DOCUMENT RESUME

ED 353 283 TM 019 001

AUTHOR	Marsh, Herbert W.; Roche, Lawrence
TITLE	The Use of Students' Evaluations and an Individually Structured Intervention To Enhance University Teaching Effectiveness.
SPONS AGENCY	Australian Dept. of Employment, Education and Training, Canberra.
PUB DATE	8 Jun 92
NOTE	25p.
PUB TYPE	Reports - Evaluative/Feasibility (142)
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*College Students; Control Groups; Experimental Groups; Feedback; Foreign Countries; Higher Education; *Instructional Improvement; *Intervention; *Professors; Self Evaluation (Individuals); *Student Evaluation of Teacher Performance; Teacher Effectiveness; *Teacher Improvement
IDENTIFIERS	Australia; Feedback Consultation; *Students Evaluations of Educational Quality

ABSTRACT

The effectiveness of students' evaluations of teaching effectiveness (SETs) as a means of enhancing university teaching was studied, with emphasis on the multidimensionality of SETs, an Australian version of the Students' Evaluations of Educational Quality instrument (the ASEEQ), and the 1986 feedback/consultation intervention of R. C. Wilson. Ninety-two teachers completed self-evaluation surveys and were evaluated by their students at the middle of semester one and at the end of semesters one and two. Three randomly assigned groups received the feedback consultation intervention at the midterm of semester one (MT), the end of semester one (ET), or received no intervention (control). Each MT and ET teacher targeted specific ASEEQ dimensions that were the focus of his/her individually structured intervention. The ratings for all groups improved over time, but only ratings for the ET group improved significantly more than those of the controls. For both ET and MT groups, targeted dimensions improved more than non-targeted dimensions. Results suggest that SET feedback coupled with consultation is an effective means of improving teaching effectiveness and provide one model for feedback/consultation. One figure and five tables present study findings. (Author/SLD)

The Use of Students' Evaluations and an Individually Structured Intervention to Enhance University Teaching Effectiveness

Herbert W. Marsh and Lawrence Roche
University of Western Sydney, Macarthur

June 8, 1992

Running Head: Students Evaluation Feedback Effects

Requests for further information should be sent to Herbert W. Marsh, Professor of Education, University of Western Sydney, Macarthur, PO Box 555, Campbelltown NSW 2560, Australia. The authors would like to acknowledge the financial support of the Australian Department of Employment and Educational Training for the conduct of this research, the teachers at the University of Western Sydney, Macarthur who volunteered to participate in this research, and Raymond Debus who made helpful comments on an earlier draft of this article..

ABSTRACT

The present investigation evaluates the effectiveness of students' evaluations of teaching effectiveness (SETs) as a means for enhancing university teaching. We emphasize the multidimensionality of SETs, an Australian version of the Students' Evaluations of Educational Quality (Marsh, 1987) instrument (ASEEQ), and Wilson's (1986) feedback/consultation intervention. All teachers (N=92) completed self-evaluation surveys and were evaluated by students at the middle of semester 1 and at the end of semester 1 and 2. Three randomly assigned groups received the feedback/consultation intervention at midterm of semester 1 (MT), at the end of semester 1 (ET), or received no intervention (control) Each MT and ET teacher "targeted" specific ASEEQ dimensions that were the focus of his/her individually structured intervention. The ratings for all groups improved over tim:, but only ratings for the ET group improved significantly more than those the control group. Foi both ET and MT groups, targeted dimensions improved more than nontargeted dimensions. The results suggest that SET feedback coupled with consultation is an effective means to improve teaching effectiveness and provide one model for feedback/consultation.

Students' evaluations of teaching effectiveness (SETs) are variously collected to provide: (1) diagnostic feedback to faculty that will be useful for the improvement of teaching; (2) a measure of teaching effectiveness to be used in administrative decision making; (3) information for students to use in the selection of courses and teachers; and (4) an outcome or a process description for research on teaching. Nearly all SET programs cite the first reason as one basis for collecting SETs and it is typically seen as the most important; none of the other reasons is nearly so universal. Consistent with these priorities, the purpose of the present investigation is to evaluate the effectiveness of feedback from multidimensional SETs and feedback/consultation as a means for enhancing university teaching, based on a new Australian version of the Students' Evaluations of Educational Quality (SEEQ; Marsh, 1987) instrument called ASEEQ.

The Multidimensionality of SETs and the SEEQ Instrument

Effective teaching is a multidimensional construct (e.g., a teacher may be organized but lack enthusiasm). Thus, it is not surprising that a considerable body of research shows that SETs are also multidimensional (see Marsh, 1987). Information from SETs depends upon the content of the items. Poorly worded or inappropriate items will not provide useful information. If a survey instrument contains an ill-defined hodgepodge of different items and SETs are summarized by an average of these items, then there is no basis for knowing what is being measured. Particularly when the purpose of the ratings is to provide teachers with formative feedback about their teaching effectiveness, it is important that careful attention be given to the components of teaching effectiveness that are to be measured. Surveys should contain separate groups of related items that are derived from a logical analysis of the content of effective teaching and the purposes that the ratings are to serve, and that are supported by theory, previous research, and empirical procedures such as factor analysis and multitrait-multimethod (MTMM) analysis. The strongest support for the multidimensionality of SETs apparently comes from research using Students' Evaluations of Educational Quality (SEEQ; Marsh, 1987) instrument.

In the development of SEEQ: 1) a large item pool was obtained from a literature review, forms in current usage, and interviews with teachers and students about what they saw as effective teaching; 2) students and teachers were asked to rate the importance of items; 3) teachers were asked to judge the potential usefulness of the items as a basis for feedback; and 4) open-ended student comments were examined to determine if important aspects had been excluded. These criteria, along with psychometric properties, were used to select items and revise the instrument, thus supporting the content validity of SEEQ responses. Marsh and Dunkin (in press) subsequently demonstrated that the SEEQ dimensions are consistent with principles of effective teaching and learning established on the basis of accepted theory and research. Based on their review, they concluded that SEEQ factors conform to principles of teaching and learning emerging from attempts to synthesize knowledge of teaching effectiveness.

Factor analytic support for the SEEQ scales is particularly strong. To date, more than 30 published factor analyses of SEEQ responses have identified the factors that SEEQ is designed to measure (e.g., Marsh, 1982a; 1983; 1984; 1987; 1991a; Marsh & Hocevar, 1984, 1991). Factor analyses of teacher self-evaluations using SEEQ also identified the SEEQ factors, demonstrating that the factors generalize beyond responses by students. Multitrait-multimethod analyses of student/teacher agreement on SEEQ factors provided support for the convergent and discriminant validity of SEEQ responses (Marsh, 1982b; Marsh, Overall & Kesler, 1979). More recently, Marsh and Bailey (in press) examined the consistency of profiles of SEEQ scores (e.g., high on Enthusiasm and low on Organization) for a cohort of teachers who had been evaluated continuously over a 13-year period. They reported that each teacher has a relatively unique profile of SEEQ scales that generalizes over different courses, over graduate and undergraduate level courses, and over an extended period of time.

Whereas the value of multidimensional ratings is widely accepted for purposes of diagnostic feedback, there is heated debate about the relative usefulness of multidimensional profiles and overall summary ratings for purposes of personnel decisions (e.g., Abrami, 1989; Abrami & d'Apollonia, 1991; Marsh, 1987, 1991b). Although this issue is not a specific focus of the present investigation, a possible compromise arising from this debate is to summarize SETs as a weighted average of specific SEEQ dimensions. Marsh and Dunkin (in press; also see Marsh & Bailey, in press) noted that one approach to operationalizing a weighted average approach is to weight specific SEEQ components according to the relative importance of each scale as judged by the teacher who is being evaluated. This strategy has the added benefit of providing the teacher with a systematic role in the interpretations of the ratings used to summarize his/her teaching effectiveness, but to our knowledge this weighted average approach has not been previously employed (but see a related application by Hoyt, Owens, & Grouling, 1973). Because all teachers were asked to judge the relative importance of each SEEQ dimension as part of the feedback intervention used in the present investigation, we were able to construct a (teacher rated importance) weighted average of SEEQ dimensions and use it as one of our criterion measures.

SETs are commonly collected and frequently studied at North American universities, but not in most other parts of the world. Because of the extensive exposure of North American research, there is a danger that North American instruments will be used in new settings without first studying their applicability. In order to address this issue, Marsh (1981) described the applicability paradigm for studying the initial suitability of SEEQ that was in several Australian studies as well as studies in Spain, New Zealand, Papua New Guinea, and elsewhere (e.g., Marsh, 1987). Of particular relevance, Marsh and Roche (in press) conducted one of these studies at the newly established University of Western Sydney that served as a pilot study for the present investigation.

Utility Of Student Ratings

Braskamp, Brandenburg, and Ory (1985), using a broad rationale based on organizational research, argued that it is important for universities and individual teachers to take evaluations seriously. Summarizing this perspective Braskamp, et al. (p. 14) stated that: "the clarity and pursuit of purpose is best done if the achievements are known. A course is charted and corrections are inevitable. Evaluation plays a role in the clarity of purpose and determining if the pursuit is on course." In a related perspective, Marsh (1984, 1987) argued that the introduction of a broad institution-based, carefully planned program of SETs is likely to lead to the improvement of teaching. Teachers will give serious consideration to their own teaching in order to evaluate the merits of the program. Clear support of a program by the central administration will serve notice that teaching effectiveness is being taken seriously. The results of SETs, as one indicator of effective teaching, will provide a basis for informed administrative decisions and thereby increase the likelihood that quality teaching will be recognized and rewarded, and that good teachers will be given tenure. The social reinforcement of gaining favorable ratings will provide added incentive for the improvement of teaching, even for teachers who are tenured. Finally, teachers report that the feedback from student evaluations is useful to their own efforts for the improvement of their teaching. Murray (1987) presented a similar logic in making the case for why SETs improve teaching effectiveness and offered four reasons: (a) SETs provide useful feedback for diagnosing strengths and weaknesses, (b) feedback can provide the impetus for professional development aimed at improving teaching, (c) the use of SETs in personnel decisions provides a tangible incentive to working to improve teaching, and (d) the use of SETs in tenure decisions means that good teachers are more likely to be retained. In support of his argument, Murray (1987) summarized results of published surveys from seven universities that asked teachers whether SETs are useful for improving teaching and, across the seven studies, about 80% of the respondents indicated that SETs led to improved teaching. None of these logical arguments, however, provides an empirical demonstration of improved of teaching effectiveness resulting from SET feedback.

Feedback Studies.

In most studies of the effects of feedback from SETs, teachers are randomly assigned to experimental (feedback) and one or more control groups; SETs are collected during the term; ratings of the feedback teachers are returned to teachers as quickly as possible; and the various groups are compared at the end of the term on a second administration on SETs. Earlier versions of SEEQ were employed in two such feedback studies using multiple sections of the same course (also see related research by McKeachie, et al., 1980). In the first study results from an abbreviated form of the survey were simply returned to teachers, and the impact of the feedback was positive, but very modest (Marsh, Fleiner, & Thomas, 1975). In the second study (Overall & Marsh, 1979) researchers actually met with teachers in the feedback group to discuss the evaluations and possible strategies for improvement. In this study students in the feedback group subsequently performed better on a standardized final examination, rated teaching effectiveness more favorably at the end of the course, and experienced more favorable affective outcomes (i.e., feelings of course mastery, and plans to pursue and apply the subject). These two studies suggest that feedback, coupled with a candid discussion with an external consultant, can be an effective intervention for the improvement of teaching effectiveness.

In his classic meta-analysis, Cohen (1980) found that teachers who received midterm (MT) feedback were subsequently rated about one-third of a standard deviation higher than controls on the Total Rating (an overall rating item or the average of multiple items), and even larger differences were observed for ratings of Instructor Skill, Attitude Toward Subject, and Feedback to Students. Studies that augmented feedback with consultation produced substantially larger differences, but other methodological variations had little effect. The results of this meta-analysis support the SEEQ findings described above and demonstrate that SET feedback, particularly when augmented by consultation, can lead to improvement in teaching effectiveness.

L'Hommediu, Menges, and Brinko (1990; also see L'Hommediu, Menges, & Brinko, 1988) noted the need for meta-analyses of the influence of design and contextual effects. They updated Cohen's (1980) meta-analysis and critically evaluated the methodology used in the 28 feedback studies. They concluded that the overall effect size (.342) attributable to feedback was probably attenuated by threats to validity in existing research and developed methodological recommendations for future research. Among their many recommendations, they emphasized the need to: use stratified random assignment and covariance analyses in conjunction with a sufficiently large number of teachers to ensure the initial equivalence of the groups; more critically evaluate findings within a construct validity framework as emphasized by Marsh (1987); more critically evaluate the assumed generalizability of MT feedback to ET feedback; and to base results on well-standardized instruments such as SEEQ. They also noted the apparently inevitable threat of a John Henry effect in which the anticipation of being rated may lead to more effective teaching by teachers in randomly assigned control groups, thus making it more difficult to measure the true effects of the intervention. Although not a particular focus of their review, L'Hommediu et al.(1990) also noted that teachers initially rated lowest tended to be more positively influenced by the feedback intervention than other experimental teachers -- an aptitute-treatment interaction. Consistent with Cohen (1980) they concluded that "the literature reveals a persistently positive, albeit small, effect from written feedback alone and a considerably increased effect when written feedback is augmented with personal consultation" (1990, p. 240), but that improved research that incorporated their suggestions would probably lead to larger, more robust effects.

Marsh (1987; Marsh & Dunkin, in press) summarized important issues that remain unresolved in SET feedback research. Of particular relevance to the present investigation, it was noted that nearly all of the studies were based on MT feedback. This limitation probably weakens effects in that many instructional characteristics cannot be easily altered within the same semester (in some studies the period between receiving the feedback and subsequent data collection is as little as 4 or 5 weeks). Also, because students may be substantially influenced by what happens in the first half of the course, even substantial changes in teaching effectiveness in the last half of the term may have only modest effects on SETs. Thus, for example, Marsh, Fleiner and Thomas (1976) found significant differences for an overall teacher rating in which students were asked to judge changes in teaching effectiveness between the middle of the term and the end of the term, but not on a traditional overall teacher rating. Adding to these concerns, Marsh and Overall (1980) used results from a multisection validity study to demonstrate that MT ratings were less valid than ET ratings. Because SETs are typically collected near the end of the term, the more relevant question for SET feedback research to address is the impact of feedback from ET ratings. Even if there are short-term gains due to MT feedback, it is important to determine whether these effects generalize to ratings in subsequent semesters. L'Hommediu et al. (1990, p. 238) similarly argued that "most experiments using midterm feedback are intended to generalize to a quite different situation: end-of-term summative ratings to be used by teachers for improving instruction in subsequent terms" leading them to conclude that "the legitimacy of extrapolating the results to end-of-term rating situations is questionable."

A few studies have considered long-term follow-ups of short-term interventions, but these were apparently not designed for this purpose and were sufficiently flawed in relation to this extrapolation that no generalizations are warranted (see Marsh, 1987). No research has examined the effects of continued SET feedback over a long period of time with a true experimental design, and such research may be ethically dubious and very difficult to conduct. The long-term effects of SET feedback may be amenable to quasi-experimental designs (e.g., Aleamoni & Yimer, 1973; Voght & Lasher, 1973), but the difficulties inherent in the interpretation of such studies may preclude any firm generalizations. For shorter periods, however, it may be justifiable to withhold the SETs from randomly selected teachers or not to collect SETs at all. In particular, it is reasonable to evaluate the effects of feedback from ET ratings -- augmented with consultation -- on SETs collected the next semester in relation to SETs for no-feedback controls. The failure to systematically compare the effects of MT and ET interventions is one of the two most important deficits in the SET feedback research.

The most robust finding from the SET feedback research is that consultation augments the effects of written summaries of SETs. Other sources also support this conclusion. For example, in the Jacobs (1987) survey of Indiana University faculty, 70% of the respondents indicated that SETs had helped them improve their teaching but 63% indicated that even when teachers can interpret their ratings, they often do not know what to do in order to improve their teaching. Also, Franklin and Theall (1989), based on an 153-item, multiple-choice test of knowledge about SETs that was validated by experts in the field, concluded that many users lacked the knowledge to adequately use the SETs for summative or formative purposes. Nevertheless, insufficient attention in SET research has been given to the nature of consultative feedback that is most effective. Based on a review of research in education, psychology, and organizational behavior, Brinko (1991) contrasted 5 models of interaction relevant to instructional consultation: product model (the consultant is the expert and provider of expertise), prescription model (the consultant identifies, diagnoses, and remedies problems), collaborative/process model (there is a synergistic relationship between the consultant as a facilitator of change and the teacher as the content expert), affiliative model (the consultant is both an instructional consultant and psychological counsellor and the teacher is a seeker of personal and professional growth), and confrontational model (in which the consultant is a challenger or a devil's advocate). She suggested that a skillful instructional consultant may need to master several styles and use the one that is most responsive to the needs of the teacher-client and the particular situation. Brinko (p. 48) concluded, however, that "we still have no empirical evidence to differentiate between strategies and practices that make consultation successful and those that do not." It is surprising that there is not more systematic research on this practical issue and it represents, perhaps, the other most important deficit in SET feedback research.

Wilson's feedback/consultation process

Wilson (1986; also see Wilson, 1984, 1987) described a feeback/consultation process that appears to have considerable potential. A key element in this process was a set of 24 teaching packets that were keyed to the 24 items on the SET instrument used in his research. Each packet contained suggestions from teachers who had received Distinguished Teaching Awards or received multiple "best teacher" nominations by graduating seniors. In an application of this program conducted over a three-year period, participants were volunteer teachers who had been evaluated previously in the same course they would again be teaching. Based on SETs and self-evaluations of their own teaching, participants nominated specific evaluation items on which they would like assistance at a preliminary consultation session. The main consultation was held shortly before the second time the teachers were to teach the same course. The consultant began the session by noting items on which the teacher received the highest ratings, and then considered 3 to 5 items which the teacher had selected or had received the lowest ratings. For each selected item the three to six strategies from the corresponding teaching packet were described and the teacher was given copies of the two or three that were of most interest to the teacher. During the next week the consultant summarized the main consultation and strategies to be pursued in a letter to the teacher and subsequently telephoned the teacher during the term to ask how things were going. This process clearly fits the "collaborative/process model" in Brinko's (1991) typology, but may be sufficiently flexible to incorporate aspects of other models as appropriate.

Wilson's results indicated that ratings were systematically better at time 2 for the targeted items -- particularly those items that referred to concrete behaviors (e.g., states objectives for each class session) -- and an overall rating item. He also recognized the need for a nonintervention comparison group. For this purpose he considered SETs for 101 teachers who had not volunteered to be in the study but who had been evaluated on two occasions for the same course during the period of his study. For this large comparison group, there were no systematic changes in either specific or global SETs, supporting Wilson's contention that SETs without a consultation intervention are not likely to lead to improved teaching. Wilson suggested that the key elements in the consultation intervention were providing teachers with information on how to improve teaching in areas in which they are weak and the interpersonal expectations that created for some teachers a desire to fulfill an implied contract with their consultant.

Despite the obvious appeal of Wilson's feedback/consultation process and its successful application, empirical support for its effectiveness is weak -- based on a non-experimental design that does not rule out alternative explanations. The interpretation of pretest/posttest gain scores is a weak basis for causal inference, particularly since subjects were self-selected volunteers and more than half of the original participants had either dropped out or had not completed the intervention within three years -- the time when the study was terminated. Although Wilson reported that another group of teachers who did not volunteer to participate in the study showed no improvement, the comparability of the two groups may be dubious. A more subtle problem is the comparison of pretest/posttest gain scores on the "targeted" items that was the major focus of Wilson's conclusions. The results suggest that the intervention is most effective for those items that teachers target. Extending the logic of Marsh's (1987) construct validity approach, this finding apparently supports the construct validity of interpretations of the intervention in that gains are larger for those areas that were the focus of the intervention and smaller for those areas that were not the focus of the intervention. Wilson, however, did not actually report comparisons of these gains on targeted items with gains on other, untargeted items or with gains in the same items by other teachers who did not target them. Also, because the targeted items were typically selected that had particularly low ratings at time 1, regression to the mean alone would result in some positive gains and must be controlled more adequately. Whereas it might be possible to randomly assign the items to be targeted by each teacher in order to strengthen the experimental design, such a strategy would pervert the intended purpose of the intervention to provide the teacher with a personal stake in the intervention by selecting the areas to focus self-improvement efforts. Thus, for the sake of experimental design, some teachers would be asked to focus on SET areas which they judged to be the least important or were already strong, and to ignore areas of weakness which they thought were most important.

The pretest/posttest gains on overall rating items for all participating teachers provides a stronger basis of inference (although the lack of a randomly assigned control group is still a concern). Because the overall ratings reflect all the different SET areas to some extent, changes in the overall ratings reflect an implicit average across targeted and nontargeted items. The essence of the feedback/consultation, however, is that teachers target particular items and the intervention is specific to these targeted items. Logically, the design of the study requires that improvement should be larger for targeted items than for nontargeted items. If there is not differential growth on targeted and nontargeted items, then the specific and individual nature of the intervention is called into question. Thus, gains on overall rating items do not adequately capture the multidimensionality of SETS or the content specificity of the intervention as embodied in the construct validity approach.

In summary, Wilson developed an apparently valuable feedback/consultation process, described its systematic application, and provided evidence suggestive of its effectiveness. In the present investigation we provide a methodologically stronger paradigm for testing this intervention and an apparently stronger evaluation of its potential usefulness.

The Present Investigation

The purpose of the present investigation is to evaluate the effectiveness of feedback from a new Australian version of the SEEQ instrument (ASEEQ) and an adaptation of Wilson's feedback/consultation process for purposes of improving university teaching . The study is apparently unique in incorporating earlier proposals (e.g., L'Hommediu, et al., 1990; Marsh, 1987; Marsh & Dunkin, in press) to systematically compare the effectiveness of MT and ET feedback. In particular, different randomly assigned experimental groups received the feedback/consultation in the middle of semester 1 or at the end of semester 1 and were compared to a randomly assigned control group that received no feedback (until the end of semester 2 after the end of the study). Results for all three groups were compared on ratings collected at the middle of semester 1 (T1), the end of semester 1 (T2), and the end of semester 2 (T3).

The feedback intervention was based substantially on the work of Wilson (1986) and incorporated slightly modified versions of his idea packets that are designed to parallel the SEEQ dimensions. We extended his research by evaluating the effectiveness of the feedback intervention with a stronger experimental design that incorporated a randomly assigned control group. Furthermore, we more systematically evaluated his suggestion that areas of teaching effectiveness selected by teachers to be targeted in the intervention are the areas most substantially influenced by the intervention.

Other components considered in the study are: the use of a well-standardized SET instrument (L'Hommediu, et al. (1990) specifically recommended SEEQ); a systematic evaluation of aptitude-treatment interactions to test L'Hommediu, et al.'s (1990) suggestion that initially less effective teachers benefit more from the intervention; the use of a weighted average total score in which teacher self-ratings of the importance of each ASEEQ factor are used to weight the factors; and the application of a construct-validity approach in order to test the underlying rationale of the intervention.

Methods

Background.

Prior to 1990, the recently established University of Western Sydney, Macarthur (UWSM) was an autonomous institution within the College of Advanced Education sector that was the middle tier of Australia's three-tier system of higher education. In 1990, however, the formal distinction between research universities and Colleges of Advanced Education was abolished and all institutions from the middle tier were either amalgamated into one of the old universities or formed new universities. What was to become UWSM combined with two other institutions from the College of Advanced Education sector to establish the three UWS campuses.

As part of a large "priority reserve" source of funding for areas of national importance, the Australian Department of Employment and Education Training provided funding for projects that would improve the quality of teaching in Australian universities. Projects were selected that would improve the quality of teaching at a specific university but that were sufficiently general to provide a model for use at other universities. The present investigation describes results from one of these grants. (Subsequent matching funds, based in part on the success of this program, were provided by the Australian Department of Employment and Education Training to establish the Centre for Teaching Development that provided a permanent home for this program.) Because there was no campus-wide SET program at UWSM prior to the initiation of this project, the formal collection of SETs using a standardized form was unfamiliar to many staff and students, although many teachers used a variety of informal means to evaluate their teaching effectiveness.

Sample and Procedures.

Teachers were recruited to participate in the study through a variety of sources including letters sent to each teacher, brief descriptions of the study in a university newletter, and presentations to faculty staff meetings. The final sample of 92 teachers who volunteered to participate in the study represented all the UWSM faculties and all the different academic ranks. Reasons for volunteering included the desire to improve teaching effectiveness, to formally evaluate teaching effectiveness for purposes of personnel decisions, to support a good cause, or simply to satisfy curiosity about the program. Not surprisingly, given the variety of reasons for participation, students subsequently judged these teachers to differ widely in terms of their initial teaching effectiveness.

Prior to the initiation of the study, teachers were told that they would be asked to evaluate their own teaching effectiveness and the relative importance of different components of teaching effectiveness, and to be evaluated by their own students in the middle of the first semester (T1), at the end of the first semester (T2), and at the end of the second semester (T3). Prospective volunteers were told that the confidentiality of all individual responses would be strictly maintained and that results of the ratings for each teacher would only be sent to the individual teacher. They were also told that randomly selected teachers would be asked to participate in a feedback/consultation program in which a consultant (one of the authors) would meet with the teacher to discuss the results of the SETs and strategies to improve ratings in areas selected by the teacher, whereas other teachers randomly assigned to the control group would not receive any feedback until the end of the second semester after the completion of the study. Prospective participants were asked to nominate two instructional sequences -- typically separate classes but occasionally an independent component that was part of a larger program -- in which to be evaluated in the first and second semester. Although all teachers were encouraged to nominate similar settings in which to be evaluated, this was not always possible and no one was excluded from the study for this reason.

At the middle of the first semester 92 teachers volunteered to be in the study, completed a self-evaluation survey, and were evaluated by students. At T1, T2, and T3 the ratings were collected by the teacher or a nominated student form his/her class. Standardized administration instructions were read aloud to students, students completed the ASEEQ forms, and forms were put into a sealed envelope that were returned to the faculty office. The completed forms were subsequently sent to the principal investigator of the study to be processed. Although teachers were encouraged to collect ET ratings during the last week or two of regularly scheduled classes, teachers selected when the ratings were actually collected. (Not all teaching sequences that were evaluated corresponded to university calendar.)

Participants were stratified on the basis of overall teacher ratings by students at T1 and were randomly assigned to the MT feedback group, the ET feedback group, or the control group. MT teachers were immediately sent relevant materials (ASEEQ instruments completed by their students, a computerized summary sheet, and a guide for interpreting the ratings) and were contacted to set up individual feedback/consultation sessions. All other participants were merely told that they had not been selected to be in the MT group (i.e., they were not told whether they were in the control or ET group). Similarly, at the end of the first semester, ET teachers were contacted to set up their feedback/consultation, and MT teachers were contacted to set up their second feedback/consultation. Finally, at the end of the second semester, all previously unreturned materials were returned to all participants -- including the control teachers.

The feedback/consultation protocol used for MT and ET groups was based substantially on earlier work by Wilson (1986). Each session began by the consultant providing a general overview and specifically stating that "I do not have sufficient background in your area to know what is 'best.' Instead, I will discuss the ratings and work with you to develop some strategies in particular areas selected by you. In this sense, my role is to be a facilitator." The teacher was then asked to describe the special or unique characteristics about the class being evaluated, the students, or the circumstances. The consultant ascertained that the teacher had read the materials previously sent to him/her. Focusing on the ASEEQ scale scores rather than responses to individual items, the consultant first emphasized the ASEEQ areas of relative strength and then noted areas in which the ratings were relatively lower. Student written comments were then examined for themes and relevant information. The consultant then suggested that the teacher select 2 or 3 ASEEQ dimensions that were important to the teacher (based on responses to the self-evaluation instrument that had previously been completed by the teacher) and that had received relatively lower ratings by students (based on decile ranks that compared the ratings by that teacher with those of all other teachers in the study -- noting potential limitations in these normative comparisons). In some cases the consultant suggested ASEEQ dimensions that satisfied these criteria, but the final selection of target areas was always made by the teacher. As part of the process, the consultant asked the teacher if "these are appropriate areas to target improvement efforts." The consultant and teacher then considered the ratings of individual ASEEQ items in each targeted dimension and the student written comments relevant to the targeted areas.

The consultant then introduced the teaching idea packet relevant for each ASEEQ dimension that had been targeted. There were between 17 and 32 suggested strategies for each ASEEQ dimension that were largely adapted from materials developed by Wilson (1986). The consultant noted that each strategy was only a potential suggestion and that some would be inappropriate in a particular situation, but that some strategies -- or derivations of them -- might be appropriate for the teacher to test out. The teacher read the suggested strategies, discussed how they might be applied in his/her situation, and noted the strategies (or variations thereof) that he/she would pursue. The teacher then recorded the particular strategy and any variations in his/her copy of the teaching idea packet. In concluding the session, the consultant noted the ASEEQ areas selected by the teacher that would be the focus of the intervention and the strategies selected for this purpose. The teacher was asked if he/she felt that they would be able to carry out the suggestions and whether the strategies were likely to lead to improvement. In closing, the consultant noted that he would send a brief letter summarizing the feedback/consultation session (particularly the targeted areas and selected strategies) in 2 weeks and in 4 weeks would telephone the teacher to check on progress. All materials, including the teaching idea packets for all (targeted and nontargeted) ASEEQ dimensions, were left with the teacher and the teacher was encouraged to telephone the consultant if he/she subsequently wanted to discuss any aspect of the study.

Statistical Analyses.

An important problem in any applied field research -- particularly a longitudinal study involving multiple waves of data collection -- is how to deal with missing data. Of the 92 teachers who began the study, a total of 9 had missing SETs for at least one of the three waves; 5 in the MT group, 3 in the control group, and 1 in the ET group. The reasons for the missing data were that the teacher was not teaching any instructional sequence at either T2 or T3 (reflecting a change in scheduling or a misunderstanding of the requirements for participation in the study), was unable to allocate class time for the administration of the SETs, or forgot to administer the forms despite intending to do so. In order to facilitate analyses and presentation, all results presented in the results section are based on the 83 teachers with complete data.

Supplemental analyses were conducted for teachers with partially complete data. The 9 teachers with some missing data did not differ significantly from the remaining 83 teachers on T1 ratings that were complete for all participating teachers. Four of the 5 MT teachers and 2 of the 3 control teachers with missing data had T2 or T3 responses, making possible some comparisons between the MT and control groups. Each of these comparisons was pursued in unreported analyses, but did not differ from the results that are presented in the results section in terms of effects being statistically significant or nonsignificant. Effect sizes in these supplemental analyses were also similar to those subsequently reported in the results section. Similarly, comparisons of ET and control groups were possible for teachers who had missing T2 data by using T1 responses -- instead of the average of T1 and T2 responses -- as the pretest covariate. Again, however, none of the differences based on these unreported analyses differed from those reported in the analyses in terms of being significant or nonsignificant, and the effect sizes were similar to those subsequently reported.

An interesting feature of the present investigation is that teachers completed self-evaluation surveys that included ratings of the importance of each ASEEQ area. These importance ratings were, of course, a central component of the intervention process. In addition, however, we used the importance ratings to construct importance weighted total scores based on the 8 ASEEQ dimensions. Following procedures described by Marsh (1986) the importance ratings were "ipsatized." For each teacher the mean importance rating for the 8 ASEEQ dimensions was computed and then each individual importance rating was divided by this mean. This resulted in a set of 8 "ipsatized" scores that had a mean of exactly 1.0 for each individual teacher. This provided an index of the relative importance of each ASEEQ factor -- relative to the importance ratings assigned by the same teacher to other ASEEQ factors. The importance weighted total score was then computed by taking the mean crossproduct of each ipsatized importance rating multiplied by the student rating of the corresponding ASEEQ factor. Separate weighted averages were computed using the importance ratings and SETs at T1, T2, and T3 so long as there were no missing values. If SETs were missing, the weighted average total was deemed to be missing (and these teachers were excluded from the final analyses). If the importance ratings were missing, however, the relative importance of each ASEEQ factor was determined by taking the average importance assigned to that ASEEQ factor on the remaining self-evaluation surveys. Because all teachers completed the self-evaluation survey at least once, this procedure allowed us to compute weighted averages for all teachers included in the final analyses. For purposes of comparison, the corresponding unweighted average of the 8 ASEEQ scales were also computed.

Although a wide variety of analytic techniques are appropriate, a multiple regression (general linear model) approach to analysis of variance (ANOVA) was selected because of its flexibility. In the general analytic strategy, each outcome (SETs at T2 or T3 depending on the particular comparison) was related to a dichotomous grouping variable (ET vs. control or MT vs control), a pretest covariate (SETs at T1 or the average of T1 and T2 responses depending on the comparison), and the group x covariate crossproduct reflecting the aptitude-treatment interaction. In order to facilitate comparisons, all independent and dependent variables were first standardized (mean = 0, SD = 1), crossproduct terms were based on products of z-scores (but were not subsequently standardized), and results are presented in terms of unstandardized beta weights. Except for the interaction terms these results are exactly the same as the standardized beta weights resulting from the analysis of untransformed independent and dependent variables that are typically easier to interpret than unstandardized beta weights. The standardized beta weights for interaction terms, however, are not generally comparable to those based on nonproduct terms. The procedure used here is an effective compromise in which all effects -- including interaction terms -- are appropriately presented in relation to the standard deviations of the underlying variables (see Aiken & West, 1991, for more discussion). In subsequent analyses, a traditional repeated measures ANOVA was used to compare relative changes in the ratings of targeted : nd nontargeted ASEEQ factors over time. Because of potential problems related to the "sphericity" assumption in repeated measures analysis, all tests of statistical significance were conducted using the Greenhouse-Geisser epsilon and the Hyuynh-Feldt epsilon correction factors (SPSS, 1991). Because both these estimates of epsilon were close to 1.0, there were no differences in effects judged to be statistically significant using either of these approaches or the traditional (uncorrected) approach.

Preliminary analyses.

Factor analysis. Particularly because the ASEEQ has not been used previously, it is important to demonstrate that SEEQ dimensions identified in North American settings (e.g., Marsh, 1987; Marsh & Hocevar, 1991) can be replicated. As argued elsewhere (e.g., Marsh, 1987), the most appropriate unit of analysis for such factor analyses is the class average. In order to obtain as large a sample as possible, all classes evaluated with ASEEQ were included -- those formally considered in the feedback intervention and those that were not. Also, each set of evaluations for each teacher -- those based on T1, T2, and T3 responses were considered as separate cases. Hence, the evaluations were based on a total of 305 sets of ratings of 118 different teachers (92 of whom also participated in the feedback/consultation intervention). The factor analysis -- as in earlier SEEQ research -- consisted of a principal axis factor extraction, following a Kaiser normalization, and an oblique rotation using the commercially available SPSS (1991) procedure. The 32 "target loadings" (see Table 1)., the factor loadings of items designed to measure each factor are consistently high (median loading = .64) and none is less than .37. The remaining 256 "nontarget loadings" are consistently much smaller (median loading = .07) and none is greater than .38. Not surprisingly, the remaining 27 factor loadings associated with the 3 overall rating items tend to be substantial for several different SEEQ factors -- particular the Teacher Enthusiasm, Organization, Learning/value, Assignments, and Individual Rapport factors. In summary, the results of this factor analysis demonstrate a clear "simple structure" that is consistent with previous SEEQ research.

Insert Table 1 About Here

Although not presented, separate factor analyses were conducted for classes evaluated at T1, T2, and T3. Particularly, the evaluations collected at the end of each semester (T2 and T3) provided very good solutions -- slightly better, perhaps, than the one based on all three times (Table 1). The factor solution based on T1 (midterm) ratings, was not quite so clean. It is not clear whether this was because some items could not be adequately evaluated at this time (e.g., some students responded "not appropriate" to items about examinations and assignments) or because many students had not previously completed a SET instrument. Whereas the factor analyses generally support the a priori factor structure, there may be some evidence in support of the Marsh and Overall (1980) suggestion that MT SETs may not be as valid as ET SETS.

Reliability. The reliability of SETs is most appropriately evaluated in studies of interrater agreement (i.e., agreement among ratings by different students in the same class; for further discussion see Feldman, 1977, Gilmore, Kane, & Naccarato, 1978, Marsh, 1987). The correlation between responses by any two students in the same class (i.e., the single rater reliability) is typically in the .20s but the reliability of the class-average response depends upon the number of students rating the class. For example, the estimated reliability for SEEQ factors (Marsh, 1987) is about .95 for the average response from 50 students, .90 from 25 students, .74 from 10 students, .60 from five students, and only .23 for one student. Given a sufficient number of students, the reliability of class-average SETs compares favorably with that of the best objective tests.

For present purposes the intra-class correlation was used to assess the reliability of class-average responses to each ASEEQ item and each ASEEQ scale. This was accomplished with a oneway ANOVA that divides variability in individual student scores into within-class and between-class components. If class-average differences are no larger than expected by chance (i.e., the F-ratio is 1.0), then the reliability of class-average scores is -- by definition -- zero. Estimates of reliability for class-average scores are higher when there are larger differences between classes, smaller differences within classes, and larger numbers of students within each class. The reliability should be higher when the average number of students in each class is larger (all other things being equal -- always a worrisome assumption). Based on results from the total sample of 305 classes, the median interrater reliability is .89 (Table 2) for an average class size of 23 students and is comparable to the median of .90 for an average class size of 25 student reported in earlier (North American) SEEQ research. Reliability estimates of ASEEQ scale scores are consistently somewhat higher than the items that comprise the scales. Consistent with expectations, the reliability estimates vary systematically with class size (median estimates are .80, .83, .86, and .96 for groups of classes in which the average class sizes are 10, 16, 21 and 48). In summary, these results demonstrate that responses to ASEEQ are reliable for classes of even moderate size and are consistent with previous SEEQ research conducted in North America.

Insert Table 2 About Here

The SEEQ Workload factor and responses to the experimental item asking students to rank overall teaching effectiveness in relation to a hypothetical "representative sample" of 100 teachers (on a 1 to 100 scale) are not considered further. The Workload factor was treated as a background factor (see Marsh, 1983, 1987); information was presented to teachers and discussed as part of the feedback/consultation -- along with other background information such as expected grades, class size, etc -- but it was not specifically considered as a target dimension and there were no "strategies" for this area. Also, interpretations of the workload ratings are complicated in that scores vary along a nonlinear scale in which some intermediate value is optimal (i.e., a class that is neither too easy nor too difficult). The experimental overall teacher "ranking" (Q32 in Tables 1 and 2) was included in an unsuccessful attempt to counter the typical negative skew in SETs. Also, some students apparently misunderstood the 1-100 response scale and responded on the 1-9 scale used for the other SEEQ items. For purposes of the preliminary analyses presented here, values between 1 and 9 were multiplied by 10. Also, whereas the extended response scale was intended to produce more reliable responses, reliability analyses in Table 2 indicate that it is slightly less reliable than the traditional overall teacher rating (Q31).

Results

In order to facilitate presentation of the results, separate analyses of the effects of ET and MT feedback are presented. We begin with the ET results in which the analyses are summarized most easily, and then move to the more complicated analyses based on the MT feedback. Finally, we compare results based on those ASEEQ scales that teachers specifically selected to target for purposes of the intervention with those based on the remaining nontargeted areas.

End of Term Feedback

A multiple regression approach to analysis of covariance (Table 3) was used to assess differences between ET and control group ratings at the end of the second semester (T3). For purposes of this analysis. each T3 ASEEQ score was related to its covariate (the mean of the corresponding T1 and T2 score after standardizing each), a group contrast variable (ET vs. control), and their interaction. Not surprisingly, the effects of the covariate were substantial (betas of .5 to .7) indicating that SET ratings are stable over time (i.e., semester 1 to semester 2). Of central importance to the present investigation, the ET feedback group has higher ratings for all 12 ASEEQ scores and 8 of these differences are statistically significant (see group effects in Table 3). Only one of the covariate x group interactions is statistically significant, indicating that the generally positive effects of the intervention generalize reasonably well across teachers with initial differences in their teaching effectiveness (i.e., there were no aptitude-treatment interactions). The one statistically significant interaction -- as well as the largest of the nonsignifiant interactions -- suggested that initially less effective teachers according to T1 responses benefited most from the intervention.

Insert Table 3 About Here

In interpreting these ET feedback results, it is important to note that the use of the overall ratings may be more defensible to use than ratings of the specific ASEEQ scales. Because each teacher in the ET group targeted only a few (typically 2 or 3) of the ASEEQ dimensions, the experimental group means for specific ASEEQ scales includes ratings by teachers -- typically a majority of the teachers -- who did not target that specific dimension. Thus, it is not surprising that the differences are apparently smaller and sometimes do not reach statistical significance for the specific ASEEQ dimensions. Whereas it would be possible to base comparisons on only those experimental teachers who targeted each dimension, this subset of self-selected teachers no longer constitutes a randomly assigned group so that comparisons with the control group may be dubious. (This issue of the distinction between targeted and nontargeted dimensions is addressed in subsequent analyses). It is, however, reasonable to expect the intervention to significantly influence overall ratings and total scores no matter what specific ASEEQ dimensions were targeted, and there were statistically significant effects for all 4 summary ratings. Whereas it was anticipated the effects would be larger for the importance weighted total scores (Table 3), the effect sizes for the different summary scores are reasonably similar (i.e., the effect size, d statistic varies between .4 and .5).

Mid-Term Feedback

A multiple regression approach similar to that used with the ET intervention was used to evaluate the MT intervention. The intervention, however, is complicated by the fact that both T2 (end of first semester) and T3 (end of second semester) ratings are outcome measures. As in the typical (mid-term) feedback study, the T2 ratings provide a basis for evaluating the short-term, immediate effects of the intervention administered in the middle of the first semester. Because MT feedback teachers received an additional feedback/consultation at the end of the first semester, the T3 scores provide a basis for evaluating the continued and cumulative effectiveness of the intervention process.

Insert Table 4 About Here

Each T2 and T3 ASEEQ score was related to the effects of the covariate (the corresponding T1 score), a group contrast (MT vs. control group), and their interaction. The T2 (Table 4) results indicate that none of the group differences are statistically significant and that this lack of difference does not interact with initial levels of teaching effectiveness (i.e., the group x covariate interactions were nonsignificant). The T3 results also indicate that none of the group differences are statistically significant. In these analyses, however, 4 of the 12 group x covariate interactions are statistically significant. As with the ET group comparisons, the nature of these interactions (as well as the nonsignificant interactions that approach statistical significance) indicate that the intervention is more beneficial for the initially less effective teachers. Thus, whereas evidence for the effectiveness of the MT intervention is weak, there is some support that it works with teachers who are initially less effective.

SETs in Targeted and NonTargeted ASEEQ Dimensions

The distinction between targeted and nontargeted ASEEQ dimensions is a critical feature of the intervention that has not been adequately captured in the analyses presented thus far. For any particular ASEEQ dimension, the so-called intervention effect was based on results of some teachers who actually targeted that dimension but the majority of the experimental group did not (i.e., they selected other dimensions to target). In this respect, results for the specific dimensions presented thus far may not give an adequate representation of the intervention effect. In contrast to the ratings of specific ASEEQ dimensions, the overall ratings -- particularly the overall teacher rating -- and total scores provide a fairer representation of the intervention effects in that all teachers in the experimental groups attempted to enhance their overall teaching effectiveness. Even these summary scores, however, do not adequately represent the multidimensional emphasis in previous SEEQ research that was the basis of this intervention.

Unfortunately, there appears to be no fully satisfactory approach to the analysis of the target/nontarget ratings. Whereas is would be possible to compare ratings of experimental teachers who did and did not target a specific ASEEQ dimension and to compare these with those of the control group, interpretations of these results would be dubious. Because each experimental teacher typically targeted only 2 or 3 of the 8 ASEEQ dimensions, such comparisons would be based on small samples. More importantly, the self-selected group of teachers selecting any one ASEEQ dimension is clearly not a "random" sample of teachers. Thus any observed group differences confound the effects of the intervention with initial group differences. Furthermore, to the extent that teachers initially selected ASEEQ scales on which they initially had the poorest ratings, apparent gains in these scales relative to teachers who did not target these scales and the control group would be expected on the basis of regression to the mean.

The approach used here is to consider 6 scores for each teacher: the mean of targeted and nontargeted ASEEQ dimensions at T1, T2, and T3. In this sense, the nontargeted dimensions form one basis of comparison for evaluating changes in the targeted dimensions that most accurately reflect the intervention effects. Whereas teachers in the control group did not actually target any dimensions, the targeted dimensions for experimental groups usually consisted of ASEEQ scales that were relatively high in importance (as perceived by the teacher) and relatively low in terms of SETs at T1. Using these criteria, we selected ASEEQ factors that we would have recommended to be targeted by control teachers. Although not totally satisfactory, this approach provides a basis for comparing differences in the ratings of targeted and nontargeted dimensions in the three groups (Figure 1) that provides an apparently defensible control for regression effects. A preliminary inspection of Figure 1 reveals that targeted dimensions are rated substantially lower than nontargeted dimensions for all three groups at T1 and T2. At T3, targeted dimensions are still rated substantially lower than targeted dimensions in the control group. In the two experimental groups, however, ratings of the targeted dimensions are marginally better than those of the nontargeted dimensions at T3. Over the course of the study, ratings of targeted dimensions improved substantially relative to nontargeted areas for both experimental groups, but not for the control group.

Insert Table 5 and Figure 1 About Here

In order to evaluate the statistical significance of these apparent effects, a 3 group (MT, ET, control) x 3 time (T1, T2, T3) x 2 target (target, nontarget) analysis of variance was conducted in which time and target are within-subject factors (repeated measures factors) and group is a between-subject factor. The results (Table 5) demonstrate significant main effects of time and target, and a significant time x target interaction. Overall, ratings went up over time, nontarget ratings were lower than target ratings, and the target/nontarget difference changed over time. The critical effect for present purposes, however, is the statistically significant time x target x group interaction. In order to more fully evaluate the nature of this interaction, a polynomial contrast was applied to the time variable and a "simple" (SPSS, 1991) contrast was applied to the group variable that contrasted the control group to each of the experimental groups. Consistent with the observation that target/nontarget differences for the three groups changed at T3, both group x nonlinear time x target interactions are statistically significant. These results, then indicate that the intervention effects had more effect on ASEEQ dimensions that were targeted for intervention than for nontargeted dimensions.

Discussion

The most important results of the present investigation were to provide varying degrees of support for a priori predictions that feedback from ASEEQ coupled with Wilson's (1986) feedback/consultation provide an effective means of improving university teaching, that the benefits are stronger for the initially least effective teachers, that improvement is largest for the specific areas each teacher targeted as the focus of the intervention, and that the effects of end-of-term feedback are stronger than those based on midterm ratings. In addition, we replicated the strong psychometric properties reported in North American SEEQ research with the Australian ASEEQ and demonstrated the use of the weighted average total score (based on teacher importance ratings) proposed in Marsh & Dunkin (in press) and Marsh and Bailey (in press).

Previous research by Wilson (1986) demonstrated the application of his feedback/consultation process as part of an on-going SET program. Results based on his nonexperimental design suggested its effectiveness, but there were numerous threats to the validity of the interpretations. Thus, a potentially important contribution of the present investigation is to provide one paradigm for evaluating this intervention as well as showing that it was effective when assessed with a more rigorous experimental design. Of particular relevance was the demonstration that teachers in the intervention group demonstrated significantly more improvement in the specific ASEEQ dimensions that they targeted for purposes of the intervention. This finding supports the construct validity of interpretations of the intervention and supports the importance of asking teachers to specifically target particular dimensions. In this way the intervention is individualized to the needs of each teacher and may provide teachers with a stronger commitment to improving their effectiveness in areas of particular relevance to them.

A particularly important -- and apparently unique -- feature of the present investigation is the comparison of the MT and ET feedback interventions. Whereas nearly all SET feedback research is based on MT feedback, reviewers (e.g., L'Hommediu, et al., 1990; Marsh, 1987; Marsh & Dunkin, in press) have questioned the implicit assumptions that effects based on MT feedback generalize to ET feedback and, apparently, that MT feedback is more effective. These reviewers noted a number of concerns with the MT feedback that apparently detract from its effectiveness, but did not identify any research that actually compared MT and ET feedback effects. In our study the effects of ET feedback effects were stronger than those of MT feedback. Despite the fact that teachers in the MT group received the intervention at both middle of term 1 and -- like the ET teachers -- at the end of term 1, their improved teaching effectives was weaker than that of the ET teachers who only received the intervention at the end of term 1. Furthermore, the modest improvements for the MT group -- those for targeted as opposed to nontargeted ASEEQ dimensions and those for the initially less effective teachers -- were observed following the end-of-term intervention and not the mid-term intervention. There are, however, some features about the present investigation that may influence the generalizability of these results. In particular, the MT feedback may have been less effective in the present investigation for a variety of reasons that are idiosyncratic to this study. In particular, this study was the first time that many teachers and students had participated in a broadly based SET program using a standardized SET instrument. Also, there were more "not appropriate" responses by students and teachers for MT ratings than ET ratings, particularly in areas such as assignments and examinations. This suggests that MT feedback was perceived to be and may well in fact have been less appropriate, thus undermining confidence in the intervention for the MT group. Finally, even though the intervention was administered in relation to a standardized protocol, it may be that consultants were more effective at delivering the intervention at T2 as a consequence of previously administering it at T1. Thus, an important direction for future research is to test the generalizability of the apparent superiority of the ET feedback compared to MT feedback.

It is also important to evaluate reasons why the intervention effects in the present investigation were no larger than they were. Despite an apparently stronger intervention than typically employed, the effects (e.g., effect size of d = .50 for overall teacher ratings for the ET group) were only somewhat larger than the average effect size of .34 reported in earlier meta-analyses. We suspect that the novelty of the SET program in this university and a lack of familiarity with SETs by both students and teachers detracted from the intervention's effectiveness. In particular, the John Henry effect identified by L'Hommediu, et al., 1990 as apparently being inevitable in all SET feedback studies was likely to be even stronger in the present investigation because of the novelty of the program and the fact that control teachers actually completed self-evaluation instruments that forced them to scrutinize their teaching effectiveness more than would be the case in a true no-treatment control. Consistent with these suggestions, ratings for the control group -- as well as the experimental group -- actually increased from T1 to T2, and from T2 to T3 even though other research has found that ratings tend to decrease between midterm and end-of-term (e.g., Overall & Marsh, 1980) and over time generally (e.g., Feldman, 1983). Coupled with this is the observation that at least one of the reasons for participating in the study was the desire to obtain standardized SETs that would be beneficial to include in applications for promotion. Thus, we suspect that the act of volunteering to participate in the program, completing self-evaluation instruments, administering the ASEEQ forms, and trying to get obtain positive SETs that would support promotions may have led to improved teaching effectiveness of control teachers that detracted from the size of experimental/control comparisons.

Wilson (1986) noted critical features of his intervention that contributed to its effectiveness were the availability of concrete strategies to facilitate efforts to improve teaching effectiveness in relatively less effective areas that the teacher perceived to be important, the facilitator role adopted by the consultant in this intervention, and the personal commitment obtained from the teacher that was facilitated by the face-to-face interaction between teacher and consultant. Based on our experience we concur that these are important components of the intervention. To this list of critical features we add the multidimensional perspective embodied in Wilson's intervention, the multidimensional SET instruments used by Wilson (Hildebrand, Wilson, & Dienst, 1971) and the ASEEQ instrument used here, and previous SEEQ research on the construct validity of multidimensional SET responses. Fundamental assumptions underlying the logic of the intervention are that teaching effectiveness and SETs are multidimensional, that teachers vary in their effectiveness in different SET areas as well perceptions of the relative importance of the different areas, and that feedback specific to particular SET dimensions is more useful than feedback based on overall or total ratings or that provided by SET instruments that do not embody this multidimensional perspective. In this respect, the results of the present investigation contribute to the growing body of research supporting the conclusion that SETs should be considered from a multidimensional perspective.

REFERENCES

Abrami, P. C. (1989). SEEQing the truth about student ratings of instruction. Educational Researcher 43: 43-45.

Abrami, P. C., and d'Apollonia, S. (1991). Multidimensional students' evaluations of teaching effectiveness: Generalizability of N = 1 research: Comment on Marsh (1991). Journal of Educational Psychology, 30, 221-227.

Aiken, L. S., & West, S. G. (1991). Multiple regression: Testing and interpreting interactions. Newburry Park CA: Sage.

Aleamoni, L.M., and Yimer, M. (1973). An investigation of the relationship between colleague rating, student rating, research productivity, and academic rank in rating instructional effectiveness. Journal of Educational Psychology 64: 274-277.

Braskamp, L. A., Brandenburg, D. C. and Ory, J. C. (1985). Evaluating teaching effectiveness: A practical guide. Beverly Hills, CA: Sage.

Brinko, K. T. (1987). The interactions of teaching improvement. In M. Theall and J. Franklin (ed.), Effective practices for improving teaching (pp. 39-49). New Directions for Teaching and Learning, no. 48. San Francisco, CA: Jossey-Bass.

Cohen, P. A. (1980). Effectiveness of student-rating feedback for improving college instruction: a meta-analysis. Research in Higher Education 13: 321-341.

Feldman, K. A. (1977). Consistency and variability among college students in rating their teachers and courses. Research in Higher Education 6: 223-274.

Feldman, K. A. (1983). The seniority and instructional experience of college teachers as related to the evaluations they receive from their students. Research in Higher Education 18: 3-124.

Gilmore, G. M., Kane, M. T., and Naccarato, R. W. (1978). The generalizability of student ratings of instruction: Estimates of teacher and course components. Journal of Educational Measurement 15: 1-13.

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

Hoyt, D. P., Owens, R. E., and Grouling, T. (1973). Interpreting student feedback on instruction and courses. Manhattan, KN: Kansas State University.

Jacobs, L. C. (1987). University faculty and students' opinions of student ratings. Bloomington IN: Bureau of Evaluative Studies and Testing. (ERIC Document Reproduction Service No. ED 291 291).

L'Hommedieu, R., Menges, R. J., and Brinko, K. T. (1988). The effects of student ratings feedback to college teachers: A meta-analysis and review of research. Unpublished manuscript, Northwestern University, Center for the Teaching Professions, Evanston, IL.

L'Hommedieu, R., Menges, R. J., and Brinko, K. T. (1990). Methodological explanations for the modest effects of feedback. Journal of Educational Psychology 82: 232-241.

Marsh, H. W. (1981). Students' evaluations of tertiary instruction: Testing the applicability of American surveys in an Australian setting. Australian Journal of Education 25: 177-192.

Marsh, H. W. (1982a). SEEQ: A reliable, valid, and useful instrument for collecting students' evaluations of university teaching. British Journal of Educational Psychology 52: 77-95.

Marsh, H. W. (1982b). Validity of students' evaluations of college teaching: A multitrait-multimethod analysis. Journal of Educational Psychology 74: 264-279.

Marsh, H. W. (1983). Multidimensional ratings of teaching effectiveness by students from different academic settings and their relation to student/course/instructor characteristics. Journal of Educational Psychology 75: 150-166.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, Reliability, Validity, Potential Biases, and Utility. Journal of Educational Psychology 76: 707-754.

Marsh, H. W. (1986). Global self-esteem: Its relation to specific facets of self-concept and their importance. Journal of Personality and Social Psychology, 51, 1224-1236.

Marsh, H. W. (1987). Students' evaluations of university teaching: Research findings, methodological issues, and directions for future research International Journal of Educational Research 11: 253-388. (Whole Issue No. 3)

Marsh, H. W. (1991a). A multidimensional perspective on students' evaluations of teaching effectiveness: A reply to Abrami and d'Apollonia (1991). Journal of Educational Psychology, 83, 416-421.

Marsh, H. W. (1991b). Multidimensional students' evaluations of teaching effectiveness: A test of alternative higher-order structures. Journal of Educational Psychology, 83, 285-296.

Marsh, H. W., & Bailey, M. (in press). Multidimensionality of students' evaluations of teaching effectiveness: A profile analysis. Journal of Higher Education.

Marsh, H. W., & Dunkin, M. (in press). Students' evaluations of university teaching: A multidimensional perspective. Higher education: Handbook on theory and research (vol. 9). New York: Agathon.

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

Marsh, H. W., and Hocevar, D. (1984). The factorial invariance of students' evaluations of college teaching. American Educational Research Journal 21: 341-366.

Marsh, H. W., and Hocevar, D. (1991). The multidimensionality of students' evaluations of teaching effectiveness: The generality of factor structures across academic discipline, instructor level, and course level. Teaching and Teacher Education 7: 9-18.

Marsh, H. W. and Overall, J. U. (1980). Validity of students' evaluations of teaching effectiveness: Cognitive and affective criteria. Journal of Educational Psychology 72: 468-475.

Marsh, H. W., Overall, J. U., and Kesler, S. P. (1979). Validity of students' evaluations of instructional effectiveness: A comparison of faculty self- evaluations and evaluations by their students. Journal of Educational Psychology 71: 149-160.

Marsh, H. W. and Roche, L. (in press). The use of students' evaluations of university instructors in different settings: The applicability paradigm. Australian Journal of Education.

McKeachie, W.J., Lin, Y-G, Daugherty, M., Moffett, M.M., Neigler, C., Nork, J., Walz, M., and Baldwin, R. (1980). Using student ratings and consultation to improve instruction. British Journal of Educational Psychology 50: 168-174.

Murray, H. G. (April, 1987). Impact of student instructions ratings on quality of teaching in higher education. Paper presented at the 1987 Annual Meeting of the American Educational Research Association, Washington, DC. (ERIC Document Reproduction Service No. ED 284 495).

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

SPSS (1991). SPSS user's guide. Chicago: SPSS, Inc.

Voght, K.E. and Lasher, H. (1973). Does student evaluation stimulate improved teaching? Bowling Green, OH: Bowling Green University (ERIC ED 013 371)

Wilson, R. C. (1984). Using consultation to improve teaching. (ERIC document number ED 242 271).

Wilson, R. C. (1986). Improving faculty teaching: Effective use of student evaluations and consultants. Journal of Higher Education 57: 196-211.

Wilson, R. C. (1987). Toward excellence in teaching. In L. M. Aleamoni (ed.), Technique for evaluating and improving instruction. New Direction for Teaching and Learning, no. 31. San Francisco, CA: Jossey-Bass.

Figure 1. Targeted and Nontargeted Dimensions: The Differential Improvement Over Time For Instructors in the Midterm Intervention, End-of-term Intervention, and Control Groups (also see Table 5).

Figure 1. Targeted and Nontargeted Dimensions: The Differential Improvement Over Time For Instructors in the Midterm Intervention, End-of-term Intervention, and Control Groups (also see Table 5).

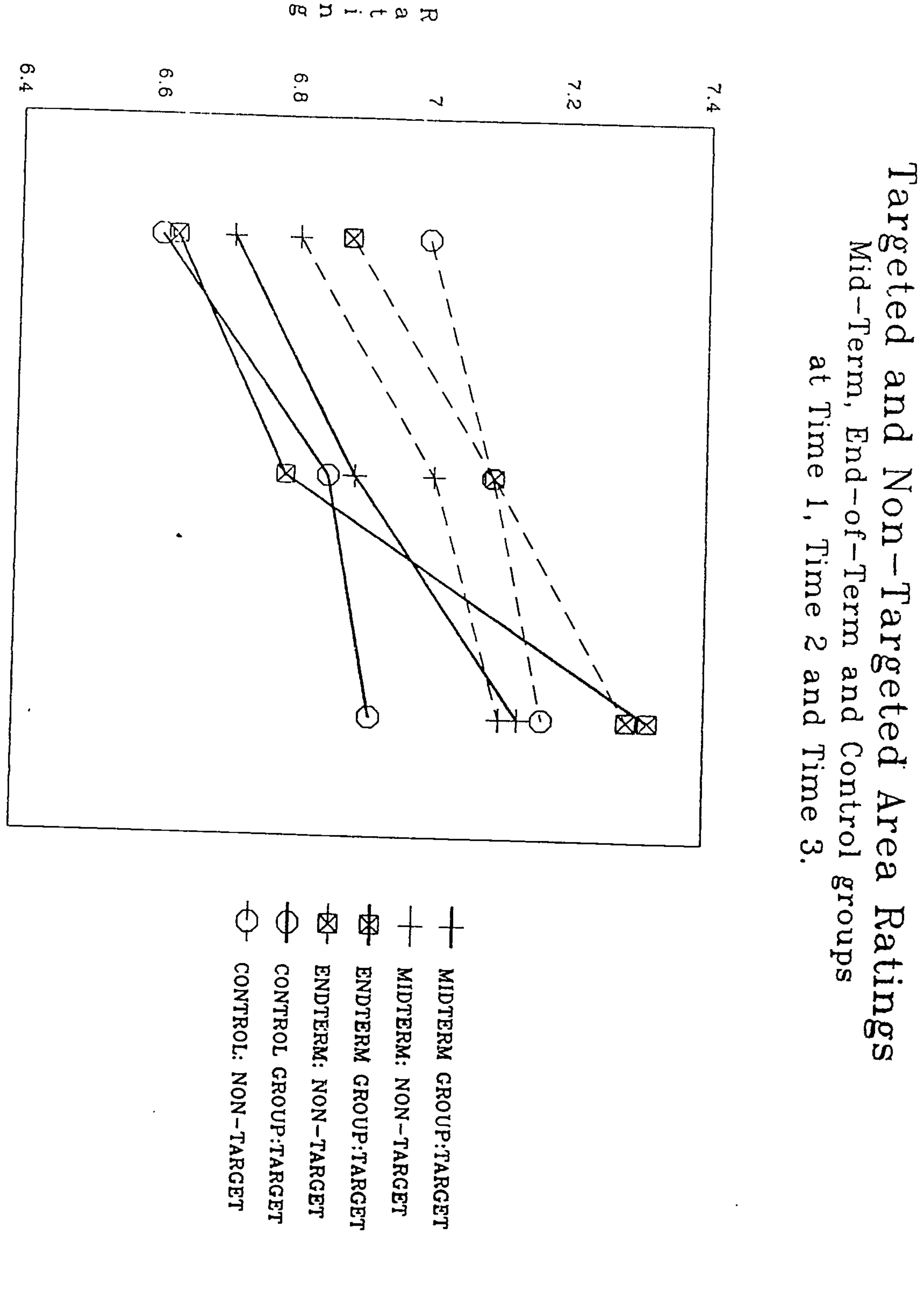

Table 1
Summary of Factor Analysis of ASEEQ responses

	Factor Pattern Loadings										
	Learn	Enthu	Organ	Group	Indiv	Brdth	Exams	Assgn	Work	Comm	MSA
Learning											
Q1	.373	.181	.217	.099	.024	.172	.046	.201	.176	.840	.969
Q2	.668	-.005	.128	.070	.027	.133	.045	.123	.051	.857	.954
Q3	.654	.105	.067	.052	.069	.119	.047	.181	.031	.932	.964
Q4	.496	.001	.154	.122	.070	.003	.142	.114	-.352	.838	.957
Enthusiasm											
Q5	.107	.539	.096	.070	.240	.139	.045	.060	-.081	.870	.946
Q6	.109	.640	.095	.090	.135	.141	.089	.039	-.040	.941	.946
Q7	.156	.590	.032	.120	.140	.170	.120	-.020	-.026	.885	.968
Q8	.160	.565	.222	.130	.087	.087	.078	.049	-.025	.955	.965
Organization											
Q9	.152	.208	.516	.132	.077	.065	.079	.078	-.139	.888	.973
Q10	.086	.058	.640	.062	.104	.132	.087	.088	-.110	.903	.972
Q11	.183	.026	.502	.040	.099	.137	.131	.186	-.032	.881	.975
Q12	.116	-.048	.557	-.097	-.070	.216	.099	.098	.206	.650	.952
Group Interaction											
Q13	.100	.063	-.003	.755	.017	.148	.101	.064	.025	.923	.934
Q14	.095	.034	-.014	.776	.066	.137	.074	.103	-.002	.977	.940
Q15	.105	.075	.184	.543	.214	.065	.099	.122	.047	.930	.968
Q16	.086	-.043	.074	.653	.222	.084	.080	.099	.000	.948	.951
Individual Rapport											
Q17	.073	.183	.010	.112	.671	.040	.084	.107	-.029	.918	.949
Q18	.123	.107	.056	.149	.637	.072	.095	.082	.015	.921	.947
Q19	.018	.077	.070	.075	.729	.090	.125	.097	.030	.945	.951
Q20	.069	.053	.064	.008	.474	.289	.174	.087	.055	.670	.957
Breadth of Coverage											
Q21	.011	.019	.230	.062	.039	.606	.067	.175	.044	.855	.949
Q22	.077	.098	.105	.092	-.001	.681	.051	.155	-.014	.923	.959
Q23	.049	.124	.082	.157	.072	.588	.083	.157	-.067	.898	.974
Q24	.380	.015	.033	.073	.115	.472	.085	.006	-.089	.798	.949
Exams/Grading											
Q25	.007	.019	.087	.081	.028	.083	.749	.106	.027	.844	.968
Q26	-.015	.021	.121	.101	.106	.022	.718	.115	-.058	.853	.958
Q27	.109	.047	-.016	-.021	.065	.071	.781	.110	.023	.865	.964
Homework/Assignments											
Q28	-.003	-.025	.026	.033	.011	.098	.050	.860	-.025	.886	.922
Q29	.092	-.017	.017	.015	.042	.027	.111	.785	.059	.842	.931
Workload/Difficulty											
Q33	-.086	-.091	.080	.026	.040	.061	-.043	-.007	.947	.946	.972
Q34	.017	-.120	-.037	.030	-.001	-.048	.011	.109	.790	.676	.970
Q35	-.218	.164	-.014	-.174	-.105	-.078	.023	.086	.611	.576	.982
Overall Rating Items											
Q30	.289	.358	.245	.051	.119	.026	.067	.218	-.051	.901	.624
Q31	.129	.437	.304	.091	.197	.025	.077	.142	-.014	.932	.627
Q32	.116	.352	.291	.075	.175	.060	.093	.168	.047	.808	.857

Factor Pattern Correlations

Learn	1.000								
Enthus	.361	1.000							
Organ	.485	.412	1.000						
Group	.371	.278	.232	1.000					
Indiv	.341	.411	.301	.394	1.000				
Brdth	.452	.301	.476	.365	.355	1.000			
Exams	.284	.240	.340	.272	.354	.328	1.000		
Assign	.378	.227	.397	.267	.295	.394	.380	1.000	
Work	-.192	-.085	-.005	-.093	-.062	-.023	-.019	-.118	1.000

Note. The factor analysis consisted of a principal axis factor extraction, following a Kaiser normalization, and an oblique rotation (SPSS, 1991). Comm = final communality estimated. MSA = Measures of sampling adequacy (SPSS, 1990). The Kaiser-Meyer-Olkin Measure Of Sampling Adequacy (SPSS, 1990) is .958.

Table 2
Interrater Reliability Estimates for Total Sample and For Classes Differing in Class Size

ASEEQ Item/Scale	Total	Very Small	Small	Large	Very Large
Learning	.914	.813	.859	.872	.965
Q1	.885	.785	.809	.843	.951
Q2	.880	.754	.820	.854	.946
Q3	.888	.757	.821	.816	.956
Q4	.889	.760	.807	.847	.953
Enthusiasm	.953	.888	.916	.934	.981
Q5	.922	.833	.879	.926	.961
Q6	.941	.859	.908	.923	.975
Q7	.953	.872	.915	.924	.982
Q8	.941	.867	.891	.901	.977
Organization	.923	.867	.858	.887	.968
Q9	.939	.815	.837	.886	.979
Q10	.907	.824	.810	.869	.962
Q11	.864	.786	.799	.806	.939
Q12	.886	.854	.872	.887	.927
Group Interaction	.933	.835	.862	.888	.975
Q13	.918	.811	.840	.886	.967
Q14	.921	.810	.843	.879	.969
Q15	.916	.805	.850	.854	.968
Q16	.921	.781	.833	.854	.971
Individual Rapport	.902	.748	.844	.869	.959
Q17	.874	.672	.820	.843	.945
Q18	.878	.715	.803	.844	.948
Q19	.893	.729	.824	.854	.956
Q20	.853	.753	.787	.799	.933
Breadth of Coverag	.901	.833	.848	.844	.959
Q21	.856	.806	.778	.795	.934
Q22	.876	.787	.833	.808	.946
Q23	.857	.711	.817	.793	.938
Q24	.902	.828	.840	.860	.959
Exams/Grading	.896	.790	.866	.875	.951
Q25	.883	.779	.859	.862	.942
Q26	.874	.744	.813	.859	.940
Q27	.891	.811	.858	.887	.943
Homework/Assignmen	.816	.688	.756	.789	.907
Q28	.810	.732	.739	.794	.898
Q29	.798	.679	.731	.767	.895
Workload/Difficult	.948	.932	.940	.928	.974
Q33	.908	.808	.866	.874	.960
Q34	.911	.814	.889	.878	.960
Q35	.817	.578	.737	.798	.914
Overall Rating Items					
Q30	.925	.840	.880	.890	.969
Q31	.948	.863	.893	.906	.981
Q32	.892	.861	.730	.830	.963
Summary Statistics					
N of Students	7038	739	1151	1903	3245
N of Classes	305	72	74	92	67
Min Class Size	3	3	14	18	26
Max Class Size	151	13	17	25	151
Mean Class Size	23	10	16	21	48

Note. Interrater reliability estimates vary systematically with class size. For present purposes, separate estimates were computed for the total sample and for classes varying in class size. SEEQ scale scores are the mean of the items in each scale and are typically more reliable than the items within each scale.

Table 3
Comparison of End-of-term feedback and Control Group: Effects of Covariate (Cov; time 1 and time 2 ratings), Feedback Intervention, and their Interaction on Time 3 Ratings.

Score	Group	Time 1		Time 2		Time 3						
		Mean	SD	Mean	SD	Mean	SD	R^2	Cov	Group	Inter	d
Scale Scores												
Learn	C	6.70	.79	6.90	.71	6.95	.77	.365	.57	.14	-.15	.27
	E	6.65	.76	6.90	.70	7.13	.65					
Enthus	C	7.09	1.18	7.16	1.00	7.23	.98	.489	.70	.11	.02	.22
	E	6.88	1.07	7.01	1.01	7.32	.99					
Organ	C	6.90	.91	7.00	.83	7.03	.88	.492	.62	.23*	-.16	.47
	E	6.89	.83	7.00	.66	7.42	.75					
Group	C	7.16	1.12	7.40	.90	7.40	.84	.459	.66	.23*	-.02	.48
	E	6.94	1.00	7.20	.86	7.66	.82					
Individ	C	7.24	.90	7.32	.79	7.43	.71	.324	.56	.09	-.07	.18
	E	7.24	.68	7.31	.72	7.55	.61					
Breadth	C	6.82	.78	6.89	.74	6.99	.68	.391	.58	.24*	-.07	.50
	E	6.57	.71	6.85	.57	7.22	.60					
Exams	C	6.35	.92	6.75	.72	6.95	.89	.344	.54	.09	-.15	.19
	E	6.47	.93	6.82	.75	7.16	.66					
Assign	C	6.81	.76	6.79	.74	6.82	.68	.328	.52	.22*	-.04	.46
	E	6.81	.74	6.84	.73	7.13	.63					
Summary Scores												
Total[a]	C	6.88	.78	7.03	.67	7.10	.70	.493	.67	.20*	-.07	.40
	E	6.80	.69	6.99	.62	7.32	.62					
Wt Total[a]	C	6.92	.79	7.05	.68	7.10	.70	.487	.66	.21*	-.09	.43
	E	6.85	.70	7.02	.65	7.36	.62					
Course	C	6.99	1.00	7.07	.93	7.05	.98	.432	.58	.22*	-.21*	.44
	E	6.84	.98	7.01	.90	7.38	.82					
Teacher	C	7.22	1.09	7.35	.94	7.31	.99	.516	.69	.24*	-.10	.50
	E	7.06	1.17	7.24	.95	7.68	.97					

Note: Each Time 3 (end of semester 2) ASEEQ factor and summary score was related to its covariate (the mean score from Time 1 and 2), Group (End-of-Term vs. Control), and their Interaction. Presented are the multiple R^2, the beta weights associated with the covariate, group effect, and their interaction, and the effect size d statistic (based on the part correlation between the grouping variable and each outcome variable). All multiple R^2s and covariates are statistically significant ($p < .05$).

[a] The Total score is an unweighted mean of the eight scale scores (excluding workload). The weighted total score is weighted by importance ratings from the teacher self-ratings. For present purposes, the importance ratings were "ipsatized" by dividing each importance rating by the mean of the importance rating for each teacher so that the ipsatized importance ratings summed to a constant (see Marsh, 1986).

Table 4
Comparison of Mid-term feedback and Control Group: Effects of Covariate (time 1 ratings), Group (feedback Intervention), and their Interaction on Time 2 Ratings and on Time 3 Ratings.

								Time 2 Ratings					Time 3 Ratings				
		Mean	SD	Mean	SD	Mean	SD	MR^2	Cov	Group	Inter	d	MR^2	Cov	Group	Inter	d
Scale Scores																	
Learn	C	6.70	.79	6.90	.71	6.95	.77	.654	.81	-.01	.01	-.02	.286	.54	.01	-.09	.02
	M	6.63	.90	6.83	.94	6.92	.92										
Enthus	C	7.09	1.18	7.16	1.00	7.23	.98	.798	.89	.00	.06	.00	.390	.62	-.01	.02	-.02
	M	7.07	1.30	7.15	1.26	7.21	1.13										
Organ	C	6.90	.91	7.00	.83	7.03	.88	.739	.86	-.03	.06	-.06	.358	.56	.06	-.20	.12
	M	6.85	.93	6.90	1.02	7.11	.89										
Group	C	7.16	1.12	7.40	.90	7.40	.84	.767	.88	-.08	.03	-.16	.377	.60	.02	-.07	.04
	M	7.20	.96	7.28	.90	7.44	.80										
Individ	C	7.24	.90	7.32	.79	7.43	.71	.664	.82	.06	.01	.12	.300	.54	.04	-.06	.04
	M	7.13	.79	7.33	.87	7.44	.71										
Breadth	C	6.82	.78	6.89	.74	6.99	.68	.738	.86	.04	.05	.04	.308	.56	.02	-.08	.02
	M	6.57	.82	6.74	.84	6.89	.75										
Exams	C	6.35	.92	6.75	.72	6.95	.89	.209	.46	.01	.07	.01	.161	.27	.00	-.27*	.00
	M	6.54	.87	6.84	.86	6.99	.79										
Assign	C	6.81	.76	6.79	.74	6.82	.68	.490	.68	.03	-.08	.06	.200	.30	.05	-.27*	.10
	M	6.77	.62	6.80	.71	6.89	.76										
Summary Scores																	
Total[a]	C	6.88	.78	7.03	.67	7.10	.70	.779	.88	-.01	.09	-.02	.381	.60	.02	-.11	.04
	M	6.85	.76	6.98	.83	7.11	.75										
Wt Total[a]	C	6.92	.79	7.05	.68	7.10	.70	.777	.88	-.02	.07	-.04	.399	.62	.02	-.09	.04
	M	6.91	.77	7.01	.83	7.14	.75										
Course	C	6,97	1.01	7.07	.93	7.05	.99	.704	.84	-.01	-.01	-.02	.348	.59	-.02	-.22*	-.04
	M	6.89	1.25	6.98	1.15	6.97	1.05										
Teacher	C	7.21	1.11	7.35	.94	7.30	1.01	.712	.83	-.05	.05	-.10	.407	.64	.10	-.26*	.20
	M	7.23	1.44	7.25	1.38	7.50	.94										

Note: Each Time 2 (end of semester 1) and Time 3 (end of semester 2) ASEEQ score was related to its covariate (the Time 1 rating from the middle of semester 1), Group (Mid-Term vs. Control), and their Interaction. Presented are the multiple R^2, the beta weights associated with the covariate, group effect, and their interaction, and the effect size d statistic (based on the part correlation between the grouping variable and each outcome variable). All multiple R^2s and covariates are statistically significant ($p < .05$).

[a] The Total score is an unweighted mean of the eight scale scores (excluding workload). The weighted total score is weighted by importance ratings from the teacher self-ratings. For present purposes, the importance ratings were "ipsatized" by dividing each importance rating by the mean of the importance rating for each teacher so that the ipsatized importance ratings summed to a constant (see Marsh, 1986).

Table 5
Difference in Groups (Mid-Term, End-of-Term, and Control Groups) over Time (Mid-term Semester 1, End-of-term Semester 1, End-of-term Semester 1) For Ratings of Targeted and Non-Targeted ASEEQ Factors.

Source Of Variation	SS	DF	MS	F-Ratio	p-value [a]
Group	.42	2	.21	.08	.919
Error	199.61	80	2.50		
Time	12.36	2	6.18	18.52	.000
Group By Time	1.81	4	.45	1.35	.253
Error	53.42	160	.33		
Target	3.93	1	3.93	15.81	.000
Group By Target	1.08	2	.54	2.18	.120
Error	19.87	80	.25		
Time By Target	.82	2	.41	9.85	.000
Group By Time By Target	.48	4	.12	2.86	.025
Error	6.67	160	.04		
Analysis of the Group (G) By Time (T) By Target Interaction[b]					
G(1) By T(1) By Target	.00	1	.00	.02	.898
G(2) By T(1) By Target	.15	1	.15	2.69	.105
Error	2.32	80	.03		
G(1) By T(2) By Target	.12	1	.12	4.08	.046
G(2) By T(2) By Target	.29	1	.29	10.01	.002
Error	2.32	80	.03		

[a] p-values adjusted for the Greenhouse-Geisser and Huynh-Feldt estimates of epsilon (SPSS, 1991) are very similar to the unadjusted values presented here and in no instances is an effect reported to be statistically significant here not statistically significant when evaluated with the more conservative epsilons. [b] For purposes of these follow-up analyses, a polynomial contrast was use for the time variable (T(1) = linear, T(2) = quadratic), and a simple contrast was used for the group variable (G(1) = Mid-Term vs. Control, G(2)= End-of-Term vs. control).