

NPS55REZW72031A

# NAVAL POSTGRADUATE SCHOOL

## Monterey, California



ON THE QUANTIFICATION OF TEACHER PERFORMANCE

USING STUDENT OPINION

by

R. R. Read

H. J. Zweig

30 March 1972

Approved for public release; distribution unlimited.

20091105027

NAVAL POSTGRADUATE SCHOOL  
Monterey, California

Rear Admiral A. S. Goodfellow  
Superintendent

M. U. Clauser  
Provost

ABSTRACT:

Several scaling methodologies were applied to the data of an experiment which polled the opinion of graduating students. Inter-comparisons of the results point to problems in producing a stable one dimensional scale or grouping. Opinions on such items as organization, knowledge, mannerisms, etc., were solicited in an attempt to relate these to overall performance ratings.

Partial support of the Naval Postgraduate School ONR Foundation Research Program is given for this report.

Prepared by:



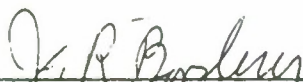
\_\_\_\_\_  
R. R. Read  
Department of Operations Research  
and Administrative Sciences



\_\_\_\_\_  
H. J. Zweig  
Department of Operations Research  
and Administrative Sciences

Approved by:

Prepared by:



\_\_\_\_\_  
J. R. Borsting, Chairman  
Department of Operations Research  
and Administrative Sciences



\_\_\_\_\_  
C. E. Menneken  
Dean of Research Administration

NPS-55REZW72031A

30 March 1972

#### ACKNOWLEDGMENTS

The experiment was suggested by Chairman J. R. Borsting, who appointed a departmental committee to administer it. The committee was chaired by N. K. Womer with members D. Courtney, R. S. Elster, P. R. Milch and R. R. Read. The use of a ladder was suggested by R. S. Elster. The data processing was performed at the school's W. R. Church Computer Center and the programming was done by Miss Patricia Combs. Professors J. K. Arima and A. F. Andrus assisted with the studies of auxiliary data. The project had the partial support of the school's ONR Foundation Research Program.

## TABLE OF CONTENTS

I.	Introduction and Description of the Experiment	1
II.	General Considerations of Ladder Data	5
III.	Comparison of Four Analyses of the Ladder Data	9
	a) Standard Scaling	10
	b) Range Scaling	11
	c) Scoring by (Modified) Paired Comparisons	14
	d) Factor Scoring	20
IV.	Appropriateness of a Single Scale	25
V.	Factor Analysis of the Nine Characteristics	30
VI.	Summaries of Auxiliary Data	35
	a) Quality Index of Response	35
	b) Frequency of Opinions on Selected Items	37
	c) Preference of Subject Areas	37
	d) Special Categorization Data	39
VII.	Results and Recommendations	41
	a) Summary	41
	b) Comments on the Students	43
	c) Future Research and Development	44
	Appendix A - Example Showing Nonsensical Results When Popular Scaling Techniques are Used	45
	Appendix B - The Questionnaire	50
	References	57

## LIST OF TABLES

3.1	Distribution of the Number of Respondents that Know A Teacher	9
3.2	Comparison of Four Scoring Techniques	13
3.3	Comparison of Range Scaled Scores of the Seven Instructors Seen by Both Groups of Students	18
3.4	Two Teachers Compared by Their Mean Scores on the Nine Characteristics	23
4.1	Testing the Significance of the Teacher Scores	27
5.1	Correlations of the Nine Characteristics	32
5.2	Factor Loadings for a Three Dimensional Solution	34
5.3	Frequency of Opinion on Selected Items	38
5.4	Win-Loss Weights for Preference of Subject Areas	39
A 1	Artificial Evaluations of Ten Objects by Six Judges	47
A 2	Win-Loss Matrix Based on Evaluations Shown in Table A 1	48

## LIST OF FIGURES

2.1	Exemplary Ladder Histograms	7
3.1	Two Hypothetical Ladder Distribution	11
3.2	Scatter Diagram of Standard and Range Scaled Scores	15
3.3	Scatter Diagram of Scores by Standard Scaling and Win-Loss Weights	17
3.4	Scatter Diagram of Factor Scores and Standard Scores	22
3.5	Scatter Diagram of Factor Scores and Win-Loss Weights	24
4.1	Histograms of Scores for Two Extreme Teachers	25
6.1	Percentage Response of Each Quality Index as a Function of Time from Graduation by Characteristic Statement	36



## I. INTRODUCTION AND DESCRIPTION OF THE EXPERIMENT

The systematic collection of data on student opinion on teaching is becoming widespread. At least two goals for such analysis are easily identified:

1. Instructors can gain insight into how they are viewed by the students in comparison to others.
2. Administrators receive a more uniform and systematic basis for their review of faculty.

Needless to say, administrative review of courses and instruction is by no means limited to student opinion.

The development of such a data collection system and the associated summaries and interpretations should be done with great care. There are a large number of potentially important variables that affect the student's reaction to an instructor and his course. Some of these variables are student oriented, e.g. his natural attraction to the subject matter, or his need to make up for poor grades received elsewhere. Some are course or subject oriented, e.g. a student's reaction to a basic requirement or service course may be different from that to an advanced or elective course; some courses are inherently more difficult to teach. Scheduling and allocation constraints can be important: Some instructors are frequently asked to teach outside of their area of specialization; small classes are inherently more easily handled than are larger ones. Most of all, these variables can be expected to interact with one another and with the teacher.

For the purpose of the first goal above, students typically are asked to rate their instructors by giving them scores on a list of characteristics that are related to the teaching-learning process. Such a list can take a myriad of forms. The number of characteristics, their specificity, and their importance to the process are important considerations in questionnaire design. The experiences of other institutions can be very valuable, but the current state of the art is such that each school probably needs to develop its own questionnaire.

With regard to the second goal the information from a questionnaire is certainly useful but not easily summarized. One can understand why the reviewers are tempted to place the instructors on a single scale. It is further tempting to have this done directly by the students rather than by projecting their responses from the questionnaire. The appropriateness of either of these is commonly challenged and some experimental work is called for.

This report deals with the results of an experiment conducted by the Operations Analysis Department and given to two graduating classes in September of 1971. The data are summarized and analyzed. The question of the appropriateness of a single scale is given the greatest attention. Recommendations for further developmental work are made and some areas of new research are identified.

The participating sections were the graduating classes in Operations Research (53 respondents) and Communications Management (15 respondents) of September 1971.



A copy of the instructions given to the students is contained in Appendix B. Although it was hoped to limit the experiment to 90 minutes, it proved to be much longer than that, and many students upon leaving indicated that they were not considering their answers carefully. They rushed to finish by 5 pm after beginning the questionnaire immediately following a full two-hour final examination. Also, interactions among groups of students within the room indicated that we were not getting independent opinions from each man.

Although the explicit details of the questionnaire appear in Appendix B, an overview follows. The students were asked to rate each professor that they had on each of nine characteristics:

1. Daily preparation.
2. Organization of the course.
3. Apparent knowledge of the subject matter.
4. Possession of distracting mannerisms.
5. Appreciation and understanding of the student's background and profession.
6. Adaptability and fairness.
7. Ability to communicate difficult ideas.
8. Ability to stimulate interest in the subject matter.
9. The amount which the student feels he learned from this instructor in this course.

The rating was done on an integer scale from 1 to 5. The characteristic statements were worded in such a way that the students were asked to indicate the intensity of their agreement or disagreement with those statements, a score of one being great disagreement and five representing great agreement. Characteristic statements 3, 4, and 7 were worded negatively so that high scores would be associated with poor performance. (This apparently caused some confusion.)

Parallel to these ratings, the student was asked to indicate the quality of his ratings. A quality index of one indicated that the rating was very sharp and could be supported with some classroom specifics. A rating of two indicated a fair amount of confidence in the rating, and a rating of three indicated that the student's recollection was vague. Also, the students were asked to indicate those instructors who made them struggle the most. The students were also asked which three courses they considered the easiest, which three the hardest, which one the most "relevant," and which one they would like most to repeat (and why).

After the questionnaire was completed, the students were asked to fill out a ladder sheet (see Appendix B) in which on a scale from 1 to 15 the students were asked to place their instructors in order of preference, and presumably the place on the ladder indicated the relative position with regard to other instructors that a student had seen. Also on the ladder, the student was asked to mark the minimum

level of acceptable instruction that in his opinion should be found at NPS. Finally, the students were asked to rate in order the subareas of their respective curricula. The Operations Analysis students were asked to state their preferences among the subareas of probability and statistics, computer programming and war gaming, linear programming and optimization techniques, human factors, and systems analysis. The Communications Management students were asked to rate the subareas of management, electrical engineering, computer science, and communications.

## II. GENERAL CONSIDERATIONS OF LADDER DATA

In trying to arrive at an overall rating scale for a set of professors based upon the judgment of a group, or several groups, of students, it is important to make the rating procedure for the judges, that is, the students, as simple as possible. If one insists that they rate the professors according to some ranking scheme or according to a scheme which forces a certain proportion into different categories, it can easily happen that one is building more structure into the ratings than may, in fact, be there inherently. Generally, psychometric research supports the position that one should leave the job of building structure into the data analysis rather than into the questions or task asked of the judges, see [9].

The procedure used was to ask each student to assign a number between one and fifteen to each of the professors that he is scoring. This allows ties. It allows, also, for a majority of professors to end up at the top, or at the bottom of the scale, and consequently,

it gives the student enough room to be comfortable in assigning numbers to the set of professors that he is rating. There was concern that the students might not spread out their data, but most students used essentially the full range.

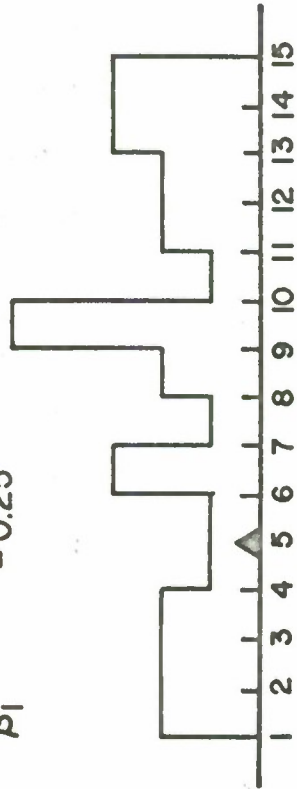
Once the ratings are obtained from the students, some transformation of the data is necessary to standardize the scales. If one student rates professors X, Y and Z respectively, at 5, 7 and 9, and another student rates the same professors 6, 8 and 10, it is legitimate to consider that they are really assigning them "identical" ratings, and that only their subjective judgment of the value of the numbers is in question. Carrying this thought to an extreme, suppose two other students rate these professors at 2, 4, 6 and 10, 12, 14 respectively. The magnitudes tell us something about the students' intensity of feeling towards the instructors, but their relative positions remain the same. Similarly, if a fifth student assigned the same professors the rating 5, 6, 7 one is very tempted to say that he is choosing a psychologically shorter scale, but is making the same judgment about the same professors.

Each student rated about 30 instructors. The ladders were summarized by computing means, standard deviations and coefficients of skewness,  $\beta_1$ , see [7]. Based on these, four were selected for exemplary presentation and they appear in Figure 2.1 in the form of histograms. The most typical ones are represented by students A and B, but with A having one of the larger standard deviations



30 Teachers Rated

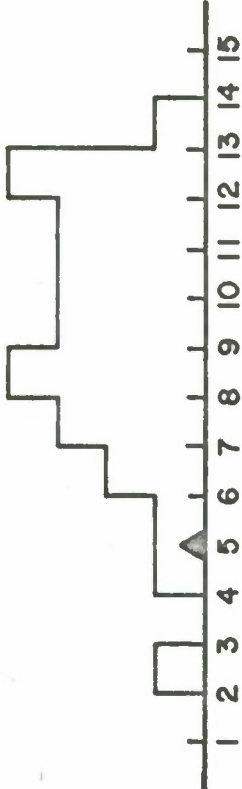
Mean 8.3  
Std. Dev. 4.3  
 $\beta_1$  -0.25



A

26 Teachers Rated

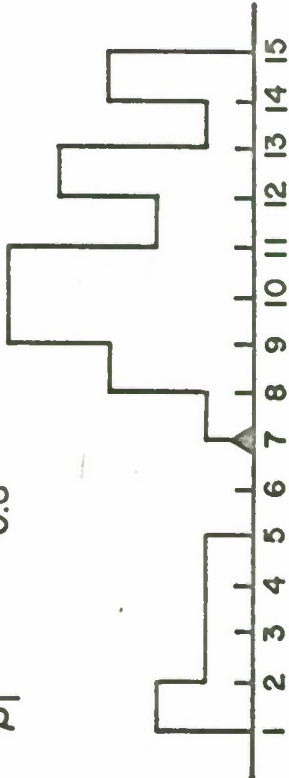
Mean 8.7  
Std. Dev. 2.7  
 $\beta_1$  -0.48



B

29 Teachers Rated

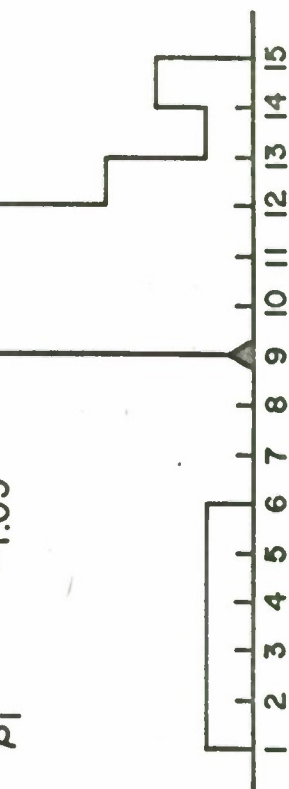
Mean 9.5  
Std. Dev. 3.9  
 $\beta_1$  -0.8



C

31 Teachers Rated

Mean 9.5  
Std. Dev. 3.4  
 $\beta_1$  -1.05



D

Figure 2.1

Exemplary Ladder Histograms  
▲ - Marks Minimum Acceptable Level

and B having a small one. About 20% of the ladders were noticeably skewed and these nearly always had negative coefficients  $\beta_1$ . Representing these are students C and D. Notice that student D placed his unacceptable instructors in single file, placed his minimum acceptable level at 9 and distributed the remaining to the right of that with greatest frequency near 9. Student C also emphasized his unacceptable instructors, but placed his minimum level at 7, giving himself more room to show some discriminating power for the acceptable ones.

The question of using all of the students' ladder data to arrive at a composite ranking of instructors is very delicate business, not only because the ladders themselves do not have a common frame of reference, but also because of the great imbalances in the experimental design. That is, there is great variability in the number of students that see each professor, and in the direct comparison of two professors, the number of students that see both is usually small and often zero.

The reader's attention is drawn to Appendix A where it is shown by example that common scaling methods can produce nonsensical results simply because the experimental design is out of balance. Objects are judged by different observers, each observer judges only a subset of the objects, and (contrary to the rating of teachers) there is very little disagreement of any two observers in judging the objects they both see.



### III. COMPARISON OF FOUR ANALYSES OF THE LADDER DATA

Having a general view of the set of ladders in hand, and being aware of the difficulties in analyzing such data, we turn to four different, but equally defensible ways to convert the ladders into scores for the instructors. The strengths and weaknesses of each are discussed and compared. This serves to illustrate the degree to which the difficulties are present in our data.

Of the 118 instructors evaluated in this experiment, 43 were eliminated from consideration because they were known by fewer than 5 of the respondents. Thus, 75 remain. The distribution of  $\{n_j; j = 1, \dots, 118\}$  where  $n_j$  = number of students that rated the  $j^{\text{th}}$  professor ( $t_j$ ), has some independent interest and is given in Table 3.1.

<u>n</u>	<u>Frequency</u>	<u>Tail Frequency</u>	<u>Percent</u>
0,4	43	118	100
5,9	10	75	64
10,14	17	65	55
15,19	11	48	41
20,24	10	37	31
25,29	7	27	23
30,34	4	20	17
35,39	6	16	14
40,44	1	10	8
45,49	3	9	7
>50	6	6	5

Table 3.1

Distribution of the Number of Respondents that Know a Teacher

As a note of caution in interpreting the results in this report it should be noted that nearly half the teachers were rated by fewer than ten students. The choice of 5 for the cutoff was arbitrary. The question of finding a better value is open and needs to be researched. It will be seen (Table 3.2) that a disproportionate number of high scoring teachers were associated with the smaller values of  $n$ . A similar result was mentioned in [4].

### Standard Scaling

Let us apply the common practice of removing the mean and dividing by the standard deviation to each ladder. This will be called the standard scaling in contrast to the range scaling that appears later. Then each professor can be scored by averaging the standard scores over the set of ladders that lists that professor. Such procedures would be most appropriate if each ladder contained a normal distribution.

The results of all four scoring techniques are listed and compared in Table 3.2. The standard scale ladder scores of each instructor were averaged to produce his standard score,  $S$ . The instructors are arranged in order of decreasing  $S$  so  $t_j$  will denote the  $j^{\text{th}}$  highest scoring teacher by standard scoring. The number of students that listed each professor is in the column headed  $n$ . Columns 3 and 4 contain the scores  $S$  and their standard deviations  $\sigma(S)$ . This latter quantity can be used as an index of controversy --- a small value indicating general agreement among the students in regard to the instructor's standard score and a

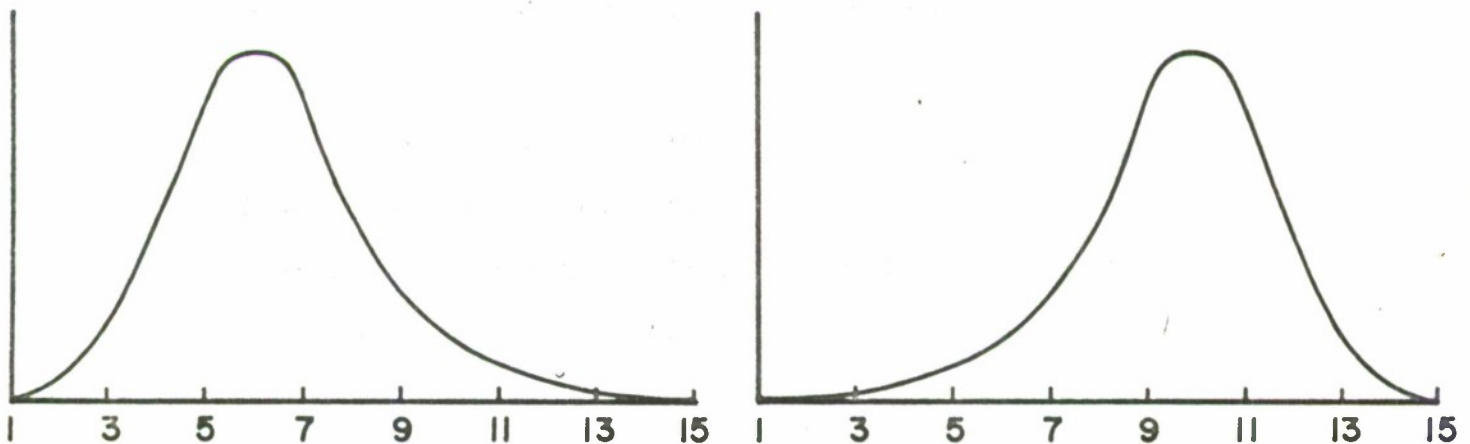
large value indicating disagreement. Note the singular disagreement of the 60<sup>th</sup> ranked instructor ( $t_{60}$ ) whose  $\sigma(S) = 1.44$  and the closest value is .72 for  $t_{74}$ .

### Range Scaling

An alternative scaling was also applied for the following considerations. Consider the two hypothetical ladder distributions (smoothed) of Figure 3.1.

Figure 3.1

#### Two Hypothetical Ladder Distributions



Student 1

Student 2

Student number 1 sees the professors as in the majority rather low since the peak of the curve is well below 8, but is impressed by one or two whom he rates high. The opposite holds for student 2 who thinks highly of most of his professors but has a low opinion of one or two. The mean for student 1 is about five, and for student 2 it

is about 10. Under standard scaling, instructors near the mean for each student will have both their scores transformed to about zero. On the other hand, a transformation of the data which tends to preserve more of the feelings shown in these two curves and at the same time does not alter their shapes is one which forces the top rating, whatever it is, to be a one, the bottom score to be zero, and assigns proportional numbers between zero and one to all the other professors that the student is scoring. This will be called range scaling. Thus, for example, if one student rates professors X, Y, Z, with 1, 3, 11, his range scaled scores would end up being 0, .2, and 1. If the second student rated the same professors 2, 8, and 10, his range scaled scores would end up being 0, .75, and 1. Any student who uses the complete scale from one to fifteen would end up having them essentially left alone. Whereas, a student who uses only a part of this range will have his ratings stretched to make them correspond to one who uses the complete scale.

Both standard and range scaling methods are "reasonable" procedures for normalizing the scales for a set of students. In so far as the results have any generality, they should be largely unaffected by which transformation is used, and in so far as there are any effects which depend on the transformation in question, these should be carefully studied to see to what extent they are in fact meaningful.



Comparison of Four Scoring Techniques

i	n	S	$\sigma(S)$	R	$\sigma(R)$	W	FS
1	19	1.39	.20	.92 (1)	.086	7.20 (1)	18.0 (3)
2	10	1.34	.14	.90 (2)	.096	6.42 (2)	17.2 (7)
3	36	1.29	.37	.88 (5)	.157	3.85 (5)	17.6 (4)
4	37	1.26	.23	.90 (3)	.128	5.60 (3)	18.1 (2)
5	7	1.24	.15	.89 (4)	.105	3.64 (6)	18.3 (1)
6	11	1.14	.19	.85 (8)	.160	4.11 (4)	16.1 (13)
7	19	1.14	.17	.86 (7)	.126	2.80 (7)	14.8 (19)
8	14	1.10	.15	.86 (6)	.135	1.50 (12)	16.5 (11)
9	12	1.06	.19	.85 (9)	.126	1.10 (16)	16.3 (12)
10	14	1.05	.29	.84 (11)	.137	1.15 (14)	17.4 (6)
11	17	1.05	.28	.83 (12)	.137	2.70 (8)	17.5 (5)
12	15	.91	.15	.81 (14)	.114	2.05 (11)	15.3 (16)
13	6	.90	.23	.83 (13)	.077	2.58 (10)	17.0 (8)
14	7	.87	.48	.80 (15)	.153	.77 (23)	14.6 (23)
15	6	.87	.14	.85 (10)	.086	2.63 (9)	16.7 (9)
16	8	.76	.54	.73 (19)	.234	1.41 (13)	15.4 (14)
17	47	.72	.61	.74 (17)	.215	1.12 (15)	13.9 (29)
18	14	.67	.23	.74 (16)	.142	.47 (33)	16.6 (10)
19	25	.65	.25	.72 (20)	.142	1.05 (17)	14.7 (20)
20	14	.64	.46	.73 (18)	.221	.42 (34)	14.3 (24)
21	27	.58	.24	.71 (21)	.136	1.02 (18)	15.1 (18)
22	32	.57	.31	.70 (22)	.165	.93 (20)	14.7 (21)
23	20	.57	.37	.69 (25)	.199	1.00 (19)	14.1 (25)
24	28	.52	.47	.69 (23)	.181	.79 (22)	12.4 (33)
25	21	.50	.61	.69 (24)	.237	.84 (21)	15.2 (17)
26	12	.47	.21	.68 (26)	.154	.29 (41)	14.6 (22)
27	6	.46	.28	.65 (31)	.171	.75 (24)	14.0 (27)
28	8	.38	.11	.65 (32)	.113	.28 (42)	14.0 (26)
29	22	.37	.51	.68 (27)	.215	.58 (28)	11.0 (42)
30	23	.37	.32	.66 (29)	.176	.59 (27)	11.1 (41)
31	14	.36	.31	.66 (28)	.162	.24 (45)	13.9 (28)
32	46	.36	.22	.65 (30)	.140	.67 (25)	13.0 (30)
33	40	.30	.61	.64 (33)	.227	.60 (26)	11.7 (36)
34	38	.30	.62	.63 (34)	.199	.54 (29)	12.6 (32)
35	53	.23	.72	.61 (36)	.252	.51 (30)	10.9 (43)
36	39	.23	.20	.61 (37)	.144	.48 (31)	13.0 (31)
37	5	.17	.46	.62 (35)	.179	.47 (32)	15.4 (15)
38	39	.07	.48	.57 (39)	.184	.36 (36)	10.8 (44)
39	54	.06	.37	.57 (38)	.174	.08 (61)	7.9 (55)
40	23	.05	.41	.56 (41)	.187	.33 (38)	12.1 (34)
41	51	.03	.33	.56 (42)	.169	.09 (55)	9.4 (49)
42	8	.01	.42	.57 (40)	.163	.33 (37)	11.7 (35)
43	12	-.03	.20	.53 (43)	.142	.29 (40)	11.4 (39)
44	21	-.07	.45	.53 (44)	.199	.25 (44)	8.9 (51)
45	52	-.11	.37	.52 (45)	.178	.26 (43)	11.1 (40)
46	16	-.16	.38	.50 (56)	.179	.08 (58)	7.4 (57)
47	28	-.17	.17	.50 (47)	.132	.15 (50)	9.7 (48)
48	53	-.18	.44	.50 (46)	.194	.23 (47)	10.5 (46)
49	24	-.22	.32	.48 (49)	.176	.20 (48)	10.3 (47)
50	29	-.22	.55	.48 (48)	.224	.23 (46)	8.5 (54)
51	24	-.36	.46	.45 (50)	.199	.13 (51)	9.3 (49)
52	11	-.39	.34	.45 (51)	.201	.16 (49)	8.9 (52)
53	13	-.49	.16	.38 (55)	.146	.04 (66)	7.0 (59)
54	21	-.55	.66	.41 (52)	.250	.37 (35)	11.5 (38)
55	6	-.56	.02	.40 (53)	.081	.10 (53)	10.7 (45)
56	28	-.56	.43	.39 (54)	.186	.12 (52)	8.6 (53)
57	16	-.59	.22	.36 (59)	.146	.03 (68)	4.4 (67)
58	46	-.60	.44	.37 (58)	.192	.08 (56)	7.5 (56)
59	12	-.63	.38	.32 (64)	.190	.04 (67)	4.8 (63)
60	51	-.66	1.44	.38 (57)	.319	.08 (60)	6.7 (60)
61	15	-.70	.32	.33 (63)	.184	.02 (70)	3.8 (71)
62	34	-.70	.53	.35 (61)	.211	.09 (54)	4.5 (64)
63	13	-.71	.41	.36 (60)	.194	.08 (59)	6.3 (61)
64	29	-.83	.61	.32 (66)	.228	.32 (39)	11.6 (37)
65	15	-.87	.57	.34 (62)	.189	.07 (62)	9.3 (50)
66	17	-.88	.69	.32 (65)	.220	.08 (57)	6.0 (62)
67	19	-.90	.40	.31 (67)	.177	.05 (64)	4.3 (69)
68	12	-.96	.06	.25 (68)	.097	.01 (73)	3.9 (70)
69	36	-1.12	.64	.24 (70)	.220	.05 (63)	4.5 (65)
70	34	-1.14	.60	.24 (69)	.215	.05 (65)	4.3 (68)
71	11	-1.31	.18	.13 (72)	.117	.01 (74)	1.2 (73)
72	33	-1.32	.46	.18 (71)	.168	.03 (69)	4.4 (66)
73	19	-1.61	.25	.11 (73)	.115	.01 (72)	1.9 (72)
74	19	-1.73	.72	.09 (74)	.199	.02 (71)	-1.2 (74)
75	14	-1.86	.04	.01 (75)	.045	.00 (75)	-2.9 (75)

## Legend

n	number of students that rated the $j^{\text{th}}$ instructor
S	average score by standard scaling
$\sigma(S)$	standard deviation of score by standard scaling
R	average score by range scaling
$\sigma(R)$	standard deviation of score by range scaling
W	score by Win-Loss weights
FS	factor scoring

The results of range scaling appear in columns 5 and 6 of Table 3.2. The average range scaled score is denoted by  $R$  and the standard deviation by  $\sigma(R)$ , also an index of controversy. The ranks of the range scaled scores are in parentheses next to the scores themselves.

The scatter plot of standard and range scaled scoring appears in Figure 3.2. The agreement is rather high, indicating that the decision as to which transformation is used will not greatly affect the outcome. The intensity of like and dislike is present however, as illustrated by two extreme cases selected from Figure 3.2, namely  $t_{15}$  and  $t_{59}$ . Under range scaling  $t_{15}$  is boosted to the 10<sup>th</sup> ranked instructor and  $t_{59}$  is drawn into the 64<sup>th</sup> position.

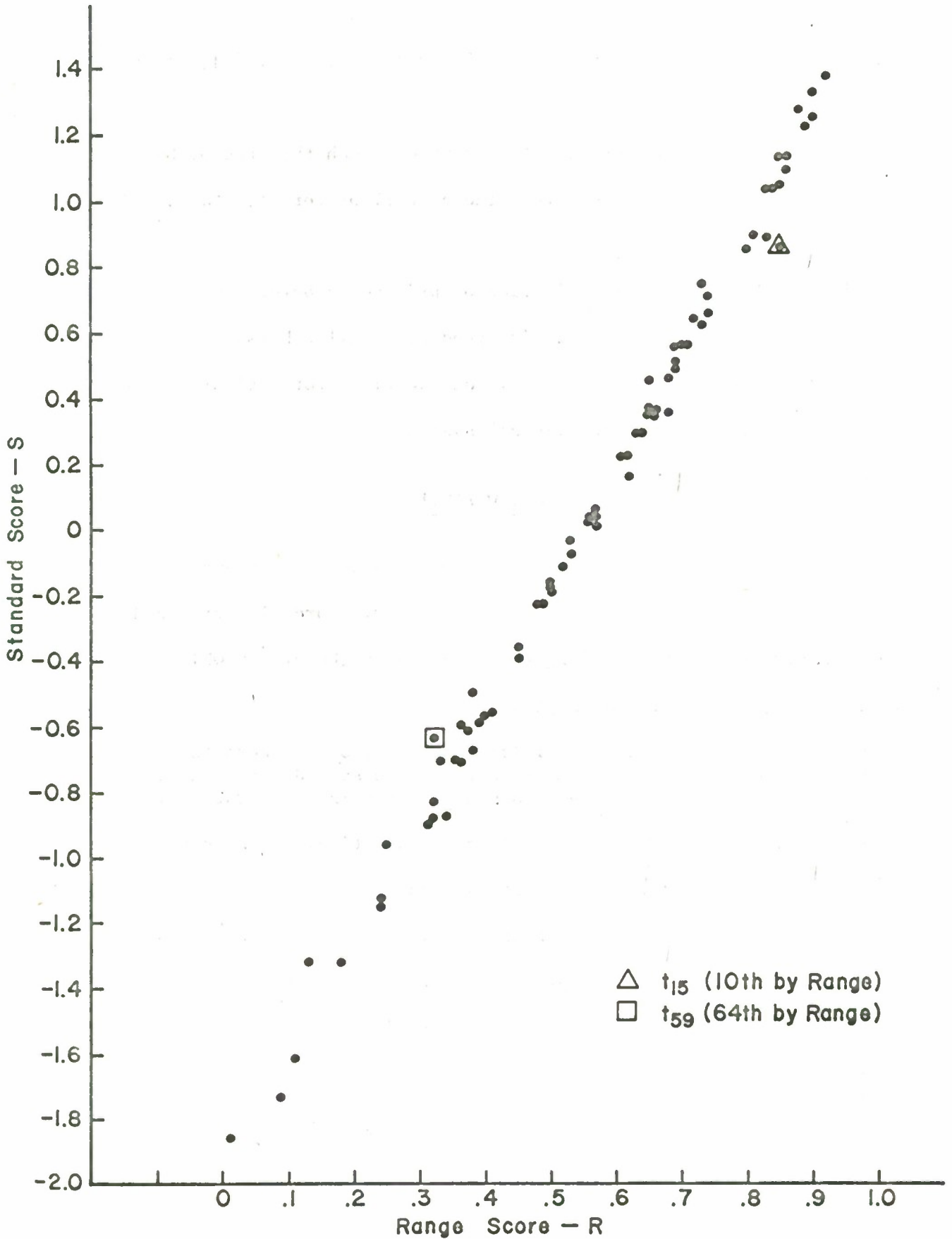
#### Scoring by (Modified) Paired Comparisons

Let us compare each pair of instructors by recording only whether a student prefers one to the other. Such data can be extracted from the ladders. A win-loss matrix is generated whose  $(i,j)$ <sup>th</sup> element is the number of students who prefer the  $i$ <sup>th</sup> professor to the  $j$ <sup>th</sup> one (1/2 is tallied for each tie).

Analysis of such preference data by paired comparisons is widespread (see [9], [10]) largely because it makes no assumptions concerning the criteria by which objects (teachers) are compared. Hence, it eliminates entirely the intensity of feelings. The usual technique requires a balanced experimental design. This being absent, some modifications are called for.



Figure 3.2  
Scatter Diagram of Standard and Range Scaled Scores



A method for converting a win-loss matrix into weights that can be used for scoring has been developed by Ford, Ref. [2], under the following postulates:

- (i) There exists a system of weights  $w_i$  such that the probability that a randomly selected student will prefer  $t_i$  to  $t_j$  is given by  $w_i/(w_i+w_j)$ .
- (ii) Each student's determination of preference between each pair of instructors is independent of all others.

Ford showed that the maximum likelihood equations for estimating the weights are given by the (nonlinear) system

$$\sum_j A_{ij} = \sum_j A_{ij} w_i / (w_i + w_j)$$

where  $A_{ij}$  is the  $i, j^{\text{th}}$  element of the win-loss matrix and  $\sum_j A_{ij}$  = total number of "wins" of the  $i^{\text{th}}$  instructor. Ford also produced an iterative scheme for finding the solution of the system which will be unique under the assumption:

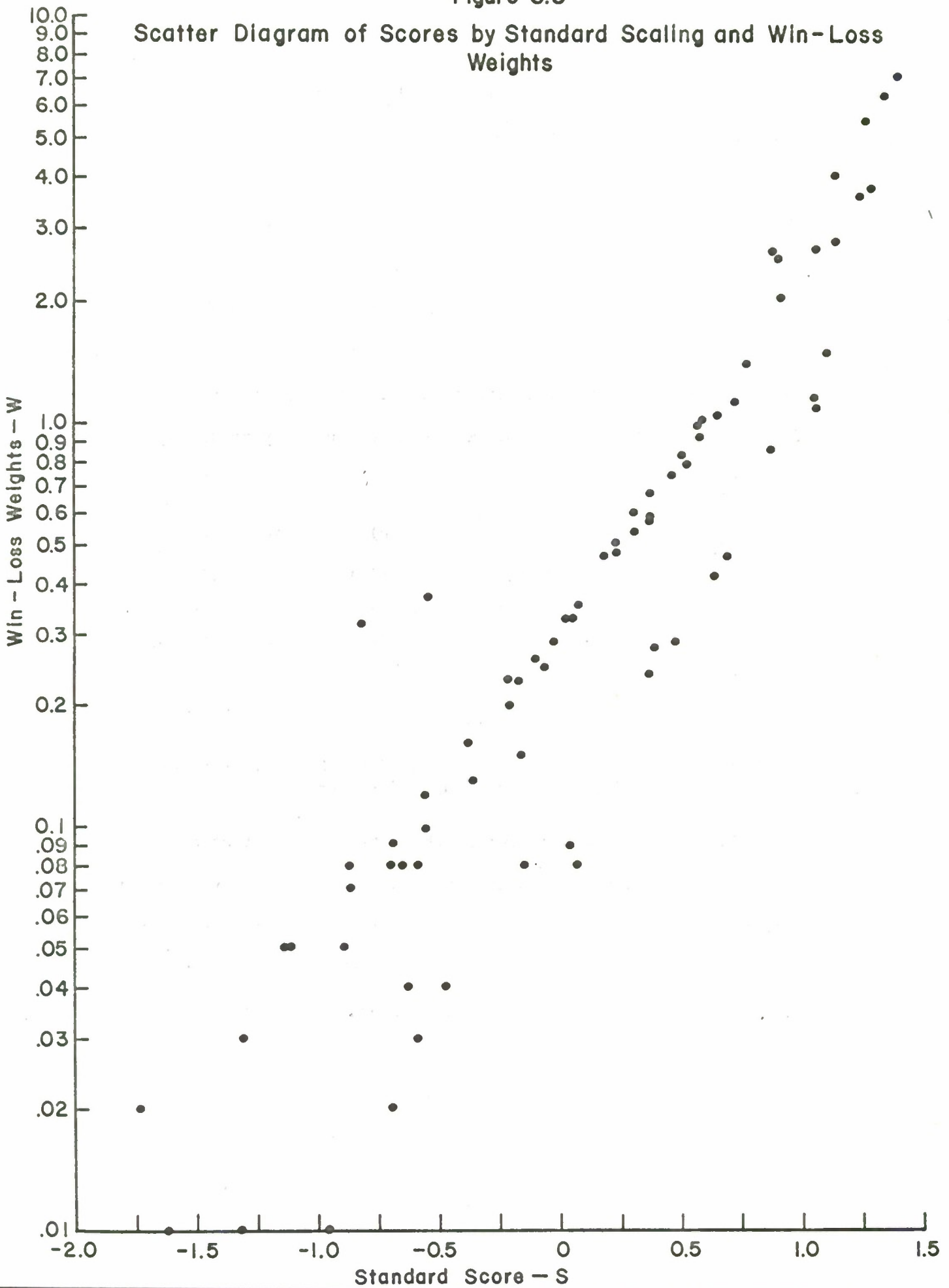
- (a) For every possible partition of the set of teachers into non-empty subsets, some teacher in the second set will be preferred at least once to some teacher in the first set.

In general, it is difficult to meet this assumption, but no serious violation of it occurred in the data at hand.

This procedure has been proposed for data such as ours, [8]. We note in passing that the application of Ford's procedure to the artificial evaluations of Appendix A resulted in total recovery of the original ordering.

Figure 3.3

Scatter Diagram of Scores by Standard Scaling and Win-Loss Weights



The estimated scoring weights  $W$  appear in Column 7 of Table 3.2. Again, the associated ranks are in parenthesis.

Figure 3.3 contains the scatter plot of these weights with the standard scores on a semi-logarithmic scale. These data divide into at least two populations. Of the nineteen points in the lower portion of Figure 3.3, sixteen represent teachers who were seen exclusively by the communications management students. Seven teachers were seen by both communications management and operations analysis students, and hence, fifty-two were seen exclusively by the operations analysis students.

$R(n)$	$R_{OA}(n)$	$R_{CM}(n)$	$R(n)$	$R_{OA}(n)$	$R_{CM}(n)$
.38 (51)	.19 (36)	.86 (15)	.45 (24)	.49 (13)	.40 (11)
.86 (19)	.82 (14)	.96 (5)	.37 (46)	.38 (32)	.34 (14)
.41 (21)	.35 (18)	.79 (3)	.48 (24)	.49 (20)	.42 (4)
.86 (14)	.80 (1)	.87 (13)			

Table 3.3

Comparison of Range Scaled Scores of the Seven Instructors  
Seen by Both Groups of Students

The comparative opinions by these two groups of students of the set of seven instructors they had in common was sharp enough to separate the data. On balance the communications management students awarded these seven higher scores than did the operations analysis students as may be seen in Table 3.3. There the range scaled scores are listed and decomposed into scores awarded by the two groups. The parentheses contain the number of students contributing. The four teachers on the left were rated significantly more highly by

the communications management students. The three teachers on the right were rated lower by this group, but not as significantly.

Two outlier points appear above the main body of data of Figure 3.3, namely,  $t_{54}$  and  $t_{64}$  whose win-loss weights are great enough to boost their ranks to 35 and 39 respectively. The three remaining points of the set of nineteen below the main body are also regarded as outliers  $t_{39}$ ,  $t_{41}$ ,  $t_{73}$ .

A further anomaly is illustrated by comparing the win-loss weights of  $t_{64}$  and  $t_{62}$ . From Table 3.2 those values are seen to be .32 and .09 respectively. Direct comparison of these two via the win-loss matrix  $A$  yields  $A_{62,64} = 10$  and  $A_{64,62} = 11$ . Thus, the indirect comparisons have separated these two teachers either because the set of judges that saw both of them also rated rather different sets of teachers or because the judges themselves were not in general agreement or both.

In conclusion, it is seen that the Ford procedure appears to have power of resolution. The role played by the set of seven teachers is very important. They set the bench marks. If all students could see a certain handful of teachers and there was great agreement about the teaching ability of each in this handful, then we could go ahead and score all the other teachers. Of course, our present set of seven does not play this role. The two groups of students do not agree about them and the question of who sees whom continues to confound us.



### Factor Scoring

The idea of Factor Scoring is to assign weights to the values given to the professors on each of the nine characteristic statements and assign a score by totaling the products of weight times value. There is a great deal of choice in the details of doing this and the one used in the current illustration was chosen almost entirely for expedience in programming.

First, let us describe the general rationale. The judges (students) give a numerical value to each of the objects (teachers) on each component of a set of specific dimensions (characteristics). Generally, it is not known in advance which dimensions are important either in terms of discriminating among the objects or in relating to an external function, such as quality of teaching. However, one can force the relationship with the external function to be as large as possible by restricting attention to the extremal groups, in this case, the best and worst teachers.

Suppose such a set has been defined. The correlation matrix of the characteristics is computed for this set. The components of the principal eigenvector of this matrix will serve as the weights whose use will produce the greatest possible separating power of a one dimensional score. (Ref. [3]).

Specifically, the following was done to the data. The mean value for each of the nine characteristics was computed for each of the seventy-five instructors. A set of best and worst professors was defined (see section IV for how this was done) and the correlation



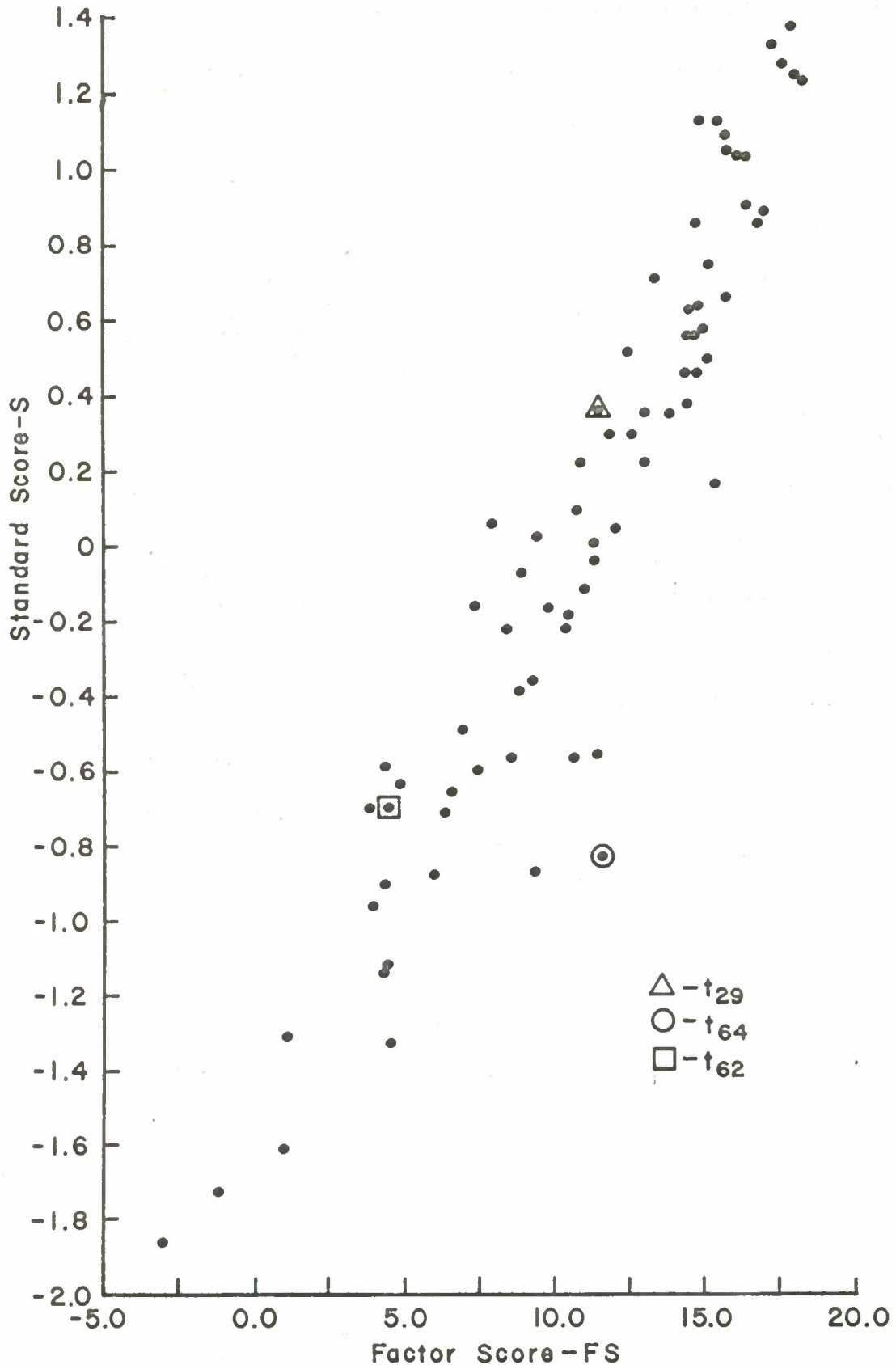
matrix for this set was computed. A Factor Analysis routine was applied to this matrix and the components of the principal eigenvector are used as weights which are applied to the normalized values of the nine characteristics for all seventy-five instructors. This list of factor scores, FS, appears in column 8 of Table 3.2 and the ranks of the scores appear in parentheses.

All of the nine characteristics used in this experiment are obviously related to the students' view of good teaching. It is not surprising that the "best" teachers rate very high in all nine characteristics and the "worst" rate low in all of them. We would not expect such uniformity would prevail except at the extremes of the scale. This is discussed further in Section V.

Figure 3.4 gives the scatter plot of these factor scores with the standard scores. The nine characteristics may be missing something of general importance because  $t_{29}$  and  $t_{64}$  have about the same factor score. It is noted that these two teach in the same area. Indeed the latter would rank 37 and the former 42 on that scale. It is seen that many other pairs of teachers having the same factor score are nearly as far apart in their standard scores. Thus, a general ingredient of teaching may be missing from our list.

Turning to the horizontal variability in Figure 3.4, let us compare  $t_{62}$  and  $t_{64}$  who are about the same on the standard scale, but are ranked 64 and 37 (respectively) on the factor scale. This great change occurred largely because of superior knowledge of

Figure 3.4  
Scatter Diagram of Factor Scores  
and Standard Scores



the subject and greater adaptability on the part of  $t_{64}$  as seen in Table 3.4. Both instructors taught the operations analysis students.

Char.	1	2	3	4	5	6	7	8	9
$t_{62}$	2.4	2.4	3.6	3.0	3.0	2.8	3.3	2.8	2.5
$t_{64}$	2.1	2.2	1.3	3.5	3.6	3.8	3.0	3.1	2.7

Table 3.4

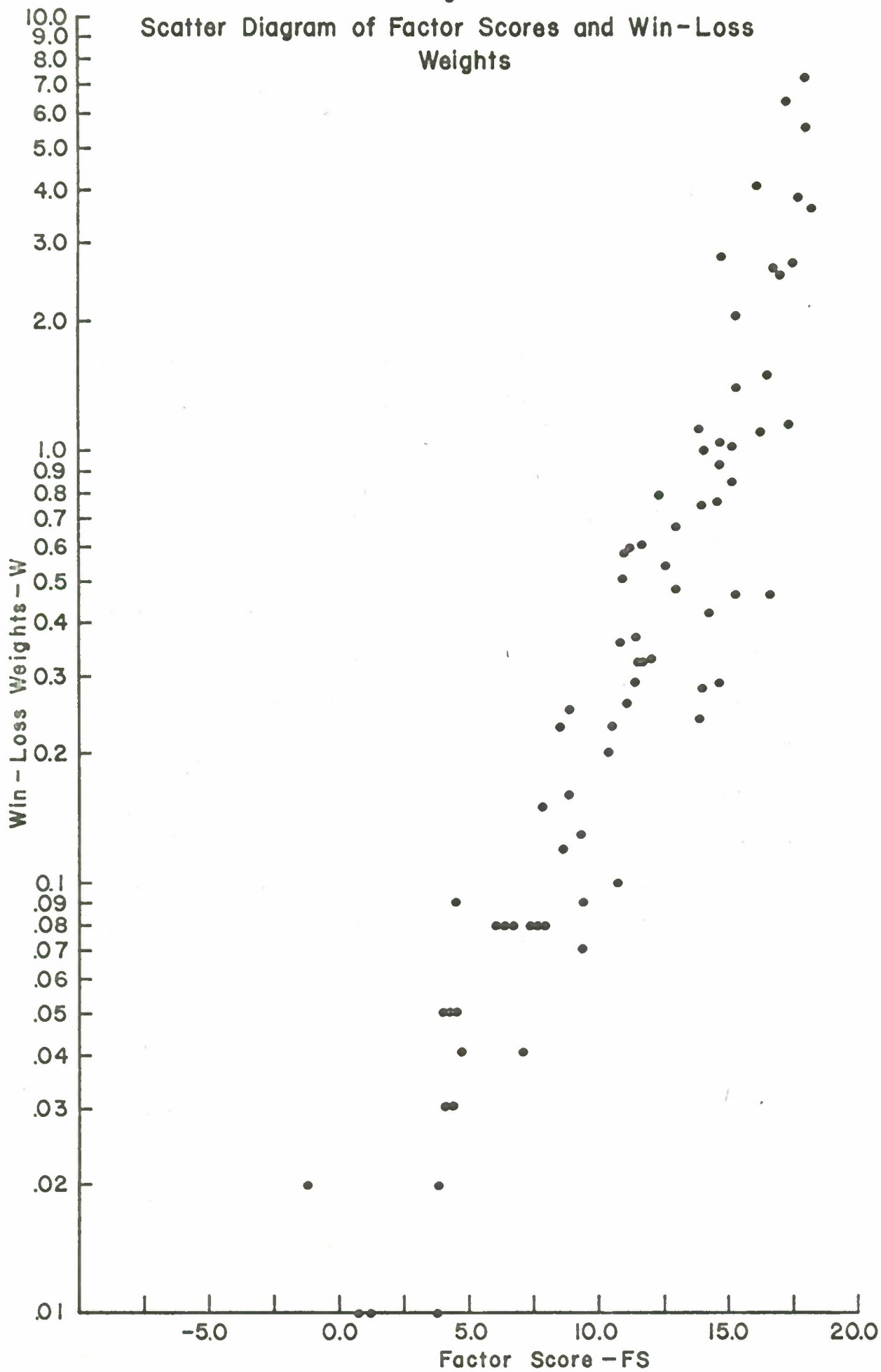
Two Teachers Compared by Their Mean Scores  
on the Nine Characteristics

Finally, Figure 3.5 shows the scatter plot of factor scores and win-loss weights. The most striking feature is that these data are not resolved into two groups as occurs in Figure 3.3. Perhaps this is to be expected since the factor scores reflect only performance on the specific nine characteristics. But we do not know if this would have been preserved had further general characteristics been added. If so, then the resolutions of Figure 3.3 would show a lack of consistency of the judges and not represent a measurable quality of the objects.

The results of this section show that four equally defensible scoring systems applied to the data actually yield rather different results. The degree of disagreement is sufficient to make one very wary of using any of them. The standard and range scores both reflect the intensity of the teacher-student interaction (one slightly more than the other). The win-loss weights scoring method does not require

Figure 3.5

Scatter Diagram of Factor Scores and Win-Loss Weights



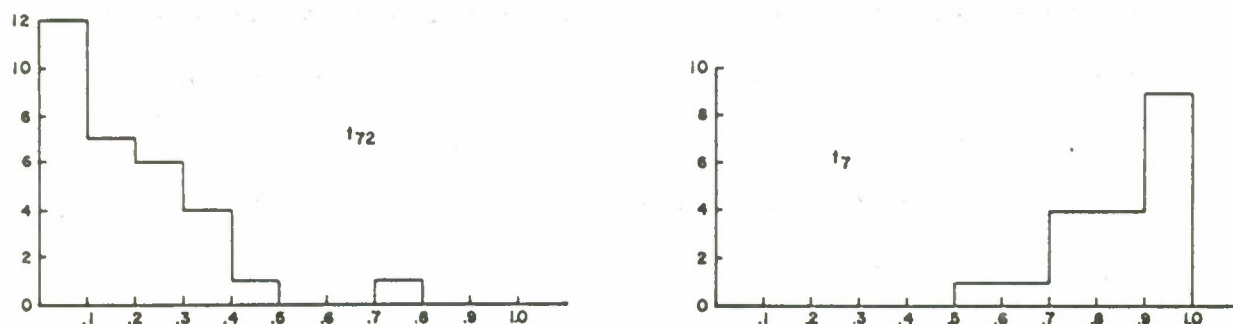
any common basis of judgment but does require general agreement about a subset of teachers seen by all students. Such is unlikely to occur. All three of these would be more reliable with a more balanced experimental design. The factor scoring yields a performance scoring system for a list of specifics. The current list is not regarded as adequate.

#### IV. APPROPRIATENESS OF A SINGLE SCALE

Let us consider the histograms for the range scaled scores obtained by different professors. Some extreme cases are shown in Figure 4.1. Teacher  $t_{72}$  exhibits a set of scores for a professor who is ranked primarily as being the lowest in comparison with the other professors that the students who scored this professor have seen, while  $t_7$  is a professor whose score distribution puts him at the high end of the range normalized scale. These two are clearly

Figure 4.1

Histograms of Scores for Two Extreme Teachers  
(Range Scaled)





different and sets of extreme teachers are not hard to identify. The follow-on question is concerned with how well one can discriminate between nonextreme teachers.

Although the underlying distributions are not all normal, an analysis of variance was performed to determine to what extent the mean scores differed significantly from each other. Part (a) of Table 4.1 shows the result of this one-way analysis of variance, and indicates that the  $F$  value of 27 which was obtained is highly significant and indicates (a not too surprising result) that there really are significant differences between the scores obtained by the different professors.

Next let us identify a set of extreme teachers and remove them. In looking at the histograms of scores (range normalized) which are assigned by any one student to the set of professors that he is rating, it appears that most students used a distribution which is a little flatter than normal, and in some cases quite uniform in assigning their rating for professors. See Figure 2.1. One thing that is noticeable in scanning such histograms is that often the student crowds most of the professors he is scoring into the center of the range, and allows 4 or 5 professors to occupy the positions near the extremes. This suggests that, perhaps, any given student has a distinct preference for 4 or 5 professors, a distinct negative attitude towards 4 or 5 professors, and is perhaps indifferent to the rest. In order to test this hypothesis, each student's scores were taken and the top 5 professors were set aside as well as the bottom 5.



	<u>Source</u>	<u>Sum of Squares</u>	<u>d.f.</u>	<u>Mean Square</u>	<u>F</u>
(a)	Between Teachers	71.70	74	0.97	27.7
	Within Teachers	<u>57.89</u>	<u>1655</u>	0.04	
	Total	129.59	1729		
(b)	Between Teachers	14.313	55	0.26	6.5
	Within Teachers	<u>51.16</u>	<u>1269</u>	.04	
	Total	65.47	1324		

(a) All 75 instructors.

(b) Residual set of 56 instructors.

Table 4.1

#### Testing the Significance of the Teacher Scores

If the majority of a man's scores were set aside, then he is eliminated (as being extreme). In this way the number was reduced to 56 from the original 75. A one-way analysis of variance was again done on this set of professors, and the results of this are shown in Part (b) of Table 4.1. The F value has now been reduced to approximately six (from 27), and although it is still statistically significant, it is dramatically reduced. Most of the differences have been largely washed out. For practical purposes, one could say that most of the remaining set of professors do not differ significantly from each other.

It is interesting to focus upon the set of teachers that ends up in the extreme group (i.e. the group about which the students are not indifferent). If there were 50 students, each of whom scored a set of 50 teachers, and if there were perfect concordance, we would

expect 5 at the top and 5 at the bottom to be eliminated by the above discussed procedure. And thus, 10 out of the 50 teachers would be contained in the extremal group. On the other hand, if the students scored all the teachers entirely at random, there would be a total of 50 times 10 or 500 extreme scores to be distributed among 50 teachers so that on the average each teacher would end up having 10 extreme scores. Since the total number of scores received by a teacher is 50, no teacher is likely to end up with a majority of extreme scores and, hence, no teacher would end up on the extreme group. On the other hand, if there were two groups of 25 teachers, which were separately rated by 25 students each, then under perfect concordance, twice the number of teachers as previously would end up in the extreme groups. In the present instance, the criterion used was that the majority of the scores received by a teacher had to be in the extreme group in order to qualify for membership in this group (and, in addition, there had to be a minimum of 5 ratings). With these criteria, a total of 19 teachers was found to be contained in the extreme group, ten on the high side, and nine on the low side. One teacher who had been rated by 12 students had exactly six extreme values and six in the indifference region, so that his scores were eliminated from consideration in both the indifference group and the extreme group. For most of the remaining 19 professors, the total number of ratings they received put them in the extreme group by well in excess of 50% indicating a considerable degree of concordance among the students with respect to this group. That the students are fairly unanimous

in their decisions is born out by the fact that of this group of 19, eleven taught the operations research students and eight taught the communications management group. Although there was some overlap, it is clear that within each curriculum there was a small group of teachers that was liked or disliked by a majority of the students within that curriculum, and the students are essentially indifferent to the remainder.

The hypothesis of indifference is also reinforced by comparing the scores of two professors within this group. Consider  $t_{30}$  and  $t_{44}$  from Table 3.2. They had 12 students in common so that a direct paired comparison could be obtained from these 12 students. On the range scale, the scores of these professors differed by more than two standard deviations (using the square root of .26 from Table 4.1 (b)). On the other hand, a direct comparison showed that 6 students preferred one and 6 the other when they were directly intercompared.

While further work to substantiate this point needs to be undertaken, it appears reasonably clear that the scale position of the professors within an indifference group should not be taken as hard and fast numbers, but subject to such random variations as to make a scaling within this group of doubtful value. At this point, a tentative conclusion is that the student will tend to put the professors in one of three groups. He is either impressed by him, indifferent to him, or is repelled by him. Just how many are in each group, for any set of students, remains to be more carefully

investigated. From intercomparing different scales it appears that the indifference group may be decomposable but further experimentation will be needed to examine this.

#### V. FACTOR ANALYSIS OF THE NINE CHARACTERISTICS

Consider the data space of the scores (ranging 1 to 5) on each of the nine characteristic statements given by each student to each instructor that he rates. This data may be viewed as a set of vectors in nine space, and two questions naturally arise: (i) are nine dimensions really needed to describe this data, and (ii) is a rotation called for in the data space in order to reduce the strength of the inter-correlations of the characteristics. The techniques of factor analysis provide a body of linear methods for approaching these problems. Although there is great choice in the techniques available, we will use the popular standard approach for the first approximation. To do this, let us proceed to discuss the correlation matrix.

Let  $X_{ija}$  be the scores (1 to 5) given by the  $i^{\text{th}}$  student to the  $j^{\text{th}}$  instructor on characteristic  $a$ ;  $a$  runs from 1 to 9,  $i$  ranges over the set of students, and  $j$  belongs to a set  $T$  of teachers under consideration. Further let

$n_j$  = number of students that rate the  $j^{\text{th}}$  instructor

$$X_{.ja} = \sum_i X_{ija} / n_j$$

$$X_{..a} = \sum_i \sum_j X_{ija} / \sum_i n_j$$



Because of the following identity it is not clear how one should develop this system of covariances.

$$\sum_i \sum_j (X_{ija} - X_{..a})(X_{ijb} - X_{..b}) =$$

$$\sum_i \sum_j (X_{ija} - X_{.ja})(X_{ijb} - X_{.jb}) + \sum_j n_j (X_{.ja} - X_{..a})(X_{.jb} - X_{..b})$$

The left-hand side of this expression, when divided by  $\sum_j n_j$  is the covariance of characteristics a and b in the scatter space introduced above. Each student-teacher pair is treated as a separate data point. If one desires to screen out the student-teacher interactions, then one should work only with the centroid  $\{X_{.ja}\}_{a=1}^9$  of data for each teacher. Thus the second term on the right hand side would be divided by  $\sum_j n_j$  to serve as the covariance estimate. It is difficult to choose between the two. For current purposes the latter choice was made arbitrarily. Of course the correlation matrix is formed from the covariances in the usual way.

Another choice must be made, namely the set T of teachers to use in the factor analysis. If the goal is to understand the data space and hope to describe it more simply, then T should be the entire set of seventy five teachers. On the other hand, if the goal is to discover the importance (as perceived by the students) to good teaching of the various components then one should be selective and limit T to a list of good and poor teachers. The list of extreme teachers developed in Section IV is used. Since both goals



are of interest, both the set of seventy-five and the set of nineteen were subjected to a factor analysis.

Two correlation matrices (symmetric) appear in Table 5.1, one for each choice of T. When T is the set of nineteen best and worst instructors it is seen that all nine characteristics are highly

Ch.No.	1	2	3	4	5	6	7	8	9
1 a	1.000	.973	-.963	-.847	.810	.839	-.950	.936	.930
1 b	1.000	.961	-.789	-.805	.616	.588	-.898	.811	.880
2 a		1.000	-.927	-.893	.845	.852	-.965	.974	.971
2 b		1.000	-.760	-.842	.650	.616	-.909	.864	.917
3 a			1.000	.759	.698	-.735	.877	-.870	-.855
3 b			1.000	.594	-.459	-.446	.751	-.686	-.714
4 a				1.000	-.942	-.936	.909	-.918	-.948
4 b				1.000	-.734	-.702	.835	-.825	-.860
5 a					1.000	.978	-.876	.902	.925
5 b					1.000	.937	-.783	.821	.814
6 a						1.000	-.877	.887	.915
6 b						1.000	-.749	.773	.773
7 a							1.000	-.951	-.960
7 b							1.000	-.884	-.934
8 a								1.000	.992
8 b								1.000	.952
9 a									1.000
9 b									1.000

a) For the 19 extreme instructors

b) For all 75 instructors

Table 5.1

Correlations of the Nine Characteristics

correlated. The negative signs merely reflect the reverse wording of characteristics 3, 4 and 7. Thus we expect the scatter plot of data points in 9-space to be more or less one dimensional. Evidence that this took place is given by the fact that 91% of the total variability (sum of the eigenvalues) is represented by the largest eigenvalue. The direction of this subspace is given by the associated eigenvector. Since these values are nearly equal, it follows that the (normalized) scatter plot is close to the main diagonal of 9-space and there is little variability in the importance of the characteristics. This principal eigenvector served to supply the weights (applied to  $X_{.ja}/\text{std. dev.}(a)$ ) in forming the factor score column of Section III.

Turning to the correlation matrix for T equal to the set of all seventy-five instructors, it is seen that the correlations are high, but not as high as before. The largest eigenvalue represents 81% of the total variability and the associated eigenvector is a little farther from the main diagonal than the previous one. The scatter plot would be a little thicker than before and with a slightly different direction.

The three largest eigenvalues of this matrix are 7.25, 0.91, 0.33 with accumulated percentages 81%, 91%, 94%, respectively. The latter two of these values are small and not nearly as stable as the former. Hence, the basic structure of the data space is still one dimensional. For purposes of illustration, however, it is tempting to go further and see what other dimensions of teaching they suggest. So let us consider the data as projected into a long flat

cigar shaped set of points in the three dimensional subspace spanned by the first three eigenvectors. Factor analysts address the question of how this set should be described. The general guideline is to perform successive rotations such that each basis vector has the direction of maximum variability in the subspace remaining.

	Factors		
	F <sub>1</sub>	F <sub>2</sub>	F <sub>3</sub>
1	.75	.27	.55
2	.79	.32	.49
3	-.35	-.19	-.90
4	-.80	-.47	-.19
5	.35	.90	.18
6	.28	.92	.19
7	-.64	-.52	-.49
8	.60	.61	.41
9	.67	.57	.42

Table 5.2

Factor Loadings for a Three Dimensional Solution

The result of this leads to the 9 by 3 matrix of Table 5.2. The entries tell us how heavily each of three components (factors) of teaching load on each of the nine characteristics. The first factor loads most heavily on characteristics 1, 2, 4, 7 and 9 (daily preparation, course organization, mannerisms, ability to communicate, amount learned) so this component might be labeled "lecturing skill." The second factor might be labeled "rapport" since it loads heavily

on 5 and 6 (understanding of students background, adaptability and fairness), while the third factor might be called "competence" since it loads largely on 3 (knowledge of subject).

The above is intended only to illustrate how one may construct important dimensions of teaching. Little importance should be attached to the specific areas found here because the projected data set is so long and thin. The direction of the minor dimensions may be quite variable. Work similar to this appears in [1], [4], and [6].

## VI. SUMMARIES OF AUXILIARY DATA

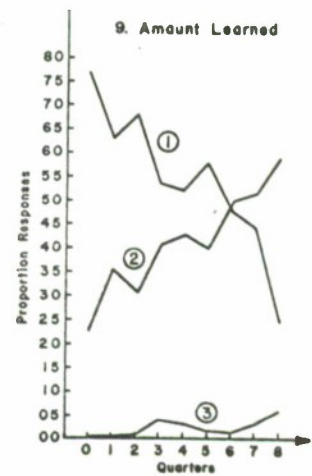
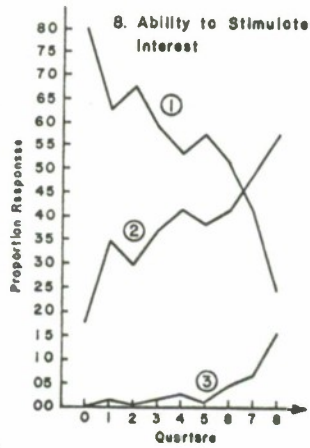
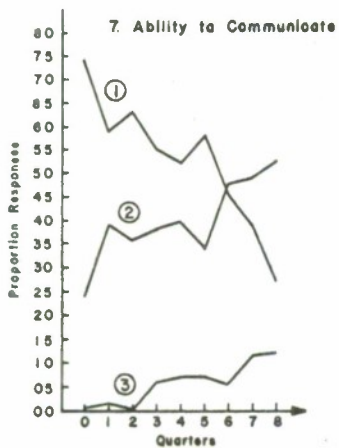
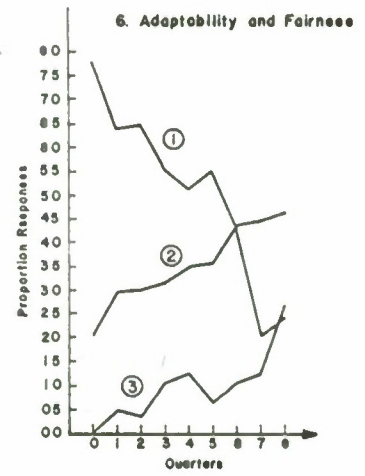
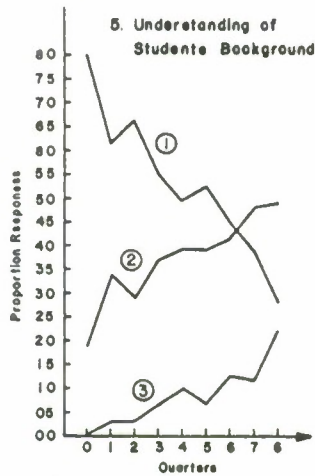
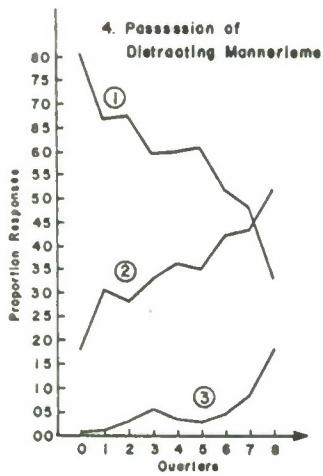
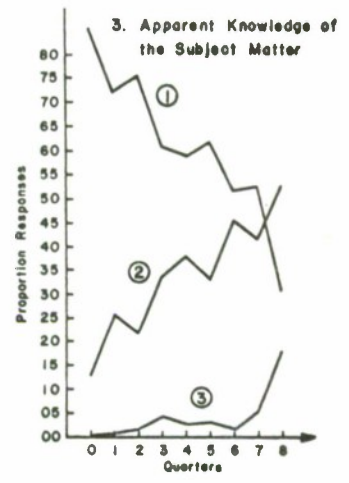
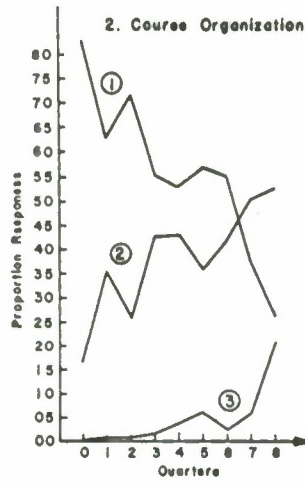
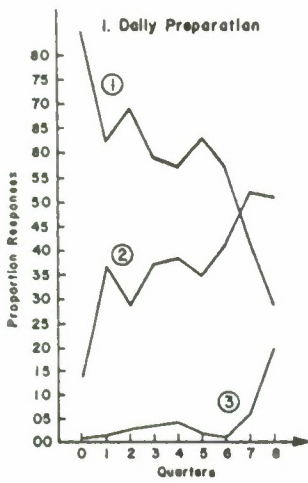
### Quality Index of Response

For each instructor and each of the nine characteristics the students indexed the quality of their responses with a value of 1, 2 or 3 with interpretations as described in the instructions (Appendix B). Since the curricular quarter that they had each teacher was also listed it was decided to find out if the sharpness of the rating diminished with time into the past.

The results appear in Figure 6.1. Time is measured in quarters with zero being the current quarter. For students who had a teacher more than once, the most recent quarter was used. Clearly the greatest percentage (75% or more) of sharp responses occurs for the current quarter and this falls off rapidly as one moves into the past. The total response of class one and two ratings can be obtained by complementing the class three ratings. Except for characteristics 5, 6 and 7 (understanding of background, fairness, ability to communicate)



Percentage Response of Each Quality Index as a Function of Time from Graduation by Characteristic Statement





the vague ratings never go above the 5% level for the last four quarters, and only achieve the 10% level for those three.

Thus it appears that questionnairing must be done quarterly if sharp responses are required, but yearly can suffice if the less stringent avoidance of vague responses is an acceptable goal.

#### Frequency of Opinions on Selected Items

The students were asked to indicate the teachers under whose direction they felt they learned the most and the teachers who made them struggle the most. This data and the counts of the number of times each teacher was considered to be below the minimum acceptable level appear in Table 5.3. It is not surprising that a teacher's standard score rank is positively correlated with the students' feelings of having learned a lot. Similarly, the negative correlation with the rate of being considered unacceptable is obvious. It appears that the question of struggling is not associated with the students' view of learning or good teaching. Indeed several of the respondents said they "didn't know how to answer that question."

#### Preference of Subject Areas

The courses were grouped into subareas and the students were asked to state their preferences among these areas as described in the introduction and Appendix B. This data was converted to win-loss matrices in the same way as before (Section III) and the Ford scoring method was applied. The resulting weights for the two groups of students appear in Table 5.4. Such weights may be useful for separating the student's attitude toward the subject area from his view of the teacher. It was not possible to pursue this effort further.

TABLE 5.3

## Frequency of Opinion on Selected Items

<u>i</u>	<u>n</u>	<u>A</u>	<u>A/n</u>	<u>B</u>	<u>B/n</u>	<u>C</u>	<u>C/n</u>
1	19	6	.32	0	0.	0	0
2	10	4	.40	1	.10	0	0
3	36	12	.33	2	.06	0	0
4	37	17	.46	0	0	0	0
5	7	3	.43	0	0	0	0
6	11	3	.27	0	0	0	0
7	19	5	.26	5	.26	0	0
8	14	6	.43	2	.14	0	0
9	12	3	.25	0	0	1	0.08
10	14	8	.57	3	.21	0	0
11	17	2	.12	0	0	0	0
12	15	2	.13	3	.20	0	0
13	6	1	.17	0	0	0	0
14	7	2	.14	0	0	0	0
15	6	2	.33	1	.17	0	0
16	8	4	.50	1	.13	0	0
17	47	12	.26	15	.32	0	0
18	14	1	.07	0	0	0	0
19	25	0	0	0	0	0	0
20	14	4	.29	0	0	2	0.14
21	27	2	.07	0	0	0	0
22	32	3	.09	0	0	1	0.03
23	20	1	.05	0	0	0	0
24	28	2	.07	4	.14	0	0
25	21	2	.10	2	.10	0	0
26	12	2	.17	0	0	0	0
27	6	0	0	0	0	0	0
28	8	0	0	1	.13	0	0
29	22	2	.09	5	.23	0	0
30	23	4	.17	1	.04	0	0
31	14	0	0	0	0	1	0.07
32	46	8	.17	2	.04	1	0.02
33	40	5	.13	4	.10	2	0.05
34	38	7	.18	0	0	0	0
35	53	4	.08	13	.25	3	0.06
36	39	0	0	0	0	0	0
37	5	0	0	0	0	0	0
38	39	3	.08	9	.23	1	0.03
39	54	2	.04	3	.06	2	0.04
40	23	2	.09	2	.09	0	0
41	51	0	0	2	.04	2	0.04
42	8	0	0	1	.13	0	0
43	12	0	0	1	.08	0	0
44	21	0	0	7	.33	1	0.05
45	52	1	.02	2	.04	2	0.04
46	16	2	.13	2	.13	2	0.13
47	28	0	0	2	.07	3	0.11
48	53	3	.06	1	.02	5	0.09
49	24	0	0	0	0	4	0.17
50	29	0	0	3	.10	3	0.10
51	24	1	.04	0	0	9	0.38
52	11	0	0	0	0	1	0.09
53	13	0	0	0	0	6	0.46
54	21	1	.05	3	.14	2	0.10
55	6	0	0	0	0	0	0
56	28	3	.11	0	0	6	0.21
57	16	0	0	1	.06	10	0.63
58	46	0	0	2	.04	15	0.33
59	12	1	.08	3	.25	8	0.67
60	51	7	.14	8	.16	18	0.35
61	15	0	0	2	.14	9	0.60
62	34	0	0	7	.21	10	0.29
63	13	0	0	1	.08	6	0.46
64	29	0	0	3	.10	10	0.34
65	15	0	0	0	0	2	0.13
66	17	0	0	4	.24	5	0.29
67	19	1	.05	0	0	13	0.64
68	12	0	0	2	.17	10	0.83
69	36	1	.03	2	.06	17	0.47
70	34	0	0	12	.35	18	0.53
71	11	1	.09	4	.36	10	0.91
72	33	2	.06	0	0	21	0.64
73	19	0	0	4	.21	14	0.74
74	19	0	0	2	.11	16	0.84
75	14	0	0	9	.64	14	1

Legend

- A Number of times rated as a teacher under whose direction the student learned the most.
- B Number of times rated as a teacher who made the student struggle the most.
- C Number of times rated below the minimum acceptable level.

Operations Analysis		Communications Management	
Prob & Stat	137	Management	271
Math Program	72	Comp Sc	27
Comp Prog	54	Elec Eng	21
Systems Anal	54	Communications	35
Human Factors	10		

Table 5.4

## Win-Loss Weights for Preference of Subject Areas

Special Categorization Data

In an attempt to determine whether there is any correlation of teacher rankings with the courses which were taught by the teacher, the students were asked at the end of the questionnaire to identify the three hardest courses, the three easiest courses, the course that they thought the most relevant, the least relevant and the one that they would most like to repeat. The results of this portion of the questionnaire were interesting although it failed to reveal any correlation between teacher rankings and course material except that the course they would most like to repeat was associated with low scoring instructors.

The students in the OR curriculum were almost unanimous in their choice of the Stochastic Models sequence OA 3704 and OA 4705 as the most difficult courses with which they had to deal. OA 3704 was named 21 times and OA 4705 was named 19 times as the most difficult course whereas the next most difficult statistics course, PS 3303, was named only 5 times. The only other course given 5 votes was OA 3610, Utility Theory. Although the Stochastic Models

courses were considered the hardest, the instructors who taught these courses did not end up on the low end of the scale. There was less unanimity as to which was the easiest course for the OA curriculum. Approximately the same number of votes were given to MA 1101, OA 3630 and OA 4633, each of them receiving 8 votes. OA 3671 received 6 votes and many courses received one or two. When asked to name the most relevant course, it is surprising that the votes were scattered over many courses with no one course receiving more than five votes (and this was PS 3303, Statistics). Also surprisingly, the courses that were voted the most difficult, OA 3704 and OA 4705, received hardly any votes as being most relevant. Similarly, there was no concurrence as to the least relevant course. OA 3765, Human Factors, received the most votes, 6, and a close runner up was PH 2122 with five votes. On the subject of which courses they would most like to repeat, a great number of students voted for "none of the above." Only two courses received five votes. These were OA 3610 and OA 4705. Among the reasons mentioned for wanting to repeat a course was to learn more provided a different instructor were teaching the course.

For the Communications Management curriculum, the results were only slightly different. In this case, the students were not so unanimous in their choice of the hardest course, three courses receiving four votes each. These were MA 2121, MN 3106 and MN 3030. Eight of the fifteen students picked OA 3211 as the hardest (and this is a survey course). There was even more agreement as to which were the easiest three courses, CO 2111 receiving 11 votes and CO 2112



receiving nine. The third runner up was MN 4182 with six votes. Of a total of 16 votes cast for the most relevant course, five went to MN 3171 and three each to MN 4183 and MN 4181. Of the 17 votes cast for the least relevant, only CO 2111 received as many as three votes. For the course they would most like to repeat, five students picked PS 3000 and four picked MN 3171. Again, the reason most often cited for wanting to repeat a course was to try to learn more in case a different instructor were teaching the course.

## VII. RESULTS AND RECOMMENDATIONS

### Summary

We have analyzed the data of an experiment designed to measure teacher performance as seen by a group of graduating students. The results of this analysis, while tentative, point to the following conclusions.

Standard and ranged scaled scores are highly correlated. The latter does a little better job of preserving the students intensity of feeling about instructors. Win-loss data eliminate entirely any information about the intensity of feeling about instructors and preserves only the statement of preference. It appears that the judges are in too great disagreement to produce a persistent linear scale. The two groups have different criteria. Factor scoring correlates with the others but not as highly. It should produce a score based only on some definitive items of performance and the resulting weights should measure the importance of the items.



The number of students that see a teacher may be an important variable. The scores of the top 20 teachers in Table 3.2 were determined by 338 votes while the scores of the bottom 20 were determined by 473 votes.

Different techniques of analyzing the data are consistent in identifying a teacher whom the student considers to be either extremely good or extremely poor. One method for producing such an extremal set was to collect all those teachers who were in the top or bottom five of any ladder on which they appeared. If the number of such listings was in the majority of the total times listed then that teacher remained in the extremal group. This resulted in a set of 19 teachers. The remaining 56 teachers (those who were judged not to be in the extreme group and rated by at least five students) were subjected to further scrutiny. An analysis of variance done on this group indicated that the difference in scores for most of these teachers was not very significant ( $F = 6.5$ ) compared to  $F = 27.7$  for all 75 teachers. This appears to indicate that the placing of teachers on a strict one-dimensional continuum may not be justified.

Further analysis on the teachers in this group led to the discovery that two types can be identified, which may be characterized as the bland and the controversial. The controversial teachers are characterized by a large variance in their scores, as obtained from individual students. In comparing difficult scaling methods we found that in some cases the controversial nature of the teacher

could be attributed to the different types of students being taught. For example, a professor teaching both management and operations analysis students in nominally the same course was awarded a score of .86 by the 15 students in the former group and a score of .19 by 36 students in the latter.

Another factor contributing to making a teacher controversial occurs when students have to reconcile the fact that a teacher who may know his subject extremely well may have distracting mannerisms or other displeasing personal habits. This was brought out by an analysis of the nine characteristics with respect to which students were asked to rate teachers.

#### Comments on the Students

The students have filled out many questionnaires and many seemed to have a rather dulled attitude toward them. Even so, they were quite cooperative and the following items have been noticed.

- (i) Most students regard the ladder as valuable.
- (ii) If given a free form linear scale (ladder) they will spread out their data with a rather flat distribution. Roughly 20% of the distributions will be skewed and most of these trailing to the left. This has the effect of holding most teachers in high regard while giving greater discriminating power to those teachers with which they were dissatisfied.
- (iii) Detailed information about teachers that the students last "saw" more than one year ago should not be sought.
- (iv) The students' responses to questions about teaching performance will reflect "second guesses" on how they will be used.

### Future Research and Development

There is need for a thorough study which would accomplish the following:

1. Identify the ways that students look at instructors.
2. Number and characterize the significant dimensions of the teaching-learning process as viewed by our students.
3. Identify the patterns or stylistic differences among teachers.
4. Identify and quantify the importance of exogenous variables (e.g. subject matter, class size, core course, service course, etc.)

The techniques of factor analysis, multidimensional scaling, discrimination, cluster and classification analysis, and regression can serve in such an effort. Some of these techniques are rather new. Research into them is required and it may be necessary to generate new methodology.

## Appendix A

Example Showing Nonsensical Results when Popular  
Scaling Techniques are Used

The following realistic but fictitious example is taken from Reference [8]. Several judges rank several objects and the idealistic assumption is made that the objects possess a true ordering. Also there is little disagreement among judges. Variability occurs only because each judge appraises only a subset of the whole.

Three common data summaries are applied, ten objects comprise the set being considered (A,B,...,J) and the table shows the categories assigned by each of six judges. They were permitted to use as many qualitative categories as they chose, and to place as many objects in each category as they thought appropriate.

One may note that the judges showed high, but not perfect, agreement in their rankings. Judges I, V, and VI considered objects relatively low on the dimension being considered. Judges II and IV ranked only objects at the upper end. Judge III categorized some of each. The number of categories used by the judges ranged from three (Judge I) to six (Judge IV). Although Judge III used only four categories, he evaluated more objects than any other judge.

Given the simplicity of Table 1, one can identify the "true" order of the objects. This is shown at the left. Note that the order between objects B and C is undefined (Judge II placed C above B, but Judge IV placed B above C, and no other judge compared them). It is clear, however, that these two objects fall below A and above



D. In addition, there was minor disagreement among the judges concerning the relative positions of F versus G (two judges out of three placed F ahead) and I versus J (three out of four placed I ahead).

#### Three Unsatisfactory Systems

In some kinds of data the average rank assigned by several judges provides a satisfactory way of combining judgments. The fact that in the present data judges evaluated different sets of objects, and used different numbers of categories, however, makes this system inappropriate. Note in Table A 1 that this first method gives results very different from the true rank order.

Allowance could be made for the judges' using different numbers of categories by converting to a "percentile equivalent" for the rank (if one were willing to assume equal distances between categories). A combined score could then be based upon an average of the "percentile equivalents." This second method was applied to the data of Table A 1. Note that it also fails to represent the true rank order. The reason is that this method, like the first, neglects the fact that judges ranked different sets of objects.

A somewhat more sophisticated system is to set up a "win-loss matrix," as shown in Table A 2. For each possible pair of objects, this matrix indicates the number of times the first was ranked above the second, and vice versa. Such a matrix can be used to determine each object's "win percentage"--the number of times it scored above others relative to the total number of times it won or lost. The



True Order	Object	Evaluations by Judges						Mean Rank	Order	Mean percentile	Order	Win %	Order
		I	II	III	IV	V	VI						
1	A			1	1		1.0	1	89	1	100	1	
2.5	B		2		2		2.0	3.5	72	3	78	3	
2.5	C		1	1	3		1.7	2	78	2	86	2	
4	D		3	2	4		3.0	6	50	6	50	6	
5	E		4		5		4.5	10	28	9	22	8	
6	F		5	2	6	2	3.2	7	47	7	45	7	
7	G			3		1	2.0	3.5	63	4	61	4	
8	H		1		3	3	2.3	5	61	5	60	5	
9	I		2	3	5	4	3.5	8	30	8	20	9	
10	J		3	4	4	5	4.0	9	16	10	6	10	

Note: Judges were free to rank only objects with which they were familiar, and were allowed flexibility in the number of categories used and the number of objects assigned to each. Methods I to III for combining evaluations give misleading results.

Table A 1

Artificial Evaluations of Ten Objects by Six Judges

Losing objects	Winning Objects										Times lost (=b)	Win % [=a/(a+b)]	Rank
	A	B	C	D	E	F	G	H	I	J			
A											0	100%	1
B	1										2	78%	3
C	1	1									2	86%	2
D	2	2	3								7	50%	6
E	1	2	2	2							7	22%	8
F	2	2	3	2	2	1					12	45%	7
G	1	1	1	1	2						5	61%	4
H					2	2					4	60%	5
I	1	1	1	1	3	2	3	1			12	20%	9
J	1	1	1	1	3	3	3	3	3		15	6%	10

Table A 2

Win-Loss Matrix Based on Evaluations Shown in Table A 1.

"win percentages" derived by this third method is also shown. It gives an order somewhat different from either of the other two and still fails to represent the true order.

Considering this data further, note the role played by Judge III, the only judge to view a large range of objects. Without him it would be more difficult to preserve the true ordering, indeed impossible if Object F were removed. Such seemingly minor changes have great influence, especially if there is controversy over the true ordering of objects.

Finally, the fact that the scoring was by ranks is immaterial. The same effects can be produced with finer, "quantitative" scoring such as appears in the ladders. The Ford procedure (see Section IIIc) recovers the true order for this example.

## Appendix B

## QUESTIONNAIRE

To the student: The purpose of this experiment is to gather student opinion information on the quality of instructors and instruction at NPS. It is emphasized that this is an experiment and that some of the questions are designed to help us understand the student's point of view and to gain introspective. Other questions ask you to do performance rating - a task that you may find difficult, especially with regard to courses and instructors that you have had some distance in the past. With these aspects in mind, we sincerely solicit your best considered opinion, both in filling out the form and in commenting at the end on your problems in so doing. We have tried to design this experiment so that you can finish in 90 minutes, and had to limit severely the type of information collected. If you feel there are important areas of information left out, please tell us in your written comments.

Most of your replies will be entered on an IBM Key punch form so that data cards can be made directly. Care is needed to be certain the entries are placed in the correct columns. Also, we need your assistance in matching the courses you have had, by quarter, with the instructor's name and numerical code. This task must be done at the outset. Directions for this are contained in the sheet entitled "Instructions for the course instructor information sheet."

LIST OF MATERIALS

1. Two IBM Key punch (KP) Forms
2. Course-Instructor Information Sheet (and instructions)
3. Questionnaire
4. Ladder Sheet

## INSTRUCTIONS

1. Enter your name in the space marked "special instructions" on both KP forms.
2. Each line (card) will correspond to an instructor. Use columns 1-10 to enter the names (or abbreviated name(s)) of your instructors; one per line.
3. In columns 11-14 insert the instructor's four-digit code.
4. Columns 15 thru 20 are to be used two at a time for the following pairs:

quarter you had the instructor (0-8)

1 or 0 according to whether the course was required or  
elective (resp.)

Thus, if you had Prof. A for one course (required) in quarter 6, columns 15 and 16 should show "61," and columns 17-20 should be blank. There is space for three such pairs only. If you had an instructor for more than three courses, enter the information for the three most recent.

5. You are asked to rate each instructor for each of nine characteristic statements which are lettered A thru I, and you are to indicate your degree of agreement with that statement. Again the responses are paired: the first entry being the rating itself and the second being the quality of the rating. Use the following scales:

## RATING SCALE

<u>Strongly</u> <u>Disagree</u>	<u>Disagree</u>	<u>Neutral</u>	<u>Agree</u>	<u>Strongly</u> <u>Agree</u>
1	2	3	4	5

## QUALITY OF RATING SCALE

- 1 - This is a sharp rating and can be supported with some classroom specifics.
- 2 - I have a fair amount of confidence in this rating.
- 3 - My recollection is vague.



Use the following columns of the KP form for your responses:

Characteristic statement	Columns	
	Rating	Quality
1	22	24
2	28	30
3	34	36
4	40	42
5	46	48
6	52	54
7	58	60
8	64	66
9	70	72

#### LIST OF CHARACTERISTIC STATEMENTS

1. His daily lectures were very well organized and prepared.
2. His organization of the course(s), i.e., the choice and sequencing of topics, and so on, was very good.
3. His apparent knowledge of the subject matter covered in the course(s) was very limited.
4. He had many distracting mannerisms which impaired his effectiveness as a teacher.
5. He showed an appreciation and an understanding of the students' background and profession.
6. He was adaptable and fair in his reactions to student comments and questions.
7. He was unable to communicate difficult ideas.
8. He stimulated my interest in the subject matter covered in the course.
9. I learned as much or more from him in this (these) course(s) as I would have from the typical teacher.

6. Overall rating. Go to the Ladder Sheet and fill it out. Transfer the scaled number for each instructor to columns 74-77 (in floating point with up to one decimal place) on the KP forms.
7. a) Place a 1 in column 79 opposite the teacher(s) under whose direction you learned the most. Ties are allowed.  
b) Place a 1 in column 80 opposite the teacher(s) who made you struggle the most. Ties are allowed.
8. The curriculum can be viewed as the union of the following subareas:

PS - Probability and Statistics	CS - Computer Science
CP - Computer programming and war gaming	EE - Electrical Engineering
LP - Linear Programming and mathematical optimization methods	MN - Management
SA - Systems Analysis, utility, and cost studies	CO - Communications
HF - Human Factors	

Please rank these subareas with regard to their interest to you. Use the symbol ">" to mean "is preferred to" and the "=" for ties. For example

$$CP > PS = LP > HF = SA$$

means CP is preferred to PS which is tied to LP, but all of these are preferred to HF and SA which are tied. Place this on a new line beginning with column 20.

9. Free Style Comments. On the back of your dittoed Course-Instructor Sheets we solicit your candid reactions to
- teachers and teaching
  - the curriculum
  - this questionnaire
  - any other pertinent items.

INSTRUCTIONS FOR THE  
COURSE-INSTRUCTOR INFORMATION SHEET

1. Enter your name and section at the top.
2. Enter the name of the instructor for each course in the space indicated. Also you will have to enter the courses taken by you that are not listed. If a course is listed in the wrong quarter, please correct.
3. (OA STUDENTS ONLY). For the courses listed in Qtr 5 enter (6) or (12) to the left of the course number according to whether you took an accelerated (6 week) or regular (12 week) course.
4. Under Req/Elec/Valid enter:
  - 1 - if course was required
  - 0 - if course was elected
  - v - if course was validated for you.
5. Under code place the 4 digit index code for the instructor. Most of the codes can be found on the supplementary sheet. Leave blank any codes that you cannot find on the sheet.
6. On the last page list:
  - a. The three most difficult courses  
(i) \_\_\_\_\_ (ii) \_\_\_\_\_ (iii) \_\_\_\_\_
  - b. The three easiest courses  
(i) \_\_\_\_\_ (ii) \_\_\_\_\_ (iii) \_\_\_\_\_
  - c. Most relevant course \_\_\_\_\_.
  - d. Least relevant course \_\_\_\_\_.
  - e. The course you would most like to repeat \_\_\_\_\_.

Why? \_\_\_\_\_.

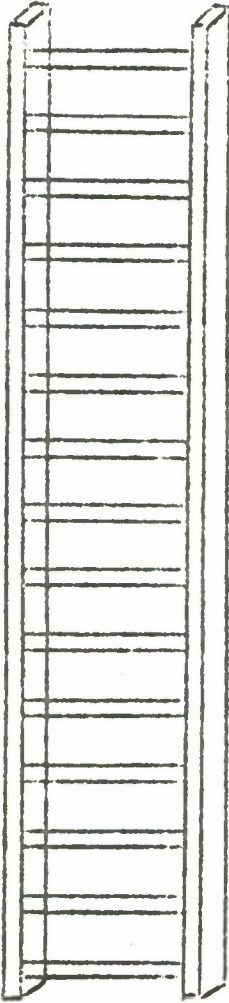
## LADDER SHEET

Instructions

1. On the last page is a vertical scale numbered "one" through "fifteen." Please look over that vertical scale at this time, and then go on to the next instruction.
2. Think of the best classroom teacher you have had. Write that teacher's code at the place where you believe it belongs on the ladder scale. (You don't have to place the teacher's code beside step "15" on the scale.)
3. Think of the poorest classroom teacher you have had. Write that teacher's code at the place where you believe it belongs on the ladder scale. (You don't have to place the teacher's code beside step number "1" on the scale.)
4. Now, think of two other classroom teachers you have had, and write each of their codes alongside the appropriate points on the ladder.
5. Please circle the 4 teachers' codes you have placed alongside the ladder.
6. Draw a horizontal line at the point indicating the minimum acceptable quality of teacher you think should be found at the Naval Postgraduate School in your curriculum.
7. Now, consider the overall teaching ability of each of the professors you have had while at the Naval Postgraduate School and write each of these professors' codes at the appropriate points alongside the ladder. Ties are allowed, but are not encouraged.

TEACHER RATING

LADDER SCALE  
OF OVERALL CLASSROOM  
TEACHING EFFECTIVENESS

<u>Absolute Best</u>	<u>Ladder Step</u>		<u>Teachers' Codes</u>
	15		_____
	14		_____
	13		_____
	12		_____
	11		_____
	10		_____
	9		_____
	8		_____
	7		_____
	6		_____
	5		_____
	4		_____
	3		_____
	2		_____
<u>Absolute Worst</u>	1		_____



## REFERENCES

- [1] Field, T. W., Simpkins, W. S., Browne, R. K. and Rich, P., "Identifying Patterns of Teacher Behavior from Student Evaluations," *Journal of Applied Psychology*, 1971, Vol. 53, No. 5, pp. 466-469.
- [2] Ford, L. R., Jr., "Solution of a Ranking Problem from Binary Comparisons," *American Mathematics Monthly*, 1957, 64 (8,II), 28-33.
- [3] Harmon, H. H., *Modern Factor Analysis*, University of Chicago Press, Chicago, 1967.
- [4] Hildebrand, M. and Wilson, R. C., "Effective University Teaching and its Evaluation," Center for Research and Development in Higher Education, University of California, Berkeley, 1970.
- [5] Horst, P., *Factor Analysis of Data Matrices*, Holt, Rinehart & Winston, New York, 1965.
- [6] Isaacson, R. L., McKeachie, W. J., Milholland, J. E., Lin, Yi.G., Hofeller, M., Baerwoldt, J. W. and Zinn, K. L., "Dimensions of Student Evaluations of Teaching," *Journal of Educational Psychology*, 1964. Vol, 55, No. 6, pp. 344-351.
- [7] Pearson, E. S. and Hartley, H. O., *Biometrika Tables for Statisticians*, V. I, University Press, Cambridge, 1966.
- [8] Pelz, D. C. and Andrews, F. M., *Scientists in Organizations: Productive Climates for Research and Development*, Wiley, New York, 1966.
- [9] Stultz, K. F. and Zweig, H. J., "Relation Between Graininess and Granularity for Black-and-White Samples with Nonuniform Granularity Spectra," *J. of Optical Society of America*, 1959, Vol. 49, pp. 695-702.
- [10] Stultz, K. F. and Zweig, H. J., "Roles of Sharpness and Graininess in Photographic Quality and Definition," *J. of Optical Society of America*, 1962, Vol. 52, pp. 45-50.
- [11] Waugh, F. V., "Factor Analysis: Some Basic Principles and an Application," *Agricultural Economics Research*, Vol. 14, No. 3 (July 1962), pp. 77-80.

## INITIAL DISTRIBUTION LIST

	No. Copies
Defense Documentation Center (DDC) Cameron Station Alexandria, Virginia 22314	12
Librarian, Code 0212 Naval Postgraduate School Monterey, California 93940	2
Dean of Research Administration Code 023 Naval Postgraduate School Monterey, California 93940	2
M. U. Clauser, Provost Naval Postgraduate School Monterey, California 93940	1
W. F. Koehler, Dean Naval Postgraduate School Monterey, California 93940	1
Professor J. R. Borsting	1
Professor C. R. Jones	1
Professor W. M. Woods	1
Professor J. K. Arima	1
Professor A. F. Andrus	1
Professor P. R. Milch	1
Professor W. P. Cunningham	1
Professor R. S. Elster	1
Professor N. K. Womer	1
W. D. Free, Instructor	1
Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
D. Courtney, Visiting Professor Department of Operations Research and Administrative Sciences Naval Postgraduate School Monterey, California 93940	1
Professor O. B. Wilson Physics Department Naval Postgraduate School Monterey, California 93940	1

Professor D. E. Kirk 1  
Department of Electrical Engineering  
Naval Postgraduate School  
Monterey, California 93940

Professor J. W. Schultz 1  
Material Science and Chemistry Department  
Naval Postgraduate School  
Monterey, California 93940

Professor P. J. Marto 1  
Department of Mechanical Engineering  
Naval Postgraduate School  
Monterey, California 93940

Professor R. R. Read 25  
Department of Operations Research  
and Administrative Sciences  
Naval Postgraduate School  
Monterey, California 93940

Professor H. J. Zweig 5  
Department of Operations Research  
and Administrative Sciences  
Naval Postgraduate School  
Monterey, California 93940

Security Classification

## DOCUMENT CONTROL DATA - R &amp; D

*(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)*

1. ORIGINATING ACTIVITY (Corporate author)		2a. REPORT SECURITY CLASSIFICATION	
Naval Postgraduate School Monterey, California 93940		Unclassified	
		2b. GROUP	
3. REPORT TITLE			
On the Quantification of Teacher Performance Using Student Opinion			
4. DESCRIPTIVE NOTES (Type of report and, inclusive dates)			
Technical Report			
5. AUTHOR(S) (First name, middle initial, last name)			
Robert R. Read Hans J. Zweig			
6. REPORT DATE	7a. TOTAL NO. OF PAGES	7b. NO. OF REFS	
30 March 1972	67	11	
8a. CONTRACT OR GRANT NO.	9a. ORIGINATOR'S REPORT NUMBER(S)		
b. PROJECT NO.			
c.	9b. OTHER REPORT NO(S) (Any other numbers that may be assigned this report)		
d.			
10. DISTRIBUTION STATEMENT			
Approved for public release; distribution unlimited.			
11. SUPPLEMENTARY NOTES		12. SPONSORING MILITARY ACTIVITY	
13. ABSTRACT			
<p>Several scaling methodologies were applied to the data of an experiment which polled the opinion of graduating students. Inter-comparisons of the results point to problems in producing a stable one dimensional scale or grouping. Opinions on such items as organization, knowledge, mannerisms, etc., were solicited in an attempt to relate these to overall performance ratings.</p>			



UNCLASSIFIED  
Security Classification

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Measurement of Teaching						