RICE UNIVERSITY

CONCEPTIONS OF EFFECTIVE TEACHING HELD BY FACULTY AND STUDENTS FROM FOUR ACADEMIC DIVISIONS

by

TODD E. MARQUES

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

MASTER OF ARTS

APPROVED, THESIS COMMITTEE:

Peter W. Dorfman, Assistant Professor
Chairman

David M. Lane, Assistant Professor

HOUSTON, TEXAS

MAY 1979

Conceptions of Effective Teaching Held by Faculty And Students From Four Academic Divisions

Abstract

The purpose of the study was to specify conceptions of "effective teaching" in terms conducive to the future development of a generally accepted, reliable, and valid system of instructional evaluation. Multiple regression was used to model individual conceptions held by male faculty (N=40) and male undergraduates (N=40) at Rice University. Faculty and student judges reviewed and rated profiles of 100 hypothetical instructors. The profiles consisted of a course subject matter designation and seven quantified cues referring to the instructors' performances on the following dimensions: lecture and/or presentation style (LECT), general rapport with students (RAPR), amount of information imparted in the course (INFO), arousal of student interest (AROU), clarity of course requirements and grading procedures (PROC), intellectual demand of the course (DEMD), and instructor's general knowledge of the field (KNOW). The judgmental policies of eight participants varied according to the subject matter designation. However, they did not vary in any normative or systematic manner. The non-configural raters (N=72) were included in a factorial analysis of group-related (i.e., Status X Discipline) differences in judgmental policy. The relative importance of the content (INFO, DEMD, KNOW) to style (LECT, RAPR, AROU, PROC) dimensions was greater for faculty judges. There was no evidence that policies are related to the raters' respective academic disciplines. Considering all raters, INFO received the highest average weighting,

followed by AROU, LECT and KNOW, RIGR and RAPR, and PROC dimensions. Four clusters of raters were identified by HIER-GRP (Human Resources Laboratory, USAF). The composition of each cluster was heterogeneous in terms of the status and academic discipline of the members. The policies characterizing the cluster memberships varied in two respects: (1) in the dimensions weighted most heavily, and (2) in the number of dimensions receiving substantial weight (i.e., policy complexity). Modifications of conventional student rating scales were suggested in view of findings from the present study.

ACKNOWLEDGMENTS

I would like to thank Drs. Peter W. Dorfman and David M. Lane for their valued guidance and insight. Thanks are also due to the members of the Rice University faculty who agreed to participate in the study. Without their cooperation the present study could not have been undertaken.

TABLE OF CONTENTS

INTRODUCTION

Student ratings of college teaching effectiveness are used for a variety of instructional and administrative purposes. Proponents of student ratings argue that they provide teachers with information conducive to instructional improvement. In addition to instructional purposes, student ratings are often used to aid in administrative decisions pertaining to tenure, promotion, salary, and other considerations (Greenwood, 1977; Luthans, 1967; Seldin, 1975).

The widespread use and general importance of student ratings have provided the impetus for a great deal of research. Questions concerning the utility, reliability, and validity of these ratings have been addressed frequently (Costin, Greenough & Menges, 1971; Rodin & Rodin, 1971; Spencer & Aleanoni, 1970; Sullivan & Skane, 1974). Although this literature has imparted useful information regarding certain technical aspects of student rating scales, it has not led to the development of a generally accepted, reliable, and valid system of instructional evaluation. To achieve this ideal system, "teaching effectiveness" first must be conceptualized in terms of behavioral dimensions which are agreed upon by faculty and students from various academic disciplines. In short, consensus must be reached regarding the nature of effective teaching before adequate evaluation procedures can be developed and implemented. Alternatively, separate conceptions may have to be identified for distinct academic disciplines if interdisciplinary consensus cannot be obtained. At any rate, it is extremely important that individual and group-

related conceptions of college teaching effectiveness be identified and described.

The primary objective of the present study was to determine the extent to which there is agreement among faculty and students from a variety of fields regarding the nature of effective teaching. A secondary objective was to determine whether or not individual conceptions of effective teaching vary according to course subject matter. Basically, the study was designed to ascertain which dimensions of teaching are most influencial in determining overall ratings of teaching. The present research served to model the processes by which individuals aggregate information about instructors to arrive at overall ratings. In the human judgment literature, a rater's (or judge's) strategy for aggregating evaluative data is often referred to as a "judgmental policy."

The present study was directed at several questions pertaining to possible differences in the judgmental policies of faculty and student raters who represented a variety of academic disciplines. For example, are there group related differences in the way judges in the humanities and social sciences view teaching effectiveness? Also, do the judgmental policies of raters vary according to the subject matter presented by the instructor? Finally, do the judgmental policies of faculty and students differ in any substantive way?

While the present study was largely exploratory in nature, two a priori hypotheses were entertained. First, a hypothesis pertaining to the relationship between rater status and judgmental policy was examined. This was students favor instructors high on

dimensions relating to student arousal, lecture style, rapport with students, and clarity of course requirements and grading procedures. By contrast, faculty favor instructors rated highly in terms of the intellectual demand of the course, the quality and quantity of information imparted, and the instructor's general level of expertise. The differences between faculty and student raters were hypothesized primarily on the basis of informal discussions with faculty and students and on the basis of literature reviews pertaining to student ratings of college teaching effectiveness (Costin et al., 1971; Follman, 1975; Wittrock & Lumsdaine, 1977).

The second hypothesis pertained to the effects of the rater's area of academic interest. It was thought that faculty and students representing the areas of engineering and natural sciences would favor instructors high in the amount of information imparted in the course and other content oriented variables. Raters representing the areas of humanities and social sciences were expected to view effective teaching more in terms of the instructor's general rapport with students, lecture style and arousal of student interest. The hypothesized differences in rating strategies were thought to be related to the nature of the subject matter which characterized the rater's primary area of academic interest. For example, consider physics, a course within the natural sciences. The existence of complicated paradigms relating to the essence of matter and energy places certain constraints on an instructor's teaching style. In many cases, the instructor must adhere to a carefully structured course format. This is true because the student's conceptual grasp of a particular

paradigm may be highly dependent on the grasp of more fundamental ideas and principles. Certain principles in physics must be mastered if a student is to become proficient in the field. Therefore, the amount of information imparted in the course, and the expertise with which the information is delivered constitute important considerations for evaluating instructors in the natural sciences.

As the hypothesis would suggest, courses in other disciplines may draw heavily on other dimensions of effective teaching. Consider an introductory course in art history. Most would agree that art is an extremely diverse form of human expression which is seldom amenable to the analytic procedures and theoretical orientations which characterize some other disciplines (e.g., physics). The educational objectives of an art history course may differ markedly from those of courses in other disciplines. In the case of art history, a highly effective instructor may try to cultivate an appreciation for a variety of art forms, or encourage class discussion of the social significance of particular works. This type of course may draw heavily on "stylistic" dimensions of teaching such as teacher-student rapport and arousal of student interest. The critical point is that faculty and students tend to evaluate teaching quality in terms of the traits which are most appropriate for their own respective disciplines.

In addressing these questions, it was necessary to invoke a two-stage procedure. The first stage entailed the compilation of a set of evaluative dimensions to be used in comparing and contrasting the judgmental policies of the various raters. The second stage

consisted of two parts: (1) the identification and description of individual judgmental policies, and (2) a factorial investigation of group related (e.g., faculty vs. student) differences in the judgmental policies of the participants. Research pertinent to both stages will be discussed in the following pages.

Dimensions of Effective Teaching

Numerous studies have been conducted with the intent of identifying the critical dimensions or hypothetical constructs that underly students' perceptions of effective teaching (e.g., Bendig, 1954; Coffman, 1954; Crawford & Bradshaw, 1964; Harari & Zedeck, 1973; Hildebrand et al., 1971; Holmes, 1971; Isaacson et al., 1964; Linn et al., 1974; Marques, Note 1; Solomon, 1966). The identification of these dimensions has been accomplished several ways. One widely used approach involves "factor analyzing" items taken from student rating forms. Based upon the interrelationships among responses to various items on a particular rating scale, a smaller set of dimensions or factors can be extracted. The resulting dimensions can be interpreted as hypothetical constructs which account for the observed interrelationships in the data. Examples of this approach include the Illinois Course Evaluation Questionnaire (Spencer & Aleamoni, 1970), the Purdue Rating Scale of Instruction (Bendig, 1954), the Student Instructional Report (Centra, 1973) and the Rice Teaching Effectiveness Questionnaire (Marques, Note 1). Results of the Rice Teaching Effectiveness Questionnaire (RTEQ) analysis indicated that the original scale of 25 evaluative items could be reduced to five independent

dimensions of teaching: (1) general or halo factor, (2) grading procedures, (3) course difficulty and workload, (4) personal opinion of the instructor, and (5) clarity of course requirements and grading procedures. The findings of representative factor analytic studies are summarized in Table 1 together with studies that used other techniques to identify critical dimensions of effective teaching.

Table 1

Findings of Representative Studies Dealing With the Dimensions Underlying Perceptions of Effective Teaching

Investigator(s)	Dimensions Identified
Bendig (1954)	General or halo factor
	Instructional competence
	Instructional empathy
Coffman (1954)	Empathy
	Organization
	Instructor's personal qualities
	Verbal fluency
Isaacson et al.(1964)	General skill level of the instructor
	Course workload and difficulty
	Organization and planning of course
	concern for quality of students' work
	Group interaction
	General rapport with students

Table 1 (continued)

Investigator(s)	Dimensions Identified
Solomon (1966)	Style of information presentation
	Communication skills
	Tolerance
	Permissiveness in course format
	Warmth
	Clarity of presentations
	Charisma
	Organization
	Self-confidence
	Impersonal vs. personal expressions
Crawford & Bradshaw (1968)	Knowledge of subject matter
	Organization and preparation of lectures
	Enthusiasm
	Willingness to help and interact with students
Hildebrand et al. (1971)	Analytic/synthetic approach to course material
	Interaction with groups
	Interaction with individual students
	Organization/clarity
	Dynamism/enthusiasm

Table 1 (continued)

Investigator(s)	Dimensions Identified
Holmes (1971)	Presentation quality
	Student-instructor interactions and evaluation processes
	Arousal and motivation of students
	Clarity of tests
Harari & Zedeck (1973)	Depth of knowledge
	Presentation style
	Organization
	Rapport with students
	Relevance of course material
	Testing and grading procedures
	Course workload
	Inspiration and motivation of students
Linn, Centra & Tucker (1974)	Teacher-student relationships
	Course objectives and organization
	Lecture quality
	Reading assignments
	Course difficulty and workload
	Examinations

Table 1 (continued)

Investigator(s)	Dimensions Identified
Marques (Note 1)	General factor (or halo)
	Grading procedures
	Course difficulty and workload
	Personal opinion of instructor
	Clarity of course requirements and grading procedures

Crawford and Bradshaw (1968) used another technique to isolate critical dimensions of teaching. They asked students to describe the salient behavioral characteristics of the most effective teachers they had ever had. The characteristics mentioned most frequently were (a) a thorough knowledge of the course material, (b) good preparation and delivery of lecture material, (c) enthusiastic and energetic teaching style, and (d) a friendly and helpful student orientation. In a similar study, Harari and Zedeck (1973) established a conference setting where small delegations of students could operate collectively to define important aspects of effective teaching. In addition to the dimensions named by Crawford and Bradshaw, these researchers reported three other aspects of teaching: (a) the general relevance of the course material, (b) adequateness of testing and grading procedures, and (c) course workload.

Examination of Table 1 will show that studies which have dealt

with the dimensions of effective teaching have varied in terms of the number of dimensions identified and in the naming of these dimensions. However, several dimensions seem to appear consistently. General rapport with students, lecture style, arousal of student interest, and course difficulty and workload are cited frequently. Taken collectively, the findings of these studies provide a "tentative consensus" of student perceptions of effective teaching.

A few studies have made indirect comparisons between faculty and student perceptions of effective teaching. Guthrie (1949) reported a correlation of .48 between "faculty-jury" scores and student-rating scores based on a sample of university professors. The faculty juries were small committees that were formed to make tenure decisions. Among the criteria they considered were: teaching effectiveness, research activity, departmental and campus involvement, value to the community at large, rapport with departmental members, knowledge of subject matter, general knowledge and range of interest, rate of personal growth, and professional recognition. Presumably, the students rated the instructors on conventional criteria discussed earlier, although no specific mention was made of the criteria used.

In another study, Lovell and Haner (1955) asked senior undergraduates to write two short essays: one describing the best teacher they had ever had, and the other describing the worst teacher they had ever had. From the two essays, 107 of the most frequently cited descriptive items were extracted. Fifty-three items pertained to the "best" teachers whereas 54 items pertained to the "worst"

teachers. Later, seniors (N=234) were given these items and asked to specify whether they were indicative of "best", "average" or "worst" teachers. Four years later, faculty were given essentially the same task. A correlation of .74 was obtained between student and faculty categorizations of the descriptive items.

More recently, Hildebrand et al. (1971) investigated student and faculty perceptions of effective teaching. These researchers found that the five most prominent dimensions for students were: (a) analytic/synthetic approach to the presentation of material, (b) organization/clarity of course material, (c) instructor-group interaction, (d) instructor-individual interaction, and (e) dynamism/enthusiasm of the instructor. Analysis of colleague ratings revealed that faculty characterize effective teachers by (a) research ability, (b) intellectual breadth, (c) participation in the academic community, (d) rapport with students, and (e) concern for teaching. Hildebrand et al. reported a high correspondence between faculty and student identifications of "best" and "worst" teachers.

It is important to note that many of the studies aimed at the critical dimensions of effective teaching (either from faculty or student perspective) have had two inherent weaknesses. First, some viewed the concept of teaching effectiveness as transituationally invariant (e.g., Crawford & Bradshaw, 1964; Lovell & Haner, 1955), a view which supposes that an effective teacher can be characterized by a combination of attributes which is desirable regardless of the educational setting, course, or any number of situational factors. The logical extension of this view is that a "model teacher" can be

identified, perfect for all schools, courses, students, and disciplines. Harari and Zedeck (1973) suggested the need to examine some of these situational variables when developing an instrument to measure teaching effectiveness. A second weakness which applies primarily to factor analytic studies of teaching is the failure to demonstrate the relationship between the various dimensions and students' perceptions of effective teaching (French-Lazovik, 1974). This failure has been attributed to the fact that many factor analytic studies did not include a criterion measure among the other items in the pool being analyzed. That is, many student (and colleague) rating scales do not have an item which can be taken as an overall rating of teaching effectiveness. The absence of this item makes it nearly impossible to determine the importance of the various items (e.g., lecture style, organization, etc.) in the raters' overall perceptions or conceptions of effective teaching.

French-Lazovik examined evaluation data collected at two universities over a period of 15 years in an attempt to determine the dimensions which were most predictive of students' overall ratings of teaching effectiveness. Her analysis involved a combination of factor analytic and multiple regression techniques. French-Lazovik noted that multiple regression analysis of student rating data is often impractical due to the large number of questionnaire items relative to the number of responses (i.e., observations). Alternatively, she utilized a reduced-rank regression model (Horst, 1941). Following the procedure developed by Horst, regression coefficients were based on the principal components factor matrix extracted from the complete

scale of questionnaire items. Because the factor matrix ordinarily has a much smaller rank than the correlation matrix, fewer variables are used to predict overall ratings of teaching effectiveness. The data collected from the two university samples were analyzed separately. High multiple correlations were obtained from each sample (.97 and .94) when items pertaining to "clarity of exposition", "arousal of student interest", and "stimulation or motivation to intellectual activity" were included in the reduced-rank regression equation. Items pertaining to the physical appearance and general demeanor of the instructor were of much less importance. Inspection of Table 1 will show that the dimensions of effective teaching revealed by French-Lazovik were typical of those described by other researchers. However, this study was important in the sense that it suggested a useful methodology for assessing the relative importance of various dimensions in determining overall ratings of teaching effectiveness.

The accurate description of judgmental policies used in evaluating teaching is largely dependent on the research related to the critical dimensions of effective teaching. Results of these studies, used in conjunction with multiple regression techniques to be described in the following pages, provide a methodology for the direct comparison of the judgmental policies of faculty and students from a variety of academic disciplines.

Policy Capturing and Clustering

Multiple regression is often used to model the judgmental processes used by individuals to arrive at overall ratings or judgments

in a wide variety of situations. Judgmental tasks which have been modelled by multiple regression techniques include: the evaluation of corn quality (Wallace, 1923), rating the job performance of nurses (Zedeck & Kafry, 1977), prediction of graduate school performance (Dawes, 1974), prediction of the existence of malignancies based on X-ray examination (Hoffman, Slovic & Rorer, 1968), and the motivational factors that lead students to seek employment at certain organizations (Zedeck, 1977). According to Dudycha and Naylor (1966, p. 583), the judgmental process can be expressed in terms of a least-squares prediction equation:

$$\hat{Y}_i = b_1X_{1i} + b_2X_{2i} + \ldots + b_kX_{ki} \qquad (1)$$

where: $\hat{Y}_i$ = predicted standardized response of a judge when presented with case i

$b_1, b_2, \ldots b_k$ = least-square regression weights for each of k cues or dimensions

$X_{1i}, X_{2i}, \ldots X_{ki}$ = standardized values for each of the cues in case i

When the least-squares prediction equation is based upon the evaluative responses of one or more raters, the equation is thought to represent the raters' judgmental policy. The regression procedure associated with the quantification of judgmental policies is typically referred to as "policy capturing." Naylor and Wherry (1965, p. 969) noted that a policy is captured "to the extent to which one can predict the actions of a rater from known characteristics of the stimuli he is being required to evaluate." A number of mathematical

models have been used to capture the judgmental policies of raters. However, they have been described most frequently in terms of an additive model. That is, on any given task, a judge's ratings are assumed to be based on a linear combination of whatever cues are available to him (Slovic & Lichtenstein, 1971).

Typically, a rater participating in a policy capturing study is presented with several quantified cues reflecting different qualities or dimensions of the entity being evaluated. The rater's task is to aggregate the information conveyed by the cues and arrive at a global rating or judgment of some sort. A regression equation is then formed to model the rater's judgmental policy if the following conditions are met: (1) each target of evaluation is evaluated in terms of a common set of cues or stimulus dimensions, and (2) the number of judgments made by a single rater is large relative to the number of cues on which the ratings are based. Perhaps an example will clarify matters. Assume that a rater has been asked to evaluate the overall job performance of an automotive assembly-line worker on the basis of three cues; working speed, error rate, and rate of absenteeism. The objective is to establish the relative importance of each cue in determining performance ratings of the workers. To accomplish this, the rater must evaluate many workers with respect to the same three cues. Using the obtained rating as a criterion variable, and the three cues as independent variables, it is possible to form a regression equation. This equation represents the judge's policy for rating the job performance of assembly-line workers. The beta weights obtained from these

procedures provide a basis by which to judge the relative importance of each cue.

Slovic & Lichtenstein (1972, p. 24) noted that "the linear model does a remarkably good job in predicting human judgments." This observation is borne out in the literature by the abundance of policy capturing studies pertaining to a wide variety of human judgment tasks. Many of these studies have used policy capturing purely as a vehicle for the elucidation of certain specialized judgment tasks (e.g., predicting graduate school performance). That is, they were content studies. However, for other studies (e.g., Anderson, 1977), the judgment task under study was ancillary to the methodological refinement of policy capturing techniques.

Policy capturing is idiographic in nature as it models the individual differences in judgmental policies. Just as it is sometimes useful to describe individual differences in judgmental policies, it is sometimes useful to describe the similarities among groups of raters. Theoretical parsimony is the primary rationale for "clustering" judgmental policies.

Basically, policy clustering is a technique for describing systematic or group-related similarities in the judgmental policies of raters. Generally, regression equations representing judgmental policies are grouped according to the homogeneity of the regression weights (Christal, 1968; Dudycha, 1970; Naylor & Wherry, 1966). While there are numerous algorithms for clustering regression data, the widely used computer program JAN (Judgment Analysis, Christal, 1968) and the lesser known HIER-GRP (used in the present study)

utilize an algorithm designed to systematically reduce a set of N regression equations (or systems) to a single equation which best represents the combined policies of all N judges. This is accomplished through the replacement of two systems with one compromise system at each of N-1 stages. At each stage, the two equations which exhibit the greatest homogeneity are combined. The resulting compromise system represents the joint-policy of two raters. When two compromise systems are replaced, a new compromise system results which represents the combined policies of members from both old systems. Although a number of grouping criteria exist, most JAN users utilize an option which minimizes the loss in R^2 associated with the combination of two systems (Christal, 1968).

Policy capturing and clustering techniques are widely used methods for assessing individual and group related differences (or similarities) in judgmental policies. As mentioned earlier, these techniques coupled with the findings of research related to the critical dimensions of teaching, constitute an important step toward the elucidation of different conceptions of effective teaching.

The present research plan was developed to test the effects of rater status and rater discipline on the weighting of critical dimensions of teaching effectiveness. The basic idea was to provide each person (judge) with a series of profiles describing different hypothetical instructors. The judges provided overall ratings of teaching effectiveness based on the information contained in each profile. The utilization of these procedures presented an opportunity

to obtain ratings of teaching effectiveness free from the biasing factors (Wittrock& Lumsdaine, 1977) and the extenuating circumstances (Scott, 1977) that have plagued the teaching evaluation process. That is, with no prior knowledge of the hypothetical instructors to be rated, the judges had only the information provided in the stimulus materials at their disposal. The absolute control of the experimental stimuli offered by the policy capturing approach was conducive to the straightforward representation of the inferential processes involved in the rating of college teaching effectiveness.

Profile Development

A number of factors were considered in the development of the profiles. Among these factors were: (a) the selection of cues or stimulus dimensions used to discriminate among the hypothetical instructor, (b) the method of quantifying the cues, (c) the empirical interrelationships among the cues, (d) properties of the scale used to assign the overall ratings, and (e) the number of profiles to be reviewed by each judge.

Clearly, it is essential to provide a judge with enough information to make a reliable assessment of teaching effectiveness. Yet, it is also important not to burden the judge with unnecessary details which may serve to blur or otherwise detract from more critical dimensions that warrant attention. The amount of information imparted in the individual profiles was determined on the basis of two considerations: (1) the number of independent dimensions thought to be related to effective teaching, and (2) the judge's capacity to

process information of this type. The decision to use the seven cues chosen for this investigation was based partly on a review of studies dealing with the dimensions of teaching (Table 1), the examination of written subjective ratings of instructors obtained at Rice, the factor analysis of the RTEQ, and personal discussions with students and faculty at Rice. Additionally, certain empirical issues were considered.

Empirical research dealing with the appropriateness of a given number of cues typically centers around the ability of judges to use the cues consistently. The consistency is often indexed by R^2. Hoffman and Blanchard (1961), Einhorn (1971), and others have noted that R^2 tends to decrease as the number of cues increases. However, there is little evidence to suggest that this reduction is directly attributable to increasing information processing demands. Rather, the decrease may be indicative of some configural combinatorial strategy imposed by the judge (Einhorn, 1971). However, Anderson (1977) found no significant differences in R^2 when profiles containing four, six, and eight cues were compared. Based on this evidence, and other findings by Phelps and Shanteau (1978), it was concluded that seven quantified cues would not exceed the information processing capabilities of the judges.

The profile cues (or dimensions of effective teaching) used in the study were quantified in terms of percentiles. Anderson (1977) demonstrated that numerical cues (e.g., a rating of 1.2 on a -1.5 to +1.5 scale) were aggregated more consistently than verbal cues (e.g., above average, very good). Moreover, Knox and Hoffman (1962) found

that judges combining cues quantified in percentiles were more reliable and yielded higher R^2 than judges dealing with another metric (T-score with fixed $\bar{X}$ & SD). More importantly, percentile values were used to make the judgmental task more explicit and population specific.

The empirical interrelationships among the stimulus dimensions were considered in designing the profiles. The approximate empirical independence of the cues was obtained by the random generation of percentile values for each cue. Thus, the cues were intercorrelated only to the extent of random sampling. The empirical independence of the cues was useful in assuring the interpretability of indices pertaining to the relative weight (assigned to each cue) in determining overall ratings of teaching (Darlington, 1968). There were additional advantages in maintaining an orthogonal cue structure. For example, Dudycha (1970) found that the clustering of judgmental policies (i.e., regression equations) was most successful when the profile cues were orthogonal. However, it should be noted that cue orthogonality can lead to the occasional occurrence of highly unrealistic profiles. Unrealistic profiles have been shown to effect raters' judgmental policies (Dudycha, 1966; Schenck & Naylor, 1968). For the present investigation, it was felt that the advantages accrued from an orthogonal cue structure outweighed the potential disadvantages. Furthermore, participants were advised of the hypothetical nature of the study in the instruction set. It was hoped that by being cognizant of the hypothetical nature of the study, the raters would be tolerant of any profiles viewed as unrealistic.

A 10-point scale was used in the overall ratings. The task required the judges to aggregate cues quantified in terms of percentiles. Essentially, the cues varied along a 99-point scale. Given the sensitivity of the cue scales, the use of a 10-point scale was thought to be more appropriate than the conventional four-to seven-point scales.

Each judge was given 100 different profiles to read and rate. No two judges rated an identical profile except by chance. That is, the departmental designation (coding for subject matter) and the cue values for every profile were generated randomly for each of the 80 participants in the study. The ordering of the cues remained fixed within each profile for all participants. Cue order was fixed to facilitate rapid information processing. Presumably, this reduced any monotony associated with the procedure by reducing the length of time needed to complete the task. The number of profiles was limited to 100 because it was felt that this was the approximate limit that would be processed conscientiously by the raters. Also, Dudycha's (1970) Monte Carlo study of capturing and clustering techniques provided evidence that 100 profiles should give stable weights.

METHOD

Subjects

Participants in the study included 40 male faculty members and 40 male undergraduates from Rice University. Equal numbers of faculty and students were distributed among four academic divisions: (a) engineering, (b) humanities, (c) natural sciences, and (d) social sciences.

Thirty-four of the undergraduates were enrolled in introductory or intermediate level psychology courses. Six students majoring in the humanities were recruited from other courses. With the exception of two engineering students, all undergraduates had either junior or senior class standing.

All faculty participants were actively engaged in undergraduate teaching. Faculty members were recruited individually for the study.

Materials

The participants received a packet of materials which consisted of an instruction set, a brief questionnaire, and a set of profiles representing 100 hypothetical instructors. The profiles contained seven cues. Each cue referred to a particular set of behaviors thought to be related to teaching effectiveness (see Figure 1). The seven cues were: (a) lecture and/or presentation style (LECT), (b) general rapport with students (RAPR), (c) amount of information imparted in the course (INFO), (d) arousal of student interest (AROU), (e) clarity of course requirements and grading procedures (PROC), and (g) instructor's general knowledge of the field (KNOW). A

randomly generated percentile value was placed to the right of each cue in the profile. The instruction set specified that the values were indicative of the hypothetical instructor's ratings on the various cues relative to other actual instructors at Rice University. In addition to the seven cues, each profile contained one of the following designations: (a) mechanical engineering, (b) art history, (c) physics, (d) psychology, and (e) undefined--no identification provided. Since the four departmental names as well as the undefined condition were assigned randomly to the profiles, each occurred with approximately equal frequency. The department names were chosen as representatives of the four academic divisions from which the participants came (i.e., engineering, humanities, natural sciences, social sciences).

Task

The participants were instructed to study the information contained in a profile and then to assign an overall rating of teaching effectiveness to the instructor depicted. The information contained in the profiles consisted of seven cues quantified in percentile values and, in most cases, the name of the department supposedly offering the course. Recall that approximately 20% of the profiles were without departmental identification. The judges rated the hypothetical instructors on a 10-point scale where the value "1" denoted the lowest possible rating and "10" denoted the highest possible rating.

Profile 1

Mech. Engineering

Attributes:	Percentiles:
Lecture and/or presentation style	87
General rapport with students	74
Amount of information imparted in the course	34
Arousal of student interest	85
Clarity of course requirements and grading procedures	36
Intellectual demand of the course	55
Instructor's general knowledge of the field	73

Overall rating of teaching effectiveness

1 2 3 4 5 6 7 8 9 10

Figure 1. Profile representing one hypothetical instructor.

RESULTS

Relationship Between Course Subject Matter and Judgmental Policy

The relationship between subject matter and judgmental policy was assessed individually for each subject as follows. The five departmental designations (mech. engineering, art history, physics, psychology, and undefined) which appeared in the profiles were coded in terms of four variables. (See Kerlinger & Pedhazur (1973) for a discussion of contrast coding). Twenty-eight additional

predictors were generated by multiplying each of the seven profile cues by each of the four coded variables. This procedure yielded a "complete" model based upon 39 predictors. The Subject Matter X Dimension interaction was assessed by comparing the R^2 obtained from the complete model of 39 predictors with the R^2 obtained from the reduced model comprised of seven dimension predictors and four categorical predictors. The significance of the difference between R^2 values was tested by a F-ratio of the following form:

$$F(PC\text{-}PR, N\text{-}PC\text{-}1) = \frac{R^2_C - R^2_R}{P_C - P_R} \Big/ \frac{1 - R^2_C}{N - P_C - 1} \qquad (2)$$

where $R^2_C = R^2$ for the complete model and R^2_R pertains to the reduced model; P_C = the number of predictors in the complete model ($P_C = 39$), and P_R = the number of predictors in the reduced model ($P_R = 11$); N = the total number of judgments made by each rater. The F-ratios obtained from each rater are presented in Appendix 1.

Of the 80 subjects in the experiment, eight showed strong evidence of using a configural aggregation strategy in rating the effectiveness of the hypothetical instructors. That is, the judgmental policies of these individuals varied according to the departmental identification contained in the profiles.

As a single set of regression weights could not adequately describe the policies of those raters who adopted configural aggregation strategies, separate equations were obtained for each departmental designation. The formation of separate equations was useful in examining the form of the interaction between subject

matter and dimension weight.

Among the eight raters who showed configural strategies, there were three faculty and five students. Of the faculty, two were natural scientists and one was in the humanities. Of the students, two were engineering majors, two were natural science majors, and one was a social science major. There was considerable variability in the form of the Subject Matter X Dimension weight interaction among these configural raters; these raters did not respond to the various course designations with similar modifications of their respective judgmental policies.

A brief discussion of the policies captured from three configural raters will exemplify the point. An engineering student assigned a negative weight to INFO (-.27) and KNOW (-.20) in art history courses but strong positive weights of .68 and .28 for INFO and KNOW when rating a physics course. For another engineering student, the most dramatic influences of course designation were seen in the weightings of INFO and RAPR dimensions. INFO ranged from a low of -.39 in art history to a high of .50 in psychology. At the same time, RAPR received the lowest value (-.14) in art courses and the highest value (.33) in physics courses. A social science rater weighted INFO most heavily (.99) when psychology instructors were evaluated. The next highest weight assigned for INFO was .48. This same student gave LECT the lowest value (-.27) for psychology and engineering courses and the highest value (.72) for physics courses.

Overall, while the configural raters varied the weight assigned

to a dimension as a function of subject matter, they apparently did not do so in any normative or systematic manner.

As stated earlier, the configural raters utilized strategies which were qualitatively different from those who adopted linear strategies. Therefore, the configural raters were not included in subsequent analyses which involved individual and group-related differences in the weighting of the seven dimensions of teaching effectiveness.

Group-Related Differences in Judgmental Policies

As an initial step in the analysis, a least-squares regression equation was formed on the basis of each non-configural judge's ratings of the 100 hypothetical instructors. This regression was done individually for each of the 72 subjects who showed no evidence of using a configural policy. The equation obtained from each rater was of the form:

$$Y_i = \sum_{j=1}^{k} \beta_j x_{ij} \qquad (3)$$

where: Y_i = the predicted standardized rating for a judge when presented with profile i ; β_j = the standardized regression coefficient (beta) associated with cue j ; x_{ij} = the standardized value of cue j in profile i . The equation defined the judgmental policy.

The mean beta weights and R^2 pertaining to the seven cues or dimensions are reported in Table 2 for each status-discipline grouping. The R^2 values corresponding to the raters' judgmental policies are found in Appendix 2.

Overall, INFO was the most important determinant of teaching effectiveness ratings. INFO was followed closely by AROU in terms of importance. LECT and KNOW were somewhat less important. RAPR and DEMD were less important than the dimensions already mentioned but were relatively more important than PROC.

Table 2

Mean Beta Weights and R^2 Associated With the Judgmental Policies of Faculty and Students From Four Academic Divisions

RATER	$\bar{R}^2$	LECT	RAPR	INFO	AROU	PROC	DEMD	KNOW
Faculty Engi.(N=10)	0.73	.14	.09	.60	.37	.12	.23	.22
Humanities (N=9)	0.66	.24	.11	.35	.38	.09	.20	.28
Natural Sci.(N=8)	0.69	.26	.16	.46	.20	.08	.24	.26
Social Sci. (N=10)	0.71	.28	.22	.35	.29	.13	.19	.31
Student Engi. (N=8)	0.68	.28	.14	.50	.20	.13	.11	.22
Humanities (N=10)	0.72	.20	.16	.44	.33	.07	.12	.19
Natural Sci.(N=8)	0.65	.38	.16	.25	.50	.15	.18	.20
Social Sci.(N=9)	0.56	.26	.14	.34	.37	.05	.09	.21
Overall Mean	0.68	.26	.15	.41	.33	.11	.17	.23

To test the hypothesis that faculty and students differ in the relative importance they attach to style and content dimensions of teaching, a measure consisting of the difference between the mean beta on the style dimensions (LECT, RAPR, AROU, PROC) and the mean beta for the content dimensions (INFO, DEMD, KNOW) was computed for

each subject.[1] The cell means derived from this procedure are reported in Table 3. As predicted, the relative importance of the content to style variables was greater for faculty, $F(1,64) = 5.96$, $p = .017$. The mean beta weight for faculty and student on each of

Table 3

Mean Beta for Style and Content Dimensions for Each Status X Discipline Grouping

Status/Discipline		Style	Content
Faculty	Engineering	.18	.35
	Humanities	.21	.28
	Natural Science	.18	.32
	Social Science	.23	.28
Student	Engineering	.19	.28
	Humanities	.19	.25
	Natural Science	.30	.21
	Social Science	.21	.21

the style and content dimensions are shown in Figure 2.

[1]Rather than actually compute these difference scores, a more efficient and algebraically equivalent procedure was used. Basically, an orthogonal linear contrast was applied to the cell means of each profile cue. The contrast compared the difference between two combinations of weights; one combination consisting of style dimensions and the other combination consisting of content dimensions. This was computed using the computer program described by Wright and Lane (1978).

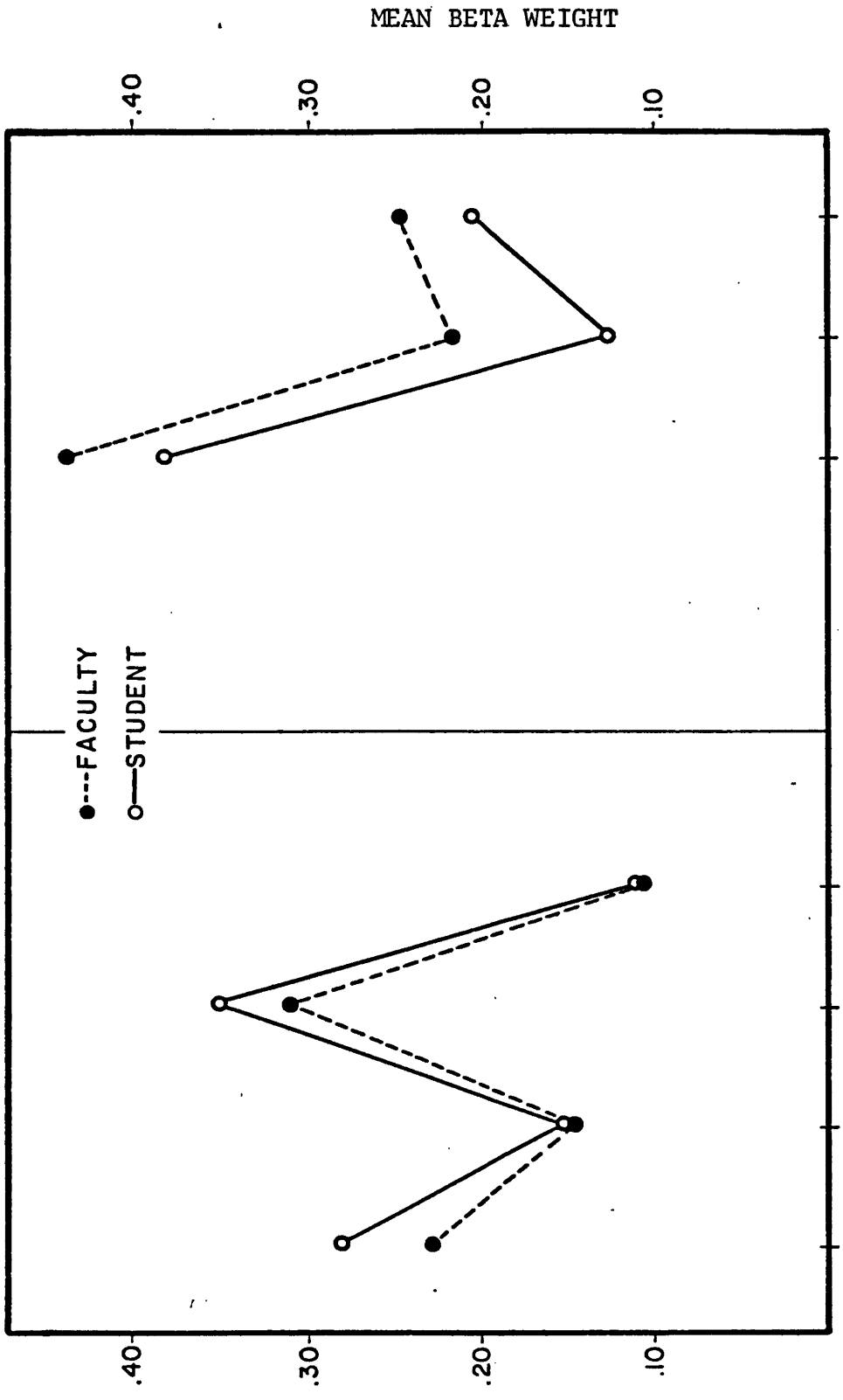

Figure 2 Faculty and student weightings of style and content dimensions of teaching

To estimate the degree of correspondence between faculty and students when rating an actual population of instructors, the following measure was derived:

$$r_{fs} = \frac{b'_f b_s}{\sqrt{(b'_f b_f)(b'_s b_s)}} \qquad (4)$$

where: b_s = the vector of standardized regression weights which defined the overall judgmental policies of students, and b_f = the vector of standardized regression weights associated with the overall faculty policy. Basically, r_{fs} is the correlation which would be obtained, in the limit, between ratings determined by the application of mean student weights and mean faculty weights to instructor profiles as the number of profiles approaches infinity. The derivation of r_{fs} is shown in Appendix 3.

A value of .98 was obtained for r_{fs}. This meant that ratings assigned to instructors by students and faculty would be correlated .98. The same measure was computed to assess the correspondence between equal weightings of profile cues and both faculty and student weightings. The correlations between ratings determined by equal weighting of profile cues and the empirically determined weighting schemes were .85 and .86 for students and faculty respectively.

A second hypothesis predicted specific discipline related differences in the judgmental policies of the raters. Contrary to the prediction, there was no evidence of a relationship between rater discipline and judgmental policy, $F(3,64) = 1.48$, $p = .229$. The interaction between rater status and rater discipline also failed to reach statistical

significance, $F(3,64) = 1.60$, $p = .20$. (See Appendix 4 for complete source table).

Policy Parallelism

Games (1973) has suggested the use of the omnibus test to see if unexpected effects are operating in a given study. Following this suggestion, profile analysis (Morrison, 1976) was used to probe for group related differences in judgmental policies which were not tapped by the planned comparisons discussed earlier. Specifically, multivariate tests of profile parallelism and flatness were conducted. These tests were used because of their sensitivity to potential group related differences in the configurations of multiple dependent measures (or beta weights in this case). Both tests of parallelism and flatness utilized the Pillai-Bartlett trace statistic (Lane & Bechtel, 1978). The test of flatness showed conclusively that judges do not weight the seven dimensions equally when rating teaching effectiveness, $F(6,59) = 26.12$, $p = .000$. A test of profile parallelism revealed no significant status related configurations of beta weights, $F(6,59) = 1.42$, $p = .221$. Similarly, there was no evidence to suggest that rater discipline was related to the overall configurations of beta weights, $F(18,61) = 0.87$, $p = .615$. The three-way interaction between rater status, discipline, and dimension was also non-significant, $F(18,61) = 1.11$, $p = .367$ (the Pillai-Bartlett trace statistics are included in the source table in Appendix 5). In short, the multivariate procedures demonstrated that the overall shapes of the judgmental policies were

not strongly related to the status or discipline of the rater.

Hierarchical Clustering of Equations

The program HIER-GRP (available from the Human Resources Laboratory, USAF) was used to systematically reduce 72 regression equations to a single system best represented the joint judgmental policy of 72 raters. An overall R^2 of .73 was obtained from 72 separate equations. The predictive efficiency dropped off to .36 when a single system was used to represent the policies of all raters. The R^2 associated with each reduction in the number of systems or clusters is shown in Table 4.

Four systems were examined in the present study. The systems were identified as Cluster 1 (n=6), Cluster 2 (n=27), Cluster 3 (n=29) and Cluster 4 (n=10). The R^2 associated with the four systems was .52. The decision to examine and report four systems was based on the contention that roughly 50% of the variance in real-world situations is explained by linear models (Slovic & Lichtenstein, 1971). Also a convenient dropoff in predictive efficiency occurred between the four and three system solutions (Table 4).

Table 5 shows the beta weights corresponding to each of the four systems. The composition of each system is described in terms of member status and discipline in Table 6.

Four of the six members of Cluster 1 were engineering faculty. The policies exhibited by the members of this cluster were characterized by strong concerns for INFO and KNOW dimensions of teaching.

Table 4

R^2 for Systems

Systems	R^2
72-54	.73
53-44	.72
43-37	.71
36-32	.70
31-27	.69
26-23	.68
22-20	.67
19-17	.66
16-15	.65
14	.64
13-12	.63
11	.62
10	.61
9	.60
8	.59
7	.58
6	.57
5	.55
4	.52
3	.48
2	.43
1	.36

Table 5

Beta Weights for the Dimensions in Each Cluster

Dimension	Cluster 1 (N=6)	Cluster 2 (N=27)	Cluster 3 (N=29)	Cluster 4 (N=10)
LECT	.15	.31	.18	.15
RAPR	.09	.17	.13	.12
INFO	.52	.19	.56	.42
AROU	.21	.43	.22	.26
PROC	.07	.13	.08	.09
DEMD	.15	.13	.19	.10
KNOW	.30	.26	.17	.21

AROU, PROC, and LECT dimensions were uniformly less important, while RAPR and PROC were regarded as unimportant.

Of the 27 members comprising Cluster 2, 16 were students. In fact, all non-configural raters majoring in the natural sciences were members. In contrast to Cluster 1, members of this cluster regarded AROU as the single most important determinant of effectiveness ratings. LECT and KNOW were of secondary importance, and RAPR, INFO, PROC and DEMD each received much lower weightings.

The membership of Cluster 3 was divided almost equally among faculty and students (14 and 15 members respectively). The membership was evenly distributed across disciplinary lines. Just as

Table 6

Composition of Cluster Memberships

Status and Discipline		Cluster 1	Cluster 2	Cluster 3	Cluster 4
Faculty	Engineering	4	0	5	1
	Humanities	0	3	3	3
	Natural Science	1	3	2	2
	Social Science	0	5	4	1
Student	Engineering	0	2	5	1
	Humanities	1	2	6	1
	Natural Science	0	8	0	0
	Social Science	0	4	4	1
Total		6	27	29	10

Cluster 2 focused intensely on a single dimension (AROU), Cluster 3 was focused on INFO. INFO was weighted twice as heavily as the dimensions of secondary import. The secondary dimensions included all remaining dimensions except for PROC which was found to be of nominal import.

Seven of the 10 members of Cluster 4 were faculty. While these individuals weighted INFO most heavily (as did Cluster 1 and Cluster 3 members), only slightly less weight was distributed among AROU, KNOW, and LECT dimensions.

DISCUSSION

It appears that few raters modify their conception of effective teaching according to the subject matter presented; 90% of those who participated in the present study showed no evidence of changing their judgmental policies as a function of this variable. The subject matter manipulation was used to operationalize a host of situational variables which may place constraints on, among other things, the method and rate of information dissemination, the flexibility of course curriculum, the general intellectual demand of the course, and grading and testing procedures. Most raters apparently do not feel that subject matter places these constraints on teaching. Alternatively, the raters simply may not take the information into account. The results suggest that both faculty and students tend to view the concept of teaching effectiveness as transituationally invariant. However, 10% of the participants in the study did vary the weights they attached to certain dimensions of teaching in response to changes in course subject matter. There was no apparent agreement in the manner in which the raters modified their conceptions according to course matter. Consequently, the data are difficult to interpret. Unfortunately, the relationship between course subject matter and judgmental policy has not been clarified in the existing literature. According to Costin et al. (1971), relationships have been examined between ratings and certain other qualities of the course and rater. For example, the ratings of psychology and non-psychology majors have been compared for certain courses. Other

comparisons have been made between student ratings and (a) the elective- or requirement-status of a course, (b) the level of the course (i.e., introductory vs. advanced), and (c) class standing. But, little work has been done with specific intent of explicating the weighting schemes which could be associated with different types of course subject matter. Additional research is needed to focus on individual perceptions of different academic areas and how they may be taught most effectively. Biglon (1973) has begun work in this area by comparing scholars' judgments about the similarities of subject matter from different academic disciplines. The findings of the present research suggest that factors (presumably) associated with different types of subject matter tend to be irrelevant in determining the perceived effectiveness of an instructor.

The hypothesis that faculty and students differ with respect to the weighting of style and content dimensions was strongly supported: students weighted each of the style dimensions more heavily than the faculty whereas the reverse pattern was found with the content dimensions. However, these differences were quite small. Although status differences were slight, they may be indicative of differing perspectives or orientations with regard to what should transpire in a classroom situation. Morstain (1977) examined the relationship between student ratings and the fit between student and instructor educational orientations. By educational orientation, Morstain (1977, p. 390) meant "orientations regarding the nature, purpose, and process of a college education." Basically, he found that when the educational orientation of the instructor was congruent with the

averaged orientations of the students, higher student ratings resulted. In a similar vein, perhaps there are group-related (i.e., faculty vs. student) differences in educational orientation which warrant exploration.

Clearly, students enroll in courses for many reasons; not solely for the purpose of accruing knowledge. For instance, students may enter a course with the idea of fulfilling university "distribution requirements" or perhaps to complete a prerequisite of some other desired course or goal. When this is the case, students may lack an inherent interest in the subject matter and may be preoccupied with the successful completion of the course. Therefore a good lecture or presentation style and the arousal of student interest are understandably important features of a course, and hence related to students' perceptions of effective teaching. However, faculty presumably have some inherent interest in the subject matter presented, and the reasons for their presence in the course may be less variable than students. Furthermore, some faculty members may see themselves primarily in the role of information disseminator, rather than as a stimulator of student arousal and intellectual curiosity. From this perspective, content dimensions appear to be relatively more important than style dimensions in determining the perceived effectiveness of an instructor.

Overall, faculty weighted the dimensions in the following order of importance: INFO (.49), AROU (.31), KNOW (.26), LECT (.23), RIGR (.22), RAPR (.15) and PROC (.11). Similarly, students weighted the dimensions: INFO (.38), AROU (.35), LECT (.28), KNOW (.20),

RAPR (.15), RIGR (.13), and PROC (.11).

The similarity between the policies raises an important question. That is, can the magnitude of the differences between faculty and student weighting schemes be regarded as important in any practical sense? Unfortunately, there are no hard and fast rules for assessing the practical importance of a particular finding in the present study. Ratings based on the mean student weightings would correlate .98 with those based on the mean faculty weightings if applied to an infinitely large number of profiles (see Equation 4). It is important to note that these findings are in basic agreement with Hildebrand et al. (1971), Gaff and Wilson (1971), and Lovell and Haner (1955) who reported a close correspondence between faculty and student ratings.

The prediction of discipline-related differences in the judgmental policies of raters was not supported. On the basis of the univariate and multivariate tests conducted, it is relatively safe to conclude that the concept of effective teaching is not closely related to the academic discipline of the rater. However, a failure to find large group related differences in the conceptions of teaching effectiveness should not be taken as an indication of large-scale agreement among individual raters. Actually, there was substantial variability in the weights attached to the dimensions. The mean and standard deviation for each dimension are presented in Table 7. The tabled values were based on the entire sample of nonconfigural raters (N=72). As can be seen, INFO, AROU, and KNOW had the greatest variability whereas there was considerable agreement in the weights assigned to RAPR and PROC dimensions. As an example of the

Table 7

Unweighted Means and Standard Deviations
For Each Dimension of Effective Teaching

Dimension	$\bar{X}$	S
LECT	.25	.15
RAPR	.15	.13
INFO	.41	.22
AROU	.36	.21
PROC	.11	.12
RIGR	.17	.18
KNOW	.27	.20

widely discrepant conceptions of effective teaching, consider two instructors in the humanities. The cluster identification, R^2, and dimension weights for these raters are presented in Table 8. From the data shown it is evident one of the humanities instructors relied almost exclusively on the INFO dimension when determining overall ratings. By contrast, the second instructor attached a great deal of weight to both AROU and KNOW dimensions. The remaining dimensions received nominal weightings. An obtained r value (Equation 4) of .26 indicated that the correspondence between these two policies was quite low. It should be emphasized that these discrepant policies were obtained from two raters who were representatives of the same Status X Discipline grouping. Considering the variability obtained

Table 8

Cluster Identification, R^2, and Dimension Weights Associated With the Widely Discrepant Policies of Two Instructors From the Humanities

Rater	Cluster Ident.	R^2	LECT	RAPR	INFO	AROU	PROC	DEMD	KNOW
Subject 1	Cluster 3	.93	-.02	-.02	.95	.26	.12	.10	.02
Subject 2	Cluster 2	.73	.11	.06	.05	.54	.11	.13	.65

within the groupings, the absence of significant between-group differences is not surprising.

Cluster analysis (Christal, 1965) was used to consolidate the various individual policies into parsimonious and meaningful groupings. In effect, this procedure released the individual raters from the a priori groupings defined by status and academic discipline and allowed them to congregate purely on the basis of policy homogeneity.

Inspection of the clusters extracted by the hierarchical grouping procedure (HIER-GRP) revealed four types of raters: (1) individuals who based ratings of teaching primarily on INFO and KNOW dimensions, (2) raters that are primarily concerned with AROU, (3) raters who weight INFO most heavily, and (4) individuals that attach substantial weight to a variety of dimensions which includes INFO, AROU, KNOW, and LECT. It is interesting to note that these raters differ in at least two respects: (1) the particular dimensions

weighted heavily, and (2) the complexity of their conceptions of effective teaching or, in other words, the number of dimensions receiving substantial weight. For example, the members of Cluster 2 tended to show singular concern for AROU while Cluster 4 members were concerned with INFO, KNOW and LECT as well.

Recall that the primary objective of the present investigation was the identification of various individual and group-related conceptions of effective teaching that exist on a college campus. The major premise being that the elucidation of views from faculty and students in a variety of academic areas will lead to the development of an accepted, useful, and valid system of instructional evaluation. Clearly, it is important to devise a method of instructional evaluation that will discriminate among instructors on the basis of those dimensions which are considered most important in college teaching. The dimension weights obtained in this study were indicative of raters' idealized conceptions of effective teaching. That is, the weights represent what the raters feel ought to be important in determining the effectiveness of an instructor. Overall, there was general agreement among faculty and students from all areas on the dimensions which carry the most weight in characterizing conceptions of effective teaching. INFO, AROU, LECT, and KNOW were considered most important by the majority of the raters. Of course, the relative weight assigned to these dimensions often differed dramatically among individual raters. However, tentative consensus has been reached with regards to what constitutes effective teaching. It follows that any adequate instrument designed to assess the effectiveness of instructors

should focus directly on measures which tap INFO, AROU, LECT and KNOW dimensions.

There are at least two essential features of an adequate evaluation instrument based on these dimensions: (1) a sensitivity to individual conceptions of effective teaching, and (2) a mechanism that will reduce the influx of irrelevant or biasing information in the aggregation of evaluative data.

Fortunately, it may be possible to satisfy both criteria with a series of refinements of the conventional objective rating scale. The refinements would include: (a) the elimination of items representing "overall" measures requiring aggregation of evaluative data, (b) the inclusion of several items to tap each dimension thought to be strongly related "effective teaching" (e.g., INFO, AROU, LECT, KNOW), and (c) the addition of a judgmental task requiring the rater to indicate the relative importance of each dimension in terms of his own conception of teaching effectiveness. If an instrument of this type were implemented, overall ratings of teaching could be determined empirically by the application of importance weightings of each dimension to the corresponding responses on the scale items. An overall rating obtained in this manner would not be unduly influenced by factors which are considered irrelevant (e.g., PROC) or biasing. Moreover, this instrument would be highly sensitive to idiosyncratic conceptions of effective teaching. However, further research is needed to verify the expected superiority of the proposed methodology. One important question which must be

addressed concerns the potential influx of biasing factors (e.g., personal opinion of the instructor) into an evaluative system that does not involve the subjective aggregation of evaluative data. That is, a rater's personal opinion of an instructor may lead him to assign importance weightings and questionnaire responses in a manner such that the derived overall rating of teaching effectiveness is actually indicative of a pre-conceived personal preference rather than performance on a given set of dimensions. It is crucial that the evaluation process measures performance and not personal preference. The present study has shed light on the meaning of "effective teaching." Hopefully, the goal of future research will be the adequate measurement of this concept.

Appendix 1

F-Test for the Effect of Course Discipline

on Each Rater's Judgmental Policy

Rater Status	Rater Discipline			
	Engineering	Humanities	Natural Science	Social Science
Faculty	1.49	2.06**	1.18	0.49
	1.11	1.56	1.09	1.32
	1.45	.159	0.61	1.09
	0.86	1.21	3.72**	0.90
	0.97	0.74	1.77*	0.79
	1.12	1.07	1.29	1.06
	0.95	1.12	1.35	0.82
	1.30	1.10	1.28	0.49
	1.50	1.02	1.03	0.72
	1.42	1.05	1.58	0.69
Student	0.74	0.99	1.09	0.92
	2.03*	1.06	1.19	1.99*
	0.86	0.78	0.88	0.75
	1.28	0.46	1.75*	0.44
	1.97*	0.58	1.08	1.15
	1.24	0.77	1.91*	1.04
	1.53	0.90	1.33	0.78
	1.35	1.02	1.03	0.97
	1.31	1.41	1.15	0.76
	1.39	1.18	1.22	0.75

* $p < .05$
** $p < .01$

Appendix 2

R^2 Associated With Each Rater's Captured Policy

Status	Rater Discipline			
	Engineering	Humanities	Natural Science	Social Science
Faculty	.62	.80	.82	.56
	.68	.73	.70	.46
	.86	.33	.67	.91
	.78	.67	.85	.91
	.59	.93	.66	.78
	.90	.62	.66	.91
	.88	.73	.78	.36
	.46	.66	.45	.68
	.87	.34	.61	.75
	.70	.79	.79	.71
Student	.71	.79	.45	.76
	.81	.73	.78	.59
	.80	.85	.57	.70
	.51	.69	.66	.20
	.60	.74	.53	.42
	.85	.58	.85	.58
	.54	.41	.77	.70
	.65	.76	.66	.73
	.74	.74	.64	.39
	.67	.85	.74	.58

Appendix 3

Recall that r_{fs} is the correlation which would be obtained, in the limit, between ratings determined by the application of mean faculty weights and mean student weights to instructor profiles as the number of profiles approaches infinity. Let X be an N X M matrix of N profiles and M cues, bf a vector of standardized regression weights associated with the overall faculty policy, and bs a corresponding vector associated with the overall student policy. The substitution of these terms into the well-known expression for Pearson's r gives:

$$r_{Xbf\ Xbs} = \frac{(Xbf)'(Xbs)/N}{\sqrt{\left(\frac{(Xbf)'(Xbf)}{N}\right)\left(\frac{(Xbs)'(Xbs)}{N}\right)}}$$

$$= \frac{bf'X'Xbs/N}{\sqrt{\left(\frac{bf'X'Xbf}{N}\right)\left(\frac{bs'X'Xbs}{N}\right)}} .$$

In standardized score form, $X'X/N$ is equivalent to the intercorrelation matrix R. Consequently:

$$r_{Xbf\ Xbs} = \frac{bf'Rbs}{\sqrt{(bf'Rbs)(bs'Rbs)}} .$$

Because the columns of X were generated independently, $R = I$, and therefore:

$$r_{bf\ bs} = \frac{bf'bs}{\sqrt{(bf'bf)(bs'bs)}} .$$

Appendix 4

A Priori Comparison of Group-Related

Judgmental Policies

Source	df	SS	ms	F	P
CT (Contrast) X St (Status)	1	0.2677	0.2677	5.9624	0.0174
CT X DS (Discipline)	3	0.2043	0.0681	1.4767	0.2294
CT X ST X DS	3	0.2160	0.0720	1.6035	0.1974
Error	64	2.8753	0.0449		

Appendix 5

Multivariate Test of Profile Parallelism

Source	Pillai-Bartlett Trace	df (NUM)	df (DEN)	F	P
DM(Dimension)	0.727	6	59	26.124	0.0000
ST (Status) X DM	0.127	6	59	1.424	0.2207
DS (Discipline) X DM	0.236	18	61	0.869	0.6154
ST X DM X DS	0.295	18	61	1.108	0.3671

REFERENCES

Anderson, B. L. Differences in teachers' judgment policies for varying numbers of verbal and numerical cues. Organizational Behavior and Human Performance, 1977, 19, 68-88

Bendig, A. W. A factor analysis of student ratings of psychology instructors on the Purdue Scale. Journal of Educational Psychology, 1954, 45, 385-393.

Centra, J. The student instructional report (Report No. 3). Princeton, N.J.: Educational Testing Service, 1973.

Christal, R. E. JAN: A technique for analyzing group judgment. Journal of Experimental Education, 1968, 36, 24-27.

Coffman, W. E. Determining students' concepts of effective teaching from their ratings of instructors. Journal of Educational Psychology, 1954, 45, 277-285.

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

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

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Dawes, R. M. Graduate admissions: A case study. American Psychologist, 1971, 26, 180-188.

Dudycha, A. L. A Monte Carlo evaluation of JAN: A technique for capturing and clustering raters' policies. Organizational Behavior and Human Performance, 1970, 5, 501-516.

Dudycha, A. L. & Naylor, J. C. The effect of variations in the cue R matrix upon the obtained policy equation of judges. Educational and Psychological Measurement, 1966, 26, 583-603.

Einhorn, H. J. Use of nonlinear, noncompensatory models as a function of task and amount of information. Organizational Behavior and Human Performance, 1971, 6, 1-27.

Follman, J. Student ratings of faculty teaching effectiveness: Rater or ratee characteristics. Research in Higher Education, 1975, 3, 155-167.

French-Lazovik, G. Predictability of students' evaluation of college teachers from component ratings. Journal of Educational Psychology, 1974, 66, 373-385.

Gaff, J. G. & Wilson, R. C. Faculty values and improving teaching. In G. K. Smith (Ed.), New teaching new learning, San Francisco: Josey-Bass, 1971.

Games, P. A. Type IV errors revisited. Psychological Bulletin, 1973, 80, 304-307.

Greenwood, G. E. Problems with an alternative to student rating systems. Paper presented at the ERIC/Higher Education Conference, Orlando, Florida, March, 1977.

Guthrie, E. R. The evaluation of teaching. The Educational Record, 1949, 30, 109-115.

Harari, O. & Zedeck, S. Development of behaviorally anchored scales for the evaluation of faculty teaching. Journal of Applied Psychology, 1973, 58, 261-265.

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

Hoffman, P. J. & Blanchard, W. A. A study of the effects of varying amounts of predictor information on judgment. Oregon Research Institute Research Bulletin, 1961.

Hoffman, P. J., Slovic, P. & Rorer, L. G. An analysis-of-variance model for the assessment of configural cue utilization in clinical judgment. Psychological Bulletin, 1968, 69, 338-349.

Holmes, D. S. The teaching assessment blank: A form for the student assessment of college instructors. Journal of Experimental Education, 1971, 39, 34-38.

Horst, P. Approximating a multiple correlation system by one of lower rank as a basis for deriving more stable prediction weights. In P. Horst (Ed.) The prediction of personal adjustment. New York: Social Science Research Council Bulletin, 1941, 48, 437-444.

Isaacson, R. L. et al. Dimensions of student evaluations of teaching. Journal of Educational Psychology, 1964, 55, 344-351.

Kerlinger, F. N. & Pedhazur, E. J. Multiple regression in behavioral research, New York: Holt, Rinehart & Winston Inc., 1973.

Knox, R. E. & Hoffman, P. J. Effects of variation of profile format on intelligence and sociability judgments. Journal of Applied Psychology, 1962, 46, 14-20

Lane, D. M. & Bechtel, L. R. A FORTRAN IV program for profile analysis. Behavior Research Methods and Instrumentation, 1978, 10, 49-50.

Linn, R. L., Centra, J. A., & Tucker, L. R. Between, within, and total group factor analyses of student ratings of instruction. Princeton, N. J.: Educational Testing Service, 1974.

Lovell, G. C. & Haner, C. F. Forced-choice applied to college faculty rating. Educational and Psychological Measurement, 1955, 15, 291-304.

Luthans, F. The faculty promotion process: An empirical analysis of the administration of large state universities. Iowa City: The University of Iowa, 1967.

Morrison, D. F. Multivariate statistical methods, (2nd ed.). New York: McGraw-Hill, 1976.

Morstain, B. R. Relationship of student and instructor educational orientations with course ratings. Journal of Educational Psychology, 1977, 69, 388-398.

Naylor, J. C. & Wherry, R. J. The use of simulated stimuli and the "JAN" technique to capture and cluster the policies of raters. Educational and Psychological Measurement, 1965, 25, 969-986.

Nie, N. H., et al. Statistical package for the social sciences (2nd ed.). New York: McGraw-Hill, 1975.

Phelps, R. H. & Shanteau, J. Livestock judges: How much information can an expert use? Organizational Behavior and Human Performance, 1978, 21, 209-219.

Rodin, M. & Rodin, B. Student evaluations of teachers. Science, 1972, 177, 1164-1166.

Schenck, E. A. & Naylor, J. C. A cautionary note concerning the use of regression analyses for capturing the strategies of people. Educational and Psychological Measurement, 1968, 28, 3-7.

Scott, C. S. Student ratings and instructor-defined extenuating circumstances. Journal of Educational Psychology, 1977, 69 744-747.

Seldin, P. How colleges evaluate professors. New York: Blythe-Pennington, Ltd., 1975.

Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. In L. Rappoport & D. A. Summers (Eds.). Human judgment and social interaction. New York: Holt, Rinehart & Winston, Inc., 1973.

Slovic, P. & Lichtenstein, S. Comparison of Bayesian and regression approaches to the study of information processing in judgment. Organizational Behavior and Human Performance, 1971, 6, 649-744.

Solomon, D. Teacher behavior dimensions, course characteristics, and student evaluations of teachers. American Educational Research Journal, 1966, 3, 35-47.

Spencer, R. E. & Aleamoni, L. M. A student course evaluation questionnaire. Journal of Educational Measurement, 1970, 7, 209-210.

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

Wallace, H. A. What is in the corn judge's mind? Journal of the American Society of Agronomy, 1923, 15, 300-304.

Wittrock, M. C. & Lumsdaine, A. A. Instructional psychology. In M. R. Rosenzweig & L. W. Porter (Eds.). Annual Review of Psychology (Vol. 28), Palo Alto: Annual Reviews Inc., 1977.

Wright, J. & Lane, D. M. A FORTRAN IV program for linear contrasts in designs with repeated measurements. Behavior Research Methods and Instrumentation, 1978, 10, 433-434.

Zedeck, S. An information processing model and approach to the study of motivation. Organizational Behavior and Human Performance, 1977, 18, 47-77.

Zedeck, S. & Kafry, D. Capturing rater policies for processing evaluation data. Organizational Behavior and Human Performance, 1977, 18, 269-294.

REFERENCE NOTE

Marques, T. E. Dimensionality and reliability of the Rice teaching effectiveness questionnaire. Unpublished manuscript, 1977.