The Endeavor Handbook[1]

Introduction

The educational process is most effective when the teacher interacts actively with students. In order to establish a meaningful relationship, the instructor must be willing to expend time and energy in preparing for and conducting each class. This is often difficult, however, since the faculty have many other demands on their time: committee meetings, speaking obligations, research activities, and professional writing. At many institutions, these non-teaching activities are as important or more important than teaching in determining rank and salary. In order to advance their careers, faculty may have to structure their priorities in a way which is not conducive to good teaching. In doing this, instructors are simply responding in an adaptive manner to the incentive system established by their institutions. It is a sad commentary that the faculty evaluation process often does very little to encourage teaching (see Faia, 1976, Shapiro, 1978).

The emphasis a college or university places on teaching is a direct measure of its concern for students and learning. The institution s commitment to instruction can be evaluated by determining whether teaching 18 given concrete recognition in routine administrative decision making. Even though teaching excellence may be essential for survival, many institutions fail to evaluate it objectively and give it a secondary role in determining promotion and salary.

Whether one recognizes it or not, each faculty member is evaluated on a regular basis. These evaluations are a serious business. At stake are self-esteem, Job security, and the individual s chosen career. The evaluation process should be fair to both the individual and the institution. It should be explicit and deliberate. This is especially important with respect to teaching. If an objective, systematic procedure is not devised, important decisions will be made on the basis of rumor and hearsay. Social conversations and corridor exchanges are a poor substitute for systematic data collection from first-hand witnesses.

On most campuses, students are the only individuals who directly observe the day-to-day conduct of each course. Probably for this reason, a questionnaire, usually consisting of 10 to 40 items, is commonly given to students at the end of the term. Each statement on the questionnaire describes some characteristic of the teacher or the course and students are asked to assess the degree to which the item applies. Usually the items are selected to provide information for the instructor on the strengths and weaknesses of the course from the student s perspective. Questionnaires of this type are not intended for use in evaluating instructors for merit, promotion, or tenure. When the ratings are gathered for formative purposes (I.e., the instructor s self-improvement), items may be deliberately chosen to spotlight and aid in diagnosing suspected weaknesses. This information is not appropriate for making decisions on rank or salary. For this latter purpose (i.e., summative evaluations), a standardized form is needed which can be applied uniformly to all teachers.

It is unlikely that a single form, with a single set of questions, will simultaneously serve both formative and summative purposes. In recent years, several institutions have developed a flexible procedure for providing formative feedback for individual instructors. The leader in this approach has been the Cafeteria system developed by the Measurement and Research Center at Purdue University. Purdue s approach features a catalog of approximately 200 items from which the individual instructor can tailor his or her own 40-item rating form. After the instructor has selected the desired items, questionnaires are printed by a computer. Student responses to these forms are also read, tabulated, and printed by machine. This system simultaneously combines instrument flexibility with machine scoring and thereby provides an economical technique for obtaining diagnostic information from students.

If a cafeteria-like system is used for summative purposes, it is reasonable to expect that instructors

---

[1]Excerpted, with permission, from *The Endeavor Instructional Rating System User's Handbook*, ©Endeavor Information Systems Inc., 1979.

will select items on which they anticipate favorable responses rather than ones that would be most helpful in diagnosing instructional problems. Thus the cafeteria approach is seriously compromised when administrators attempt to use it for summative purposes. Because of this conflict, it is appropriate to develop two separate rating forms, one for formative purposes which is individually tailored to the needs of each instructor and one for summative purposes which is standardized for all instructors.

The Endeavor Instructional Rating System is an efficient and economical system which has been designed for summative evaluations. The rating form is a brief, objective questionnaire which has been subjected to extensive analysis. The reliability and validity of this instrument are examined in a series of publications which have appeared in professional journals. This research has also provided useful information on a number of important variables which are potential sources of bias. The Endeavor system also has several special features. To aid in interpretation, the response distribution for each rating item is presented graphically in histogram form. In addition, the response summaries are organized in terms of specific course clusters which are designated in advance by the user institution. Average ratings are computed for each course cluster so that instructors have an appropriate set of norms for each course. Scatter plots are also provided which summarize the ratings for all of the instructors within each cluster.

Summative Evaluations

There appear to be several general qualities which distinguish the superior teacher. To be effective, instructors must be knowledgeable in their subject matter, be skilled in organizing and presenting material, be concerned about individuals, and be sensitive to the student s level of understanding (see, e.g., Crawford & Bradshaw, 1968; Schein & Hall, 1967). To evaluate teaching, therefore, it is reasonable to measure performance in respect to each of these traits. A careful assessment requires the use of three different data collection procedures. A thoughtful inspection of course materials (e.g., course outlines, examinations, reading lists, etc.) by appropriate colleagues can determine whether the instructor is keeping up with recent developments in his or her field. This limited form of peer review can elicit valuable information which students are not qualified to provide. Letters from recent alumni can indicate whether the instructor has a long-term impact on students. Alumni can often provide a mature overview which is not available from other potential witnesses.

A third source of information can be provided by the instructor s present students. These individuals have a relationship with the instructor which is not matched by any other segment of the academic community. The student is probably the most reliable source of information on important traits such as presentation skill, daily organization, and student-instructor rapport. The systematic collection of student observations provides a valuable resource for evaluating teaching. By combining information from these three different sources, the evaluator will have sufficient knowledge to reach an informed decision. This strategy has the interesting consequence of removing students from the jury and recasting them as witnesses (see Menges, 1974).

Formative evaluations of teaching are not always well received by the faculty but they are considerably more popular than summative evaluations. In fact, educators have usually emphasized the formative aspects of teacher evaluations while avoiding a discussion of their summative role. A common theme is that student ratings will provide valuable feedback which is conducive to improved teaching. Unfortunately, Centra (1973), Kulik and McKeachie (1976), Andy Rotem (1978) have found little evidence that faculty improve their teaching after receiving formative instructional ratings. If an evaluation system is to improve the educational process, it would appear that it needs to be an integral part of the promotion and salary review process. Thus, summative evaluations seem to be an unpleasant necessity.

Constructing a Rating Form

When a questionnaire is designed for diagnostic purposes, it should use items which are tailored to the specific needs of the instructor (as with a cafeteria

systems or should have open-ended items. When the rating form is designed for administrative purposes, it should have a uniform format for all courses such that evaluators can make direct comparisons among instructors. It is also helpful if the summative questionnaire can be administered easily and quickly and can be scored efficiently by machine. To provide reliable data, summative ratings should be collected on a regular basis. Many institutions feel that it is important to collect ratings every term for each and every course. Because of this, a brief questionnaire has several important advantages.

Because class time is a valuable commodity, it should not be squandered unnecessarily. Why devote 20 to 30 minutes of class time to evaluation activities it 10 minutes will suffice? Second, the students enthusiasm often diminishes when they are asked to respond to a lengthy questionnaire in every class. Careful attention to each item is more likely when the rating form is brief. A third advantage is that short forms are compatible with graphical response summaries. With 25 or 30 items, individual histograms for each item become impractical. Last, and most important, a carefully designed short form provides essentially the same summative information as much longer forms. Extensive research involving various factor analysis procedures has demonstrated that even very long forms (e.g., 30 or more items) seldom tap more than 4 to 6 primary dimensions of teaching (see Kulik and Kulik, 1974, for a review). Thus, a questionnaire which has each dimension represented by one or two items can be both brief and comprehensive.

One of the most important decisions in constructing a questionnaire is that concerning item selection. Committees have often developed a consensus form by accepting all items suggested by interested parties. This strategy usually leads to a lengthy instrument which may or may not have good psychometric properties. In the absence of adequate research, this procedure is often unavoidable since every item appears superficially to be just as good as every other. Unfortunately there is a tendency to confuse the quantity of items with the quality of the instrument. In point of fact, the inclusion of a large number of items generally reflects an attempt to compensate for the absence of appropriate empirical knowledge. An item should be selected only if it has been shown to be reliable and to provide information not already provided by other items.

Several guidelines are relevant. The kinds of questions students can answer most reliably are those which fall directly within their power of observation. Thus items which ask for descriptive information are generally preferred over those which require judgment or interpretation. For this reason it is inappropriate to ask students to judge the competence of faculty members in their professional specialties when colleagues are in a much better position to make these judgments. A second important point is that each item should tap information which is relevant to an important aspect of the course. Thus an item about the teachers grooming habits or wardrobe is not particularly germane. Many questionnaires ask about the teacher s enthusiasm or dynamism. There is considerable evidence, however, that this trait is neutral in respect to pedagogical competence (see the literature on the Dr. Fox effect, e.g., Williams 8 Ware, 1977; Ware & Williams, 1975). It is also common to include items which provide demographic information about the student, e.g., class level, grade point average, sex, major field. This information is helpful for research purposes and for norming. In most cases however, this information is never used and therefore the time and expense involved in generating it are wasted. The student s task can be simplified by simply omitting these demographic items.

Another important decision involves the choice of a particular response field. There are several commonly used options. The response scale can be continuous (e.g., make a check mark at some point along a line) or be divided into 4 to 11 discrete categories (e.g., mark the appropriate box). In either case, it is also necessary to select descriptive labels to distinguish the high and low ends of the response dimension. Psycho metric research provides information which is helpful in making these decisions.

A recent report by McKelvie (1978) based on original research and an extensive review of the literature concludes that categorized scales are as reliable and valid as continuous scales. Ratings made by subjects using a continuous scale indicate that they

are operating essentially with five or six discrete categories. This is an important conclusion because continuous scales do not lend themselves to machine scoring. Given these observations, a categorized scale seems most appropriate for summative instructional ratings.

A second finding by McKelvie (1978) was that there is no advantage in a large number of scale categories (greater than nine) and that there may be a loss of discriminative power with fewer than five categories. These conclusions are compatible with the selection of a response field containing five to nine discrete categories. This choice should permit students to communicate their observations in an accurate and detailed manner.

Another important decision relates to the labels which are chosen for the response dimension. It is common to ask students whether they agree or disagree with each item. This procedure has an important drawback. Disagreement can be expressed for two very different reasons. The rater may disagree with the item either because it does not fit the instructor or because the wording is not suitable. For example, in our early research we employed degree-of-agreement response fields. One student responded strongly disagree to the global rating item, overall this was a good course. Since most students rated the course very positively, the nonconformist was asked about his response. He responded that the course was the best he had ever taken and therefore the word good was not accurate. He would have responded strongly agree if the item had been written overall this was a terrific course.

We concluded that this type of response field requires that each item be worded very carefully and even then, it is likely that some responses will be misinterpreted.

Response fields can also be labeled with frequency indicators. In this case, the student considers how often the specified behavior occurred, e.g. never, seldom, sometimes, often, or always. If this procedure is adopted, it is often necessary to reword many items to emphasize behaviorally observable aspects of the classroom experience. This often increases the reliability of the instrument.

When a questionnaire is intended for regular, institution wide use, it should be designed so that the results can be processed by machine. Human analysis is slow, error- prone, and relatively expensive. There are several machine compatible procedures which are commonly used. One involves transferring the responses from the questionnaires to computer cards by key punching. This is slow and expensive.

A second procedure involves the use of special forms which can be optically scanned. This procedure is relatively inexpensive and permits rapid data processing. It has two important disadvantages. Pencils have to be provided when the form is administered and these have a habit of disappearing and thus require frequent replacement. This can be a major unanticipated expense. Optical scanning procedures also require that each answer be placed accurately within a designated response area. If the student is a bit careless, his response may ice missed or misinterpreted. This problem reduces the reliability of the instrument.

A third procedure involves the use of a specially prepared computer card which is perforated in selected locations so that small portions of the card can be punched out. This "porta-punch" procedure can be quite effective when the rating card is designed and processed properly. A key design feature is the use of a small number of questions and widely spaced response foliage so that the perforations do not structurally weaken the card. Processing of the questionnaires must also be preceded by a careful examination of each card to remove any excess chad.

Development of the Endeavor System

The Endeavor Instructional Rating System is the outgrowth of an academic research program that was initiated to examine the reliability and validity of student instructional ratings. The founder of the system, Dr. Peter W. Frey, was asked in 1971 by the Provost of Northwestern University to serve as research coordinator for an instructional rating organization which was initiated and run by students. He organized a special seminar for graduate and undergraduate students. After studying the research litera-

ture and examining many different rating forms, this seminar group devised its own 38-item form and administered it to about 200 students in half a dozen classes. The resulting ratings were then analyzed in detail. Items which displayed a large within-instructor variance were either discarded or reworded. A factor analysis was also performed. Half of the items loaded on a single dimension and many of these were discarded.

This process was subsequently repeated. A new form was prepared (Endeavor 11) consisting of proven old items and a few new items that had been suggested by colleagues or students after the first form had been distributed. Item variances were again computed and another factor analysis was performed. This new information was then used to restructure the questionnaire. A third-generation questionnaire was constructed and then a fourth. After many revisions, and well after the term had ended and the seminar group had disbanded, Endeavor iX was developed. This 18-item questionnaire was employed for a validation study which was conducted with the cooperation of the mathematics apartment. The results of this research were subsequently published (Frey, 1973).

Subsequent modifications led to two additional forms, Endeavor X and Endeavor XI, each having 21 items. Endeavor XI was used for three validity studies conducted at three different universities (Frey, Leonard & Beatty, 1975). As a part of our research effort, we had regularly solicited suggestions from students and colleagues for improving the form. Common complaints were that the form was too long and that some of the items were redundant. We had also noted that students often placed their marks outside of the designated response field and that the optical scanner was missing these responses. To correct these problems, we made a major revision of the form and devised our present seven-item porta-punch questionnaire (Endeavor XII). Each of the items was selected by identifying the most representative item on each of the seven dimensions represented on our previous 21-item form. As a result, the 7-item form provides comprehensive coverage despite its brevity. In addition, the porta-punch form was designed with widely spaced perforations and with only 49 punch positions instead of the conventional 200. This appears to reduce the incidence of weakened or torn forms which is a common problem with standard porta-punch cards.

The seven-item Endeavor questionnaire is unusual in that it combines a 7-category response field with five labels on a frequency dimension, i.e., never, seldom, sometimes, often, always. Although it has been suggested that there be a label for each response category, this would require a reduction in the number of categories or an increase in the number of labels. At the time the questionnaire was designed, we systematically examined the possibility of using two additional labels. Twenty-nine students from 3 different classes were asked to rate 20 adjectives in terms of their frequency connotation. The number of times an event would be expected to occur in 100 opportunities was rated on the average of 97.4 for always, 76.5 for often, 42.0 for sometimes, 16.9 for seldom, and 0.9 for never. Our attempt to find two additional labels which would represent the 50% to 70% range and the 20% to 40% range was not successful. The results for the words we tested were: 6.4 (almost never), 10.0 (hardly ever), 10.5 (rarely), 18.5 (not usually), 21.9 (infrequently), 24.0 (once in a while), 34.3 (occasionally), 70.8 (commonly), 78.0 (frequently), 78.2 (ordinarily), 79.0 (customarily), 79.8 (usually), 88.3 (very often), and 90.8 (almost always). The adverb occasionally fell in the appropriate range but it was judged to be undesirable because it had the largest standard deviation of any of the words we examined. As a result of this research, we decided to employ the 5 conventional frequency labels for our form.

At the request of two educational institutions, the Endeavor system was first offered for commercial use in 1973. The original service was based on our 21-item form. In 1974, this questionnaire was retired and the present 7item form was adopted. Comments from our users quickly indicated that the conventional tabular summaries (i.e., mean, standard deviation, and response distribution for each item) were not particularly informative for many instructors. Because of these complaints, we devised a graphical technique for representing the pattern of responses to each item. This effort has resulted in our present histogram displays (see Figure 4, page 14) which have been received with enthusiasm.

During the 1975-1976 academic year, a major research study was conducted using the 7-item form (Frey, 1978). This study involved 1298 classes and 26,787 forms. Two separate factor analyses (one for the Fall quarter and one for the Winter quarter) indicated that most of the between-course response variability was accounted for by two global response factors. The first item cluster, which we have labeled pedagogical skill, is influenced most heavily by the items on presentation clarity, organizational skill, and the student's estimate of how much he or she had learned. The second cluster, called rapport, centers on the instructors willingness to help students and to encourage class discussion. Additional analyses indicated that these two global factors are quite different on measures of reliability and validity and show differing relationships to important demographic variables such as class size, instructor experience, and instructor grading leniency (see Frey, 1978). Because of this, our rating summaries, starting in 1976. provide gum Mary information on both of these global performance dimensions. To aid in interpreting this information, a graphical display of these two indices is provided for all of the instructors within each thematic group in the form of a scatter plot (see Figure 3, page 13).

Administering the Endeavor Questionnaire

The major considerations which are important in administering an instructional rating form are those of objectivity, coverage, timing, and efficiency. If the data are not collected in a systematic manner, this aspect of the process can influence the validity of the results as much as the choice of items. Minimum necessities for gaining a measure of objectivity include: a specially designated time; a third party in charge of distributing and collecting forms; assurances that the student s anonymity will be protected. If the students are asked to complete the form under the watchful eye of the instructor, their responses may be inhibited or distorted. Proper timing is important to insure that a representative sample of students is present. Ideally one would like to have the participation of every member of the class because the results may be biased if only a small segment of the class responds.

The most common time to collect instructional ratings is at the end of the term. Research with the Endeavor form indicates that the students ratings are highly similar whether collected during the next-to-last week of the term or a month after the term is over (Frey, 1976). This outcome suggests that the time of collection is not particularly critical. We do recommend, however, that users avoid administering the questionnaire at the same time that an examination is given or on a day when important papers or exams are being returned to the class. It also seems prudent to avoid the last week of the term when students are apt to be anxious about final exams.

The efficiency of the administration procedure is particularly important when a large number of courses are involved. The Endeavor form, because it is brief, can be administered quickly and easily. The organization of the data collection process can be handled by a single person, the campus coordinator. This person should be familiar with the institution s course offerings and be compulsive about details. With the Endeavor system, the campus coordinator can determine the eventual organization of the summary booklet by assigning appropriate code numbers to each class. This flexible arrangement permits individuals at each institution to define an appropriate norming group for each course. One of our staff will be happy to discuss the details of this procedure with campus coordinators in order to individualize the course summaries to meet the specific needs of their campuses.

Evaluating the Endeavor System

a) Reliability

Reliability refers to the accuracy of measurement in terms of its consistency from one situation to another and its stability over time. Several methods are commonly used to assess the reliability of a measuring instrument. One is the test-retest method in which the instrument is administered twice to the same individual. This procedure is not particularly useful with the Endeavor form since with only seven items it would

not be clear whether the scores on the second administration reflected the consistency and stability of the ratings or simply the student s ability to remember his or her previous ratings. Another common method involves the computation of a split-half or odd-even correlation coefficient to determine if the different items are consistently measuring the same thing. This procedure is widely criticized for producing overly optimistic estimates of reliability. In addition, the split-half method is not appropriate for instructional rating forms since the different items measure several different traits (see Kulik and Kulik, 1 974).

To assess the reliability of the Endeavor form we have employed several other procedures which we believe are more appropriate. To assess the stability of the mean ratings over time we conducted a research study (Frey 1976) in which the students in seven different classes were contacted by mail. These classes had an average enrollment of 29 students. A random sample of half of the students in each class were contacted during the next-to-last week of the term. The remaining half were contacted approximately 6 weeks later, after the course was over and the students had received their final grades. An analysis of the ratings on each item for the two groups of students in each of the seven classes indicated no significant change over time (a separate one-way analysis of variance was computed for each item). The correlation between the mean classroom rating from the first measurement period and the second measurement period was .87, .81, and .66 for item 5 (presentation clarity), item 2 (advance planning), and item 7 (increased knowledge) respectively. The correlation was much lower for the other four items: .43 for item 3 (class discussion), .38 for item 6 (grading), .34 for item 1 (work hard), and .30 for item 4 (personal help). The Endeavor ratings can also be summarized in terms of two global rating scores. Pedagogical skill, representing items 2, 5, and 7 primarily, showed a correlation of .90 between the first and second data collection periods. The second global factor, rapport, which is influenced most heavily by items 3 and 4, showed a correlation of .38. This analysis indicates that the ratings on several of the items, notably 2, 5, and 7, are quite stable over time.

Another study (Pasen, 1977) also examined the stability of ratings on the Endeavor form in a large English class (n = 553). Ratings were made immediately after the course was completed and one year later. Pasen's analysis indicated that the mean scores on the two global factors were virtually identical for the two different data collection periods. Thus with a large sample, both ratings remained stable for one year.

To examine the consistency of the ratings to the Endeavor items, an analysis-of-variance design was employed on the ratings from two large classes (organic chemistry and human biology) which had many (n=72) students in common (Frey, 1974). Responses to item 5 (presentation clarity) were analyzed with a mixed-design two-way analysis of variance (2 instructors x 72 raters). This analysis determined the consistency of the ratings by assessing whether the observed variation among the scores could be attributed to differences among the students or to differences between the two teachers. The results indicated that 44 percent of the variance could be attributed to the teachers, 5 percent was attributable to the tendency for some students to use high or low ratings for both instructors, and 51 percent of the variance was attributable to differences among the students in the way they perceived the two teachers (i.e., error variance). Since it is rare for an independent variable in a behavioral science study to account for more than 10 percent of the total variance, this outcome clearly indicates a substantial level of agreement among the students in the way they perceived the two instructors.

A more recent reliability analysis (Frey, 1978) examined the variability of a teachers ratings from one class to another and from one term to the next. The study examined the ratings for 60 instructors who had taught 3 or more classes during the Fall and Winter quarters. Ebel (1951, p. 410) provides a formula for computing inter-rater agreement which makes use of the relationship between the within-instructor variance estimate and the between-instructor variance estimate. Our two global rating factors were examined using Ebel s formula. The coefficient for pedagogical skill was .61. For rapport, it was .32. This outcome is consistent with our previous results in indicating that

ratings on the skill items (2, 5, and 7) are more reliable than ratings on the rapport items (3 and 4).

These results, bringing together information on the stability of the ratings over time, the within-class consistency of the ratings, and the consistency of ratings for the same instructor in different classes, suggest that the Endeavor ratings have good reliability characteristics. This is especially true for the global rating on the skill factor and the ratings on three of the individual items, numbers 2, 5, and 7.

b) Validity

Within the context of instructional ratings, validity refers to an instrument s accuracy in describing (i.e., measuring) the teacher s classroom skills. Two different questions are often raised in respect to the validity of an instructional form. The first concerns content validity. Do the items on the questionnaire adequately sample the traits and behaviors which are relevant to effective teaching? The Endeavor form was designed using a factor analysis approach and therefore it encompasses essentially all the traits and behaviors which are commonly covered by much longer forms. There are two notable exceptions to this statement. The Endeavor form does not ask students to make judgments about their instructor s professional knowledge. This trait is purposely omitted because we believe that the instructors colleagues are more qualified to make this assessment. The Endeavor form also does not ask about the instructors enthusiasm or dynamism since there is good evidence that this trait is neutral in respect to pedagogical skill (see, for example, the studies on the Dr. Fox effect, Ware and Williams, 1975; Williams and Ware, 1 977)

A second form of validity is criterion-related validity. Ratings on the instrument are compared to one or more external measures whose validity is already established or at least generally accepted. A major difficulty in applying this test to instructional ratings is that academicians seldom agree upon an acceptable external criterion of good teaching. Many instructors insist that the real goal of teaching is to impart general philosophical values or to develop a love and enthusiasm for learning. Mastery of the subject matter may be considered of secondary importance. Even if faculty will accept subject mastery as a reasonable criterion of effective teaching, there is considerable disagreement over how one should determine mastery, even among instructors teaching the same course. Only by defining teaching in a relatively narrow way, such as preparing students to solve calculus problems, is it feasible to develop an objective criterion for effective instruction.

The original motivation for developing the Endeavor form was based on an extensive research project directed at examining the criterion-related validity of instructional ratings. The ideal design is one in which students are randomly assigned to different instructors who each cover the same subject matter. Each class uses the same textbook and course syllabus and all are similar in terms of class size, type of classroom, and time and place of meetings. A common final exam is used in all classes to determine how much each student has learned. The average final exam score for each instructor s students provides the external criterion indicating which teachers are most effective. Student ratings of the different instructors are then validated against this criterion. This research design is then repeated many times to test the generality of the outcome across different subject matters and different institutions.

Research with the Endeavor form has followed this design only in general format. It has not been possible in any of our studies to assign students randomly to class sections. In some cases, post-hoc adjustments have been made in the criterion scores to adjust for initial ability levels. The number of sections in each course has always been quite small, ranging from a minimum of 5 to a maximum of 12. Only two different subject matters have been examined, calculus and educational psychology. A major reason for this is that it has been difficult to get faculty to agree on a common final exam in most subject areas. The research encompasses three midwestern universities North Dakota State, Northwestern, and Purdue. Detailed descriptions of the procedures and results of this research have been published in professional journals (Frey, 1973; Frey, Leonard, and Beatty, 1975; Frey, 1976). The results have been uniformly positive (see Frey, 1978, for a summary table). When the ratings of the 7 studies are summarized in terms of our two global factors, the pedagogical skill measure

shows a median correlation of .81 with the external criterion. The rapport measure shows a much lower median correlation, .29. These outcomes indicate that the ratings on the pedagogical skill items (2, 5, and 7) correlate strongly with content mastery. The correlation with the rapport items is much less impressive.

There is an extensive literature devoted to the criterion related validity of student instructional ratings. Investigations which have used experienced instructors and have made a separate analysis of items on presentational skill and/or organizational ability have reported results which are consistent with our findings (e.g., Cohen and Berger 1970; Gessner, 1973; Sullivan and Skanes, 1974; Marsh, Fleiner, and Thomas, 1975; Centra, 1977). This literature is reviewed by Frey (1978).

Some faculty have questioned this research because they believe that the students may be trading good ratings for high grades and poor ratings for low grades. Since the pattern of results obtained in the validity studies could be produced by this hypothesized grade-rating trade-off, we made an empirical examination of this possibility. Relevant data are reported in two publications. In our Science study (Frey, 1973) and Journal of Higher Education study (Frey, 1976), we examined the correlation between ratings and final exam performance both between-sections and within sections. If the reported validity data are produced by the students trading good ratings for a high grade, one should observe a positive relationship between grades and ratings within each section. In the Science study, which involved 13 sections, the within-class correlations between the exam grade and the presentation clarity item ranged from a -.33 to a +.43 with an average value of -.02. The between-section correlation for this item was .75. A similar analysis was made in a more recent study (Frey, 1976). The average between-section correlation for six items (excluding work load) was .75. The average-within-section correlation for these same items was .19. These data indicate that the between-section correlations are not an artifact of a simple grade-rating trade-off.

c) Potential sources of bias

There has been considerable faculty resistance to the use of instructional ratings because of widely-held beliefs that students are easily influenced by irrelevant factors. When ratings have an influence on salary and promotion (4) decisions, it is important to use a questionnaire which is backed by an extensive research program. It is inappropriate to assume, in the absence of empirical studies. that a rating instrument is unaffected by potential sources of bias.

Research on the Endeavor form has examined a number of relevant variables: average grade in the class, individual students grades, grade point average, quantitative score on the Scholastic Aptitude Test, student s class level (freshman, sophomore, junior, or senior), student s sex, instructor s academic rank, the number of citations to the (5) instructors publications, student s major, class size, and the instructions for administering the form.

(1) **The students' grade** — The evidence indicates that there is no consistent relationship between the grade (6) individual students receive and their ratings on the Endeavor questionnaire. Relevant data appear in two publications (Frey, 1973; Frey, 1976). In each case, within-class correlations had average values in the neighborhood of zero. These results contradict the common belief that students who are about to receive a low grade or who have already received a low grade will use instructional ratings as a means to get even (7) with the instructor.

When the relationship between grades and ratings is examined with the classroom as the unit, the results are different (Frey, 1978). Classes in which the average grade is high are rated no differently than classes with a low average grade on items related to the skill factor (2, 5, and 7 on our form). The ratings on items related to the rapport factor (3 and 4 on our form) do appear to be influenced by the instructors leniency in grading. The correlation between the average grade and the rapport factor was .46 as compared to only .02 for the skill factor (Frey, 1978). This result indicates that instructors who grade conservatively (i.e., mostly B's and Cs) will receive lower ratings, on average, on the rapport (8) items (class discussion and personal help). This research study also found that conser-

vative graders are more likely to be perceived as unfair graders (item 6 on the Endeavor form).

It is interesting that the effect of the grade variable depends on the method of analysis. When the individual student is used as the unit of analysis, grades have no consistent effect on instructional ratings. When the unit of analysis is the classroom, however, instructors who grade conservatively receive lower ratings on items related to student-instructor rapport. It may be that an instructors grading style is one of the attributes that students employ to define student instructor rapport. An individual who gives low grades on average is perceived as being cold and distant. A different interpretation is that instructors who grade more conservatively than their colleagues are often individuals who do not have a close, personal relationship with their students. The existing data are consistent with either interpretation. Until this issue is resolved, it may be prudent to avoid unadjusted comparisons of ratings on rapport items among instructors who grade differently.

(2) **Class size** —The number of students in a class does not affect ratings on the skill items, but does influence ratings on the rapport items (Frey, 1978). Rapport ratings decrease as a direct function of class size and this effect is fairly large in magnitude (r= -.40).

(3) **Student's grade point average** —There is no consistent relationship between ratings on any of the Endeavor items and the student s grade point average (Frey, Leonard, and Beatty, 1975).

(4) **Student's class level** —There is a positive relationship between class level (freshman, sophomore, junior, or senior) and ratings on each of the Endeavor items except number 1 (work load). This relationship is fairly strong for these items, ranging from .39 to .61 (Frey, Leonard & Beatty, 1 975) . Apparently juniors and seniors are either more lenient or more tolerant than freshman and sophomores. Thus ratings from classes composed mostly of freshman should generally be lower than ratings from classes composed primarily of juniors and seniors

(5) **Sex of the rater** —There appears to be no overall difference between male and female students in the way they rate a male instructor (Pasen, 1977). There does seem to be a tendency for male students to react more negatively to a low grade (Pasen, 1977).

(6) **Quantitative SAT score** —Students with high math aptitude scores rate calculus instructors essentially the same on most items as students with low scores. On the work load item, high aptitude students tend to give lower ratings (r= —.32). On the item reflecting increased knowledge, they give higher ratings (r= .27). All other items show correlations between .00 and .12 (Frey, Leonard, and Beatty, 1975).

(7) **Instructor's academic rank** —The academic rank of the instructor bears an interesting relationship with the ratings on the Endeavor form. Ratings on the rapport factor are higher in general for low-ranking instructors (Frey, 1978). This may indicate that older instructors have more difficulty relating to students and may also have less time to engage in informal discussions. Ratings on the skill factor are higher in general for high-ranking instructors (Frey, 1978). This is consistent with the idea that teaching skill improves with experience. Since academic rank is strongly correlated with both chronological age and years of experience, it is not clear which of these factors is responsible for each relationship.

(8) **First-author citations** — Some faculty believe that teaching and research are incompatible. An analysis of the ratings on the Endeavor form for a sample of 42 science faculty at a midwestern university in relation to the number of first-author citations indicated different relationships for the skill and rapport factors (Frey, 1978). Ratings on the skill factor showed a positive relationship to citation frequency (r=0.37). Ratings on the rapport factor showed a weak negative relationship with citation frequency (r= -0.23).

(9) **Instructions during administration** — If the students are told that their ratings will be used to determine promotion and salary they tend to rate their instructors more favorably than if they are

told that the ratings are for the instructors personal use. This effect is small on items related to the skill factor but quite substantial on items related to the rapport factor (Pasen, Frey, Menges, and Rath, 1978). Apparently students become more lenient in their judgments when they are asked to bear an important responsibility.

10) **The student s major**—Do chemistry majors rate instructors more harshly than do education majors? This is a difficult question to answer empirically because there are very few courses which are taken by students from all parts of the university. Pasen (1977) examined this question in a large humanities course attended by students from the social sciences, the physical sciences, the humanities, the school of journalism, the school of education, and the engineering school. Although his results are not conclusive because of small sample sizes, Pasen (1977) noted that students from different segments of the university were- probably using different standards when they rated the instructor. For example, it was quite clear that students in chemistry and students in education had very different ideas concerning what constituted hard work. Given this observation, one should be cautious about comparing ratings made by students majoring in different - areas, especially if these groups appear to differ on one or more important academic dimensions.

Our research, as well as that of others, indicates that instructional ratings are not free from bias. Students are no different than other raters; they are influenced by factors which may be incidental to the relevant decision criteria. The most important factors to consider when using the Endeavor form appear to be class size, the student s class level, the student s major, and the instructor s grading style. Because these variables often influence students, many academicians have concluded that instructional ratings are not a useful source of information for evaluating faculty. There are two problems with this conclusion. One is that other potential measures of teaching have as many or more problems than instructional ratings. A second is that the procedures used for evaluating scholarship and service are open to at least as much bias as those used to evaluate teaching. All evaluation measures are imperfect, and therefore it is necessary to use them carefully. Knowledge about each potential source of bias is needed if the evaluation is to be fair and valid.

The Endeavor system has been designed in a way which minimizes the inherent limitations of fallible ratings. Because the user institution can designate a specific thematic grouping (i.e., control group) for each course, it is possible to provide norms which avoid the most common sources of bias. In view of our research, it is advisable to establish thematic course groupings which separate courses as much as possible on the basis of discipline, class size, type of student (i.e., primarily underclassmen vs. primarily upperclassmen), and typical grading styles. The first and the last variables (discipline and grading) are often confounded in the sense that grading style correlates strongly with discipline and this simplifies the challenge of establishing appropriate thematic groups.

In establishing course clusters, the campus coordinator can assign course code numbers in a way which is responsive to these empirical relationships. For example, large enrollment courses and small enrollment courses can be placed in different clusters. Courses attended primarily by freshman or sophomores can be categorized separately from those composed mostly of upperclassmen. The assignment procedure can also insure that social science courses are not mixed with those in the physical sciences or in the humanities. By establishing course clusters in this manner, the campus coordinator can avoid many of the problems which commonly compromise the interpretation of instructional ratings.

Guidelines for Interpreting the Ratings

There are many sources of information about teaching effectiveness. Although student ratings provide important data, they should not be used in isolation. A decision on teaching competence should be based upon a number of independent judgments drawing upon various sources of information. In addition to student ratings, it is useful to have colleagues examine classroom material, such as course outlines, exams, reading lists, and lecture notes. An experienced colleague can

learn a great deal about the candidate s teaching from these materials. It is also useful to have an appropriate administrator (e.g., departmental chairman, a dean, or a representative of the developmental office) conduct in-depth exit interviews with graduating seniors. As part of these interviews, the student can describe his most effective and least effective teachers. By combining information from these three sources it is more likely that a fair and accurate decision will be made. In cases where the instructor employs nontraditional teaching methods, it is a good policy to ask the candidate for a written description of his or her teaching methods and teaching goals. This information can be helpful in interpreting student ratings and colleague assessments.

The computer summary of the Endeavor ratings includes several special features which facilitate their interpretation. The first page presents a tabular comparison of the ratings on each of the seven items for all instructors within a predefined thematic group. Means for this group are also presented for each item for normative purposes. The instructor s copy identifies only his own course. The other 3 copies give the identity of all instructors. This tabular display addresses the question How am I doing in relation to my colleagues who are teaching similar courses?

The second page of the computer summary provides a scatter plot representing the two global rating factors for each instructor in the predefined thematic group. The instructor's copy identifies only his own course. The other 3 copies provide an identification key for all instructors. A sample scatter plot is reproduced below.

Each course is represented by a letter of the alphabet. The position of the letter in relation to the two axes indicates the mean ratings which were received on the two global factors. The pedagogical skill rating is reflected by the left right dimension. The rapport rating is reflected by the up down dimension. A letter in the upper right-hand corner indicates a high rating on both global factors. A letter in the upper left-hand corner indicates a high rating on rapport and a low rating on pedagogical skill. A letter in the lower right-hand corner indicates a low rating on rapport and a high rating on pedagogical skill. Thug scatter plot provides a succinct graphical summary of the global ratings of each instructor in the thematic group and provides an excellent starting point for making individual teaching assessments.

The third page of the summery provides a detailed breakdown of the instructor 8 ratings. For each item, a frequency count is provided for each response category and a mean and standard deviation is computed. This information is similar to that provided by most rating systems. In addition, global ratings are calculated for the rapport factor and the pedagogical skill factor. These global factors are calculated in the following way:

rapport =
-0.2 x mean rating on item 2 (plan)
+0.5 x mean rating on item 3 (disc)
+0.5 x mean rating on item 4 (help)
+0.2 x mean rating on item 6 (grade)

pedagogical skill =
+0.1 x mean rating on item 1 (work)
+0.4 x mean rating on item 2 (plan)
-0.2 x mean rating on item 3 (disc)
+0.3 x mean rating on item 5 (pres)
+0.4 x mean rating on item 7 (know)

The fourth page of the computer summary presents frequency histograms for each item. A sample is presented below. The height of the column of x s above each rating category (1 through 7) is proportional to the number of students who placed their rating in that category.

The histogram displays can be very informative. The sample page, for example, presents ratings which are highly item specific. The instructor received high ratings (mostly 7s) on advanced planning (item 2) but very low ratings (mostly 1 s) on class discussion (item 3). The heading at the bottom of the page indicates an enrollment of 46 in the course and therefore the absence of class discussion is not particularly surprising.

Another advantage of the histogram is its ability to clearly depict rater disagreement. On item 4 (personal help), the instructor received both high and low

ratings. The frequency distribution is clearly bimodal. In cases like this, the mean rating (5.19) is representative of neither the high raters nor the low raters. It is clear from the histogram, however, that the majority of the students gave the instructor a high rating (6 or 7) and a small minority gave him a very low rating (2). This rating pattern is not uncommon. It reflects a situation in which the needs and expectations of the class are not homogeneous. Although inconsistencies among the raters are reflected by the standard deviation measure, the histogram provides a more detailed account of bimodal response pattern and does not require prior statistical experience. By making it easier to detect bimodal response patterns, the histograms can alert an evaluation committee to look beyond a simple assessment of mean values.

The computer summaries also contain an additional page which tables the mean ratings for each department and the institutional means for each of the seven items. This information is useful for long-term analysis in determining whether the institution is making progress over time. These institutional means are probably not very meaningful as norms for individual instructors. Because of the differences among disciplines, departmental or divisional means will probably provide better norms for individual assessments.

Two copies of the computer summary are provided in bound form. The rationale for this is that the ratings are most meaningful when examined over a period of several years. The bound summaries are easily filed for future reference. When an evaluation decision is required, the individual s ratings across several years can be easily located and given careful and deliberate consideration. Without a bound copy, it is likely that the ratings for some courses might be misplaced.

It is important to recognize that student ratings are a form of behavioral measurement. All measurement involves error and measurement of attitudes and behaviors is especially prone to error. For this reason, it is essential that student ratings not be interpreted as magic indices which provide infallible information. In addition, the instructional ratings a teacher receives will differ from one class to another, even when the subject matter remains constant. Some semesters seem to be more successful than others and certain subjects are intrinsically more difficult to teach than others. Therefore it is important to examine ratings from at least 3 or 4 classes for each instructor before reaching any conclusion. In addition, it is helpful to compare the instructor s ratings with those of other individuals who have taught the same course or a very similar course. It would be unfair to penalize an individual because he or she was teaching in an important but not particularly popular subject area.

References and Bibliography

Brown, D.L. Faculty ratings and student grades: a university-wide multiple regression analysis. Journal of Educational Psychology, 1976, 68, 573-578.

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

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

Cohen, S.H. and Berger, W.G. Dimensions of students ratings of college instructors underlying subsequent achievement on course examinations. Proceedings, 78th Annual Convention, American Psychological Association, 1970, 605-606.

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

Crawford, P.L. and Bradshaw, H.L. Perception of characteristics of effective university teachers: A scaling analysis. Educational and Psychological Measurement, 1968, 28, 1079-1085.

Doyle, K.O., Jr. Student Evaluation of Instruction, Lexington, Mass.: Heath, 1975.

Ebel, K.E. The recognition and evaluation of teaching. American Association of University Professors, One Dupont Circle, Washington, D.C. 20036

Ebel, R.L. Estimation of the reliability of ratings. Psychometrika, 1951, 16, 407-424.

Faia, M.A. Teaching and research: rapport or mes alliance. Research in Higher Education, 1976, 4, 235-246.

Feldhusen, J.F. and Starks, D.D. Bias in college students ratings of instructors. College Student Survey, 1970, 4, 6-9.

Feldman, K.A. Grades and college students evaluations of their courses and teachers. Research in Higher Education, 1976, 4, 69-111.

Frey, P.W. Comparative judgment scaling of student course ratings. American Educational Research Journal, 1973, 10, 149-154.

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

Frey, P.W. The ongoing debate: student evaluation of teaching. Change, Feb., 1974, 47-49.

Frey, P.W. Student evaluation. Science, 1 975, 187, 557-558.

Frey, P.W. Validity of student instructional ratings: Does timing matter? Journal of Higher Education, 1976, 47, 327-336.

Frey, P.W. A two-dimensional analysis of student ratings-of instruction. Research in Higher Education, 1978, 9,69-91.

Frey, P.W., Leonard, P.W., and Beatty, W.W. Student ratings of instruction: validation research. American Educational Research Journal, 1975, 12, 435-447.

Genova, W.J., Madoff, M.K., Chin, R., and Thomas, G.B. Mutual Benefit Evaluation of Faculty and Administrators in Higher Education Cambridge, Mass.: Ballinger, 1976.

Gessner, P.K. Evaluation of instruction. Science, 1973, 180, 566-570.

Hildebrand, M. The character and skills of the effective professor. Journal of Higher Education, 1973, 44, 41-50.

Hoyt, D.P. Interrelationships among instructional effectiveness, publication record, and monetary reward. Research in Higher Education, 1974, 2, 81-88.

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

Kulik, J.A. and McKeachie, W.J. The evaluation of teachers in higher education. In F.N. Kerlinger (Ed.), review of Research in Education (Vol. 3). Itasca, Ill.: Peacock, 1976.

Marsh, H.W. The validity of student s evaluations: classroom evaluations of instructors independently nominated as best and worst teachers by graduating seniors. American Educational Research Journal, 1977, 14, 441 -447.

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

McKeachie, W.J. Student ratings of faculty. AAUP Bulletin, 1969, 55, 439-444.

McKelvie, S.J. Graphic rating scales — how many categories? British Journal of Psychology, 1978, 69, 185-202.

Menges, R.J. The new reporters: students rate instruction. In Pace, C.R. (Ed.), New Directions in Higher Education: Evaluation of Learning and Teaching San Francisco: Jossey-Bass, 1974.

Miller, G.A. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological Review, 1956, 63, 81-97.

Pasen, R.M. The differential effect of grade, sex, and discipline on two global factors: a within-class analysis of student ratings of instruction. Ph.D. Dissertation, Northwestern University, June, 1977.

Pasen, R.M., Frey, P.W., Menges, R.J., and Rath, G.J. Different administrative directions and student ratings of instruction: cognitive versus affective effects. Research in Higher Education, 1978, 9, 161-168.

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

Schein, E.H. and Hall, D.T. The student image of the teacher. The Journal of Applied Behavioral Science, 1967, 3, 205-237.

Shapiro, E. The effects on teaching of changes in relative rewards for research and teaching. Research in Higher Education, 1978, 9, 43-67.

Sullivan, A.M. and Skanes, G.R. Validity of student evaluation of teaching and the characteristics of successful instructors. Journal of Educational Psychology, 1974, 66, 584-590.

Ware, J.E., and Williams, R.G. The Doctor Fox effect: a study of lecturer effectiveness and ratings of instruction. Journal of Medical Education, 1975, 50, 149-155.

Williams, R.G. and Ware, J.E. An extended visit with Dr. Fox: validity of student satisfaction with instruction ratings after repeated exposures to a lecturer. American Educational Research Journal, 1977, 14, 449-457.