concludes that his study supports the notion that such faculty workshops promote participant self-actualization.

Goldman's study has been assigned a low confidence rating because the nonequivalent control group design allows for a number of plausible explanations for the findings. For example, local history may have influenced the experimental and control groups differently. Resentful demoralization may have occurred for those who were excluded from the workshop. Small sample ~~size may have contributed to nonsignificant results~~. Furthermore, although pretest-posttest differences are significant on some of the subscales for the experimental group, the differences in absolute values may not be of practical significance. Also, only one instrument was used to assess self-actualization (mono-method bias). Two strengths of the study are, first, that it is based on a clear foundation, the model of faculty development set forth by Bergquist and Phillips (1975b). Second, it attempts to assess the complex construct of self-actualization rather than limiting itself to participant reactions.

Changes in Skill

Seminars on college teaching at a number of colleges and universities provide training for teaching assistants, prospective college instructors, and inservice faculty. For example, at Northwestern University, The Seminar on College Teaching examines such topics as selecting course objectives, task analysis, presentation techniques, discussion skills, and course evaluation. Each student prepares a "unit of instruction" for classroom use. As well as affecting one's knowledge of these topics, such seminars should enhance teaching related skills. Unfortunately, the literature holds many more descriptions of seminars in college teaching than systematic studies of their impact. The following discussion briefly samples from descriptive reports and case studies. Then, systematic studies are reviewed in more detail.

The case studies and reports detail either university-wide or departmental courses for college teachers or teaching assistants. The courses vary in length from one term to year-long programs. Some seminars focus on a particular theme; others cover a variety of issues. For example, Finger (1969) described a two-term graduate seminar entitled, "Professional Problems," offered to psychology graduate students. The seminar covered such topics as employment settings, the history of academic and professional psychology, the history of higher education, curriculum alternatives, instructional techniques, and student rights and responsibilities. Some practical teaching experience is arranged for each seminar member. Finger reports that both students and faculty have derived benefits from this seminar experience.

A year-long teaching fellow training program was described by Kapfer and Della-Piana (1974). This program includes an orientation workshop followed by several options for developing teaching techniques in the areas of proficiency testing, personalized instruction, or student testing techniques.

Rose (1972) reports a campus-wide program at the University of California at Los Angeles for increasing the effectiveness of teaching assistants. Entitled "University Level Instruction," the course was taught by Professor W. James Popham in the winter quarter of 1969. The overall objective was to help teaching assistants become competent in planning and evaluating instructional sequences. Two indicators of success of the course are reported. All students performed 90 percent or better on the final examination, and a significant shift in attitude toward the criterion-referenced approach was found.

A pilot project for teaching assistants at the University of Florida (Smith, 1974) assessed course impacts on the participants' classroom teaching behavior. One objective was to develop the skill of probing, and the material for that objective was based on an instructional module used in programs for public school teachers. Other topics included new media in higher education and the use of a systematic approach to college teaching. Teaching assistants were assigned supervisors who observed their classrooms, videotapes of their teaching, and provided feedback. Eleven of 15 teaching assistants increased the amount of time they spent in asking questions of students. Those whose questioning time declined had initially spent more time questioning and had apparently chosen to develop other skills. It was also found that at the seminar's end teaching assistants spent less time lecturing and more time responding to students' questions.

The impact of a seminar or workshop on teachers' skills may be inferred from researchers' observations of teacher behavior, from student perceptions of the teacher, from gains in student learning, or by some combination of these. First we mention two studies limited to researcher-observed changes in classroom behavior. Then eight studies are described in which student perceptions were gathered. Finally four studies are reviewed where student achievement was measured. Some studies in the two latter groups also include data from classroom observers. (Many of these are dissertation studies and, for some of them, we have had access only to the abstract. If important information is missing from the abstract, the omission is noted in Appendix A, but abstracts have usually been sufficient for deriving a confidence rating.)

Impact on classroom behavior. In one dissertation study of teaching assistants' classroom behavior (Murphy, 1972), new teaching assistants in chemistry at Ohio State were assigned by a stratified random technique to a training group or to a no treatment control group. Training included group discussions, microteaching sessions, and classroom observations followed by conferences with the observers. To evaluate the training, all participants were observed once before training and twice after training; audio recordings were made of those classes. Classroom events were coded according to the categories of Flanders Interaction Analysis and the Question Category System for Science.

Several post training differences were revealed by analysis of variance. Trained teaching assistants were more successful in drawing students into discussion. They lectured less, used more praise and encouragement, and asked more questions. But there were no differences in the type of question asked or in the proportion of correct responses elicited.

A second dissertation study limited to observations of classroom teaching skill assessed effects of 10 one-hour seminars for teaching assistants in biology at Georgia State University (Rhyne, 1973). Twelve teaching assistants were observed in the lab for one and one-half hour periods just before and just after training. Analyses were made of verbal interaction patterns, nonverbal movements, and the types of questions asked.

After training, teaching assistants spent more time with students, asked more convergent and divergent questions (but no more managerial or rhetorical questions), and engaged in more indirect talk. Absence of a comparison group and use of weak statistical techniques prevent us from drawing causal inferences about the observed changes. The findings are suggestive, nevertheless, and the study is notable as the only one we have located using a lab setting.

Impact on student perceptions. Yaghlian (1972) worked with teaching fellows in chemistry for his dissertation study. A series of five workshops on a variety of topics were attended by from eight to 15 persons. Students of the 15 participating teaching fellows gave higher ratings to the course than did students of nonparticipating teaching fellows. Changes in attitudes of participants were also discerned. The study had an applied emphasis and elements of the program were subsequently adopted by that department.

Costin (1968) assessed the impact of seminar participation on student ratings of psychology teaching assistants. Entitled "Principles and Methods of Teaching Psychology," the seminar has been taught for some years at the University of Illinois (Urbana-Champaign). During the course, students are asked to make a 30 minute presentation which is then critiqued by seminar participants. A survey of 65 seminar participants indicated that the most important course topics in their view related to practical daily work of a college teacher and to specific aspects of the following areas: (a) developing course objectives, (b) selecting and organizing course content, (c) planning and handling teaching-learning situations, and (d) evaluating the attainment of course objectives.

Two substudies were carried out, comparing teaching assistants who had participated in the seminar with those who had not yet enrolled on the following dimensions: skill, structure, feedback, group interaction, and student-teacher rapport. In one analysis, teaching assistants were rated by their students. Participants' mean ratings were significantly higher on rapport. The second analysis was limited to teaching assistants with at least two terms'experience. For that group, adjusted mean ratings after one semester revealed no significant differences between seminar participants and nonparticipants; after the second semester, differences favored participants on group interaction and feedback. Costin concludes that the seminar was reasonably successful in helping teaching assistants to develop more positive interpersonal relationships in the classroom.

At Florida State University, a program for teaching assistants was developed in the geology department and subsequently used in the chemistry department. Hockett (1972) found that, after participation, teaching assistants showed less teacher control, more individual interaction, and more high-level questions. Attitudes of the students of these teaching assistants also are reported to have changed in a positive direction. This is entitled a "pilot study" and requires cautious interpretation since the sample is nonrandom and apparently there was no control group.

Teaching assistants in business administration at Arizona State University participated in Haber's (1973) dissertation study. Twelve teaching assistants randomly selected from 19 in the department were in turn assigned at random to three groups. One group received instruction in effective questioning techniques, using the Flanders system, and also received feedback on their classroom performance. A second group received feedback and no instruction. A control group received no feedback/no instruction. At pretest teaching assistants were generally found to be "direct teachers" who favored a controlling role which limits student participation. After training, there were no differences between groups in teachers' classroom behavior or in their ratings by students. Teachers' attitudes, measured by the Minnesota Teacher Attitude Inventory, were significantly related to their observed behavior.

In another study, teaching assistants in psychology, both graduates (N=4) and undergraduates (N=15), taught weekly seminars in their areas of interest as a supplement to faculty lectures (Carroll, 1977). In a posttest only control group design, teaching assistants were randomly assigned to an experimental seminar (N=10) or a control seminar (N=9). All teaching assistants were required to attend but were unaware of their group membership and of the variables being studied. The experimental seminar included scheduled readings, individual conferences, at least one individual critique of a videotape, an unstructured group meeting, and five formal workshop sessions. The control seminar was less structured, included less input by the instructors, and provided an opportunity to view one videotape alone without a critique.

Experimental and control teaching assistants did not differ on sex, grade level, verbal aptitude, cumulative grade point average, major, or primary reason for taking the course. Interaction analysis of tapes obtained near the end of the term showed that classrooms of the experimental teaching assistants were more student centered than those of control teaching assistants ($p < .06$), although experimental classrooms did not show higher levels of student talk. As predicted, the experimental group received higher student ratings than controls on the use of objectives ($p < .07$) and on general effectiveness of instruction ($p < .10$). Use of indirect teaching skills was correlated with student ratings ($p < .02$).

Less powerful effects of teaching assistant training were found in Dalgaard's (1976) dissertation study. Her dependent variables included ratings of the teaching assistant (a) by their students, (b) by experts, and (c) on a self-evaluation form. Twenty-two inexperienced and untrained teaching assistants in economics, business administration, and geography departments at the University of Illinois (Urbana-Champaign) were randomly assigned to a training group or to a no treatment control (stratified by department). Training included six two-hour seminars on topics related to instruction. Trainees also individually viewed videotapes of their classes with a trained supervisor.

Experts rated the teaching performance of trained teaching assistants higher than that of untrained teaching assistants, but no impact was found on student ratings or self-evaluations. Participants recommended that the program be required for new teaching assistants, and the dissertation includes materials used in that training.

We have located two studies in which faculty members participated. In the first, courses at the University of North Dakota College of Nursing were rated both fall term and winter term (Kingston and Lacefield, 1979). During winter term, faculty participated in the TIPS workshops developed at the University of Kentucky. Sessions dealt with organizational skills, interpersonal communication skills, teacher behavior, and evaluation skills. A microteaching component was also included. Over half of the ratings in areas covered by the workshops increased significantly from fall to winter for the 29 instructors, despite the brief interval between training and winter ratings. No control data are reported from previous years or from nonparticipants (since all faculty participated) and so it is not possible to estimate the chance that changes would have occurred in the absence of workshop participation.

A detailed description of a 10-week workshop for faculty is given by Howard (1977). In weekly two-hour sessions, participants developed such skills as identifying their own teaching goals, discussing teaching in nonjudgmental terms, and consulting with one another about teaching. Members observed one another's classes, and videotapes of their own classes were viewed and discussed in the group. Hoyt and Howard (1978) report an evaluation of two such eight-member groups at Wichita State University. Of 68 faculty who indicated interest in the program, 16 were randomly assigned to the experimental groups and 16 to a control group. Students in one course taught by each of the 32 participating faculty completed a course evaluation form at midterm and again at end-of-term. Changes on 12 of the 13 items and on the total score favored teachers in the experimental groups. ANCOVA found the experimental groups significantly higher on four of the 13 items and on total score. Because faculty were randomly assigned to conditions, the study controls for motivation (within a volunteer group) and supports the value of these workshop activities; however, one possible source of bias is that raters, if they noticed the taping and observation activities, may have suspected that an experiment was under way.

In summary, all but two of the studies of seminars for teaching assistants found changed attitudes of participants' students, particularly with regard to students' perceptions of teachers' classroom performance. The magnitude of the impact is small, and of the three studies with experimental designs, two failed (Dalgaard, 1976; Haber, 1973) and one succeeded (Carroll, 1977) in demonstrating statistically significant impact. The two studies of workshops for faculty did not investigate participants' classroom behavior but, like the studies with teaching assistants, did demonstrate impact upon students' ratings.

Impact on student learning. Of the four studies which examine impact upon student learning, we first describe a training program for seven graduate assistants in introductory economics which was conducted during the second term of their teaching (Lewis & Orvis, 1973). Each was responsible for two sections of 25 students; all students also met together for lectures by senior faculty. During fall term no training was available. During winter term, instructors met for weekly seminars and each was videotaped three times, following which about two hours were spent in individual review and critique of the tape and of the instructors' ratings from the previous term. Student achievement, student ratings of instructors, and the instructors' classroom behavior were compared between fall (control) and winter (experimental).

Stepwise multiple regression indicated that the average student of a trained teaching assistant scored significantly higher on a standardized test of achievement in economics ($p < .05$). The following variables were also significantly associated with achievement: prior knowledge of economics, mental ability and achievement, maturation, sex, and student evaluations of instruction. Student evaluations were significantly more positive winter than fall term. Anticipating criticism of the quasi-experimental design, the authors argue that results do not represent a practice effect since such fall to winter changes had not occurred the year before.

Thirteen teaching assistants in rhetoric, participating in Koffman's (1974) dissertation research, were videotaped and completed questionnaires and tests at the start of a term. They were divided into two treatment groups and a no-treatment control group. Groups one and two reviewed their data with an instructional specialist. Group one, in addition, held subsequent meetings with the specialists who provided further suggestions and training. After eight weeks, all teaching assistants were again taped and again completed the written measures. Videotapes, student evaluations, and tests were also analyzed. Measures of teacher behavior and of student attitude and achievement favored the continuing treatment group (group one) and less so group two, compared with controls. These trends, however, did not reach statistical significance.

Training in both interaction analysis and heuristic questioning was investigated with teaching assistants in mathematics in Tubb's (1974) dissertation study. Eight teaching assistants were randomly selected from 21 who were teaching a calculus course for nonmathematics and nonengineering students. Teaching assistants were trained in Flanders Interaction Analysis or in Polya's Heuristic Teaching or in both. Although the numbers receiving each type of training are not reported in the abstract, each strategy appears to have influenced classroom behavior of those who were trained, as shown by change scores. Students of trained teachers showed higher achievement and problem-solving skill than control students and rated their instructors even higher than their "ideal expectations" for teaching ability.

Eight teaching assistants in the mathematics department at East Carolina University were randomly assigned to a group trained in interaction analysis or to a group receiving no training in Daniels' (1970) dissertation study. Some of the participants were pursuing a degree in mathematics education and others were pursuing a degree in mathematics. Flanders categories were applied to audiotapes made at several points during the term. Trained teaching assistants scored higher on four of the nine categories used in the analysis, indicating greater indirectness and flexibility. The mathematics education group scored higher than the mathematics group on six of the categories, regardless of training. Students of the mathematics education teaching assistants scored higher than those of mathematics teaching assistants regardless of training group. Thus, both training and degree objective are influential in this study.

In conclusion, the evidence from these courses and seminars spans a number of academic fields and suggests that seminar experience can affect the achievement of students of trained teachers as well as affect student attitudes and teacher classroom behavior. Not all studies find significant differences and not all studies avoid important threats to validity, but such trends are well worth pursuing. Because they are based primarily on graduate teaching assistants, their generalizability is limited. Experienced faculty may be unwilling to volunteer and may strongly resist being assigned to such activities. Further, teaching experience may interact with program activities and thus decrease (or perhaps increase) the impact of training.

Guidelines for Assessing Impact

Estimates by participants of their satisfaction and learning are the most common data for evaluating the impact of workshops in which faculty participate. There are important problems with relying on such estimates. To close this chapter, we refer to the literature in continuing medical education for illustrations of these problems.

One study evaluated intensive instruction (12-20 hours) given to practicing physicians in recognizing unknown heart sounds (auscultatory skill) (McGuire, Hurley, Babbott, & Butterworth, 1964). During instruction, heart sounds and their visual representations were simulated; participants practiced naming the sounds and received immediate feedback. Anonymous evaluations showed that participants felt they had learned a great deal and assessment of their skills showed that, compared with a control group, they made significant gains from pretest to posttest.

Six months later, a representative subgroup of participants was again tested. Two results are noteworthy. First, their mean skill score at six months was not significantly different from their mean score at <u>pretest</u>.

Second, it was expected that even if there was a decrement in skill, the course might have produced increased sensitivity to cardiac findings and a consequent increase in the frequency and variety with which cardiac information was observed and recorded. A comparison of hospital charts completed by these physicians before and after the course revealed no differences in the amount or quality of the cardiac information recorded.

Assuming that skill-oriented teaching improvement workshops are designed in some ways parallel to this one, these findings should caution us (a) against accepting end-of-course satisfaction as predictive of long-term learning, (b) against accepting end-of-course skill gains as indicating long-term skill learning (unless there is opportunity for subsequent practice with critical evaluation), and (c) against assuming that in the absence of changes in performance a workshop may, nevertheless, produce changes in a general characteristic such as sensitivity.

The relationship between self-rated learning and objectively assessed learning was also explored in the evaluation of an educational development program at Wayne State University School of Medicine. Fifty-five persons participated in two three-hour meetings for each of 12 weeks. The sessions covered a variety of topics related to learning and instruction. Participants rated their progress on statements expressing the objectives of each session. Their ratings were generally high and fairly uniform across objectives, surprising staff who had noted considerable variability in actual accomplishment. Further consideration of staff observations and of the participants' ratings suggested several conditions which affect the accuracy of participants' estimates: When participants could "engage in free discussion, when there was a comfortable rapport between teacher and participants, when relatively few demands were made on them to demonstrate their skills, and when there was little external feedback to them on their performance, there were uniformly high achievement ratings. When there were clear tests of their knowledge and external feedback, ratings of achievement varied between people and between objectives and were generally lower" (Koen, 1976, p. 855).

These illustrations imply several guidelines for workshop assessment, guidelines which are seldom followed in the research on faculty workshops. Both immediate and delayed tests of ability should be made, but it should be recognized that without opportunity for continuing practice with feedback, the post-course level of skill mastery is not likely to be maintained. Participant self-assessments, if they are to be accurate, should refer to specific behaviors, those behaviors should have been assessed during instruction, and participants should have had opportunity to compare their performance with an external criterion. Finally, if participant self-assessments are used to evaluate sessions which include goals related to attitude change, the sessions should include exercises or discussions which insure that participants have become actively involved with a variety of views.

Practice with Feedback: Microteaching and Minicourses

During the last 20 years, programs which prepare teachers for elementary and secondary schools have increased the time during which teaching is actually practiced. Expansion of practice teaching in real classrooms accounts for some of this increase. In addition, there has been an increase in brief teaching encounters focused on behaviorally specific skills and videotaped for subsequent review. One strategy for providing such practice with feedback is microteaching. Another involves self-contained instructional packages, called minicourses, prepared especially for inservice teachers.

Both microteaching and minicourses show promise for improving college teaching, although most systematic evaluations of their use have been in precollege settings.

Microteaching

Microteaching, a scaled-down teaching encounter, was originally developed for use with preservice elementary and secondary school teachers. It allows teachers to learn and practice teaching skills within "micro" conditions, that is by teaching a five to ten minute lesson to a small group of approximately five pupils. The microteaching process has four steps. First, a preservice teacher is presented with a behaviorally defined teaching skill. Second, the teacher plans a lesson which incorporates the skill and teaches the lesson to a group of approximately five pupils while being videotaped. Third, the teacher receives feedback on the lesson from peers and supervisor and by viewing the tape. Fourth, the teacher reteaches the lesson to another small group of students and incorporates feedback suggestions. A variety of skills is usually taught in the microteaching experience, and for each new skill this four-step sequence is followed.

Many elements of the microteaching format are based on research on observational learning and behavior modification. For example, Bandura and Walters (1963) have studied imitative learning and modeling and their findings have influenced the microteaching model. Cognitive discrimination training, with roots in the behavioral movement, serves to make the teacher aware of appropriate teaching behavior. In discrimination training, the learner is presented with relevant behavioral instances and then taught to discriminate between them. Learning consists of two steps: learning to attend to the relevant dimension and then to distinguish between different values of this dimension (Wagner, 1973). In the microteaching situation, teachers learn to discriminate between effective and ineffective instructional behavior by viewing samples of their own and others' teaching.

Microteaching's underlying component-skills approach requires that teacher behavior be broken down into specific components. Emphasis is on acquisition of one skill at a time. Technical skills that are often taught include stimulus variation, fluency in asking questions, and the use of higher-order questions. The selection of skills is based on the relationship between these technical skills and pupil performance (for a comprehensive review see Turney, Clift, Dunkin, & Traill, 1973, chapter 2).

Some researchers have emphasized the self-confrontation aspect of microteaching (Perlberg, Peri, Weinreb, Nitzan, & Shimron, 1972; Perlberg, Bar-On, Levin, Bar-Yam, Lewy, & Etrog, 1974; and Fuller & Manning, 1973). They suggest that microteaching provides feedback to prospective teachers by causing the teachers to confront themselves. Through self-confrontation, the teacher becomes aware of any discrepancy between intentions and outcomes. A discrepancy leads to negative feelings such as dissatisfaction and discomfort. Festinger (1957), in his theory of cognitive dissonance, proposes that the reduction of such dissonance is a motivating force in individuals, leading to a change in self-perception and/or behavior. This suggests that in microteaching, prospective teachers improve their teaching skills in order to reduce dissonant feelings produced by the self-confrontation process.

Numerous studies investigating microteaching have been conducted with prospective elementary and secondary teachers and programs have been set up on some college campuses to work with teaching assistants and faculty (for example, see Miltz, 1978), but we have located only three systematic studies that use microteaching to improve college teaching. Nevertheless, this technique appears to be easily adaptable to higher education and we will review the major and exemplary studies both at the elementary/secondary levels and at the college level.

We first discuss the earlier studies by relying, for the most part, on secondary sources, and then review and critique findings from more recent research. Although these studies investigate the relationship between microteaching and improved teaching performance, not all of them conceptualize improved teaching performance in the same way. In some studies, the microteaching skills are aimed at improving overall teacher competence by concentrating on such areas as lesson planning, discussion skills, and controlling techniques and procedures. In other studies, skills are more narrowly focused and directed toward developing specific technical skills.

It should be noted that recently, Hargie, Dickson, and Tittmar (1978) have described a variation of microteaching entitled "miniteaching." In this variation, 'reteach' has been abandoned, integration of skills is stressed, lesson length and number of pupils is gradually increased and remedial sessions are sometimes programmed. We have found no systematic studies of this technique, so a critique of miniteaching is not included in this review.

Early studies. After microteaching was developed in the early 1960's, numerous studies compared it with conventional teacher training methods. Allen and Clark (1967), in one of the first studies comparing microteaching to conventional student teaching, found microteaching to be more effective than student teaching in developing teaching competence. Subsequent studies at Stanford did not compare microteaching with conventional methods; rather, microteaching was assessed in terms of change in teacher effectiveness occurring from first to last microteaching session. For example, Fortune, Cooper, and Allen (1967), reported the results of an investigation of the effectiveness of the Stanford Micro-Teaching Clinic of 1965. They claimed microteaching to be effective in improving overall teaching performance, but their study has been assigned a low confidence rating because among other problems, it lacked a control group. A survey by Ward (1970) of microteaching in United States elementary and secondary programs noted in Turney, Clift, Dunkin, and Traill (1973) reported microteaching to have been generally effective in improving teaching competence and developing favorable attitudes toward education. Turney, Clift, Dunkin, and Traill (1973) also have reviewed the microteaching literature, drawing similar conclusions regarding the general effectiveness of microteaching.

Jensen and Young's (1972) methodologically sound comparison of microteaching with conventional methods in developing teaching skills assessed teaching performance on three different occasions using the Teacher Performance Evaluation Scale. Factor analysis identified six performance factors: personality traits, warmth of teacher behavior, general classroom atmosphere, lesson usefulness, teacher interest in pupils, and teacher interest in student achievement. Microteaching was found to be significantly better than student teaching practice for the first five of these six factors, although the superiority of microteaching was sometimes not evident until the third observation after about six weeks of teaching. Jensen and Young interpret this finding as evidence that the effects of microteaching are not temporary and may increase with time.

Not all studies find microteaching more effective than traditional methods. Kallenbach and Gall (1969) found no significant differences between the use of microteaching and student teaching. Nevertheless, they conclude that microteaching can be considered superior to conventional methods because it achieves similar results and requires less administrative work and time. This study earns a high confidence rating.

The relative merits of components of the microteaching process have been assessed in several studies. Turney, Clift, Dunkin, and Traill (1973) reviewed research findings on six areas of microteaching: (a) attitudes toward microteaching, (b) modeling, (c) pupils versus peers in the microlesson, (d) supervision, (e) feedback, and (f) the teach-reteach interval. Their findings include generally positive trainee attitudes toward microteaching, although some instances of unfavorable attitudes have been noted particularly toward the videotape recording. Skill acquisition seems more effective when positive models are used, and perceptual models seem to be superior to symbolic models. Some skills, however, are just as effectively taught through symbolic models. Discrimination training appears to be an important element of microteaching. Several presentations of model behavior are superior to a single presentation. Practice in a context similar to that of the model enhances learning. School students rather than peers are recommended for the microlesson. For feedback to be effective, it should be directly related to the model toward which trainees are molding their behavior. Videotape feedback appears to ensure the best feedback, particularly when it is varied, positive, and specific. Research on the teach-reteach interval was inconclusive.

Hargie's (1977) review of early research on microteaching organized the evidence into four categories: changes in teaching performance, pupil attitudes toward their teacher, trainee teacher attitudes toward their course of training, and increases in pupil learning. He concluded that microteaching, as measured by ratings of behavior or by counts of actual behavior, was generally effective in improving teacher performance. Studies assessing pupil attitudes toward teaching were rare but generally positive results with respect to microteaching were found. With respect to trainee attitudes toward microteaching, generally trainees consider microteaching to be an effective teacher training tool. Hargie noted that few studies had been carried out to investigate increases in pupil learning as a result of teachers trained in microteaching. However, one study does suggest that pupil learning may vary according to age and subject characteristics.

Recent studies. The studies reviewed in this section sample recent research on microteaching alone or on microteaching in combination with other techniques. Like the earlier research, these studies for the most part favor the microteaching approach. However, three (Johnson, 1977; Perlberg, Peri, Weinreb, Nitzan, & Shimron, 1972; Perlberg, Bar-On, Levin, Bar-Yam, Lewy, & Etrog, 1974) of the quantitative studies did not include control or comparison groups and have been assigned low confidence ratings. The elimination of control groups in these studies was sometimes justified by earlier studies investigating classroom teachers and showing that teacher behavior is remarkably stable from lesson to lesson. Assuming that the teaching performance of a group not receiving the intervention would remain unchanged, researchers felt no obligation to include control groups. However, some studies have found unstable behavior for control groups (e.g., Borg, 1975; Perrott, Applebee, Heap, & Watson, 1975). Furthermore, there is little evidence from higher education to support the stability of teacher behavior. Of the three studies from higher education, two (Johnson, 1977; Perlberg, Peri, Weinreb, Nitzam, & Shimron, 1972) did not have control groups and were assigned low ratings.

Among recent studies reviewed here there is evidence for changes in teacher knowledge, teacher behavior, and pupil behavior. Wagner (1973) compared two methods of influencing the knowledge and teaching skills of undergraduates studying distinctions between student-centered and teacher-centered teacher behavior. Seventy-eight undergraduates were randomly assigned to three groups: Discrimination training, microteaching, and control. All participants had 15 minutes to prepare a five minute lesson. The discrimination group then received about 30 minutes of training on discriminating student-centered from teacher-centered teacher comments: they rated 33 taped teacher comments and were given the correct answers to each as well as brief explanations. The microteaching group taught the prepared lesson, reviewed the videotape of that lesson, and discussed the tape and student ratings with a supervisor. They then retaught the lesson. The control group merely proceeded to the criterion test.

On a criterion test immediately after training, trainees in all groups prepared and taught a 10 minute lesson to three college students. Videotapes of these lessons were coded according to the six categories of student-centered teacher behavior used in the training. A week later all students completed a test in which they coded a number of teacher comments. On the written tests the discrimination group scored significantly higher than the control group, but the microteaching group did not differ from the other two groups. On the performance test the discrimination group was more student-centered as represented by such behaviors as asking for clarification, restating and using student's ideas, than either the microteaching ($p < .01$) or control group ($p < .0005$). The microteaching group was not significantly more pupil-centered than the control group. The greater student-centered behavior of the discrimination group was for the most part due to an increase in pupil-centered behavior rather than to a reduction in teacher-centered behavior.

Wagner concludes that it is the discrimination training rather than the actual practice in microteaching that results in teacher change and that without discrimination training microteaching practice is ineffective. It is suggested that the combination of discrimination training and microteaching might prove very effective. Wagner's study is well designed and executed. Such weaknesses as the time lag between the two measurements and the fact that the discrimination test may have precluded assessment of whether teachers learned to attend to relevant dimensions are noted in discussion. Although the study is limited in its generalizability to those individuals motivated to change and resentful demoralization may have occurred among those in the control and microteaching groups, we rate it with high confidence.

The critical role of discrimination training in the microteaching sequence has more recently been discussed by Hargie and Maidment (1978). They found a number of studies supporting discrimination training as a necessary component in teaching performance.

Three studies have investigated microteaching with college teachers. Johnson (1977) investigated combined training in Flanders' Interaction Analysis and training in microteaching labs for producing teacher change in interaction behavior, questioning, and reinforcement techniques. Fourteen community and junior college professors participated. Analysis of variance revealed significant change from pretest to posttest scores for all eight variables measuring teaching performance. All of the changes were increases with the exception of teacher talk which significantly decreased. Since there was no control group and a small sample was used, many plausible alternative explanations exist. It is possible that the volunteer participants were initially motivated to change their teaching behavior and would have done so with many kinds of training (Hawthorne effects). Or possibly the group improved as a result of maturation. Therefore, a low confidence rating has been assigned.

Perry, Leventhal, and Abrami (1979) also investigated the effects of a variation of microteaching experience on college teachers. The microteaching experience, called Modified Observational Learning, consisted of microteaching feedback along with cognitive discrimination training. Trainees were asked to role-play four teaching behaviors. For each behavior, participants were videotaped and provided with remedial feedback until a criterion level was reached. Subsequently, the master tape of the four videotaped role-play "takes" along with a pretraining tape was given to each subject. The subject was instructed to spend three and one half hours each week viewing both tapes as a cognitive discrimination exercise.

For the experiment, four graduate students, the "instructors," were randomly assigned to either a training or a control group. Within each group, instructors were labeled as high or low effective according to pretest ratings. Two subsamples of introductory psychology students from the same introductory psychology course participated. Students from one subsample were randomly assigned to four pretraining conditions while students from the other subsample were randomly assigned to the four posttraining conditions. Thus, separate pre and posttraining samples were used. Students completed a questionnaire for assessing teaching effectiveness and an achievement measure.

Findings indicated that training interacted with lecturer differences. That is, for initially low effective teachers, there were no differences in student ratings and achievement between the experimental and control groups. However, for the initially high effective lecturer, higher student ratings and achievement scores were reported for those trained by Modified Observational Learning. In terms of performance over time for the trainees, low effective lecturers showed no change in ratings or achievement from pre to posttraining while high effective lecturers' student ratings did not change but student achievement increased significantly between testing sessions. In the control condition, the low effective lecturers showed no change in ratings or achievement while high effective lecturers' ratings decreased from pre to posttraining. This study has been assigned fair confidence for a number of reasons. Important information relevant to the study's conclusions was not included in the brief report such as the duration of the experiment and the probability levels used to determine significance; nor were reliability of measures reported. The small number of instructors involved limits generalizability although this weakness is noted by the investigators. Also, graduate students with no teaching experience were used as instructors, thereby limiting generalizability to inexperienced college teachers.

Perlberg, Peri, Weinreb, Nitzam, and Shimron (1972) studied sixteen faculty members in dentistry to determine if microteaching techniques designed to develop classroom interaction styles and student-centered teaching would increase use of such behaviors. They also hypothesized that change produced by microteaching would be directly related to a participant's openness: the more dogmatic and authoritarian a participant's attitude toward education, the less likely the participant would change. All seven skills used to measure teaching performance (lesson organization, lecture style, providing examples, fluency in question, probing questions, higher order questions, and divergent questions) showed significant improvement ($p < .01$) from pretest to posttest. Data also indicated that there was greater improvement in questioning skills than in lecturing skills. Three measures designed to assess participant's attitudes, the Rokeach Dogmatism Scale, the Permissive-Authoritarian Scale and a bipolar adjective scale, as well as attendance at microteaching sessions (perserverance) were used to investigate the relationship between attitudes toward openness to behavioral change and acceptance of innovation. Only on the bipolar adjective scale were scale scores significantly related to post-treatment ratings. The best predictor of openness to change and willingness to accept innovation was perserverance in microteaching clinic sessions. The second best predictor was the participant's attitudes toward the microteaching concept and the third best predictor was the participant's attitude toward "dentist."

This study has been assigned a low confidence rating because it lacks an adequate control group. Faculty improvement may have been due to factors other than the treatment such as effects of history, the group's prior training over two years in teaching improvement activities, and Hawthorne effects.

Perlberg and his associates have conducted two other microteaching studies with precollege teachers. Perlberg, Bar-On, Levin, Bar-Yam, Lewy, and Etrog (1974) investigated the effectiveness of a combination of microteaching and a computerized feedback system called Technion Diagnostic System on the behavior of 60 students in teacher training programs at Technion Institute in Israel. This combined technique brought about significant changes in combined scores measuring student-centered teaching behavior (nonverbal, not lecturing, relates to) and higher cognitive questioning (analytical thinking). For the three student-centered teaching behaviors, peak performance was reached at the end of training and posttest scores showed a decrease from the last training session. However, two plausible explanations are given for this finding: (a) student fatigue, and (b) the fact that the posttest lesson was a general lesson not a specific skill lesson. This study was assigned a low confidence rating primarily because in the absence of a control group we cannot rule out alternative explanations for teaching improvement such as history and maturation.

A workshop utilizing demonstration, discussion, and microteaching to develop teacher strategies for increasing independent learning skills in pupils was investigated by Kremer and Perlberg (1979). Changes in both teacher and student behavior were assessed. Results indicated that teachers in the experimental group talked less and gave less information than control teachers. They also asked broader questions and gave more direction. This finding is explained as resulting from experimental pupils being involved in many activities thus requiring more directions. Significant pupil behavior changes favoring the experimental group were found for three of four variables representing child-centered teaching (responds to teacher, initiates talk to teacher, and initiates talk to another pupil). Increases in number of questions and problems raised by students were also noted for the experimental group pupils. However, significant differences in higher level questions in favor of students taught by the experimental group were found for only two of seven variables, divergency and analysis. Kremer and Perlberg point out that there were more changes in classroom interaction than in cognitive processes.

Overall, this study indicates that microteaching can be used to increase independent learning skills of pupils. Although the study is well designed and the analysis appears appropriate, we have rated it fair because it is not clear that random assignment to groups was carried out. Strengths of the study include the choice of instruments, its thorough literature review, its well-developed theoretical framework, and its inclusion of qualitative data.

In summary, the results of recent studies on the use of microteaching indicate that microteaching can be effective in improving actual teaching performance. More specifically, it appears that microteaching can develop student-centered teaching behavior. Generally student-centered teaching behavior results in less teacher talk and more pupil talk. More questioning goes on and less lecturing is done. Furthermore, microteaching can be used to develop higher-order questioning on the part of teachers and students as well as to increase teacher reinforcement skills.

No significant relationships have been shown between personality correlates and microteaching performance or microteaching attitude.

Of particular interest is the finding that discrimination training is a critical component of microteaching. Discrimination training is a cognitive exercise that is concept-based rather than practice-based. The findings with regard to discrimination training suggest that concept-based training may be a powerful tool not only for increasing concept acquisition but also for increasing skill acquisition. When one considers the lower cost of discrimination training in comparison to microteaching and practice teaching, one begins to realize the importance of these findings. Particularly for the college setting, discrimination training seems more feasible than practice-based models. We return to this theme in the later discussion of protocol materials.

Although positive results have been found both for microteaching alone and in conjunction with other techniques, a good number of the studies rate only low confidence. These ratings are due for the most part to the one-group designs which allow for a number of plausible alternative explanations for significant findings.

Microteaching studies conducted with college teachers have seldom been well designed. Although the evidence indicates microteaching combinations to be beneficial in improving teacher competence, better designed research directed at faculty improvement needs to be conducted before conclusions may be drawn about which aspects of the technique are effective for improvi g what skills for which college teachers.

Minicourses

Minicourses are based on the microteaching model and draw upon research on technical-skills training, modeling, feedback, and film production. Essentially the minicourse teaches the technical skills of teaching through the following process: (a) viewing films of behaviorally defined skills in a specific domain of classroom teaching, and (b) practicing those skills within a microteaching format. The minicourse differs from simple microteaching in that it was designed particularly for inservice teachers, although it has been used with preservice teachers as well. The minicourse model allows a working teacher to develop needed technical skills in a microsetting and eventually to adapt these skills to a regular classroom. By providing regular classroom experience, the minicourse model counteracts the criticism leveled against microteaching that acquisition of teaching skills in a restricted setting does not necessarily prepare a teacher for regular classroom conditions.

Minicourse titles incluce, "Developing Learning Skills," "Tutoring in Mathematics," "Thought Questions in the Intermediate Grades," and "Effective

Questioning in a Classroom Discussion (Secondary Level)." Minicourse activities are integrated into a regular school day, and may be taken by a group of teachers in that school, by a pair of teachers who review one another's tapes, or even by an individual. The minicourse cycle includes (a) reading, viewing films, and planning a lesson, (b) teaching to a small group from a regular class, (c) viewing the tape, (d) reteaching followed by feedback. Focus is on practice and feedback since "about 10 percent of the course involves telling the teacher; 20 percent involves showing him; and the remaining 70 percent involves allowing him to practice his teaching skills and watch replays of his own performance" (Borg, Kelley, Langer, & Gall, 1970, p. 31).

Although we found a few studies that adapted the microteaching model to higher education teaching improvement, no studies were located that used minicourses for improving college teaching. Therefore, minicourse studies included in this review were done with elementary and secondary school teachers. Minicourses are included because they are highly effective at those levels, and because we feel that their format may be viable for use with college teachers. Furthermore, since there is evidence that microteaching at the college level is effective in improving instruction, it seems probable that minicourses are also potentially effective at that level.

Developmental studies. Numerous minicourses have been developed by the Far West Laboratory for Educational Research and Development. All have gone through extensive field testing. Both preliminary and main field tests have been conducted for each minicourse. In these tests teachers were videotaped in their classrooms prior to the introduction of the minicourse. After completing the minicourse, teachers were again videotaped in their classroom. Pretest-posttest analyses were made of the videotape.

For the most part, minicourses have proven to be effective for improving the specific technical skills for which each was designed. Further analyses have investigated delayed post-course performance, pupil change, and the use of the minicourse with different social classes. Revisions were initiated when preliminary or main field tests indicated lack of teacher improvement on a particular skill.

Almost all of the minicourse field tests were conducted without control groups. This deficiency in design in addition to other design problems threatens the validity of these studies. For example, such effects as testing, maturation, and evaluation apprehension may have biased study results and conclusions. However, Borg, Kelley, Langer, and Gall (1970) anticipate these criticisms and are able to rule out a number of threats. For example, it has been impractical for some investigators to find appropriate control groups, and this deficiency allows for a plausible alternative explanation of effects; that is, the changes noted for teachers may have been due to maturation rather than to the intervention. They note, however, three reasons why one would expect a comparable control group's teaching behavior to remain stable. First, the average teacher in their study had nine years experience, and thus was unlikely to make any significant teaching change without intervention. Second, they cite research evidence indicating that classroom teaching behavior is remarkably stable from lesson to lesson. Finally, they cite a study that used student teachers as a control group. This control group, which could be expected to be much less stable than an experienced group, showed significant improvement in only two of 12 Minicourse I behavior areas over a two month span. In those field tests which did include control groups, little significant change was found.

Other limitations of the field test procedures are also discussed by Borg, Kelley, Langer, and Gall (1970) who note that the studies were conducted with volunteer teachers, and so generalizability is restricted. They go on to state that this limitation is not as serious as it first appears. Because inservice programs are generally voluntary, the field test data would apply to inservice conditions. Furthermore, they cite one minicourse as an example where non-volunteers and volunteers were used and changes were found for all of them.

Regarding the possible effects of a videotape recorder in the classroom, Borg, Kelley, Langer, and Gall (1970) admit that the equipment might contribute to atypical teaching behavior particularly at the pretest (evaluator apprehension and testing effects). It is also pointed out that the equipment might have been serving as a discrimulative stimulus; that is, only when the recorder was present were teachers emitting target behaviors. They rule out this possibility by stating that it is unlikely that teachers would maintain their posttest performance after a four-month interval has occurred unless those skills had been practiced during that period. Another limitation is the possibility that positive changes noted at posttest resulted merely from the teachers' awareness at posttest of the target behaviors under study. They countered this assertion by noting that only after hours of concentrated effort did teachers display the target behaviors and thus, it was unlikely that teachers were emitting those behaviors simply because they knew which skills were under study. Other findings from two studies conducted with student-teachers (Borg, 1969) did not find significant differences in behavior between a group informed of target behaviors and a group that had not been informed.

As can be seen, although a single field test for one minicourse may not have accounted for all possible threats to validity, the sum total of studies that have been carried out to investigate minicourses has for the most part ruled out a good many threats. Numerous replications have also been conducted. Overall, then, it appears that minicourses do effect positive changes in behavior of precollege teachers.

Recent studies. Aside from these field tests, other studies have been made of the basic minicourse model and its effectiveness over an extended period of time. Four of these are discussed here. Each has been assigned either a fair or high confidence rating and each supports the minicourse model in improving instructional effectiveness.

In 1972, Borg studied the effectiveness of Minicourse I ("Effective Questioning") over an extended time interval. The study was designed as a three-year follow-up of the effects of Minicourse I. Of the 48 original field-test teachers, 30 teachers were still at field test schools and 24 agreed to participate. No control group was used. At the initial evaluation of Minicourse I, 11 of 13 target teacher and pupil behaviors showed large and statistically significant improvement. Four months later, teachers showed continued improvement in three of the 11 skills that were measured and had not regressed significantly on any skill. Approximately three years later (39 months), subject performance still remained significantly greater on eight of the 10 scored behaviors. Thus, most changes induced by Minicourse I persisted over three years. Some behaviors, however, did regress. After three years, frequency of one-word student responses increased significantly and this frequency was even higher than the precourse mean. Also, teacher talk had regressed significantly; teacher talk had increased from 33 percent at the course's end to 45 percent after three years, but was still below the initial frequency level of 53 percent.

Borg's (1972) study has been given a fair confidence rating. It is subject to a number of validity threats including testing effects, selection, history, and maturation. Several of these threats are discussed; for example, he contends that maturation is not a serious threat by citing research showing that teacher behavior remains stable over time, but as we have seen this evidence is mixed. Problems not ruled out by Borg are the threats of evaluator apprehension and mortality.

Perrott, Applebee, Heap, and Watson (1975) investigated the feasibility of transfer of Minicourse I to Great Britain. In a one-group pretest-posttest design, they checked for testing effects by randomly assigning participants at pretest into two subgroups; one was informed of the target behaviors involved in the study and the other was not informed of the behaviors. There were no differences in performance between the groups on the pretest videotape, thus ruling out the possibility that positive posttest changes could be attributed to testing effects rather than to the intervention itself. The minicourse was effective in producing significant changes at posttest on eight of 14 measures. The most important change was the consistent reduction in proportion of discussion dominated by teacher talk, a change concurrent with changes in more specific teaching behaviors. This study is thorough and well planned except that it lacks a control group; it serves not cnly as a test of information transfer but as a replication of Borg's three-year follow-up. Perrott, Applebee, Heap, and Watson (1975), as noted above, also offer evidence of mixed results concerning stability of teaching behavior.

Buttery and Michalak (1978) also used Minicourse I in a study which modified the minicourse format in two ways. First, they devised the Teaching Clinic Feedback Prccess which substituted audio tape for videotape for recording behavior and providing feedback. The second modification involved a naturalistic setting, using regular classroom groups and thus eliminating the need for potentially inconvenient special microteaching conditions. Further, this study used preservice teachers as its subjects rather than inservice teachers. The teaching clinic model was used with one group and compared to a control group which received regular student teaching instruction. It is unclear whether subjects were randomly assigned to groups. The Teaching Clinic Process consisted of (a) lesson planning session, (b) observation session, (c) critique preparation session, (d) critique session, and (e) clinic review session. Results indicated that preservice teachers who completed Minicourse I with these modifications displayed more significant changes in teacher behavior than those who received regular student teaching instruction. Eleven of 13 target behaviors changed significantly for the experimental group while only two of 13 were significant for the control group. A number of design and analysis problems result in the fair confidence rating. Because it is unclear whether randomized assignment was carried out, effects of selection-maturation, regression and testing may bias the results.

Collins' study (1978) differs from the one just described in that it investigated effects of a minicourse designed by herself and her educator colleagues rather than by the Far West Laboratory. The target of Collins' minicourse was teacher enthusiasm. The study focused on two issues: (a) whether a minicourse on enthusiasm could increase the level of teacher enthusiasm of preservice teachers, and (b) whether the effects of this course would be maintained three weeks after the course's end. A pretest-posttest control group design was used with delayed posttest. Participants were preservice teachers rather than inservice teachers. Results indicated that the experimental group increased their overall level of enthusiasm and also tended to exhibit a greater amount of variance in performance during posttests. In contrast, control subjects tended to display more similar behaviors in enthusiasm during the posttests. The experimental group maintained the increased level of enthusiasm three weeks after the minicourse training while no important differences were evidenced for the control group from one test to another. An observable decrease was noted for the experimental group from posttest I to posttest II. Collins suggests that the performance of preservice teachers was leveling after the immediate effects of training and that if tested in another six weeks, the experimental group's posttest III scores would not have differed from posttest II scores. Collins supports this explanation by pointing to other research with similar results. A high confidence rating has been assigned to this study. The investigators attempted to control for a number of internal and external validity threats by using observers blind to the experimental conditions, by not informing subjects that they were involved in a research project, by using random assignment, and by using reliable measures. A repeated measures ANOVA was used appropriately.

In summary, the basic minicourse appears to be highly effective in changing teacher behavior. From recent studies it appears that the minicourse is a flexible tool that can be modified and adapted in a number of ways while remaining effective. For example, the minicourse can be used in naturalistic settings and in settings where videotaping equipment is not available, or it can be transferred from the United States to Great Britain. Minicourse-induced change in instructional effectiveness has been shown to persist over three years.

More research should be conducted on whether teaching behavior of inservice teachers not exposed to such an intervention does indeed remain stable, whether videotaping affects teachers so that nontypical teaching behavior is recorded, whether videotaping equipment serves as a discriminative stimulus to teachers in these experiments, and whether knowledge of target behaviors at the pretest makes a difference in pretest behavior. In view of the apparent effectiveness of the minicourse model with elementary and secondary school teachers, research should be extended to college teachers to determine if developing minicourse materials would be cost effective at the postsecondary level.

Feedback from Ratings by Students

In studies using student ratings to improve instruction, feedback is regarded as an impetus for change in teaching performance. These studies have included (a) the use of written student rating feedback alone, (b) the effects of student rating feedback over time, (c) the use of written student rating feedback with consultation, (d) the study of discrepan:ies between student evaluations and faculty self-evaluations, and (e) the impact of student rating feedback ~~and~~ student performance.

Ratings Feedback Alone

Most studies on written student feedback are conducted in the following manner. Rating forms are completed by students approximately three to four weeks after the beginning of the term. These ratings are analyzed and averages or percentages are computed for each item and/or dimension. About the fourth or fifth week of the term, results are returned, perhaps accompanied by normative data, to one group of instructors and withheld from others. Student ratings are again collected as a criterion measure at the term's end. Such studies investigate whether mid-term feedback contributes to change in rated teacher performance. In this case, no consultation between faculty development specialists and instructors occurs; written student feedback results alone are used.

Twelve studies were located using this approach. The results of the studies vary. Six studies found significant positive change in teaching performance (Butler & Tipton, 1976; Bledsoe, 1975; Sherman, 1978; Braunstein, Klein, & Pachla, 1973; Overall & Marsh, 1976; and Tuckman & Oliver, 1968). Three studies found no significant differences between feedback and no feedback (Centra, 1973; Miller, 1971; and Rotem, 1978). Three studies reported mixed (Marsh, Fleiner, & Thomas, 1975; and Murphy & Appel, 1978) or uncertain (Friedlander, 1978) results.

Although nine of the 12 studies provide at least some support for impact from student feedback, a critical review of the quality of the studies indicates that this conclusion may not be warranted. Several studies finding significant positive change are flawed by design and analysis problems. For example, in the study by Butler and Tipton, no control group was used, the sample size was small (N=17 instructors) and conclusions attributed to the findings seem premature. The investigators claim that six of 17 instructors showed significant improvement on post-ratings, but the design of the study does not permit us to determine the causes of these changes. Bledsoe's study (1975) also suffers from several methodological problems including participation of only one instructor and his class in the experiment, and the fact that the instructor under study was also the investigator (the threat of experimenter expectancies).

In Sherman's study (1977-78), two instructors were rated after each class meeting. Students rated the quality of instruction at that meeting and the value of the content of that class. They were also asked to give reasons for their ratings. Instructors were not present during data collection and were not told the purpose of the research until later in the term. The three conditions were no feedback (baseline), feedback in the form of average ratings only, and feedback including average ratings, range of ratings, and reasons for ratings. Results showed that under the third condition the ratings of both instructors were significantly higher than during baseline. Among the problems of this study are the absence of a condition to control for the reactive effects of testing, dropout of participants, and lack of parallel data for the two instructors. Nevertheless, the question of optimal level of feedback specificity for affecting teaching is an important one, deserving further research.

Three studies of higher quality favoring student ratings are Braunstein, Klein, and Pachla (1973), Overall and Marsh (1976), and Tuckman and Oliver (1968). Braunstein, Klein, and Pachla (1973) compared a feedback condition with a no-feedback control condition. Although randomized assignment to conditions was carried out, pretest results indicated that the two groups were not equivalent at midsemester. The no-feedback group had higher midterm ratings than the feedback group. When changes were analyzed, strong positive shifts in evaluations were found for the feedback condition while strong negative changes were noted for the no-feedback condition. Two explanations for these results are offered: (a) that feedback contributed to the end-of-semester group differences, or (b) that regression toward the mean occurred for both groups. The nonequivalence of groups at mid-term and a possible mortality bias have contributed to a confidence rating of fair for this study.

Overall and Marsh (1976) sought to clarify the mixed findings of earlier studies on student rating feedback. In those studies, including one by Marsh, Fleiner, and Thomas (1975), both positive and no-difference findings had been shown. The more recent investigation by Overall and Marsh found significant differences favoring student rating feedback. This study is well designed and executed using analysis of covariance, although unlike other studies, the unit of analysis is not instructors but the students who filled out the questionnaire.

Tuckman and Oliver (1968) found significant differences in favor of the feedback condition with high school teachers. Although this study is well designed, it is questionable whether the use of change score analysis was appropriate. Two other studies conducted with high school teachers support Tuckman and Oliver's findings (Bryan, 1963; and Gage, Runkel, & Chatterjee, 1960). These studies were located in reviews, and so we cannot comment on their quality.

The three studies (Centra, 1973; Miller, 1971; and Rotem, 1978) that found no significant differences between feedback and no feedback conditions are randomized studies with appropriate comparison groups. Miller notes that combining data from various sections of one instructor may have resulted in sampling errors due to a small _n_ per cell. The unit of analysis in Miller's study was teaching assistants. Rotem (1978) notes that the short time interval of his study may have contributed to his no-difference findings. The Rotem study is unique because it was conducted at a research-oriented university.

As stated previously, Murphy and Appel (1978) like Marsh, Fleiner, and Thomas (1975), offer mixed findings. Murphy and Appel's feedback conditions varied slightly from other studies. The design included three conditions: no feedback, rating feedback only and augmented feedback. Augmented feedback consisted of student ratings along with individual performance standards and remedial alternatives reported by each instructor prior to the midsemester evaluation. Significant differences for the feedback conditions were found, although change score analysis was used. Absolute change was small and thus implies little practical significance. In another finding, augmented midsemester feedback was no more effective than simple feedback in improving end-of-semester ratings.

Instructors in 85 management classes were invited to distribute midterm evaluations to their students (Friedlander, 1978). As part of an end-of-term evaluation, students were asked whether the instructor had distributed the midterm questionnaire and discussed its results with the class. About one-third of the responding students indicated the midterm questionnaire had been distributed. The author concludes that students attribute change in the course to the midterm evaluation to a greater extent when there was adequate class discussion of midterm results than when there was not. The report is difficult to follow, however, since it is unclear which students were included in subsequent analyses. Because of this and other design problems, the study rates low confidence.

In summary, these studies seem to provide more evidence for than against written student feedback alone, but many of the studies are poorly designed and analyzed. Three previous reviews have been conducted of this research. Kulik and McKeachie (1975) concluded that research at that time did not support differences between feedback and no feedback conditions in improving instruction. A more recent review by Abrami, Leventhal, and Perry (1979) states, "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 the findings across studies. There are also no reports of the magnitude of significant effects so it is difficult to estimate the amount of improvement which feedback can produce" (p. 361). Rotem and Glasman (1979) in reviewing a large body of research on feedback regarding teaching concluded that there is a "minimal effect at best of feedback on instructional improvement at the university level" (p. 497). It will become clear as we proceed that our conclusions are somewhat more optimistic than theirs.

Since most of the studies using student rating feedback involve volunteer subjects, their generalizability is limited. Centra (1973) notes, however, that most faculty who use instructional improvement programs are volunteers. He argues that generalizability is therefore appropriate for those most likely to use instructional improvement programs.

Effects of Ratings Over Time

For the most part, studies in the previous section investigated the effects of written feedback on teaching performance during one term. Two studies have investigated the effects of student ratings (without consultation) over two or more terms (Centra, 1973; Vogt & Lasher, 1973). Students were handed rating forms about the fourth week of the term. These ratings were tabulated and provided to the instructors as feedback. The students were asked to fill out rating forms at the end of that term and successive terms. In Centra's study (1973), the effects of rating feedback on teaching performance was investigated over two semesters. Among the conditions studied were: a feedback pre/post condition, a no-feedback pre/post condition, and a no-feedback posttest only condition. Interestingly, there were no significant differences among the groups after one semester even when sex, subject area, and college teaching experience were taken into account. However, an analysis after two semesters based on much smaller samples in each group revealed that teachers who had received feedback twice received better ratings than those who had received feedback once or not at all. Centra's study is well designed, earning high confidence. Appropriate statistical analyses were carried out and a thorough discussion of plausible explanations for the study's findings was included.

Vogt and Lasher (1973), at a college of business administration, also investigated the effects of rating feedback on instructional effectiveness over time. They analyzed ratings from 26,458 questionnaires for 63 teachers over six to eight quarters. All instructors received feedback. Their design is quasi-experimental and, hence, not as strong as Centra's. Regression analysis indicates that feedback did not contribute to improved teaching performance over time.

Since only two studies have investigated the effects of rating feedback over time and since their findings are contradictory, we await further research to settle this issue.

Ratings with Consultation

Personal consultation is sometimes provided along with rating feedback tabulations and normative data. Usually, consultations include the interpretation of ratings and suggestions for improving teaching skills.

Seven studies investigated the effects of this combination of ratings and consultation on instructional effectiveness. All of these studies appeared since the Kulik and McKeachie review mentioned above. For the most part, they support the effectiveness of a rating/consultation combination in improving instructional performance; however, confidence ratings vary from low to high for these studies.

Aleamoni (1978) used a nonequivalent control group design to assess the combined effects of consultation and rating feedback over a period of one semester to a year later. Therefore, feedback was distributed and consultations were conducted at least a semester before follow-up rating forms were collected. Ratings of the feedback recipients improved significantly on two of five dimensions. Rather than a repeated measures analysis of variance, a more adequate strategy might have been a multivariate analysis of covariance. Also, Aleamoni does not state whether his analysis adjusted for unequal N's. Aside from these problems, the nonequivalent control groups raised threats to internal validity such as selection-history and regression. With respect to regression, ten subjects were initially dropped from the experimental group because they did not qualify for remediation; the experimental group then consisted of low scorers. Consequently, this group's final higher scores may be due to regression of their low scores toward the mean. Resentful demoralization may have affected the control group which originally was to have consultation, thus inhibiting changes which might otherwise have occurred.

McKeachie and Lin (1975b) studied 37 graduate assistants and three faculty members teaching the introductory psychology course at the University of Michigan. Students completed a 32-item form about one-third through the term and again near the end of the term. At a voluntary evening session some students also provided data on academic measures, including an achievement test in psychology. Instructors were randomly assigned to three groups: no feedback (13 sections), printed feedback (13 sections), and personal feedback (14 sections).

This report provides a well-detailed description of the personal feedback condition:

> At the beginning of the feedback sessions teachers were asked to fill out forms indicating their expectation of the student ratings on each dimension, their own self-perceptions, and where they would like to be. Typically, Professor McKeachie then asked them how the class was going and in response to their reactions, suggested how the student ratings confirmed (or rarely did not confirm) their perceptions. He then pointed out factors on which the teacher differed significantly from the mean of all classes. If there seemed to be any problems, he suggested some possible alternative methods of handling the problem. All of the mean ratings, however, were relatively favorable... so that the hope that he could help teachers cope with very negative feedback was not realized. (McKeachie and Lin, 1975b, p. 6).

The group receiving personal feedback was rated significantly higher on two general items (overall value of course and general teaching effectiveness) and on one of the seven dimensions (impact on students). No clear pattern of significant effects on academic measures was found. Among other problems, the study suffers from subject mortality, but, particularly because of the random assignment of teachers, it does support the value of feedback with consultation over feedback alone.

Hoyt and Howard (1978) report two studies conducted at Kansas State University using a combination of computerized rating feedback and consultation. One study (Study 1 in Hoyt and Howard, 1978) compared the first and last student ratings of the same instructor and course that had been taught on two different occasions. Results were statistically significant for 13 of 15 measures, but Hoyt and Howard point out that they were not dramatic in the absolute sense. Since no comparison groups were used in this study, confidence in the results is limited. Hoyt and Howard replicated this study (Study 2 of Hoyt & Howard, 1978), using a single group, and found significant improvement on the objective, "progress on relevant objectives." The fact that significant improvement was not shown for individual objectives on the rating scale was discounted on the basis that most faculty had rated these as irrelevant to the course. A second analysis was conducted to examine instructional improvement relative to contact (none, some, much) with the office that provided consultation services. When posttest measures were adjusted for pretest differences, it was found that rated teaching effectiveness increased as a function of amount of contact with faculty development services. But our confidence in the findings is low due to its nonrandomization and single group design.

Studies of Erickson and Sheehan (1976) and Erickson and Erickson (1979) investigated a combination of rating feedback and consultation offered by a Teaching Improvement Clinic. In a well-designed and well-executed study, Erickson and Sheehan (1976) compared three conditions: data collection only, diagnostic (ratings feedback alone), and full process (ratings and consultation). Instructor self-ratings and student ratings indicated that, overall, the full process members changed no more or less than members in the other conditions, although all three groups made positive changes. Erickson and Erickson (1979) then designed a study with only two conditions: data collection only and full process. Significant differences favored the full process group for both student and instructor ratings. As the investigators were concerned that these findings merely reflected different group expectations of change, a follow-up study was conducted to investigate this possibility. Differences in performance between semester I and II were significant for 11 of 20 faculty members. Erickson and Erickson claim that these results show that qualitative changes do occur and are not the result of different group expectations. Overall, the Erickson and Erickson study earns high confidence, since certain initial weaknesses were tested in a follow-up study.

Two studies have failed to support the ratings/consultation treatment. One, Erickson and Sheehan (1976), was mentioned above. The second, Weerts (1978) found no significant differences from midterm to end-of-term for two feedback groups (printed feedback and verbal feedback). A two-factor ANOVA with repeated measures on one factor was used. The analysis also indicated that there were no significant differences among these groups and a no feedback control group at the term's end. Yet, Weerts points out that, although no significant differences were found, results show an interesting pattern; that is, 20 of 28 items in the verbal feedback group had higher ratings than corresponding items in the no feedback group. The chances of this occurring were less than five in 100. Similarly, on 23 of 28 items, the printed feedback group had higher ratings than the no feedback group. The chance of this occurring was one in 1000. Thus, Weerts believes that these results indicate that ratings and consultation might improve teaching performance. It is important to note that the unit of analysis was classes and that graduate teaching assistants taught these classes. This study is assigned a low rating because of several analysis problems; a multivariate analysis of covariance, for example, might have been more appropriate.

Reviewing these findings with regard to the quality of studies, we see that of the studies that found significant results in favor of this technique, three were assigned low confidence, one was given a fair rating, and one received high ratings. Although the results are not clearcut, they do indicate directions to pursue in further research. For example, even though Weerts did not support the effectiveness of this technique in a statistically significant way, positive trends were noted in favor of a rating/consultant approach.

Instructor-Student Discrepancies

If there exists a negative or positive discrepancy between the instructor's and the students' evaluation of instruction, an imbalance is created for the instructor. In order to restore the state of equilibrium, the instructor may attempt to reduce this imbalance. Such a prediction may be made from social psychological theories such as incongruity theory, dissonance theory, and balance theory. Several studies investigating discrepancies were located.

As mentioned above, Rotem (1978) found that feedback did not affect subsequent ratings compared with a no-feedback control and a posttest only control. He also found that discrepancies (a) between instructors' actual and desirable ratings or (b) between students' and instructors' ratings were no more effective than midterm ratings alone in predicting end-of-term ratings.

Braunstein, Klein, and Pachla (1973), mentioned above, assessed the effects of discrepancies between midterm perceived performance (as rated by instructors) and actual performance (as rated by students) on end-of-term evaluations. They concluded that when an instructor's expectancy was discrepant from students' ratings for a trait, a subsequent shift in the direction of the instructor's expectancy for that trait is likely. The strength of the relationship between discrepant expectation and change in ratings was .77 (phi coefficient).

In Pambookian's 1974 study, it was postulated that moderately rated instructors would improve more than those rated favorably or unfavorably. Based on his results, Pambookian claimed that the initial level of student evaluation strongly influenced the instructor and that moderately rated instructors improved more than favorably or unfavorably rated instructors. In a later study (1976), Pambookian hypothesized that the greater the discrepancy between student ratings and instructor self-rating, the greater the improvement after feedback for those instructors. It was found that unfavorably discrepant teachers improved on skill, feedback, rapport, general teaching ability, and overall value of course more than the favorably discrepant. The minimally discrepant improved significantly on one dimension, rapport, as compared to the favorably discrepant and showed strong trends in the same direction on skill. The least gain was made by the favorably discrepant. Pambookian's studies earn low confidence for several reasons. The sample sizes were small (N=13) and no control group was used. Statistical analysis appears to have been inappropriate. For example, change score analysis was used with nonequivalent control groups. Furthermore, when an analysis of variance did not reveal significant differences on certain skills, t-tests were used (inappropriately) to investigate differences between groups.

Centra's 1973 study, mentioned above, also investigated the effect of discrepant ratings. It is well designed with a multi-institution sample. Centra hypothesized that student feedback would produce change in instructors who had rated themselves more favorably than their students had rated them (unfavorably discrepant group as defined by Pambookian). The analysis generally supported this conclusion: five of 17 items showed significantly higher scores for the unfavorably discrepant group compared with the favorably discrepant group. Thirteen of the 17 items showed trends in that direction.

Twenty-eight instructors at the University of Michigan participated in a study of the effects of feedback discrepancies on subsequent ratings (McKeachie & Lin, 1975a). A 32-item questionnaire with seven dimensions was completed by students about one-third through the term and two weeks before the end of the term. Instructors also completed the form once as they expected to be rated by students and once as they "would like to teach" (ideal). All teachers received their ratings as feedback. For analysis teachers were blocked into eight groups depending on the discrepancy between student ratings and various combinations of expected and ideal self-ratings. On two of the seven questionnaire dimensions (group interaction and feedback), significantly improved ratings were found for those whose expected and ideal ratings were higher than student ratings. The group which was rated more highly by students than by themselves changed in a negative direction (on feedback dimension only). This pattern of changes and other trends in the data suggest to the authors that the discrepancies may raise (or lower) faculty motivation and thus affect behavior.

In summary, the findings of these studies suggest that instructors who rate themselves more favorably than their students are more likely to improve their teaching performance as a result of student rating feedback than those who rate themselves less favorably than their students.

As a final study dealing with discrepancies, we cite one in which instructor self-rating was used as a dependent variable. (In the studies cited above, student ratings constituted the dependent variable.) Oles and Lencoski (1973) investigated whether an instructor's own self-rating of his course and teaching would be affected in any way by receiving the results of students' evaluations. In this study, 24 graduate level instructors were assessed using a pretest-posttest control group design. All subjects were asked to complete a self-rating form 2 weeks prior to the end of the course. In addition, in 12 of the subjects' classes, students were asked to fill out a course/instructor evaluation form. These forms were analyzed and the results were returned to each instructor along with another self-evaluation form that the faculty member was requested to return as soon as he reviewed the student evaluation results. Instructors for the other 12 classes served as a control group and received no feedback but did complete a second self-evaluation form. The study's findings indicate that while the test-retest correlation coefficient for the control group was .82, the correlation for the experimental group was .54 suggesting according to Oles and Lencoski that receiving student evaluations did have some influence on the instructors' self-rating. A chi square test on the total number of changes regardless of direction of change was significant. Changes in self-ratings were not all in the direction suggested by student evaluations.

Effects on Student Performance

The relationship between the use of student rating feedback and student performance has also been investigated. The assumption underlying studies that used student achievement as an outcome measure is the following: if student rating feedback does improve instruction, that improvement should be evident in student performance. As we have seen, McKeachie and Lin (1975b) did not find clear effects of feedback on student achievement. Three other studies have investigated this notion. Both Miller (1971) and Marsh, Fleiner, and Thomas (1976) found no overall significant differences between feedback and no feedback groups on student achievement exam scores. Miller's study has been assigned a high confidence level, and we regard this aspect of Marsh, Fleiner, and Thomas' study with a fair level of confidence.

Overall and Marsh (1977) conducted a similar study a year later and found that students and faculty who received ratings feedback with consultation scored significantly higher and noted greater interest in taking future coursework in the subject area than students of instructors in a no-feedback condition. Their analysis may be regarded with a fair level of confidence. Based on their findings and the previous contradictory findings, Overall and Marsh recommend additional research on this issue. It is our recommendation as well.

To conclude this chapter on student ratings, we are pleased to note the relatively large number of studies although we are disappointed with their variable quality. The clearest finding concerns discrepancies between the instructor's self-rating and ratings by students. This discrepancy appears to be an effective predictor of who will benefit from ratings feedback. Feedback has its greatest impact on those whose self-ratings are more positive than the ratings made by their students.

The most pressing topic for further research, in our opinion, is the relative effectiveness of written feedback alone compared with written feedback plus consultation. Either written feedback alone or written feedback plus consultation has been shown by most studies to be superior to no feedback. Only three studies (Erickson & Sheehan, 1976; Weerts, 1978; McKeachie & Lin, 1975b) directly compared written feedback alone with feedback plus consultation, and only one of them (McKeachie & Lin, 1975b) found clear support for consultation as more effective. Since consultation is an expensive activity, it is important to learn for which faculty it is most useful. Greater attention should be given in this research to instructor variables such as motivation and self-other rating discrepancies.

Concept-Based Training: Protocol Materials

Protocol materials are film or videotape recordings which illustrate educationally relevant concepts. Developed for precollege teachers, they also show promise for postsecondary education. Protocols are designed to link educational theory to the teaching process. Generally, a single protocol module focuses on a set of related concepts. For that reason protocols are considered to reflect a concept-based model of teacher education in contrast to microteaching and minicourses which reflect a practice-based model.

Protocol training is carried out in the following manner. Teachers are provided with written materials and films which describe and illustrate the concepts. They learn how to apply the concepts through a sequence of visual illustrations, written exercises, and tests. To illustrate protocols, we describe materials produced at Indiana University entitled, "Concepts and Patterns in Teacher-Pupil Interaction." There are ten films in the series. Concepts basic to the series are introduced in three films, "Questioning: Reproductive and Productive," "Probing and Informing," and "Approving and Disapproving" (six concepts in all). Each film is seven or eight minutes long and provides classroom examples of the concept. Six films show classroom episodes to be analyzed according to the target concepts, thus providing practice in interpreting classroom behavior. Each of these films is approximately eleven minutes long. The tenth film, 35 minutes in length, is used as a performance test. It includes 30 brief scenes to be categorized according to the target concepts. Protocol materials are aimed at producing concept acquisition in users, facilitating skill acquisition, and (by inference) promoting desirable changes in the students of teachers who have been trained.

The protocol idea, materials protraying behavioral events relevant to instructional concepts, was first proposed by Smith (1969). In 1970, the Bureau of Educational Personnel Development of the Office of Education funded a number of projects at universities throughout the country. Partly because of the funding arrangements, more work has gone into development of the materials than into evaluation. In his survey of protocol module evaluations, Cooper (1975) notes that compared to the number of protocols produced, relatively few have been adequately evaluated. Cooper summarizes evidence from 73 sources on the effectiveness of protocols in improving teaching. He reviews these studies with respect to four issues: teacher skill acquisition, teacher concept acquisition, reactions to protocols, and pupil outcomes. Of this research, only one study was identified showing that protocol modules could change on-the-job teacher behavior (Borg and Stone, 1974), and this study is discussed below with Borg's other protocol studies. Cooper also identified a number of studies conducted at Utah State University, Michigan State University, Far West Laboratory for Educational Research and Development, and Indiana University. For the most part these studies found positive results for concept acquisition by preservice and inservice teachers. Furthermore, Cooper indicated that teachers generally had positive reactions to protocol materials. However, Cooper notes an absence of research showing impact on pupil behavior.

Since Cooper's 1975 survey, we have identified additional studies of protocol's effects on teacher and pupil behavior carried out primarily by two groups, by Borg and associates at Utah State University and by Gliessman, Pugh, and associates at Indiana University. None of these studies investigated protocol materials at the college level, but they are included here because of the potential adaptability of the technique to postsecondary education.

Concept Acquisition

Several of the Indiana studies investigated users' reactions to protocols and alternative ways of structuring protocols.

Gliessman and Pugh (1976) studied concept acquisition and teacher and student reaction to the protocol on teacher-pupil interaction. Generally, use of the protocols resulted in significant gains in acquisition of concepts basic to the series, and teachers and students reacted favorably to the series. The experimental design also allowed for checking effects of pretesting on posttest results, and pretesting did affect posttest scores. This study rates fair confidence, primarily because the comparison group also received the protocols intervention.

Gliessman and Pugh (1978b) explored the instructional rationale of protocol material. More specifically, they were interested in determining what components of a protocol sequence were necessary for and effective in producing concept acquisition. Teacher-pupil interaction protocols were used in two studies to compare a number of instructional treatments. For example, one group received names of concepts only, while another group received concept names and concept definitions. A third group received concept names, definitions, and filmed exemplifications. A fourth group received a combination of concept names and filmed exemplifications. Gliessman and Pugh concluded that receiving concept definitions alone did not yield effects equivalent to those achieved through the exemplifications of defined concepts; exemplification contributed significantly to concept acquisition. We view this study with fair confidence. Such problems as selection-history biases in study one and the use of a probability level of .081 preclude a high confidence rating.

Another study by Gliessman and Pugh (1978a) also investigated concept acquisition of teachers trained with the teacher-pupil interaction protocol. Its distinctive purpose was to investigate the effect of protocol films of contrasting structure on the acquisition of teacher behavior concepts and reactions to the filmed treatment. Three separate studies were carried out with preservice and inservice teachers enrolled in a graduate level educational psychology course. Significant gains in concept acquisition were found for groups viewing high or low structure films but no significant differences were found between these two film treatments. When high structure, low structure, and a high/low structure combination were compared, significant increases in concept acquisition were found for all three groups. A comparison of means revealed significant differences between the high- and low-structure groups favoring the low-structure group. A third substudy investigated the contradictory results of the first two substudies--the finding of both significant and nonsignificant differences between groups trained by high-structure films or low-structure films. When teacher discussion was controlled for, no significant differences were found between groups trained by low or high structure. This study rates high confidence. It is well designed and appropriate statistical analyses were used.

These studies verify the effectiveness of protocol materials for concept acquisition. The amount of structure in the films may vary without reducing learning, but learning is enhanced when concepts being taught are exemplified as well as defined.

Skill Acquisition

Other studies have assessed the impact of protocols on the classroom skills of teachers. Gliessman, Pugh, and Bielat (1979a) investigated concept acquisition for the protocol on teacher-pupil interaction. One group, the protocol training group, received protocol training. A second group, the alternate group, served as the control group and received student counseling training. Mean concept acquisition scores and mean skill acquisition scores were significantly greater for the group trained with the protocol module. The correlation between concept and skill acquisition was .51 (df = 8, p = .08) and the investigators conclude that mean skill frequency scores tend to increase with increasing levels of concept acquisition. This correlation, however, is rather low and may in part be due to low statistical power. A larger sample size might produce a higher correlation. A number of design and interpretation weaknesses have led to our low confidence rating. First, the study does not make clear whether randomization was carried out. Thus, it is possible that the two groups operated under different historical circumstances and hence, that significant differences are the result of selection-history biases. Second, differences between the groups for concept acquisition are statistically significant but their practical significance is uncertain.

Another study by Gliessman, Pugh and Bielat (1979b) failed to support the relationship between skill concept acquisition scores and skill frequencies. In this study, a one group pretest-posttest design was used to further explore and replicate the findings of Gliessman, Pugh, Bielat (1979a). Thirty inservice teachers were trained in teacher-pupil interaction skills using the teacher-pupil interaction protocol. Probing behavior was the focal criterion for both concept and skill acquisition. Three different measures were used: 1) performance of trainees on a concept acquisition test, 2) teaching behaviors as exhibited in a microteaching session, 3) trainees' interpretive written responses regarding their audiotaped interactive skills. The content of the trainees' written responses was analyzed (a) for evidence of nominal outcomes ("name" condition in Gliessman and Pugh, 1978b), conceptual outcomes ("definitions" condition in Gliessman and Pugh, 1978b), and observational influences ("exemplification" condition in Gliessman and Pugh, 1978b) in their use of interactive skills, and (b) for evidence of the ability to probe interpretively using the criteria of accuracy and application.

No significant relationship was found between skill concept acquisition scores and skill frequencies. However, trainees' written responses provided subjective evidence of both conceptual and nominal outcomes of protocol training. No observational effects were found in trainees' written responses. Trainees' written responses also indicated that ability to apply the concept of probing was positively and significantly related to the frequency of probing. However, written responses indicated that trainees' accuracy in dealing with probing concept characteristics was unrelated to skill acquisition. The data did appear to confirm previous findings that protocol training leads to concept acquisition.

Gliessman, Pugh, and Bielat (1979b) has been assigned a fair confidence rating. Although we applaud its intended purpose of replication, the short duration of training leads us to question the findings. This limitation is also noted by the investigators. The authors further point out that the mean skill frequency of probing was considerably smaller than in their previous investigation, possibly a result of the short training interval.

Both concept acquisition and skill acquisition were investigated by Kleucker (1974). She studied preservice teachers randomly assigned to four conditions: protocol training alone, skill training alone (microteaching), both protocol and skill training, and a placebo, that is training unrelated to the study. The two target behaviors were asking probing questions and offering accepting reactions. Protocol training and skill training led to concept acquisition and skill acquisition respectively when compared to control groups. But protocol trainees did not perform better on concept tasks than those trained with microteaching, and those trained by microteaching did not perform better on skill tasks than those trained with protocols. Training in both was at least equally effective and sometimes significantly more effective than training in either alone. This study rates high confidence. Discussion of findings, limitations, and implications is thorough. As limitations, it is noted that the small number of participants may have contributed to no significant differences in some of the comparisons, that the control group may have served as a treatment, and that instruction time was not held constant across conditions. Furthermore, Kleucker notes certain limitations in the criterion tests used.

Borg and Stone (1974) made a pretest-posttest comparison of behavior changes brought about by the protocol modules on extension and encouragement. These Utah State University protocols are part of a series of six related to teacher language behaviors. It is important to note that all teachers were informed of the target behaviors prior to the pretest in order to eliminate one threat to validity. The threat was that positive gains would result not from the treatment but from subjects' posttest knowledge of the target behaviors. Results showed that teachers made significant gains on five of seven specific behaviors covered in the protocol materials.

The second part of the study compared protocol modules and minicourses in effecting teacher behavior change. A nonequivalent control group design used field test data previously collected with Minicourse I, which trains behaviors similar to the extension protocol study, and with Minicourse II, which trains behaviors similar to the encouragement protocol. Although the sample used in the Minicourse I study was similar to the protocol study, the sample from Minicourse II was not. Both groups showed similar gains for most of the behaviors that were compared. Borg and Stone conclude that from a cost-benefit perspective, the protocol model might be more desirable than the minicourse model for increasing the use of simple, clearly defined teaching behaviors. This study rates low confidence because of the nonequivalent control group design and the use of change scores in analysis. The differences in sampling for the minicourse and protocol groups and the possibility of differential history effects also confound the interpretation of results.

Changes in Students

Pupil behavior as well as teacher behavior was assessed in two protocol studies. Borg, Langer, and Wilson (1975) compared teachers trained by the classroom management skills with a no-treatment control group of inservice elementary school teachers. Changes in teacher and pupil behavior were assessed. Teachers using protocols were rated more favorably on all 13 target teaching behaviors but differences were generally small and nonsignificant. For pupils taught by teachers in the experimental group, work involvement increased significantly and deviant behavior decreased significantly in recitation situations. In seatwork situations, although pupil work involvement significantly increased, deviant behavior showed no significant changes. Two reasons were given for the low teacher behavior frequencies: (a) the possibility that the observation time period was too short, and (b) the possibility that the observers became fatigued over the two-hour observation period. The results have a low confidence rating because of design deficiencies. Low statistical power (N=29) may account for the nonsignificant changes in teaching behavior and in pupil deviant behavior during seatwork. Second, data were analyzed using analysis of covariance on nonequivalent control groups. It is possible that a combination of measurement error in the pretest and differential growth patterns between the experimental and control groups may have led both to overadjustment and underadjustment of the data, washing out significant differences between the groups.

A study by Borg in 1977 investigated the impact of two protocols, teacher-pupil interaction and pupil self-concept, on changes in teacher and pupil behavior. Subjects were randomly assigned to one of the two protocols, each condition serving as a control for the other. With respect to changes in teaching performance, about one-half (seven of thirteen) of the classroom management behaviors increased and 11 of 12 self-concept teacher behaviors increased (except for four negative behaviors that had not been present prior to the treatment). For pupils of teachers using management protocols, no significant change was noted for work involvement, but significant decreases in deviant behavior were found. Students of teachers using self-concept modules significantly reduced off-task behavior but did not reduce other target behaviors. There was no significant improvement in pupil self-concept for either experimental or control groups.

The mixed results of this study are viewed by Borg as partially successful. He suggests that teacher behaviors improved more with the self-concept module because less time was necessary for training these teacher behaviors than for training classroom management skills. Furthermore, Borg suggests two reasons that the improvement in pupil self-concept was small: (a) the possibility that in fact there is no relationship between the behaviors taught in the self-concept protocol and an improved student self-concept, and (b) the possibility that overall differences were not revealed because most of the Anglo students in the classrooms had initially good self-concepts and hence, served to wash out the gains of the small group of minority students. Borg's study rates high confidence. Although small sample size possibly contributed to low statistical power, the use of ANCOVA and the thorough discussion of plausible explanations merits some confidence. A no-treatment control group would have permitted assessment of cross-protocol effects, i.e., whether the self-concept protocol contributes to improvement in classroom management and vice versa.

Borg conducted a 1975 study using the four Utah State University protocols on teacher language. He compared teaching performance between a group trained in four protocols and a no-treatment control group of inservice teachers. He also investigated the relationship between teacher behaviors covered in these four protocols and pupil achievement as well as the relationship of teacher characteristics and pupil achievement. Significant gains were made by the experimental group on all twelve measured teaching behaviors while the control group made significant gains on four of the twelve behaviors. When both groups' posttest measures were adjusted for pretest differences, it was found that the experimental group had significantly higher scores on four of the teaching behaviors. Borg notes that significant change for the control group for some of the teaching behaviors is in conflict with the premise that teachers' behavior remains stable over time without intervention. He suggests three possible explanations for his results: (a) changes in observer standards between pretest and posttest, (b) the content area taught for the posttest being more appropriate for language development, and (c) contamination (compensatory rivalry, diffusion or imitation of the treatment). Borg concludes that contamination was the most likely cause of the control group's gains on the four teaching behaviors. In addition, partial correlations were computed between pupil achievement on two achievement measures and the 12 teaching behaviors. When pupil academic ability, parents' occupation, and teacher coverage of the unit's content were partialed out, it was found that the teacher's use of defining, voice modulation, paraphrasing, and cueing were significantly related to student achievement on two measures and the teacher's use of opening review and terminal structure were significantly related to one achievement measure. However, none of the partial correlations between ten high inference teacher characteristics and student achievement were significant.

Several problems with the study reduce our confidence in its results. As in the Borg, Langer, and Wilson study (1975), an analysis of covariance was used to analyze data collected from nonequivalent control groups, and so significant results for four of the teaching behaviors may be the result of underadjustment caused by pretest measurement error rather than by the protocols themselves. Or the nonsignificant differences may be a wash-out effect. Also, the possibility that compensatory rivalry, or diffusion or imitation of treatments took place on the part of the control group complicates the interpretation of the findings even further.

Our review of research with protocols leads to several conclusions. Teachers and pupils appear to react favorably to the use of protocols. Generally, teachers show significant concept acquisition from protocol traiuing. For skill acquisition, results of protocols are not as clear, although in some studies skill gains have been documented for at least some target behaviors. Findings are also mixed when protocols' impact on pupil behavior and achievement is investigated. Each study finds some positive effects for protocols. Further research should reveal for which teacher and student behaviors effects are most reliable.

To our knowledge, no research has been conducted on the impact of protocols on college teachers and students. Since the training of teachers using protocol modules appears to lead to increased concept acquisition, colleges interested in this goal might explore the protocol format. Since protocol training requires neither videotaping nor classroom practice, it is less threatening and less disruptive of regular teaching than are practice based programs. Of course, protocol development is expensive, beginning with the identification of concepts critical to instruction. Some of that fundamental work should be repeated for higher education, since it is by no means clear that existing protocols and the concepts they exemplify are the critical ones for the college classroom.

Conclusions and Implications

We have reviewed scores of empirical studies of attempts to improve college teaching. These studies evaluate interventions aimed at assisting faculty to change their teaching activities or roles in order to enhance the educational experience for themselves and their students. Impact of the interventions has been assessed through measures of the professors' attitudes, through observations of their classroom behavior, through reports of their students about the class, and through measures of their students' learning.

Our review was undertaken to determine what guidance this literature can provide to those who conduct research and to those who design and implement instructional improvement programs in postsecondary education. In this final chapter we discuss several issues regarding research and practice. These issues constitute an agenda for our own subsequent research and writing and are discussed here only briefly.

The literature on teaching improvement in higher education is larger than we had expected when we began this review. It is also of lower quality than we had hoped. Table 1 summarizes studies charted in Appendix A according to the intervention addressed and our confidence rating. Recall that our confidence rating serves only as an approximation. Our criteria are not rigidly fixed and reliability of classification may not be perfect. Nevertheless, there are sufficient entries in most cells of that table to convey an adequate impression of the pattern of relative attention given to topics and of the quality of research from topic to topic. We also note in Table 1 (in parentheses) the number of entries which support the intervention in question. This display suggests several observations.

1. Most studies support the intervention in question. Overall, 82 percent of the entries in Table 1 support the intervention being investigated. (Please note, that studies with multiple variables are entered in more than one category of the table.)

2. Each specific intervention category receives support from at least 50 percent of the entries. For 11 of the 13 categories, support is provided by 70 percent or more of the entries.

3. The higher the methodological quality of the entry the less likely it is to support the intervention being investigated. Interventions are supported by 93 percent of entries rated low, by 86 percent of entries rated fair, and by 60 percent of entries rated high. This does not mean that only high quality studies should be taken seriously. It may be that in fine tuning methodology, investigations have become insensitive to the phenomenon being studied. It is also possible that, since lower quality studies are flawed in different ways, combining their results exploits overlapping strengths, while not doing so would overemphasize their separate weaknesses.

4. We have been particularly impressed with the research on interventions developed for precollege teachers. The precollege research on microteaching, minicourses, and protocols has in large part involved research programs rather than single studies and has shown awareness of desirable

Table 1

CONFIDENCE IN RESULTS

	Low	Fair	High	Total
Grants	-	1(1)*	-	1(1)
Attitude workshops	1(1)	-	-	1(1)
Skill workshops	5(5)	9(8)	-	14(13)
Microteaching	4(4)	2(2)	3(1)	9(7)
Minicourses	-	2(2)	2(2)	4(4)
Ratings alone	4(4)	4(4)	4(1)	12(9)
Ratings over time	-	1(0)	1(1)	2(1)
Ratings and consultation	4(3)	1(1)	2(1)	7(5)
Ratings discrepancy	4(4)	1(1)	2(1)	7(6)
Ratings on students	-	3(2)	1(0)	4(2)
Protocols concepts	1(1)	3(3)	2(2)	6(6)
Protocols skills	4(3)	1(0)	2(2)	7(5)
Protocols on students	2(2)	-	1(1)	3(3)
TOTAL	29(27)	28(24)	20(12)	77(63)

*Numbers in parentheses represent studies which support the intervention. The total number of entries in this table (77) is greater than the number of studies in Appendix A (60) because several studies apply to more than one intervention category.

design characteristics even when circumstances did not permit incorporation of all the desired features. It is worth speculating on reasons for the apparent lower quality and greater fragmentation of research in postsecondary settings. Are higher education researchers less competent, standards less stringent, problems more difficult, funding less available, or is some combination of these at work?

A well-defined field of inquiry should draw upon coherent theory, subscribe to high standards of research, and build upon previous research in a systematic way. By these criteria, research on the improvement of college teaching does not yet constitute a well-defined field. For most studies, the basis in theory is strained and for some it is non-existent. Work on major conceptual issues remains to be done; before we can validate materials or programs for instructional impact, we must clarify the nature of "instruction" and the meaning of "improvement." These concepts are seldom explicitly defined in this literature and, as we struggled with implicit definitions, they often struck us as inappropriately narrow. Further, a host of design problems plague this research. Finally, the field is fragmented because most research is only a single study effort.

Implications for Research

We shall limit our discussion here to only five implications for research. They are general in nature but progress on them is basic to the further development of the field.

1. Individual difference variables deserve greater attention. Most of this research treats participating faculty as an undifferentiated mass, distinguished only by the treatment to which they are assigned. More attention should be given to individual differences (either as independent variables or as blocking variables). The value of attention to individual differences is demonstrated by the studies of discrepancies between faculty self-ratings and student ratings. Systematic study of demographic information, motivation, and other self-described characteristics may assist in identifying those persons who are most ready to engage in change projects and for whom particular interventions are most suitable. Likewise, when the impact on students of a teaching-improvement intervention is studied, individual differences among students should be noted; otherwise significant interactions will not be documented.

2. Dependent variables require comparable definition and operationalization across studies. We hoped to aggregate the findings from studies of several of the interventions under review. For example, research on the impact of student feedback might be combined across studies according to the dimensions of the questionnaires used in each study. One hypothesis is that ratings feedback would have greater (and faster) impact on a "rapport" factor than on a "course organization" factor. Since so few studies use the same questionnaire or analyze questionnaires in a similar way, our attempt at such aggregation proved futile. Seldom are common schedules for classroom observation used and in few fields are there standard measures of student achievement. Although studies should not require uniformity in design, they cannot build upon one another until some comparability emerges.

3. Much wisdom remains undocumented and unshared. A number of figures in the faculty development movement have accumulated impressive experience in a variety of settings and projects during the last few years, but little of that experience is systematized and available to others. For instance, many people have learned a great deal from the PIRIT project, and it has informed the design of a subsequent national project; yet little generalizable knowledge emerged from the research on PIRIT. The field needs better communication channels to capture and share such wisdom. Although it may not itself be research-based, that wisdom is empirical in that it derives from experience, and it should play a critical role in the planning of subsequent research.

4. Cross campus collaboration is absent. Appendix A studies are isolated efforts of investigators on individual campuses. Inter-campus research networks are potentially powerful tools for dealing with several of the problems we have noted. Wisdom from previous efforts would be part of the planning of such studies. Experts in research methodology could be part of the research team. Practical problems of research design such as random assignment and small numbers of participants would be alleviated. The time required for planning, data analysis, and writing could be shared. Similar collaboration is not unknown in other fields. For instance, cooperative clinical trials have long been used in medical research, but that method would be new to higher education.

5. Most data reflect only superficial levels of experience. The studies rely primarily on self-report and questionnaire data. Seldom does the research go to levels of experience below the surface and reveal cognitive, emotional, political, and developmental experiences. What goes on in the mind of the professor while teaching or while watching a tape of his or her class? What feelings are experienced while reviewing a computer report of student ratings? How do perceived rewards for teaching relative to rewards for research productivity influence professors' responses to opportunities for improving their teaching? How do developmental tasks at particular stages of adult life interact with perceived teaching problems and challenges?

The dominant research strategies in this body of literature come out of the quantitative methodological tradition and are insufficient for investigating questions such as those just listed. To advance the field we need careful classroom ethnographies, disciplined case studies, sensitive clinical interviews, as well as rigorous experimentation. The literature of higher education does contain exemplary efforts using several such methods. Andrews' (1978) case study is illuminating. Cottle's (1977) essays are provocative. Axelrod's (1973) portraits of teachers provide unusual depth. Becker, Geer, and Hughes' (1968) participant-observations richly develop the context of student life. And Mann, Arnold, Binder, Cytrynbaum, Newman, Ringwald, Ringwald, and Rosenwein (1970) document the classroom using multiple sources of data. Admirable as these efforts are, none is directed toward interventions for improving teaching practice. The necessary tools have been developed and their use has been mastered, but the quantitative and qualitative approaches are not yet intertwined and applied to the study of improving college teaching.

Implications for Practice

What does this research offer those who design teaching improvement programs? What activities available to them should be supported for maximum impact and cost effectiveness?

Given the mixed quality of research design, no conclusions can be drawn without reservation, yet several generalizations do seem justified.

1. Workshops and seminars are useful instruments for motivating and consciousness raising under certain conditions. Nevertheless, most workshops and seminars, even those with specific training goals, are unlikely to bring about lasting changes in classroom behavior or student impact unless there is provision for faculty to continue practicing the skills in question and to receive critical feedback on their efforts.

2. Concept-based practice appears to be a promising tool, if educationally critical concepts are selected. Discrimination training which is central to concept-based practice is less costly, disruptive, and intimidating than is training-with-practice which is required in experience-based training.

3. End-of-course feedback from students has become institutionalized on many campuses. Little is known about how faculty "process" their feedback, but active processing can be facilitated if the ratings are accompanied by other help, particularly by personal consultation. Those faculty most likely to change are persons whose ratings by students are less positive than their ratings of themselves, and they are probably the faculty in whom the time of consultants should be invested.

4. Grants to support faculty-designed projects require considerable staff time if their impact is to be optimized. Staff involvement in refining proposals and carrying them out is likely to enhance the quality of the work. Staff assistance in evaluating the project provides a data base for making further awards. Otherwise, evaluation is unlikely to be done by the grant recipient alone.

As a general note in conclusion, we observe that the study of these interventions, at least as it is conveyed in research reports, typically fails to engage faculty as collaborators in inquiry. Instead, we make our colleagues the "objects" of our training programs and the "subjects" for our research studies. That situation is lamentable since the questions about teaching and learning which engage this field are as intellectually challenging as any a scholar might find in his or her own field of specialization. For the classroom teacher, such questions also have the attraction of day-to-day relevance. It is our hope that in the next generation research will include fewer studies where faculty are assigned to treatments and more studies which are collaborative attempts to grapple with the phenomenology of teaching and learning. From such inquiry will come fuller understanding of the operations by which effective instruction is carried out and of the impacts it hcs on learning.

References

Abrami, P. C., Leventhal, L., & Perry, R. P. Can feedback from student ratings help improve college teaching? Paper presented at the Fifth 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.

Allen, D. W., & Clark, R. J. Jr. Microteaching: Its rationale. High School Journal, 1967, 51, 75-9.

Andrews, J. D. W. Growth of a teacher. Journal of Higher Education, 1978, 42, 136-150.

Axelrod, J. The university teacher as artist. San Francisco: Jossey-Bass, 1973.

Bandura, A., & Walters, R. H. Social learning and personality development. New York: Holt, Rinehart and Winston, 1963.

Becker, H. S., Geer, B., & Hughes, E. Making the grade: The academic side of college life. New York: Wiley, 1968.

Bergquist, W. H., & Phillips, S. R. (Eds.) Handbook for faculty development, Vol. 1. Washington, D.C.: Council for the Advancement of Small Colleges, 1975.

Bledsoe, J. C. Insight into one's own teaching through feedback from students' evaluation. Psychological Reports, 1975, 37, 1189-1190.

Blumenthal, P. Watching ourselves teaching psychology. Teaching of Psychology, 1978, 5, 162-163.

Bogdanoff, E. Review and evaluation of the California State University and Colleges Minigrant Program for Academic Innovation and Instructional Improvement. Long Beach, CA: Center for Professional Development, California State University and Colleges, 1979.

Borg, W. R. Changing teacher and pupil performance with protocols. Journal of Experimental Education, 1977, 45, 9-18.

Borg, W. R. The minicourse as a vehicle for changing teacher behavior. Paper presented at the American Educational Research Association, Los Angeles, 1969.

Borg, W. R. The minicourse as a vehicle for changing teacher behavior: A three-year follow-up. Journal of Educational Psychology, 1972, 63, 572-579.

Borg, W. R. Protocol materials as related to teacher performance and pupil achievement. Journal of Educational Research, 1975, 69, 23-30.

Borg, W. R., Kelley, M. L., Langer, P., & Gall, M. The minicourse: A microteaching approach to teacher education. London: MacMillan Educational Se.vices, 1970.

Borg, W. R., Langer, P., & Wilson, J. Teacher classroom management skills and pupil behavior. Journal of Experimental Education, 1975, 44, 52-58.

Borg, W. R., & Stone, D. R. Protocol materials as a tool for changing teacher behavior. Journal of Experimental Education, 1974, 43, 34-39.

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.

Brock, S. C. Personal communication, 1976.

Brown, H. H., & Inglis, S. C. Faculty development practices in Ohio colleges and universities. Athens, OH: Ohio University Press, 1978.

Butler, J. R., & Tipton, R. M. Rating stability shown after feedback of prior ratings of instruction. Improving College and University Teaching, 1976, 24, 111-2; 115.

Buttery, T. J., & Michalak, D. A. Modifying questioning behavior via the teaching clinic process. Educational Research Quarterly, 1978, 3, 46-56.

Campbell, D. T. Qualitative knowing in action research. Kurt Lewin Award Address, Society for the Psychological Study of Social Issues. Meeting with the American Psychological Association, New Orleans, 1974.

Campbell, D. T., & Stanley, J. C. Experimental and quasi-experimental designs for research on teaching. In N. L. Gage (Ed.), Handbook of research on teaching. Chicago: Rand-McNally, 1963.

Carroll, J. Assessing the effectiveness of a training program for the university teaching assistants. Teaching of Psychology, 1977, 4, 135-138.

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

Centra, J. A. Faculty development in higher education. Teachers College Record, 1978, 80, 188-201.

Collins, M. L. Effects of enthusiasm training on preservice elementary teachers. Journal of Teacher Education, 1978, 29, 53-57.

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

Cook, T. D., & Flay, B. R. The persistence of experimentally induced attitude change. Advances in Experimental Social Psychology, 1978, 11, 1-57.

Cooper, J. E. A survey of protocol materials evaluation. Journal of Teacher Education, 1975, 26, 69-77.

Costin, F. A graduate course in the teaching of psychology: Description and evaluation. Journal of Teacher Education, 1968, 19, 425-432.

Cottle, T. J. College: Reward and betrayal. Chicago: University of Chicago Press, 1977.

Curtis, M. H. The impact of institutional grants of the National Endowment for the Humanities: An evaluation. Claremont, CA: Claremont University Center, 1978.

Dalgaard, K. A. Some effects of training on teaching effectiveness of untrained university teaching assistants. (Doctoral dissertation, University of Illinois [Urbana-Champaign], 1976). Dissertation Abstracts International, 1977, 37, 6416A. (University Microfilms Order No. 77-8968.)

Daniels, J. W. Effects of interaction analysis upon teaching assistants and student achievement in introductory college mathematics. (Doctoral dissertation, Indiana University, 1970). Dissertation Abstracts International, 1970, 31, 2768A-2769A. (University Microfilms Order No. 70-25, 186.)

Davis, R. H. Special funds for the improvement of instruction. In S. C. Ericksen (Ed.), Support for teaching at major universities. Ann Arbor, MI: Center for Research on Learning and Teaching, University of Michigan, 1979. Pp. 32-41.

Davis, R. H., Abedon, A. J., & Witt, P. W. F. Commitment to excellence: A case study of educational innovation. East Lansing, MI: Michigan State University, 1976.

Deutscher, I., & Beattie, M. Success and failure: Static concepts in a dynamic society. Akron, OH: University of Akron, 1978.

Deutscher, I., & Gold, M. Traditions and rules as obstructions to useful program evaluation. Studies in Symbolic Interaccion, 1979, 2, 107-140.

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

Erickson, G. R., & Sheehan, D. S. An evaluation of a teaching improvement process for university faculty. Paper presented at the American Educational Research Association, San Francisco, 1976.

Festinger, L. A theory of cognitive dissonance. Evanston, IL: Row, Peterson, 1957.

Finger, F. W. Professional problems: Preparation for a career in college teaching. American Psychologist, 1969, 24, 1044-1049.

Fortune, J. C., Cooper, J. M., & Allen, D. W. The Stanford Summer Micro-Teaching Clinic, 1965. Journal of Teacher Education, 1967, 18, 389-393.

Francis, J. B. An evaluation of Change's national teaching project. New Rochelle, NY: Change Magazine Report on Teaching Number 6, 1978. Pp. 1-24.

Friedlander, J. Student perceptions on the effectiveness of midterm feedback to modify college instruction. Journal of Educational Research, 1978, 71, 140-143.

Fuller, F. F., & Manning, B. A. Self-confrontation reviewed: A conceptualization for video playback in teacher education. Review of Educational Research, 1973, 43, 469-528.

Gaff, J. G., & Morstain, B. R. Evaluating the outcomes. New Directions for Higher Education, 1978, 24, 73-83.

Gleissman, D., & Pugh, R. C. The development and evaluation of protocol films of teacher behavior. AV Communication Review, 1976, 24, 21-48.

Gleissman, D., & Pugh, R. C. Acquiring teacher behavior concepts through the use of high-structure and low-structure protocol films. Journal of Educational Psychology, 1978, 70, 779-787. (a)

Gleissman, D., & Pugh, R. C. Research on the rationale, design, and effectiveness of protocol materials. Journal of Teacher Education, 1978, 29, 87-91. (b)

Gleissman, D., Pugh, R. C., & Bielat, B. Acquiring teaching skills through concept-based training. Journal of Educational Research, 1979, 72, 149-154. (a)

Gleissman, D., Pugh, R. C., & Bielat, B. The concept acquisition model in the development of teaching skills. Revision of a paper prepared for the American Educational Research Association, San Francisco, 1979. (b)

Goldman, J. A. Effects of a faculty development workshop upon self-actualization. Education, 1978, 98, 254-258.

Haber, F. B. The effect of instruction in questioning strategies on teaching assistants' classroom performance. (Doctoral dissertation, Arizona State University, 1973). Dissertation Abstracts International, 1973, 34, 2458A-2459A. (University Microfilms Order No. 73-21,888.)

Hargie, O. D. W. The effectiveness of microteaching: A selective review. Educational Review, 1977, 29, 87-96.

Hargie, O. D. W., Dickson, D. A., & Tittmar, H. G. Mini-teaching: An extension of the microteaching format. British Journal of Teacher Education, 1978, 4, 113-118.

Hargie, O. D. W., & Maidment, P. Discrimination training and microteaching: Implications for teaching practice. British Journal of Educational Technology, 1978, 9(2), 87-93.

Hockett, J. C. An examination of changes in teacher behaviors and learner perceptions associated with a program for teaching assistants at the Florida State University: A pilot study. (Doctoral dissertation, The Florida State University, 1972). Dissertation Abstracts International, 1974, 35, 302A-303A. (University Microfilms Order No. 74-15,026.)

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.

Jensen, L. C., & Young, J. I. Effect of televised simulated instruction on subsequent teaching. Journal of Educational Psychology, 1972, 63, 368-373.

Johnson, G. R. Enhancing community/junior college professors' questioning strategies and interaction with students. Community/Junior College Research Quarterly, 1977, 2, 47-54.

Johnson, D. W., & Johnson, R. T. Conflict in the classroom: Controversy and learning. Review of Educational Research, 1979, 49, 51-69.

Kallenbach, W. W., & Gall, M. D. Microteaching versus conventional methods in training elementary intern teachers. Journal of Educational Research, 1969, 63, 136-141.

Kapfer, M. B., & Della-Piana, G. M. Educational technology in the inservice education of university teaching fellows. Educational Technology, 1974, 14(7), 22-28.

Kingston, R. D., & Lacefield, W. Relationships between faculty evaluations and faculty development. Paper presented at the Fifth International Conference on Improving University Teaching, London, 1979.

Kleucker, J. Effects of protocol and training materials. Acquiring teacher competencies: Reports and studies (No. 6). Bloomington, IN: National Center for the Development of Training Materials in Teacher Education, 1974.

Koen, F. M. A faculty educational development program and an evaluation of its evaluation. Journal of Medical Education, 1976, 51, 854-855.

Koffman, M. A study of instructional analysis and feedback of the classroom behavior and student achievement of university teachir ; assistants. Unpublished Doctoral Dissertation, University of Massachusetts, 1974.

Kozma, R. B. Faculty development and the adoption and diffusion of classroom innovations. Journal of Higher Education, 1978, 49, 438-449.

Kremer, L., & Perlberg, A. Training of teachers in strategies that develop independent learning skills in their pupils. British Journal of Teacher Education, 1979, 5, 35-47.

Kulik, J. A., Kulik, C., & Cohen, P. A. A meta-analysis of outcome studies of Keller's Personalized System of Instruction. American Psychologist, 1979, 34, 307-318.

Kulik, J. A., & McKeachie, W. J. The evaluation of teachers in higher education. Review of Research in Education, 1975, 3, 210-240.

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

Mann, R. D., Arnold, S. M., Binder, J., Cytrynbaum, S., Newman, B. M. Ringwald, B., Ringwald, J., & Rosenwein, R. The college classroom: Conflict, change, and learning. New York: Wiley, 1970.

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.

Mayo, G. D. Faculty evaluation of a faculty development center. Research in Higher Education, 1979, 10, 253-262.

McGuire, C., Hurley, R. E., Babbott, D., & Butterworth, J. S. Auscultatory skill: Gain and retention after intensive instruction. Journal of Medical Education, 1964, 39, 120-131.

McKeachie, W. J., & Lin, Y. G. Do discrepancies between student ratings, teacher expectations, and teacher ideals result in changes in teacher behavior? Final Report to the National Institute of Education. (Grant No. NE-G-00-3-0110), March, 1975. (a)

McKeachie, W. J., & Lin, Y. G. Using student ratings to improve teaching. Final Report to the National Institute of Education. (Grant No. NE-G-00-3-0110), March, 1975. (b)

McMillan, J. H. The impact of instructional improvement agencies in higher education. Journal of Higher Education, 1975, 46, 17-23.

Menges, R. J. Raising consciousness about college teaching: Rationale and effects of college classroom vignettes. *Educational Technology*, 1979, *19*(5), 14-18.

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

Miltz, R. J. Application of microteaching for teaching improvement in higher education. *British Journal of Teacher Education*, 1978, *4*, 103-112.

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

Murphy, M. D. The development and assessment of an experimental teacher-training program for beginning graduate assistants in chemistry. (Doctoral dissertation, Ohio State University, 1972). *Dissertation Abstracts International*, 1973, *33*, 4223A. (University Microfilms Order No. 73-2083.)

Oles, H. J., & Lencoski, A. Changes in an instructor's self-rating resulting from feedback from student evaluations. *JSAS Catalogue of Selected Documents in Psychology*, 1973, *3*, 17. MS No. 309.

Overall, J. V., & Marsh, H. W. The relationship between students' evaluations of faculty and instructional improvement. Paper presented at the Third International Conference on Improving University Teaching, Newcastle-upon-Tyne, England, 1977. (ERIC No. ED 138 165.)

Pambookian, H. S. Discrepancy between instructor and student evaluations of instruction: Effect on instruction. *Instructional Science*, 1976, *5*, 63-75.

Pambookian, H. S. 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.

Perlberg, A., Bar-On, E., Levin, R., Bar-Yam, M., Lewy, A., & Etrog, A. Modification of teaching behavior through the combined use of microteaching techniques with the Technion Diagnostic System TDS. *Instructional Science*, 1974, *3*, 177-200.

Perlberg, A., Peri, J. N., Weinreb, M., Nitzan, E., & Shimron, J. Microteaching and videotape recordings: A new approach to improving teaching. *Journal of Medical Education*, 1972, *47*, 43-50.

Perrott, E., Applebee, A. N., Heap, B., & Watson, E. P. Changes in teaching behavior after completing a self-instructional microteaching course. *Programmed Learning and Educational Technology*, 1975, *12*, 348-362.

Perry, R. P., Leventhal, L., & Abrami, P. C. An observational learning procedure for improving university instruction. Paper presented at the Fifth Annual Conference on Improving University Teaching, London, 1979.

Rhyne, P. J. A training program to improve teaching methods and student-teacher interaction of graduate biology teaching assistants. (Doctoral dissertation, Georgia State University, 1973). Dissertation Abstracts International, 1974, 34, 4971A. (University Microfilms Order No. 74-2893.)

Rose, C. An in-service program for teaching assistants. Improving College and University Teaching, 1972, 20, 100-102.

Rotem, A. The effects of feedback from students to university instructors: An experimental study. Research in Higher Education, 1978, 9, 303-318.

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

Sherman, T. M. The effects of student formative evaluation of instruction on teacher behavior. Journal of Educational Technology Systems, 1977-78, 6, 209-217.

Smith, A. B. A model program for training teaching assistants. Improving College and University Teaching, 1974, 22, 198-200.

Smith, B. O. Teachers for the real world. Washington, DC: American Association of Colleges for Teacher Education, 1969.

Sweeney, J. M., & Grasha, A. F. Improving teaching through faculty development triads. Educational Technology, 1979, 19(2), 54-57.

Tjosvold, D., & Johnson, D. W. Effects of controversy on cognitive perspective taking. Journal of Educational Psychology, 1977, 69, 679-685.

Tubb, G. W. Heuristic questioning and problem solving strategies in mathematics graduate teaching assistants and their students. (Doctoral dissertation, Texas A & M University, 1974). Dissertation Abstracts International, 1975, 36, 235A-236A. (University Microfilms Order No. 75-15,077.)

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

Turney, C., Clift, J. C., Dunkin, M. J., & Traill, R. D. Microteaching: Research, theory and practice. Sydney, Australia: Sydney University Press, 1973.

Vogt, K. E., & Lasher, H. Does student evaluation stimulate improved teaching? Bowling Green, OH: Bowling Green State University, 1973. (ERIC ED 078 748.)

Wagner, A. C. Changing teaching behavior: A comparison of microteaching and cognitive discrimination training. Journal of Educational Psychology, 1973, 64, 299-305.

Weerts, R. R. Student Perceptions of Teaching (SPOT): VIII. The use of feedback from student ratings for improving college teaching. Research Report No. 6, Evaluations Examination Service, University of Iowa, Iowa City, Iowa. March, 1978.

Yaghlian, N. University teaching: The impact of an inservice program for teaching fellows in chemistry. (Doctoral dissertation, The University of Michigan, 1972). Dissertation Abstracts International, 1973, 33, 6192A. (University Microfilms Order No. 73-11,302.)

Appendix A: Summary of Studies Critically Reviewed

This Appendix contains schematic outlines of the studies analyzed in the text. Criteria for including studies are described in Chapter I. Detailed discussion of several categories of these charts are also given in that chapter. Symbols frequently appearing in the charts are defined below:

E - experimental group

C - control group

R - randomization

O - observation

X - intervention or treatment

(X) - alternate intervention

---- groups not randomly formed

┆ pre and post data from different persons

? - the information in question was not reported or was ambiguous in the source available to us.

Threats to validity, general categories:

SC - statistical conclusion validity

I - internal validity

C - construct validity

E - external validity

Lower case letters denoting particular threats within the categories are defined in Appendix B.

GRANTS

thor/Date	Purpose	Components of Design Code	Participants	Duration	Instrumentation	Stated Results	Threats to Validity SC	I	C	E	Weaknesses	Strengths	Confidence Rating
ma 975)	1) To assess the effects of a faculty development project on faculty use of innovations 2) To investigate the diffusion of knowledge about innovations	Study 1: E_1: 0 X_1 0 (n=20 fellows) E_2: 0 X_2 0 (n=25 IDF awardees) C_1: 0 0 (n=8 applicants) C_2: 0 0 (n=13 chairmen) C_3: 0 0 (n=137 other faculty) X_1=Faculty development project X_2=Instructional development awards (IDF)	193 faculty at University of Michigan	2 years	Questionnaire on use of instructional techniques	Study 1: Significant differences between groups due to greater use of innovations by faculty fellows, fellowship applicants and IDF recipients. Significant interaction due to greater increase in use of techniques by faculty fellows & IDF recipients.		f g j		a b	1) Problem of self-reported data (noted by investigator) 2) Crude measure of innovation used (noted by investigator)	Discussion of limitations	Fair
		Study 2: E: X 0 C_1: 0 C_2: 0 X=Faculty development project	faculty fellows, peer contacts, & peer contacts who also happened to be in survey sample	8 months	1) Questionnaire on functions of peer contacts 2) Peer contact logs	Study 2: Fellows did contact other faculty & discuss instructional matters. Discussions served to increase awareness & to influence adoption decision of fellows. Few adoption decisions made by other faculty as a result of those contacts.							

thor/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidenc Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
roll 77)	To assess the effects of a seminar on the teaching of psychology on teaching performance of teaching assistants & on student ratings	E: R X 0 C: R (X) 0 X='Teaching of Psychology' seminar (X)=control version of seminar involving less assignments, unstructured group meetings, viewing of videotape alone & without critique (participation in conditions was mandatory)	19 novice teaching assistants teaching introductory psychology course at Cornell University (705 students involved)	1 semester	1) Rating of videotaped behaviors using Flanders Interaction Analysis 2) Rating of cognitive levels of classroom & quiz questions using a modified version of Teacher-Pupil Question Inventory (Davis & Tinsley) 3) Cornell Inventory for Student Appraisal of Teaching & Courses (Moos)	1) E group made significantly greater use of objectives (p<.07) & engaged in significantly more student-centered teaching (p<.06) than C group. 2) E group student ratings of effectiveness were significantly higher than for C group (p<.10). 3) No differences between groups on student-talk ratios. 4) For both groups, use of indirect teaching skills was positively correlated with student ratings of instructional effectiveness. 5) No differences between groups on congruity among cognitive levels of classroom & quiz questions. No relationship shown between congruity variable & student ratings of instruction. (Due to small sample size, significance level of p<.10 was used)	a	j		a b c	Small N	Randomization	Fair
ecia 963)	To assess the effects of a seminar on the teaching of psychology on teaching performance of teaching assistants	Survey: X 0 X=seminar in teaching of psychology	65 former participants of seminar	opinions obtained over a 3 year period	Survey of opinions about seminar	Topics rated most important were related to practical everyday work of a college teacher.							
		Study 1: E: 0 X 0 C: 0 0 X=seminar in teaching of psychology	49 teaching assistants of psychology	1 semester	Student rating items developed by Isaacson, McKeachie, & Milholland (1964)	E group made a significant gain on 1 of 5 factors, that of rapport, although magnitude of difference was small.		g		a b a			Fair
		Study 2: E: 0 X 0 0 (n=21) C: 0 0 0 (n=11) X=seminar in teaching of psychology	32 teaching assistants in psychology	2 semesters	Student rating form developed by Isaacson, McKeachie, & Milholland (1964)	1) After 1 semester, no significant differences between E & C groups. Group interaction factor approached .05 significance level in favor of E group. 2) After 2 semesters, E group received significantly higher mean ratings on 2 of 5 factors, feedback & interaction.		g		a b			

WORKSHOPS/SEMINARS

thor/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
lgaard 976)	To assess the effects of a seminar for teaching assistance on teaching performance	E: R O X O C: R O O X=six 2-hour seminar sessions on instructional organization, techniques & materials & videotaping with feedback	22 inexperienced TAs in economics, business administration & geography at University of Illinois	1 term	1) Illinois Course Evaluation Questionnaire 2) Instructor Self-Evaluation Form 3) Teacher Performance Appraisal Scale 4) Informal questionnaire for TA evaluation of training seminar	1) ANCOVA used to adjust for initial differences between E & C on expert ratings. After adjustment, E group higher on expert ratings. 2) No significant differences between E & C on student ratings; no training or teaching experience affect on self-evaluation profiles; no teaching experience affect on expert ratings. 3) TA's ratings of training seminars were favorable.	a ?	?	?	a b c	Small N	1) Multiple measures 2) Randomization	Fair (tentative rating based on abstract)
niels 970)	To assess effects of instructing teaching assistants in Flanders interaction analysis on classroom verbal behavior & on student achievement	E: R O X O X O X O C: R O O O O X=training in interaction analysis	8 teaching assistants of the mathematics department at East Carolina University (211 students)	1 quarter	Audiotapes analyzed using Flanders Interaction Analysis (FIA) TA's were divided into: a) mathematics education TA's b) mathematics TA's	1) Significant differences in favor of E on 4 of 9 verbal behavior characteristics (I/D ratio, Steady-State cells, Area A cells, & Teacher Response to Student cells). 2) Significant differences in favor of the mathematics education TA's on 6 of 9 verbal behavior characteristics (I/D ratio, S/T ratio, Steady-State cells, Content Cross cells, Teacher Response to Student cells, & Student Talk Followed by Teacher Talk cells). 3) Significant differences in favor of mathematics education TA's on student achievement.	?		c	a b c	1) Unit of analysis=students?		Fair (tentative rating based on abstract)
Coldman (1978)	To assess the effects of a faculty development workshop on participants' level of self-actualization	E: O X O - - - - - C: O O X=6-day faculty development workshop consisting of discussions & series of micro-colleges	22 college professors (participants in 2 groups equated on age & academic division)	approximately 1 week	Personal Orientation Inventory (POI) (Shostrom)	1) E group made significant increases on 6 of 12 scales (Inner Directedness, Self-Actualizing Values, Existentiality, Feeling Reactivity, Acceptance of Aggression, Capacity for Intimate Contact) while no significant changes fur C group. 2) Pretests did not indicate significant differences between the groups.	a	g j	a e	a b c	1) Nonequivalent control group design 2) Small N	1) Use of self-actualization as dependent variable 2) Theoretical framework for faculty development	Low

thor/Date	Purpose	Components of Design					Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*			*SC*	*I*	*C*	*E*			
ber 973)	To assess the differential effects of instruction in effective questioning, & student rating feedback on teaching performance of teaching assistants	E_1: R O X_1 O E_2: R O X_2 O C: R O O X_1=student rating feedback X_2=student rating feedback & instruction in effective questioning techniques (FIA)	12 graduate teaching assistants randomly selected from the College of Business Administration at Arizona State University	?	1) Flanders System of Interaction Analysis (FIA) 2) Purdue Instructor Performance Indicator (PIPI) 3) Minnesota Teacher Attitude Inventory (MTAI)		1) No significant differences among E_1, E_2 & C in teaching performance as measured through indices derived from TA's classroom behavior matrix (FIA). 2) No significant relationship found between TA's teaching performance as measured by FIA, & PIPI. 3) Significant relationship between TA's MTAI attitude scores & 2 of 5 teaching performance FIA indices (Direct/Indirect influence & Teacher/Student talk ratios). Two other ratios suggested a strong association with MTAI.	a			b c	Small N	1) Randomization 2) Multiple measures	Low (tentative rating based on abstract)
ckett 1972)	To assess the effects of a TA training program on teaching performance	O? X O X=training program on a wide range of topics (writing behavioral objectives, art of questioning, personal interaction, sensitivity)	Teaching assistants in both geology & chemistry departments at Florida State University	?	?		1) TA training caused significant changes in teaching behavior including less teacher control, more individual interaction & more high-level questioning. 2) Use of desired teaching behaviors resulted in positive student attitudes toward class, TA as instructor, science in general & increased self-learning.	?	a b e	?	a b a	No control group		Low
oyt & oward (1978)	The Wichita State Study: to determine the effectiveness of faculty development programs built on model of 'teachers helping teachers'	E: R O X O C: R O O X=faculty development activities conducted in group sessions or dyads	32 randomly selected instructors from 82 volunteers	8-10 weeks	13 item student rating form		ANCOVA-adjusted measures significantly higher for E on 5 of 14 measures (total, overall rating as a teacher, discussed opinions & ideas other than own, encouraged class discussion, was aware if students understood subject matter). Given lack of control over other factors that might influence performance (e.g.,short intervention period, small N), results offer strong support for this faculty development procedure.	d	j		a b c	Volunteer sample	1) Motivation was controlled 2) Randomization	Fair

WORKSHOPS/SEMINARS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
Kingston & Lacefield (1979)	To assess the effects of a sequence of faculty development workshops on teaching performance	0 X 0 X=Teaching Improvement Project System (TIPS)	29 instructors of the University of North Dakota College of Nursing	approximately 2 semesters	Faculty Enrichment & Assessment of Teaching evaluation system (FEAT) (developed at tho University of Kentucky)	1) Significant gains for 14 of 26 items. The most significant gains wero associated with items with factual content. 2) Multivariate analysis found significant differences among gain scores on 4 global variables. Univariate tests found significant gains for organization, presentation, & evaluation.	f	a b a	c	a b c	No control for course content	Multivariate analysis for global variables	Low
Koffman (1974)	To assess the effects of instructional analysis & feedback from an instructional specialist on the classroom behavior & student achievement of university teaching assistants	E_1: 0 X_1 0 E2: 0 X_2 0 C: 0 0 X_1=review of data, & remedial suggestions & activities with instructional specialist over 8 week period X2=review of data with instructional specialist	13 graduate student teaching assistants teaching a required freshman rhetoric course at University of Massachusetts	8 weeks	1) Videotapes analyzed by Flanders Interaction Analysis (Amidon & Flanders) 2) 31 item student evaluation form (SCAT) of Clinic to Improve University Teaching 3) Student achievement test (parallel forms)	1) Among trends noted in data: E_1 & E2 instructors increased their using student ideas, focusing, summarizing, introducing or orienting statements & lecturing. Percentage of teacher talk in class increased & student talk decreased. C instructors showed an increase in silence in their posttest lessons. Their using student ideas increased slightly & there was a decrease in focusing, summarizing, introducing or orienting statements & lecturing. 2) Among trends in student evaluations: E_1 showed positive change in clarity, evaluation & feedback & relating to student responses & C improved in relating to student responses. 3) No differences in achievement among 3 groups.	a	g h		a b c	Small N (noted by investigator)	1) Use of multiple measures 2) Discussion of limitations	Fair
Lewis & Orvig (1973)	To assess the effects of a seminar on the teaching of economics on teaching performance & student performance	C: 0 0 (n=323) E: 0 X 0 (n=348) X=seminar on teaching economics consisting of student evaluation input, videotaped observations, & instructional seminars	761 students enrolled in principles of economics course (same 7 graduate instructors involved over 2 quarters)	2 quarters	1) Questionnaires dealing with student characteristics 2) Test of Understanding in College Economics (Part I, Forms A & B) (TUCE) 3) Postcourse use of Purdue Rating Scale for College Instructors	1) Student performance of E group increased significantly over C group. 2) Instructor ratings of E group were significantly higher than C group. 3) High association between instructor ratings & student performance on TUCE.		c g h	d	a b c			Fair

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Murphy (1972)	To assess the effects of a training program for teaching assistants on verbal interaction & questioning	E: R O X O O C: R O O O X=seminars, microteaching, observation & conferences	New teaching assistants in freshman chemistry (number not specified in abstract)	1 term	1) Audiotapes coded according to: a) Flanders Interaction Analysis Category System b) Question Category System for Science 2) Placement tests in 4 fields of chemical knowledge	1) E group more successful in drawing students into discussion. 2) E group lectured less & used more praise & encouragement. 3) E group asked more questions. 4) Training program showed no effect on type of question asked or on proportion of correct responses elicited.	?	?		a b c		1) Multiple measures 2) Randomization	Fair (tentative rating based on abstract)
Rhyne (1973)	To assess the effects of a training program for teaching assistants on teaching performance & student-teacher interaction	O X O X=ten 1-hour seminars based on rationale of Interaction Analysis for Science Teaching	12 teaching assistants in biology at Georgia State University	?	1) Teacher & student behaviors coded using Interaction Analysis for Science Teaching (IAST) 2) Nonverbal movement of TA's was recorded 3) Questions asked by TA's were analyzed for number & level 4) Rokeach Dogmatism Scale 5) Role Conflict Test 6) Teacher Concern Statements	1) Significant changes in the following IAST ratios: I/D teaching ratio, S/T talk ratio, revised I/D teaching ratio. 2) Significant change in teacher behavior block, an interaction region on the IAST matrix,but no changes in 3 other blocks. 3) Significant change in nonverbal movement of TA. 4) TA's increased amount of time spent with students. 5) Significant changes in TA's total number of questions & number of convergent & divergent questions, but no change in managerial & rhetorical questions asked. 6) No significant changes or correlations for other scales & measures.	a	a b c e	d	a b	Small N	Multiple measures	Low

WORKSHOPS/SEMINARS

uthor/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Tubb (1974)	To assess the differential affects of training teaching assistants in Polya's hearistic questioning strategies and/or Flanders Interaction Analysis on verbal interaction, problem solving sequence, achievement & evaluative perception of TAs by students	?	243 undergraduate students in introductory calculus course for majors other than math or engineering (18 calculus sections involved)	?	1) Flanders Interaction Analysis (FIA) 2) Polya's hearistic questioning strategies (PHQPS) 3) Donkey-mule problem 4) Achievement test covaried with course grade, CEEB scores, & Nelson-Denny vocabulary scores	Overall, FIA & PHQPS training of TA's significantly affected verbal interaction of TA's with their students, the problem solving sequence of the TA's & their students, the achievement of TA's students, & the evaluative perception of TA's by their students.	?	?	?	a b		Multiple measures	Fair (tentative rating based on abstract)
aghlise (1972)	To assess the effects of an in-service program for teaching fellows on attitude toward teaching as a career, job satisfaction, interpersonal style of teaching, & student satisfaction with teaching fellow	? E: X O X O ? ? C: O ? X=in-service program involving workshops and consultation	E=15 teaching fellows in chemistry department at University of Michigan C=number not reported (498 students)	2 terms	?	1) Students of E group more satisfied than students of C group at end of fall term. Winter term students of E group more satisfied than fall term students of E group. 2) Change in attitude toward teaching seems related to reconsideration on part of teaching fellows of relative advantages & disadvantages of teaching. 3) Change in job satisfaction seems to be related to level of ambivalence toward teaching. 4) Change in self description seems to be related to certain perceptions of potential for an interpersonal style.	?	g? h? i?	?	a b c			Low (tentative rating based on abstract)

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
Fortune, Cooper & Allen (1967)	To assess the effects of the Stanford Summer Microteaching Clinic (1965) on teaching performance	X=microteaching clinic O X O X O X O X O X O X O X O X O X O X O X O X O X O X O X O X O X O X	140 secondary education teacher interns	6 weeks	1) Stanford Teacher Competence Appraisal Guide (STCAG) 2) Questionnaire to evaluate student acceptance of microteaching	1) Trainees showed significant mean gain over 6 week session on 9 of first 12 STCAG items. 2) 70% of trainees indicated supervisory feedback was useful while 24% indicated pupil feedback was useful.		b c		b	1) No control group 2) Possible testing effects	Replitation study of 1963 & 1964 clinics	Low
Jensen & Young (1972)	To assess the effects of microteaching training on subsequent teaching performance	E: R X O O O C: R (X) O O O X=microteaching training (X)=conventional student teaching practice	37 subjects selected from a teacher training program	3 sessions of microteaching & 8 weeks in assigned classroom	Teacher Performance Evaluation Scale	E group received higher ratings on 5 of 6 factors (personality traits, teacher warmth, general classroom atmosphere, lesson usefulness, teacher interest in pupils) than C pupils. Microteaching is beneficial although superiority sometimes not evident until third observation.		e	d	a b c			High

Author/Date	Purpose	Components of Design					Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*			*SC*	*I*	*C*	*E*			
Johnson (1977)	To assess the effects of combined training in Flanders Interaction Analysis & microteaching labs on instructor interaction behavior, questioning & reinforcement techniques	O X O X=combined training in Flanders Interaction Analysis (FIA) & in microteaching	14 community/junior college professors	one summer	Analysis of videotapes using FIA & author's descriptions of questioning & reinforcement techniques	Trainees improved significantly on all 8 variables: a) Significant gains were shown for pupil talk, teacher question ratio, direct or indirect influence reinforcement, probing questions & higher order questions. b) Significant reduction shown for teacher talk.		a b a	d	a b c	1) Volunteer sample 2) Small N		Low
Kallenbach & Gall (1969)	To assess the effects of microteaching training on subsequent teaching performance	E: R O X O O O (n=19) C: R O (X) O O O (n=18) X=microteaching training (X)=conventional student teaching practice	37 students selected by education department to begin elementary teacher training program in summer 1966 (San Jose State College)	approximately 1 year	1) Stanford Teacher Competence Appraisal Guide (STCAG) 2) Instrument for the Observation of Teaching Activities (IOTA)	1) No differences between E & C on post-training ratings. 2) The two groups did differ on pretest measures so ANCOVA was carried out but no significant differences were found.		a	d	a b	Loss of some videotapes (problem noted by investigator)	Good discussion	High
Kremer & Perlberg (1979)	To assess the effects of microteaching training in independent learning teaching strategies on teaching & pupil performance	E: R? O X O C: R? O O X=Workshop involving demonstration, discussion, peer teaching & microteaching	22 elementary inservice teachers 448 pupils of 8-12 years of age	1 school year	Analysis of videotapes: 1) Teaching style measured by behavior counts using Verbal Inventory Category System (Amidon & Hunter, 1967) 2) Fluency of pupils' questions measured by counting their number 3) Level of pupils' questions & problems analyzed using categories suggested in Bloom's taxonomy (1976)	1) E teachers talked less, gave less information, asked broader questions & gave more directions than C teachers. 2) E pupils showed significant behavior changes compared to C group for 3 of 4 variables, that of responds to teacher, initiates talk to teacher & initiates talk to another pupil. 3) E pupils showed significant increases in number of problems & questions voiced, but significant differences in higher level questions for E pupils only found for 2 of 7 variables, divergency & analysis.		g h		a b c		1) Good literature review 2) Theoretical framework 3) Inclusion of qualitative data	Fair

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Perlberg, Bar-On, Levin, Bar-Yam, Levy & Ftrog (1974)	1) To assess the effects of microteaching training combined with Technion Diagnostic System (TDS) computerized feedback on teaching performance 2) To investigate linear relationships between the student's performance in different lessons 3) To assess differential affects of treatment on experimental subgroups	0 X 0 X=microteaching training combined with Technion Diagnostic System computerized feedback	60 students in Teacher Training program at Technion Institute enrolled in 'Principles of Teaching Methods' course	approximately 2 semesters	ratings of videotaped lessons on 13 categories	1) Trainees showed significant changes on all 4 combined scores (non-verbal, not lecturing, relates to, analytical thinking). 2) Increases in first three combined scores reached peak at end of training & posttest showed a decrease from the last training session. 3) No linear relationship found between pre and posttest scores. 4) Treatment effective both for those with low entry behavior & those with some teaching experience. Low entry participants gained more from treatment.		a b	d	a b	No control group	Use of Technion Diagnostic System computerized feedback	Low
Perlberg, Peri, Weinreb, Nitzer, & Shimron (1972)	1) To assess the effects of microteaching training in student-centered & classroom interaction styles on teaching performance 2) To investigate the relationship between changes effected by microteaching & a participant's openness	0 X 0 X=microteaching training	16 faculty members, 30-60 years old	2 sequences of 5 weeks each (each faculty member went once a week)	1) Rokeach's dogmatism scale 2) P-A (permissive-authoritarian) scale 3) Bipolar adjective scale (based on Osgood Semantic Differential) 4) Flanders Interaction Analysis used to analyze pre & post videotapes	1) Trainees showed significant improvement on all 7 teaching skills (lesson organization, lecture style, providing examples, fluency in questions, probing questions, higher-order questions & divergent questions). 2) Differences for questioning skills were greater than for lecturing skills. 3) Trainees showed substantial increase in use of all questioning skills, the increase in high order & divergent questioning being the greatest. 4) Perservarance in microteaching clinic found to be best predictor of openness to change & willingness to accept innovation.		a b c e f	d	a b c	1) No control group 2) Subject mortality	Microteaching applied to higher education	Low

MICROTEACHING

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
Perry, Leventhal & Abrami (1979)	To assess the effects of the Modified Observational Learning (MOL) procedure on teaching performance of instructors who differ in pretraining teaching ability	E: R O ⁞ X O C: R O ⁞ O X=Modified Observational Learning (MOL) involving videotape feedback & cognitive discrimination training	187 introductory psychology students at University of Manitoba (4 instructors involved)	?	questionnaire consisting of: 1) 2 single item student rating measures 2) achievement subscales: a) student competence in chemistry b) content covered in lecture material	1) For high effective lecturers, MOL training produced more favorable student ratings on teaching ability, on lecture value, & produced greater student achievement than no training. 2) For low effective lecturers, MOL training produced less favorable ratings than no training on lecture value & no differences on teaching ability & student achievement.	d	g		a b c	1) Unit of analysis=students 2) Separate pretest-posttest samples 3) Novice instructors	1) Microteaching applied to higher education 2) Multiple measures	Fair
Wagner (1973)	1) To assess the relative effects of cognitive discrimination training, microteaching & a control condition on student-centered teaching performance 2) To assess the relative effects of cognitive discrimination training, microteaching & a control condition on teachers' ability to discriminate classes of teaching behavior	E_1: R X_1 O O E_2: R X_2 O O C: R (X) O O X_1=cognitive discrimination training X_2=microteaching training (X)=conventional student teaching practice	78 undergraduates from 5 sections of introductory educational psychology course	approximately 2 weeks	1) Observer ratings of teacher responses to student comment 2) Discrimination test consisting of coding teacher responses to students' comments	1) E_1 group was significantly more student-centered (ask for clarification, restate, use of student's idea) than E_2 or C groups. 2) E_2 group not significantly more student-centered than C group. 3) E_1 group better able to discriminate teaching behaviors than C. E_2 did not differ from E_1 & C on discrimination test.	d	j	d	a b e	1) Time lag between observations (noted by investigator) 2) Short duration 3) Discrimination test precluded assessment of whether subjects had learned to attend to relevant dimension	Good discussion	High

MINICOURSES

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Borg (1972)	To investigate the persistence of behavior change of teachers completing Minicourse 1	0 X 0 0 X=Minicourse 1: Effective Questioning	24 of 48 elementary teachers who participated in initial experiment	approximately 3 years	Scoring of videotape transcripts on use of Minicourse 1 skills	<u>Initial Experiment (N=48)</u> 11 of 13 teacher & student behaviors showed significant improvement <u>4 Months Later (N=38)</u> a) 3 of 11 measured skills continued to improve b) No significant regression on any skill <u>39 Months Later (N=24)</u> a) 8 of 10 measured behaviors still significantly superior compared to precourse means b) Teacher talk regressed significantly but still below initial frequency c) 1-word pupil response frequency was up significantly & higher than pre-course mean		b e f	d	a b	1) Volunteer sample	Discussion of limitations	Fair
Buttery & Michalak (1973)	To assess two modifications to the minicourse format: a) use of "Teaching Clinic" process feedback system (no videotape equipment used) b) naturalistic setting	E:R?: 0 X_1 0 C:R?: 0 (X) 0 X_1=Minicourse 1: Effective Questioning coupled with "Teaching Clinic" feedback (X)=conventional student teaching practice	40 undergraduate elementary school majors at University of Georgia	approximately 8 weeks	Coding of audio cassette tape recordings for 13 teaching behaviors relevant to Minicourse 1	11 of 13 't' ratios significant at .05 level for E group (5 significant at .01 level) while 2 of 13 't' ratios significant at .05 level for C group.	e	c d h	d	a b c			Fair
Collins (1978)	To assess the effects of an enthusiasm minicourse on subsequent teaching performance	E: R 0 X 0 0 C: R 0 0 0 X=minicourse on enthusiasm	20 preservice elementary teachers (participants not aware of experiment)	8 weeks	1) Rating form to assess 8 variables in terms of level of performance 2) Tally sheet to record frequencies of 8 variables	1) E group showed a significant increase in teacher enthusiasm between pretest & posttest I, & posttest I & II. 2) No differences in the C group in teacher enthusiasm among 3 testing periods.				a b c		1) Observers blind to experiment 2) Randomization 3) Repeated measures ANOVA	High

MINICOURSES

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Perrott, Applebee, Heap. & Watson (1975)	1) To assess the effects of Minicourse 1 on teaching performance in Great Britain 2) To investigate the international transfer of Minicourse 1 to Great Britain	E: R X_1 O X_2 O O C: R O X_1=informing of target behaviors at pretest X_2=Minicourse 1: Effective Questioning	28 inservice junior & secondary school teachers	?	1) Scoring of videotapes on 14 aspects of teaching behavior related to Minicourse 1 2) Questionnaire on teacher's perceptions of course effects	1) No significant difference was found between E & C on effects of knowledge of target skills on pre-course performance. 2) Multivariate effect for time was highly significant while multivariate effect for centre X time interactions was not significant. 3) For planned contrasts, 8 of 14 measures showed significant differences between pre-course & both post-course sessions. 4) Findings suggest that familiarity with videotaping at posttest may be a cause of differences between pre- & post-course performance.			d	e		1) Well-planned multivariate analyses 2) Replication of Borg (1970) 3) Presents evidence for mixed findings of stability of teaching performance over time 4) Randomization	High

RATINGS

Author/Date	Purpose	Components of Design *Code*	*Participants*	*Duration*	*Instrumentation*	Stated Results	Threats to Validity *SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Aleamoni (1978)	To assess the combined effects of student rating feedback and consultation on faculty performance from one semester to the next semester in which the course is taught	E: 0 X 0 C: 0 0 X = student rating feedback & consultation (involved problem identification & suggestions for resolution)	E=20 instructors teaching 24 courses C=13 instructors teaching 18 courses (3358 students involved, & course was the unit of analysis)	1 year or 1½ years depending on when course was taught	Illinois Course Evaluation Questionnaire (CEQ)	E significantly improved on 2 (Course Content and Instructor) of 5 dimensions		d g j	c	a b	1) All subjects wanted treatment (reacntful demoralization) 2) Repeated measures ANOVA analysis may have been inappropriate		Low
Bledsoe (1975)	1) To assess the effects of mid-term student rat-feedback on end-of-term faculty performance 2) To compare instructor self-ratings with class ratings at mid-term and end-of-term	0 X 0 X=student and instructor rating feedback and student-instructor dialogue concerning ratings	1 instructor and 31 advanced graduate students at University of Georgia	1 quarter	26 item standard evaluation Faculty-Course Evaluation Form (used at University of Georgia)	1) Instructor received significantly higher end-of-term class evaluations as a result of mid-term class feedback and dialogue, but instructor decreased his self-evaluation. 2) Correlation for class means on items correlated .93 on 2 occasions. Mid-term self-ratings correlated .60 & .65 with class evaluations. 3) Greatest gains made on items rated lowest at mid-term.		a b a	a c a	a b c	Only 1 instructor involved and he was also the experimenter		Low
Braunstein, Klein & Pachla (1973)	1) To assess the effects of mid-term student rating feedback on end-of-term faculty performance 2) To explore the effects of discrepancies between mid-term faculty self-ratings and student ratings on end-of-term ratings	E: R 0 X 0 C: R 0 0 X=student rating feedback	At Oakland University in Detroit: E=15 classes (10 different professors) C=12 classes (9 different professors)	1 semester	23 item teaching evaluation instrument	1) Change score analysis was used due to nonequivalence of groups as indicated on mid-term ratings. 2) E showed a strong increase in positive changes while C showed strong increases in negative changes. 3) When an instructor's expectancy is discrepant from students' ratings for a trait, a subsequent shift for that trait is likely.		f	c	a b c		1) Theoretical framework for experiment 2) Thorough discussion 3) Randomization	Fair

RATINGS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Butler & Tipton (1976)	To assess the effects of mid-term student rating feedback on end-of-term instructor performance (and to investigate the reliability of student ratings over time)	O O X O X=student rating feedback	17 instructors from English Department at Virginia Commonwealth University (1000 students involved)	approximately 5 months	rating scale consisting of 13 items	6 of 17 instructors showed significant improvement on post-ratings (student ratings averaged across 13 items)		e e	c	a b	Volunteer sample		Low
Centra (1973)	1) To assess the effects of mid-semester student rating feedback on subsequent faculty performance across several types of post-secondary institutions 2) To assess the effects of student-instructor rating discrepancies at mid-term on end-of-term faculty performance	E: R O X O O (n=8) C_1: R O O C_2: R O O (n=13) - - - - - - - C_3: O (n=30) X=student rating feedback	Instructors from 5 institutions Mid-semester: 505 college instructors End-of-semester: 436 college instructors Spring semester: 51 college instructors	2 semesters	23 item Student Instructional Report (SIR) Based on pre-test ratings, instructors divided into: a) more favorably rated b) less favorably rated	1) E did not differ from C_1 & C_2 on end-of-semester ratings (sex, subject area, college & teaching experience were controlled). 2) 5 of 17 items showed significant improvement in favor of less favorably discrepant group over favorably discrepant group and 13 of 17 items indicated a similar trend. 3) In terms of changes over time, E received better ratings than C_2 & C_3.	e f		c	e		1) Excellent discussion ruling out plausible hypotheses 2) Randomization	High

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				RATINGS		
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Erickson & Erickson (1979)	Study 1: To assess the combined effects of mid-term student rating feedback and consultation on end-of-term faculty performance	Investigators claim these studies are quasi-experimental E: R O X O C: R O O X=student rating feedback and consultation (including interview, observation, and videotaping)	31 faculty of University of Rhode Island	1 semester	1) early semester Teaching Analysis by Students (TABS): Short form A, Part 1 2) late semester TABS: Short form A, Part 1 3) Two 15 item questionnaires on effectiveness of consultation procedure	Study 1: The E group late semester faculty and student ratings on all 3 components (Stimulation, Organization, Evaluation) were more positive than C group. E instructors indicated a positive attitude toward the procedure.			a	a b c	1) Of 700 invited, only 31 agreed to participate 2) Volunteer sample 3) Since student raters were told of study, results may reflect differing expectancies of change (noted by investigator)	Randomization	High
	Study 2: To check whether the results of Study 1 just reflect differing group expectations of change	O X O ↑ (This observation used same data as first observation of Study 1) X=student rating feedback and consultation (including interview, observation, and videotaping)	20 faculty from Study 1 who agreed to participate (14 of Study 1 were on leave)	1-4 semesters (depending on when a similar course to that of Study 1 was scheduled again)	Same as for Study 1	Study 2: Differences between semester I & II significant for 11 of the instructors.		a		a b c	Small N	Investigator's check for expectancy effects	
Erickson & Sheehan (1976)	1) To assess the relative effects on end-of-term faculty performance of a) student rating feedback with consultation b) student rating feedback alone, & c) no feedback 2) To assess satisfaction with teaching improvement process 3) To investigate faculty & student attitudes toward selves, courses, & teaching	E_1: R O X_1 O (n=13) E_2: R O X_2 O (n=13) C: R O O (n=14) X_1= full process (rating feedback & consultation, including interview, observation & videotaping) X_2=diagnostic (rating feedback only)	40 fac from [illegible] academic departments	approximately 6 weeks	1) Teaching Analysis by Students (TABS) 2) Instructor Questionnaire 3) Student Questionnaire 4) Evaluation of Teaching Clinic (Part I) (all instruments designed by clinic)	1) No significant differences among groups 2) E_1 faculty were satisfied with teaching improvement process				a b c	1) Volunteer sample 2) No investigation of whether teaching skills amenable to change would affect student learning (noted by investigator)	1) Discussion of limitations 2) Randomization	High

		Components of Design					RATINGS						
							Threats to Validity						
Author/Date	Purpose	*Code*	*Participants*	*Duration*	*Instrumentation*	Stated Results	*SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Friedlander (1978)	To examine student perceptions of instructor change as a result of: a) student rating feedback to instructors b) instructor's discussion with class about feedback	O X O O X=student rating feedback & student-teacher discussion	2,014 graduate students in 85 courses, UCLA Graduate School of Management	1 quarter	1) Mid-Quarter Course Evaluation (MQCE) 2) End-of-quarter rating form 3) Experimental Questionnaire	A greater percentage of students who reported a meaningful & helpful discussion of the MQCE attributed change in their course to the MQCE (77%), than students who reported an inadequate discussion (50%), or no discussion of the MQCE although such discussion was needed (13.6%)	d	a c	a	a b c	Optional for teachers to give out forms & for students to respond		Low
Hoyt & Howard (1978)	The Kansas State Studies -4 studies/surveys to evaluate effectiveness of Faculty Development Office activities												
	Survey 1: To investigate outcomes of contact with Faculty Development Office	X O X=contact with Faculty Development Office	381 faculty	approximately 1 year	User Satisfaction Faculty Survey	Respondents indicated satisfaction with most aspects of services. While substantial numbers became involved at a superficial level (54% tried a new approach), only a small number made serious efforts to improve (15% sought help from office).							
	Survey 2: To investigate satisfaction with Graduate Teaching Assistant Orientation Workshop	X O X=Orientation Workshop	85 graduate teaching assistants (GTA)	?	Survey of Orientation Workshop (1974)	GTA's found orientation workshops more helpful in dealing with administrative detail than in working with students or faculty members.							
	Study 1: To assess the effects of student rating feedback & consultation on subsequent faculty performance	O X O X=student rating feedback & consultation	263 faculty	2 or more terms between 1969 & 1972	Student rating form	Significant improvements shown for 13 of 15 measures. While results statistically significant, they are not dramatic in absolute sense. Results consistent with expectation that voluntary participation in student evaluation programs with feedback can help faculty improve instructional effectiveness.	d	a b c		a b	Volunteer sample		Low

RATINGS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Hoyt & Howard (1978) continued	Study 2: Part 1-To assess the effects of student rating feedback & consultation on subsequent f..ulty performance	O X O X=student rating feedback & consultation	348 faculty	2 or more terms between 1973 & 1975	Student rating form (revised)	Posttest mean for "Progress on Relevant Objectives" significantly higher than pretest mean.	d	a b a		a b	1) Volunteer sample 2) ANCOVA on nonequivalent control groups		Low
	Part 2-To examine instructional improvement relative to contact with office					Ancova-adjusted measures of effectiveness increased as a function of amount of contact with director (14 of 18 measures increased). Significant improvement resulted when consultative services made available to motivated faculty.							
Marsh, Fleiner & Thomas (1975)	To assess the effects of midterm student rating feedback on end-of-term course evaluations & achievement (validity also assessed)	E: R O O X O C: R O O O X = student rating feedback	287 UCLA students (18 different sections involved & instructors were graduate students	1 quarter	1) Pretest designed to predict final exam performance 2) Final exam 3) 46 item evaluation instrument (UCLA developed) 4) short form of 46 item form	1) E students had significantly higher responses on summary comparison item; on 8 of 46 items and on 2 (instructor approachability & value of the readings) of 7 evaluation factors. 2) No significant differences between groups in overall student performance.				a b c	Student was unit of analysis	Randomization	Fair
McKeachie & Lin (1975e)	To investigate the effects of discrepancies between mid-term student ratings & faculty self-ratings of expected & ideal teaching performance on faculty performance	O X O X=student rating feedback	28 instructors of introductory psychology classes at University of Michigan	1 semester	32 item Michigan Student Perception of Teaching form Based on mid-term student & faculty ratings, instructors were divided into 8 groups	1) Significant differences for 2 (group interaction & feedback) of 7 dimensions were found for those whose expected & ideal ratings were higher than student ratings. 2) The group rated more highly by students than by themselves changed in a negative direction (on feedback only).		a b a d	a b c	a b c	1) Unclear text 2) Unwarranted conclusions		Low

RATINGS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
McKeachie & Lin (1975b)	To assess the relative effects on end-of-term faculty & student performance of a) mid-term student rating feedback combined with consultation b) student rating feedback alone & c) no feedback	E_1: R O X_1 O E_2: R O X_2 O C: R O O X_1=student rating feedback & consultation X_2=student rating feedback	37 graduate assistants & 3 faculty teaching introductory psychology courses at University of Michigan	14 week term	1) 32 item Michigan Student Perception of Teaching & Learning (McKeachie & Lin) 2) Selected items from Introductory Psychology Criteria (Milholland 3) Attitude toward Psychology questionnaire 4) Attitude toward self questionnaire 5) Attitude toward Mental Illness questionnaire 6) Curiosity Test	1) Significant differences in favor of E_1 group for both general teaching effectiveness & overall value of course & for 1 (impact on students) of 7 dimensions. 2) E_1 was significantly higher in student achievement for 1 set of psychology classes as measured by Criteria Test & for measure of Curiosity in another set of classes. 3) Among groups initially rated low, medium or high, no significant differences on final criterion measures.	d	f j	a	a b c	1) unclear text 2) unwarranted conclusions	Randomization	Fair

Author/Date	Purpose	Code	Participants	Duration	Instrumentation	Stated Results	SC	I	C	E	Weaknesses	Strengths	Confidence Rating
												RATINGS	
		Components of Design					Threats to Validity						
Miller (1971)	1) To assess the effects of mid-term student rating feedback on end-of-term faculty & student performance 2) To investigate instructor attitudes toward value of student ratings	E: R O X O O C: R O O O X=student rating feedback	36 teaching assistants (TAs) teaching courses in religion or earth science (approximately 2000 students involved)	1 semester	1) Survey of Student Opinion of Teaching (SSOT) 2) Instructor Attitude Questionnaire based on first 10 items of SSOT 3) Student achievement on mid-term & final Based on Instructor Attitude Questionnaire, instructors divided into: 1) Feedback/Favorable Attitudes 2) Feedback/Unfavorable Attitudes 3) No Feedback/Favorable Attitudes 4) No Feedback/Unfavorable Attitudes	1) Instructors in feedback & attitude groups did not differ significantly on end-of-term ratings. 2) In 2 of 3 courses, no significant differences on final exam scores for feedback or attitude groups. 3) In 3rd course, significant difference on achievement in favor of feedback condition ($p<.01$)				a b c	Use of instructors as unit of analysis may have resulted in sampling errors due to small n per cell (data combined for sections) (noted by investigator)	1) Randomization 2) ANCOVA analysis	High
Murphy & Appel (1978)	1) To assess the relative effects on faculty performance of a) student rating feedback combined with consultation b) student rating feedback alone & c) no feedback 2) To investigate a problem-solving approach to utilizing feedback	E_1: R O X_1 O E_2: R O X_2 O C: R O O X_1=student rating feedback & consultation (augmented feedback utilizing non-expert consultants) X_2=student rating feedback (simple feedback)	70 faculty at University of Texas (each randomly selected from pool of potential subjects)	18 week semester	Adapted form of Course Instructor Survey: General Questionnaire (developed at University of Texas)	1) E_1 not significantly different from E_2 in improvement of ratings. 2) E_1 & E_2 showed more gain (statistically) than C. 3) Instructors receiving feedback did not utilize feedback in item-by-item problem-solving approach.		f g	c	a b c	1) Although statistically significant difference in gains between feedback & no feedback conditions, gain was small in absolute sense (noted by investigator) 2) Change score analysis--should ANCOVA have been used?	Randomization	Fair

RATINGS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
Oles & Lencoski (1973)	To assess the effects of student evaluations on faculty self-ratings	E: 0 X 0 - - - - - - - - - C: 0 0 X=student rating feedback	24 instructors in a graduate school of education	approximately 2 weeks	1) 12 item instructor evaluation form 2) 12 item student evaluation form	1) C group's test-retest correlation coefficient was .82 while E group's correlation was .54. 2) Chi square test on total number of changes was significant at .001. 3) Instructor self-rating changes not always in direction of student ratings	d	g	c	a b c	1) Nonequivalent control group design 2) Reliability of measures not reported		Low
Overall & Marsh (1976)	To assess the effects of student rating feedback on faculty & student performance & to assess affective consequences of such a procedure (application of subject matter & plans to pursue subject further)	E: R 0 0 X 0 C: R 0 0 0 X=student rating feedback	993 UCLA undergraduates who completed an introductory course in computer programming during Fall, Winter, or Spring 1973-74	3 quarters	1) Pretest to predict final exam performance 2) Evaluation of Instruction Program questionnaire -7 dimensions of teaching -questions on affective consequences 3) Final exam	1) Significant differences in favor of E for 2 summary items (overall rating of instructor, & of course), for perceived difference in instructional quality and for 4 (concern, learning, interaction, & examinations) of 7 dimensions. 2) E significantly higher on exam performance. 3) E gave more favorable responses to affective consequence items. E significantly higher on 3 of 5 items.	d			a b	Unit of analysis=student	1) Investigation of affective consequences 2) Randomization 3) ANCOVA analysis	Fair
Pambookian (1974)	To investigate the effects of discrepancies between mid-term student & faculty self-ratings on end-of-term performance	0 X 0 X=student rating feedback	13 teaching fellows teaching psychology at University of Michigan	1 semester (approximately 14 weeks)	21 items from Student Opinion Questionnaire (SOQ) revised by McKeachie-Lin Based on pretest ratings, subjects divided into: a) more favorably rated (F) (n=7) b) more moderately rated (M) (n=3) c) more unfavorably rated (U) (n=3)	1) Significant differences among groups on rapport & strong trends on skill (F=3.23, df=2/8, p<.08), overload (F=3.58, df=2/10, p<.07), & interaction (F=3.24, df=2/10, p<.06) 2) Individual 't' tests to compare gain scores between groups: a) Between F & M, significant differences on skill, interaction, & rapport in favor of M b) Between U & M, no significant differences c) Between M & F, significant differences on overall value of course in favor of M d) Trends in favor of M over U in rapport and toward less work overload.	a a	d g h	b c	a b c	1) Small N 2) Individual 't' tests used to further investigate no significant differences in findings using ANOVA (fishing & error rate problem) 3) Change score analysis.		Low

Author/Date	Purpose	Components of Design				*Stated Results	Threats to Validity				RATINGS		
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Pambookian (1976)	To investigate the effects of discrepancies between mid-term student & faculty self-ratings on end-of-term faculty performance	0 X 0 X=student rating feedback	13 teaching fellows teaching introductory & educational psychology	1 term	21 items from Student Opinion Questionnaire (SOQ) revised by McKeachie-Lin Based on pretest ratings, subjects were divided into: a) unfavorably discrepant (UD) (n=2) b) minimally discrepant (MD) (n=2) c) favorably discrepant (FD) (n=9)	1) Differences among groups on skill (p<.2), feedback (p<.3), & rapport (p<.01). 2) Individual 't' tests to compare gain scores between groups: a) U significantly changed more on skill, feedback, rapport, general teaching ability & overall value of course than FD. b) MD improved significantly on rapport compared to FD & showed strong trends in same direction on skill. c) Least gains made by FD.	a c	d g h	b c	a b e	1) Small N 2) Individual 't' tests used to further investigate no significant finding using ANOVA (fishing & error rate problem) 3) Change score analysis	Discussion of limitations	Low
Rotem (1978)	To assess the effects of mid-term student ratings & faculty self-ratings of actual & desirable teaching performance on end-of-term faculty performance.	R 0 X 0 R 0 0 R 0 X=student rating feedback	51 instructors at University of California at Santa Barbara (2,980 students involved)	1 term	Student rating form of 9 items selected from a set described by Isaacson, McKeachie, Milholland, Lin, Hofeller, Baerwaldt, & Zinn. 6 Factors: a) overload b) organization c) feedback d) interaction e) rapport f) skill	1) On student ratings & instructor ratings, no significant differences between group means as a result of feedback or prior experience with pretest. 2) No functional relationship between rating discrepancies & posttest ratings.		f	e	a b c	1) Volunteer sample 2) Short time interval between pre & posttests due to quarter system	1) Excellent discussion 2) Good design 3) Planned comparison contrasts 4) Multiple regression analysis for discrepancies 5) Randomization	High

		Components of Design					Threats to Validity				RATINGS		
Author/Date	Purpose	*Code*	*Participants*	*Duration*	*Instrumentation*	Stated Results	*SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Sherman (1977-78)	To assess the effects of student rating feedback ('formative' feedback) on subsequent faculty performance	Instructor 1: O O X_1 O X_2 O - - - - - - - - - - Instructor 2: O O X_2 O X_1=Feedback Low Specificity (FLS) X_2=Feedback High Specificity (FHS)	Instructor 1 =35 students (health class) Instructor 2 =23 students (educational psychology class)	?	Student rating form consisting of: a) value of instruction b) quality of instruction c) explanation of ratings	1) For Instructor 1, significant differences between baseline & FHS on value of instruction. 2) For Instructor 2, significant differences between baseline & FHS on quality of instruction.	a d	b a a f	b a	a b c	1) Cumulative effect of treatments 2) Unclear text		Low
Tuckman & Oliver (1968)	To assess the relative effects on faculty performance of: a) student rating feedback, b) supervisor feedback, c) student rating & supervisor feedback, & d) no feedback	E_1: R O X_1 O E_2: R O X_2 O E_3: R O X_3 O C_1: R O O - - - - - - - - - C_2: O X_1=Student rating feedback X_2=Supervisor feedback X_3=Student rating & supervisor feedback	286 teachers of vocational subjects at high school or technical level 15 additional teachers in posttest-only condition	1 semester (12 weeks)	Student Opinion Questionnaire (SOQ) developed by Bryan	1) E_1 & E_3 showed significantly greater change than E_2 & C_1. 2) E_1 & E_3 were statistically comparable indicating a failure for supervisor feedback to generate any change beyond that accounted for by student feedback alone. 3) E_2 produced a significantly greater negative shift (that is, opposite to feedback recommendations) than C_1. 4) C_2 served to rule out testing effects.			a	a b c	Change score analysis--should ANCOVA have been used?	1) Excellent discussion 2) Large N 3) Teacher years of experience was controlled 4) Randomization	High

													RATINGS
Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Vogt & Lasher (1973)	To assess the effects of student rating feedback on subsequent faculty performance	B | A | X | O | X | O X | X O | O X | X O | O X | X O | O X | X O | O X | X O | O X | X O | O X=student rating feedback	Group A: 50 instructors who were members of Bowling Green College of Business Administration at time that mandatory student evaluation system introduced, Winter 1969-70 (22,141 students in 1000 courses) Group B: 13 instructors who joined Bowling Green College of Business Administration in September, 1970 after introduction of mandatory student evaluation system (4317 students in 195 courses)	8 quarters	Bowling Green evaluation form (open-ended questions & student assignment of grade as index of teaching performance)	Regression coefficients of regression equations not significant--student rating feedback did not result in improved faculty performance.	d e f g		b c d	e b			Fair

RATINGS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Weerts (1978)	To assess the combined effects of mid-term student rating feedback & consultation on end-of-term faculty performance	E_1: R O X_1 O E_2: R O X_2 O C: R O X_1=student rating feedback X_2=student rating feedback with consultation & student-instructor dialogue	54 classes in Rhetoric program at University of Iowa (3 full-time faculty & 51 graduate TAs)	1 school term	28 item Student Perceptions of Teaching form (SPOT) (Whitney & Weerts) (items chosen from a pool of items)	1) No significant differences among 3 groups from mid-term to end-of-term (a 2 factor ANOVA with repeated measures on 1 factor was used & an alpha level of $p<.001$ used because of 28 separate analyses). 2) No significant differences among 3 groups at end-of-term. 3) For 20 of 28 items, E_2 had higher ratings than C. Statistically, chances for this occurring less than 5%. 4) For 23 of 28 items, E_1 had higher ratings than C. Statistically, chances for this occurring $p<.001$.	e d		e	a b c	Repeated measures ANOVA with repeated measures on 1 factor--should MANOVA or MANCOVA been used?	Randomization	Low

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Borg (1975)	To assess the effects of teacher language protocol modules on teacher skill acquisition & change, & on student performance	E: 0 X 0 (n=25) C: 0 0 (n=15) X=protocol training in teacher language protocol modules	40 fourth, fifth, & sixth grade in-service elementary school teachers	approximately 7 weeks	1) Observation form to record 12 teaching behaviors (multiple questions, defining, vague words, general praise, specific praise, use of student ideas, voice modulation, paraphrasing, cueing, opening review, terminal structure, summary review) 2) Observer ratings of 10 teacher characteristics 3) 2 achievement tests 4) SRA Short Test of Educational Ability, level 3 5) Warner, Meeker & Eells Revised Occupational Rating Scale	1) E made significant gains on all 12 teaching behaviors while C made significant gains on 5 of 12. E significanrly exceeded C on 4 of 12 behaviors. 2) When pupil scholastic ability, parents' occupation & teacher coverage of units' content were partialled out, teacher's use of defining, voice modulation, paraphrasing & cueing were significantly related to pupil achievement on 2 measures, & teacher's use of opening review & terminal structure were significantly related to 1 achievement measure (across all subjects). 3) No significant relationships shown between teacher characteristics & pupil achievement.	c	k l		a b c	ANCOVA on nonequivalent control groups	Standard content unit for final observation	Low

PROTOCOLS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Borg (1977)	To assess the effects of classroom management protocol modules & pupil self-concept protocol modules on teacher skill acquisition & on pupil behavior	E: E O X_1 O C: E O X_2 O X_1=protocol training in classroom management modules X_2=protocol training in pupil self-concept modules	28 in-service elementary school teachers	approximately 8 weeks	1) Observation of teacher & pupil behaviors 2) North York Self-Concept Inventory 3) Piers-Harris Children's Self-Concept Scale	1) E teachers made significantly greater improvement in 7 of 13 behaviors than C teachers. 2) For recitation situations, E pupils showed no significant change in work involvement but significant reduction in both mildly deviant & seriously deviant behavior. C pupils showed a significant reduction in definitely off-task behavior but no other significant changes. 3) For seatwork situations, E pupils showed significant reductions in mildly deviant & seriously deviant behavior. No significant changes for C pupils. 4) C teachers received significantly more favorable post scores on 11 of 12 self-concept behaviors. 5) No significant improvement in pupil self-concept for E or C.	a		d	a b c	Unit of analysis for self-concept changes = classroom	Excellent discussion	High
Borg, Langer, & Wilson (1975)	To assess the effects of classroom management protocol modules on teacher skill acquisition & on pupil behavior	E: O X O (n=20) - - - - - C: O O (n=9) X=protocol training in classroom management modules	29 in-service elementary school teachers (control subjects drawn from same school as experimental subjects)	approximately 10 weeks	1) observation form to record classroom management behaviors 2) Pre & post pupil observations of 5 pupil behaviors (definitely involved in class work, probably involved, definitely off task, mildly deviant, seriously deviant)	1) E teachers received more favorable post ratings on all 13 behaviors but differences generally small & nonsignificant. 2) For recitation situations, E pupils' work involvement significantly increased & deviant behavior significantly decreased. 3) For seatwork situations, E pupils' work involvement significantly increased but no significant changes for deviant behavior.	a f	d g h		a b c	Use of ANCOVA with non-equivalent control groups	Discussion of plausible alternative explanations	Low

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	Participants	Duration	Instrumentation		SC	I	C	E			
Borg & Stane (1974)	Part 1: To assess the effects of 2 teacher language protocol modules on teacher skill acquisition & change	$0\ X_1\ 0$ X=protocol training in 2 teacher language protocols, encouragement & extension	19 in-service elementary school teachers	5 hours extended over one week for each protocol module. Total: 10 hours	Ratings of audiotapes on 7 specific behaviors: a) general praise b) specific praise c) use of student ideas d) prompting e) seeking further clarification f) refocusing g) redirection	Part 1: Teachers made significant gains on all but 2 behaviors, general praise & redirection.		a b e		a b c	1) Volunteer sample	Reactive effects of testing were controlled	Low
	Part 2: To compare the effects of the protocol module with the minicourse model in changing teacher behavior	E_1: $0\ X_1\ 0$ E_2: $0\ X_2\ 0$ X_1=encouragement & extension protocol training X_2=Minicourses 1 & 2 (general praise, specific praise, use of student ideas, prompting, seeking further clarification, refocusing, redirection)	protocol = 19 elementary school teachers Minicourse 1 = 48 intermediate elementary school teachers Minicourse 2 = ? number of kindergarden teachers	protocol = 10 hours Minicourses 1 & 2 = ?		Part 2: E_1 & E_2 conditions brought about similar gains on most behaviors compared.		g h i		a b c	1) Volunteer sample for protocols 2) Change score analysis		
Cleissman & Pugh (1976)	Part 1: To assess the effects of protocol films on teacher concept acquisition	E: $0\ X\ 0$ C: $X\ 0$ X=training with teacher-pupil interaction protocols	89 masters students enrolled in educational psychology course	6-8 hours of classroom instruction over a 2-3 week period	Categorizing Teacher Behavior test, Form F1	Significant gains in concept acquisition as a result of use of this protocol series.	c			a b c	1) No real control group (noted by investigatar) 2) Difficulty in following text		Fair
	Part 2: To assess characteristics & reactions to use of protocol film series	$X\ 0$ X=training with teacher-pupil interaction protocols	15 classes taught by 14 instructors -294 undergraduates, graduates, pre-service, in-service & school administrators	1-4 class periods	1) Instructor questionnaire 2) Student rating scale	1) Instructors favorably received protocol training. 2) Pupils of instructors favorably received use of protocol training.							

PROTOCOLS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Gliessman & Pugh (1978a)	To investigate the effect of protocol films of contrasting structure (high & low) on teacher concept acquisition & teacher reactions to use of filmed treatment		Pre-service & in-service teachers enrolled in graduate level educational psychology course		1) Categorizing Teaching Behavior test 2) Likert-type items from evaluation scale (to assess reactions)					a b c		1) Randomization 2) Design 3) Statistical analysis	High
	Study 1: To compare the effects of high & low structure films	E_1: R 0 X_1 0 E_2: R 0 X_2 0	N=20	1 day		Study 1: Significant gains in concept acquisition for E_1 & E_2 but no significant differences between the two.							
	Study 2: To assess the interactive effects of using both types of films in a single training group	E_1: R 0 X_1 0 E_2: R 0 X_2 0 E_3: R 0 X_3 0	N=30	2 days		Study 2: 1) Significant differences in concept acquisition between E_1, E_2 & E_3. 2) Comparison of means revealed significantly greater concept acquisition for E_2 than E1. 3) Significant increases in concept acquisition for all groups.							
	Study 3: To assess the effect of a variation that emerged during first two studies	E_1: R 0 X_1 0 E_2: R 0 X_2 0 X_1=protocol training using high structure film X_2=protocol training using low structure film X_3=protocol training using high/low structure film	N=20	2 days		Study 3: 1) No significant differences between E_1 & E_2 for concept acquisition. 2) Significant increases in concept acquisition for both groups. 3) E_1 had significantly more favorable reactions to films than E_2.							

Author/Date	Purpose	Components of Design *Code*	*Participants*	*Duration*	*Instrumentation*	Stated Results	Threats to Validity *SC*	*I*	*C*	*E*	Weaknesses	Strengths	Confidence Rating
Gleissman & Pugh (1978b)	To assess the relative effects of different instructional treatments on concept acquisition	<u>Study 1</u>: E_1: R X_1 O (n=10) E_2: R X_2 O (n=11) E_3: R X_3 O (n=12) - - - - - - - - - - E_4: X_4 O (n=11)	44 students in a graduate elementary education course	1 day?	Categorising Teacher Behavior test	<u>Study 1</u>: Significant differences among the combined E_1 & E_2 groups, E_3 & E_4. E_3 & E_4 group means both significantly lower than mean of combined E_1 & E_2 groups.		g h		a b c	Significance level of p=.081		Fair
		<u>Study 2</u>: E_5: R X_5 O (n=10) E_6: R X_2 O (n=9) X_1=Concept names, definitions, & filmed exemplification X_2=Concept names & definitions X_3=Concept names on film test X_4=unstructured viewing of protocol films followed by concept names on film test X_5=Concept names, definitions, & filmed exemplification & direct "cues" to instances	19 students in a graduate educational psychology course	1 day?	Categorising Teacher Behavior test	<u>Study 2</u>: E_5 group mean greater (p=.081 directional) than E_6 group mean.							

PROTOCOLS

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Cleissman, Pugh, & Bielat (1979a)	To assess the effects of protocol films on teacher concept & skill acquisition	E: X_1 0 X_2 0 C: (X) 0 X_2 0 X_1=protocol training in teacher-pupil interaction (X)=instruction in individual student counseling X_2=instruction in principles of concept teaching	20 in-service teachers enrolled in masters level course in educational psychology	6 days	1) Categorizing Teacher Behavior test 2) Frequency counts of specified behaviors in microlesson	1) For concept acquisition, E had significantly greater mean scores on total, probing & informing, than C. 2) E had significantly greater mean than C for skill acquisition. 3) Correlation between concept & skill acquisition was reliable & positive: r=.51, df=8, p=.08.	e	8		a b c	Significant differences not necessarily of practical importance	Questions under investigation	Low
Cleissman, Pugh, & Bielat (1979b)	1) To assess the effects of protocol films on teacher concept acquisition 2) To assess the relationship between concept scores & frequencies of skill acquisition 3) To assess trainees' ability to use skill concepts interpretively	0 X 0 X=training with teacher-pupil interaction protocols	30 in-service teachers enrolled in a masters level course in psychology of teaching	6 hours of class time over 2 consecutive days	1) Categorizing Teaching Behavior test 2) Skill acquisition (probing) measured by coding of audiotaped microteaching session 3) Written responses to questions on audiotaped interactive skills	1) No significant relationship found between skill concept acquisition scores & skill frequencies. 2) Trainees' written responses indicated subjective evidence of both conceptual & nominal outcomes but no observational effects. 3) Also on written responses: a) Trainees' ability to apply concept of probing was positively & significantly related to the frequency of probing. b) Trainees' accuracy in dealing with probing concept characteristics was found to be unrelated to skill acquisition.		e b e		a b c	Short duration of training	Intended as a replication	Fair

Author/Date	Purpose	Components of Design				Stated Results	Threats to Validity				Weaknesses	Strengths	Confidence Rating
		Code	*Participants*	*Duration*	*Instrumentation*		*SC*	*I*	*C*	*E*			
Kleucker (1974)	To assess the relative effects of protocol training & skill-training (microteaching) on teacher concept & skill acquisition	E_1: R X_1 O E_2: R X_2 O E_3: R X_3 O C: R (X) O X_1=protocol training in teacher-pupil interaction X_2=skill training (microteaching) X_3=protocol & skill training (X)=non-related instruction	38 undergraduates enrolled in 2 sections of an educational psychology course	10 days	1) Videotape concept test 2) Printed concept test 3) Microteaching test	1) Protocol training & skill training lead to concept acquisition & skill acquisition respectively. 2) Protocol training & skill training alone do not lead to differential outcomes--both lead to concept & skill acquisition. 3) Combination of skill & protocol training is at least as effective, & frequently more effective, than either used alone.	a			a b c	1) Small N (noted by investigator) 2) 5-day duration of protocol condition; 10-day duration in other conditions	1) Discussion of limitations 2) Inclusion of study's implications	High

Appendix B: Threats to Validity*

Threats to Statistical Conclusion Validity

a) Low Statistical Power
b) Violated Assumptions of Statistical Tests
c) Fishing and the Error Rate Problem
d) Reliability of Measures
e) Reliability of Treatment Implementation
f) Random Irrelevancies in the Experimental Setting
g) Random Heterogeneity of Respondents

Threats to Internal Validity

a) History
b) Maturation
c) Testing
d) Statistical Regression
e) Selection
f) Mortality
g) Interaction of Selection and History
h) Interaction of Selection and Maturation
i) Interaction of Selection and Instrumentation
j) Resentful Demoralization
k) Diffusion or Imitation of Treatments
l) Compensatory Rivalry

Threats to Construct Validity

a) Inadequate Preoperational Explication of Constructs
b) Mono-Operation Bias
c) Mono-Method Bias
d) Evaluation Apprehension
e) Experimenter Expectancies

Threats to External Validity

a) Interaction of Selection and Treatment
b) Interaction of Setting and Treatment
c) Interaction of History and Treatment

*Derived from Cook and Campbell, 1979.