DOCUMENT RESUME

ED 118 580 TM 005 060

AUTHOR Gilmore, Gerald M.
TITLE Statistical Analyses of the Data from the First Year of Use of the Student Ratings Forms of the University of Washington Instructional Assessment System.
INSTITUTION Washington Univ., Seattle. Educational Assessment Center.
REPORT NO 76-9; EAC-P-503
PUB DATE Nov 75
NOTE 27p.; A few pages of the text and Table 6 contain light print

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS *College Students; *Course Evaluation; Data Analysis; *Data Collection; Evaluation Methods; Higher Education; Participant Satisfaction; Rating Scales; *Statistical Analysis; *Student Attitudes; Student Evaluation of Teacher Performance
IDENTIFIERS Instructional Assessment System; *University of Washington

ABSTRACT

This report presents statistical analyses of data derived from the first year's use of the Instructional Assessment System. Included are: means, standard deviations, and several reliability estimates for each item within each form; inter-item correlations; and correlations of items with non-evaluative variables. Among the major results discussed are the high item reliabilities for all but small classes, the high inter-item correlations and their implications for use of ratings results for diagnosis of instructional problems, and the causal implications of item correlations with non-evaluative variables, e.g., whether students wanted to take the course and grade expected. (Author)

EAC REPORTS

EDUCATIONAL ASSESSMENT CENTER

University of Washington, 1400 Campus Parkway, PB–30, Seattle, Washington 98195 (206) 543–1170

Instructional Evaluation

Educational, Psychological, and Survey Research

Test Administration, Construction, Scoring, and Analysis

Educational Assessment Center

University of Washington

November 1975

76-9

Statistical Analyses of the Data from the First Year of Use of the Student Ratings Forms of the University of Washington Instructional Assessment System

Gerald M. Gillmore

This report presents statistical analyses of data derived from the first year's use of the Instructional Assessment System. Included are: means, standard deviations, and several reliability estimates for each item within each form; inter-item correlations; and correlations of items with non-evaluative variables. Among the major results discussed are the high item reliabilities for all but small classes, the high inter-item correlations and their implications for use of ratings results for diagnosis of instructional problems, and the causal implications of item correlations with non-evaluative variables, e.g., whether students wanted to take the course and grade expected.

Educational Assessment Center Project: 503

Statistical Analyses of the Data from the First Year of Use of the Student Ratings Forms of the University of Washington Instructional Assessment System

G. M. Gillmore

The systematic collection and dissemination of evaluative information from students concerning the courses in which they are enrolled is a common occurrence in American Higher Education. The University of Washington (UW) as an institution, is a pioneer in this endeavor, with efforts dating back to the 1920's. During this period of time, data have been collected from students via a variety of means. For example, in 1951 a procedure was implemented in which students at registration ranked their instructors of the previous term on the basis of teaching merit.[1] More common over this time, and still prevalent on college campuses, is the single form containing a series of evaluative statements about a course and instructor, which is administered at or near the end of a course. Students either indicate the extent of their agreement to each statement, in the classical Likert Strongly Agree to Strongly Disagree format, or rate the "goodness" of the course in terms of each item.

The most recent major change in the Student Ratings program at the UW occurred in the fall of 1974, when the Instructional Assessment System (IAS) was implemented.[2] The purpose of this report is to present analyses of the data collected the first year of use. To achieve this purpose in a coherent manner, a small amount of description is necessary. The key concepts guiding the development of the system can be briefly stated as follows:

---

[1] According to T. F. Hodgson (1974), "The faculty quickly labeled this procedure with the ungraceful name of Dragnet, the program fell from favor, and after a few years was discarded by the administration" (p. 5).

[2] Shelley Tucker, Helen Smith, and John McMillin, along with the author, were responsible for the forms and the items within forms of the system. Jerry Edwards designed the optically scannable input documents, and he and Ronald Stofer designed and wrote the computer analysis system. The entire developmental project was wholly funded by the University of Washington Educational Assessment Center.

> First, there is an explicit recognition the student ratings can and do serve multiple functions, and the same evaluative questions are not necessarily appropriate for each. Secondly, there is an explicit recognition that adequate diagnostic information cannot be efficiently provided instructors with use of a common set of evaluative questions for all classes (Gillmore, 1974, p. 1).

The impetus for the IAS came in part from the delineation of types of information yielded by student ratings found in Smock and Crooks (1974). Also influential in a negative regard was my experience with use of the same evaluative items for the diversity of classes found within a large university.

The system contains five forms, each tailored for a broad course type. Form A was designed for small lecture-discussion classes; Form B for large lectures; Form C for seminar-discussion classes; Form D for problem-solving classes, and Form E for skill-acquisition classes. Each form has items within three sections, each with distinct instructions to students. Section 1 contains 4 global evaluative items whose major purpose is normative, i.e., to allow comparisons with various populations. These items are common to every form. Section 2 contains 11 items that are diagnostic in nature. They are designed to provide feedback to instructors, useful for improving the course. These items are unique to each form, although overlap is present. Section 3 contains 7 items which are published for student use, with instructor permission. These items are common to all forms. Thus, it is only Section 2, the diagnositc items, which change from form to form.[3] The items for all forms are listed in Tables 2 through 7 (pages 6 through 11).

All items use a six position response scale. The response position labels and their numerical values are as follows: Excellent (5), Very Good (4), Good (3), Fair (2), Poor (1), and Very Poor (0).

Additional information asked of students when they complete the form is as follows:

---

[3]To be more precise, the forms contain two additional sections: space for 8 optional, instructor-chosen closed items and two open-ended questions. Neither of these sections is relevant to the present discussion.

When registering, was this a course you wanted to take? Yes, Neutral, No.

Is this course: In your major, In your Minor or program requirement, A distribution requirement, An elective, Other.

Your class: Freshman, Sophomore, Junior, Senior, Graduate, Other.

Grade you expect to receive: A, B, C, D, E, Pass.

The results to be presented in this report include the number of classes in which each form was used at the UW and the average class sizes, the mean, standard deviation, and reliability of each item, the inter-item correlations of the eleven common items (Sections 1 and 3) and correlations of the common items with selected non-evaluative variables such as class size and level.

Data Source

All data presented in this report are derived from courses at UW in which the IAS was used. In most cases this is an exhaustive sample of those using the system for the academic year 1974-75, excluding summer quarter. Where the data presented are not an exhaustive sample, it will be so indicated. It is best to consider the data as coming from a non-random volunteer set of courses since, strictly speaking, student evaluations of courses using IAS are not mandatory. However, faculty are required by Faculty Senate regulations to provide some evidence of teaching effectiveness as perceived by students in any request for promotion or merit pay increase. Most faculty fulfill this obligation by using a form of the IAS for some or all of their courses. Unquestionably, some academic departments place more pressure on faculty to use the IAS than others. About all that can be said for certain is that the sample is neither exhaustive of nor a random sample from all classes taught.

Results

The number of classes in which each form was used is presented in Table 1. Also in Table 1 is the average class size for each. Class size is defined as the number of forms completed within a given class, in this case and throughout this report. These values are almost sure to be somewhat less on an average than the total enrollment due to absentees at the

time of administration. However, they may be more representative of actual attendance, which in turn may be more important than enrollment in terms of relationships to be presented.

Table 1

Number and Average Size of Classes in which Each Form Was Used

Form	Number of classes	Average class size
A	1705	19.21
B	826	46.35
C	667	14.48
D	558	18.22
E	607	14.10
Total	4373	22.76

Form A, which was designed to be the most general, was clearly the most popular choice, being selected for 39% of the classes. The remaining forms were roughly equivalent in usage. The form designed for use in large lecture classes (B) was used by larger classes on the average, as expected. Both forms C and E tended to be used in fairly small classes, which again was as expected.

In Tables 2 through 7, means, standard deviations, and 3 reliability estimates are presented for items within each of the forms and for the common items across all forms. We will discuss the means and standard deviations first, and then turn to the reliability estimates.

Means and standard deviations. Means and standard deviations were calculated for each item using the following numerical codes: Excellent = 5, Very Good = 4, Good = 3, Fair = 2, Poor = 1, and Very Poor = 0. Thus the most favorable possible mean value is 5.0, and the least is 0.0. Note that the unit of analysis is classes and thus the basic datum entered into these particular calculations is class means for given items. Hence, the means presented are in reality means of class means and the standard deviations are actually standard deviations of class means. They are

presented here mostly for their usefulness as a reference for specific questions which might arise in the reader's mind. However, some general statements can be made.

If one compares the four general items across forms, users of Form E received the highest average rating, Form C was next most favorably rated, followed by A, B, and D. One way analyses of variance show these differences to be highly statistically significant. However, this significance must be interpreted within the context of large numbers of classes entering into the analysis. In fact, the specific form used only accounts for about five percent of the total variance. Thus, although there are average differences, there is also a great deal of overlap.

In designing the system, we felt that one of the outcomes of distinguishing items for normative purposes from those for diagnostic purposes would be that the latter would elicit more critical information from students. If this were the case, one would expect means for the diagnostic items to tend to be lower than the general items. The content of the items is, of course, different, and thus, strictly speaking direct comparisons cannot be made. However, the item means within Section 2 are close in value to those in Section 1 for all forms but E, where they tend to be somewhat smaller.

In Section 3, which like Section 1 has common items across all forms, Form E users were rated highest on the average on every item. This is an interesting result considering the range of aspects of a course covered by these items.

Over all forms, item 20 received the lowest average rating (Evaluative and grading techniques) while item 3 received the highest average (The instructor's contribution to the course). The highest average rating given any item was 4.20 for item 11 on Form E (Student confidence in instructor's knowledge). The lowest average rating (3.15) was given to item 14 on Form B (Interest level of class sessions).

The standard deviations were quite consistent for items both across forms and within forms. The range was from .45 for item 2 on Form E to .70 for item 15 on Form C. Most, however, fall within a tenth of a scale point of each other.

Table 2

Mean, Standard Deviation, and Inter-Rater Reliability of All Items on Form A

Item	Mean	SD	Reliability* N=1	N=10	N=40
Section 1:					
1. The course as a whole was:	3.59	.58	.26	.78	.93
2. The course content was:	3.58	.53	.22	.74	.92
3. The instructor's contribution to the course was:	3.81	.65	.28	.80	.94
4. The instructor's effectiveness in teaching the subject matter was:	3.61	.69	.29	.80	.94
Section 2:					
5. Course organization was:	3.48	.59	.23	.75	.92
6. Clarity of instructor's voice was:	3.98	.57	.24	.76	.93
7. Explanations by instructor were:	3.66	.60	.24	.76	.93
8. Instructor's ability to present alternative explanations when needed was:	3.62	.59	.22	.74	.92
9. Instructor's use of examples and illustrations was:	3.74	.57	.22	.74	.92
10. Quality of questions or problems raised by instructor was:	3.62	.56	.21	.73	.91
11. Student confidence in instructor's knowledge was:	4.02	.59	.25	.77	.93
12. Instructor's enthusiasm was:	3.95	.64	.30	.81	.94
13. Encouragement given students to express themselves was:	3.73	.59	.21	.73	.91
14. Answers to student questions were:	3.62	.56	.22	.74	.92
15. Availability of extra help when needed was:	3.70	.57	.20	.71	.91
Section 3:					
16. Use of class time was:	3.47	.61	.23	.75	.92
17. Instructor's interest in whether students learned was:	3.69	.58	.22	.74	.92
18. Amount you learned in the course was:	3.52	.59	.20	.71	.91
19. Relevance and usefulness of course content is:	3.64	.59	.19	.70	.90
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.31	.62	.20	.71	.91
21. Reasonableness of assigned work was:	3.50	.58	.19	.70	.90
22. Clarity of student responsibilities and requirements was:	3.62	.58	.19	.70	.90

*N = number of students rating a given class.

Table 3

Mean, Standard Deviation, and Inter-Rater Reliability of All Items on Form B

Item			Reliability		
Section 1:	Mean	SD	N=1	N=10	N=40
1. The course as a whole was:	3.52	.57	.26	.78	.93
2. The course content was:	3.52	.51	.21	.73	.91
3. The instructor's contribution to the course was:	3.71	.66	.30	.81	.94
4. The instructor's effectiveness in teaching the subject matter was:	3.52	.72	.31	.82	.95
Section 2:					
5. Course organization was:	3.45	.60	.24	.76	.93
6. Sequential presentation of concepts was:	3.45	.56	.21	.73	.91
7. Explanations by instructor were:	3.49	.62	.26	.78	.93
8. Instructor's ability to present alternative explanations when needed was:	3.48	.61	.23	.75	.92
9. Instructor's use of examples and illustrations was:	3.68	.58	.25	.77	.93
10. Instructor's enhancement of student interest in the material was:	3.40	.68	.26	.78	.93
11. Student confidence in instructor's knowledge was:	3.98	.57	.22	.74	.92
12. Instructor's enthusiasm was:	3.84	.64	.28	.80	.94
13. Clarity of course objectives was:	3.30	.60	.20	.71	.91
14. Interest level of class sessions was:	3.15	.67	.27	.79	.94
15. Availability of extra help when needed was:	3.48	.54	.16	.66	.88
Section 3:					
16. Use of class time was:	3.44	.59	.23	.75	.92
17. Instructor's interest in whether students learned was:	3.56	.59	.21	.73	.91
18. Amount you learned in the course was:	3.45	.56	.18	.69	.90
19. Relevance and usefulness of course content is:	3.59	.56	.18	.69	.90
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.20	.62	.20	.71	.91
21. Reasonableness of assigned work was:	3.40	.55	.18	.69	.90
22. Clarity of student responsibilities and requirements was:	3.50	.56	.18	.69	.90

Table 4

Mean, Standard Deviation, and Inter-Rater Reliability of All Items on Form C

Item	Mean	SD	Reliability N=1	Reliability N=10	Reliability N=40
Section 1:					
1. The course as a whole was:	3.78	.59	.32	.82	.95
2. The course content was:	3.71	.55	.27	.79	.94
3. The instructor's contribution to the course was:	3.92	.64	.34	.84	.95
4. The instructor's effectiveness in teaching the subject matter was:	3.75	.66	.33	.83	.95
Section 2:					
5. Course organization was:	3.53	.59	.27	.79	.94
6. Instructor's preparation for class was:	3.89	.59	.30	.81	.94
7. Instructor as a discussion leader was:	3.81	.65	.30	.81	.94
8. Instructor's contribution to discussions was:	3.94	.57	.29	.80	.94
9. Conduciveness of class atmosphere to student learning was:	3.75	.65	.30	.81	.94
10. Quality of questions or problems raised was:	3.77	.55	.23	.75	.92
11. Student confidence in instructor's knowledge was:	4.09	.57	.31	.82	.95
12. Instructor's enthusiasm was:	4.13	.59	.32	.82	.95
13. Encouragement given students to express themselves was:	4.06	.56	.26	.78	.93
14. Instructor's openness to student views was:	4.08	.56	.28	.80	.94
15. Interest level of class sessions was:	3.54	.70	.33	.83	.95
Section 3:					
16. Use of class time was:	3.48	.62	.29	.80	.94
17. Instructor's interest in whether students learned was:	3.87	.57	.28	.80	.94
18. Amount you learned in the course was:	3.64	.60	.25	.77	.93
19. Relevance and usefulness of course content is:	3.80	.60	.24	.76	.93
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.46	.63	.26	.78	.93
21. Reasonableness of assigned work was:	3.66	.55	.23	.75	.92
22. Clarity of student responsibilities and requirements was:	3.73	.59	.23	.75	.92

Table 5

Mean, Standard Deviation, and Inter-Rater Reliability of All Items on Form D

Item	Mean	SD	Reliability N=1	Reliability N=10	Reliability N=40
Section 1:					
1. The course as a whole was:	3.46	.58	.27	.79	.94
2. The course content was:	3.44	.51	.22	.74	.92
3. The instructor's contribution to the course was:	3.67	.67	.28	.80	.94
4. The instructor's effectiveness in teaching the subject matter was:	3.49	.69	.28	.80	.94
Section 2:					
5. Course organization was:	3.38	.58	.23	.75	.92
6. Sequential presentation of concepts was:	3.39	.56	.21	.73	.91
7. Explanations by instructor were:	3.43	.62	.25	.77	.93
8. Instructor's ability to present alternative explanations when needed was:	3.42	.63	.24	.76	.93
9. Instructor's use of examples and illustrations was:	3.59	.57	.21	.73	.91
10. Quality of questions or problems raised by instructor was:	3.46	.57	.21	.73	.91
11. Contribution of assignments to understanding course content was:	3.48	.61	.20	.71	.91
12. Instructor's enthusiasm was:	3.86	.62	.25	.77	.93
13. Instructor's ability to deal with student difficulties was:	3.44	.67	.26	.78	.93
14. Answers to student questions were:	3.46	.60	.24	.76	.93
15. Availability of extra help when needed was:	3.59	.60	.20	.71	.91
Section 3:					
16. Use of class time was:	3.36	.60	.23	.75	.92
17. Instructor's interest in whether students learned was:	3.64	.61	.22	.74	.92
18. Amount you learned in the course was:	3.43	.58	.20	.71	.91
19. Relevance and usefulness of course content is:	3.51	.62	.23	.75	.92
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.29	.60	.18	.69	.90
21. Reasonableness of assigned work was:	3.39	.59	.20	.71	.91
22. Clarity of student responsibilities and requirements was:	3.45	.60	.20	.71	.91

Table 6

Mean, Standard Deviation, and Inter-Rater Reliability of All Items on Form E

Item	Mean	SD	Reliability N=1	Reliability N=10	Reliability N=40
Section 1:					
1. The course as a whole was:	3.89	.50	.25	.79	.93
2. The course content was:	3.78	.45	.20	.71	.91
3. The instructor's contribution to the course was:	4.09	.58	.29	.80	.94
4. The instructor's effectiveness in teaching the subject matter was:	3.94	.61	.29	.80	.94
Section 2:					
5. Opportunity for practicing what was learned was:	3.85	.56	.19	.70	.90
6. Sequential development of skills was:	3.66	.49	.19	.70	.90
7. Explanations of underlying rationales for new techniques or skills were:	3.69	.53	.20	.71	.91
8. Demonstrations of expected skills were:	3.67	.54	.21	.73	.91
9. Instructor's confidence in students' ability was:	3.81	.53	.22	.74	.92
10. Recognition of student progress by instructor was:	3.66	.55	.20	.71	.91
11. Student confidence in instructor's knowledge was:	4.20	.53	.25	.77	.93
12. Freedom allowed students to develop own skills and ideas was:	3.82	.58	.21	.73	.91
13. Instructor's ability to deal with student difficulties was:	3.72	.59	.21	.73	.91
14. Tailoring of instruction to varying student skill levels was:	3.50	.58	.19	.70	.90
15. Availability of extra help when needed was:	3.93	.56	.20	.71	.91
Section 3:					
16. Use of class time was:	3.76	.60	.26	.78	.93
17. Instructor's interest in whether students learned was:	4.05	.53	.22	.74	.92
18. Amount you learned in the course was:	3.84	.48	.17	.67	.89
19. Relevance and usefulness of course content is:	3.99	.48	.15	.64	.88
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.61	.59	.22	.74	.92
21. Reasonableness of assigned work was:	3.73	.61	.23	.75	.92
22. Clarity of student responsibilities and requirements was:	3.81	.62	.25	.77	.93

Table 7

Mean, Standard Deviation, and Inter-Rater Reliability for Common Items Across All Forms

Item			Reliability		
Section 1:	Mean	SD	N=1	N=10	N=40
1. The course as a whole was:	3.64	.60	.27	.79	.94
2. The course content was:	3.61	.55	.22	.74	.92
3. The instructor's contribution to the course was:	3.85	.67	.30	.81	.94
4. The instructor's effectiveness in teaching the subject matter was:	3.66	.71	.31	.82	.95
Section 3:					
16. Use of class time was:	3.50	.63	.24	.76	.93
17. Instructor's interest in whether students learned was:	3.75	.62	.24	.76	.93
18. Amount you learned in the course was:	3.58	.60	.20	.71	.91
19. Relevance and usefulness of course content is:	3.71	.61	.20	.71	.91
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	3.37	.67	.21	.73	.91
21. Reasonableness of assigned work was:	3.54	.62	.20	.71	.91
22. Clarity of student responsibilities and requirements was:	3.61	.66	.20	.71	.91

Reliability. The coefficients of reliability which are presented can be interpreted as indices of inter-rater agreement, with the raters in this case being the students within a class. Perfect reliability from this viewpoint would be achieved when all students within each class gave the same rating to that class, and there were differences between classes. Zero reliability, on the other hand, would indicate that the rating given by a student would not depend upon the class he was in.

The presence of reliable ratings is essential to a successful system, for without reliability there can be no validity. If the students who are enrolled cannot demonstrate any consistency in how they rate the course, then the resulting course mean ratings can have no meaning.

The reliability coefficients to be presented are intraclass correlations[4] (Ebel, 1951), and coefficients for a single rater were computed using the following formula:

$$r = \frac{MS_B - MS_W}{MS_B + (k-1)MS_W}$$

where MS_B is the mean square between classes

MS_W is the mean square within classes

k is the average class size.

As the number of students who rate a class increases, the reliability of the resulting class means also increases as a function of the Spearman-Brown formula:

$$r_k = \frac{k(r_1)}{(k-1)r_1 + 1}$$

where r_k is the reliability of a class with k students and

r_1 is the reliability of a single rater.

Reliabilities are presented for class sizes of one student, ten students and 40 students. The reader can start with the value for one rater and

---

[4]The reliability coefficients can also be viewed as generalizability estimates with items considered finite and raters infinite (Kane et al., 1974).

use the Spearman-Brown formula above to compute it for any size class. I have chosen to present these values for a class of size 10, to represent a typical small class, and for size 40, to represent a typical large class.

The reader is reminded that our unit of analysis is the class, which is in reality a confounding of an instructor with a course. There is no simple way to separate these two. The reliabilities to be presented are those of classes, and as such give us information about the dependability of class ratings. They do not give us information about the dependability of ratings of courses, e.g., Economics 201, or instructors, e.g., Professor Doe.

The reliabilities for single raters vary from .15 for item 19 on Form E to .34 for item 3 on Form C. For classes of size 10, the reliabilities range from .64 to .84, and for classes of 40 students, all reliabilities are .88 and above. These data would seem to indicate that items of the IAS are of adequate reliability for all but the very smallest of classes.

In general, the reliability estimates for Form C are somewhat higher than those for the other forms. The reliabilities of items within Section 1 are also somewhat higher than of those within Sections 2 and 3. General instructor effectiveness and course contribution seem to be the concepts which are most reliably rated.

Inter-item correlations. Classes with fewer than six respondents have been eliminated from this analysis. As seen above, smaller classes have less reliability. Hence, the instability of the class means in small classes might unduly influence the magnitude of the correlation coefficients.

Inter-item correlations for the 22 items could be presented for each of the forms. This yields, however, 1155 correlation coefficients which is far more than any rational human being wants to know, not to mention the awesome burden that places on a typist. We opted rather to present the correlations among the 11 common items, across the 5 forms only, which reduces the number of correlations to a more modest 55. Little information is lost as a result of the selection because items within forms tend to be highly correlated, and there does not seem to be a great deal of change in correlational patterns from one form to another.

To illustrate the high inter-item correlations, the average off-diagonal correlations between the 11 items of Section 2 for the five forms is as

follows: Form A = .73; Form B = .77; Form C = .77; Form D = .72; Form E = .66.

The correlations among the 11 general items, 4 from Section 1 and 7 from Section 3 across all forms are found in Table 9. One can immediately note that the values tend to be fairly high, ranging from .54 (items 20 and 21) to .95 (items 3 and 4). Items 3 and 4 are the general instructor items and seem to be eliciting highly similar ratings from students. If the rating of the "Course as a whole" (item 1) is viewed as the most general of items, it is interesting to note which items correlate most highly with it: The content of the course, amount learned in the course, and the two general instructor items, in that order.

A close look at the inter-item correlations reveals some clustering. Items 1, [illegible], 18, and 19 correlate with each other a bit higher than they correlate with the other items. Items 3, 4, and 17 also seem to go together. The former items deal with the course and its content, the latter the instructor. Finally there is a slight tendency for items 20, 21, and 22 to form a cluster. These items have to do with grading, assigned work, and student responsibilities.

The presence of these clusters seems satisfactory intuitively. However, one should not overlook the magnitudes of all of the inter-item correlations, which are consistently high. In other words, the three clusters which have been identified are highly correlated with each other.[5]

Correlations with Non-evaluative Variables. Previously, a list of variables was presented which represented the information solicited from students at the top of each form. To this list, we add some additional variables, which also are not directly evaluative, but could be of interest.

Class size. Actually this variable is the number of forms which were filled out within each class. This value is not identical to class size, but surely highly correlated. In fact, one could make the case that it is actually more representative of the number of students who attend a given

---

[5] One might wonder why factor analytic techniques were not applied to these data. It is my contention that the full correlation matrix is much more informative and less misleading than a factor loading matrix. In the latter case, there is a distinct leveling and sharpening effect, especially with use of orthogonal rotations, which tend to distort the importance of a general factor.

Table 8

Intercorrelations among Common Items Over All Forms (N > 4000)

Item	1	2	3	4	16	17	18	19	20	21	22
Section 1:											
1. The course as a whole was:											
2. The course content was:	.92										
3. The instructor's contribution to the course was:	.88	.80									
4. The instructor's effectiveness in teaching the subject matter was:	.88	.79	.95								
Section 3:											
16. Use of class time was:	.83	.78	.83	.83							
17. Instructor's interest in whether students learned was:	.79	.72	.83	.81	.73						
18. Amount you learned in the course was:	.91	.89	.81	.82	.80	.77					
19. Relevance and usefulness of course content is:	.83	.84	.68	.69	.66	.69	.86				
20. Evaluative and grading techniques (tests, papers, projects, etc.) were:	.74	.68	.71	.72	.67	.76	.72	.63			
21. Reasonableness of assigned work was:	.66	.60	.64	.65	.63	.65	.61	.54	.76		
22. Clarity of student responsibilities and requirements was:	.73	.67	.70	.72	.71	.70	.70	.59	.80	.75	

class, and thus of greater relevance. We also entered the square root of the above value into the analysis, thinking that the student's psychological perception of the bigness of a class might not actually be linear with the actual size, but rather increase as a function more closely representing the square root, e.g., 100 seems twice as large as 25 for instructional purposes, not four times as large.

Publication questions. At the UW, instructors have a choice as to whether they want their results sent to their chairman, their dean, and the results of items 16-22 published for student use. These decisions are made at the time the forms are administered to the class by the instructor and each decision is made independently of the other two. Each of the three variables was dichotomously scored, 0 if not published, 1 if published. For the classes used in this study, the results of 63% of the surveys were sent to the chairman, 33% were sent to deans, and 53% were published for student use.

Course level. Finally, course level was entered as a variable. We used the first digit of the course number for this value. At the UW, 100 through 300 level courses are for undergraduates, with a few graduate students enrolled in 300 level courses. Four-hundred and 500 level courses are for graduate students, with some advanced undergraduates enrolling in the former. Even though students are asked if the course is within their major, etc., at the top of each form, these data were not entered into the analysis due to difficulty in scaling the item.[6] Thus, we are left with 9 variables, the list of which, along with how each was coded for analysis, is presented in Table 9. The correlations of these variables with the 11 common

[6] The reader should again recall that the unit of analysis is the course. Thus, a single value for each variable is entered into the correlations for each course. What this value could be for the major-elective variable is not clear. Arguments could probably be made for using the modal response, i.e., let's consider the course in terms of what the greatest number of students are taking it for. However, this ignores all students who are enrolled for a "non-modal" reason. Also, we could opt to define a continuous type variable which we could define as requiredness, or something like that. However, any such construction placed upon this variable would be potentially misleading. Thus, for the present study, I chose to ignore it.

evaluative items are presented in Table 10. Close examination of the correlations of these variables with the items of the five individual forms failed to yield much information beyond that shown by the correlations with the common items across all forms. As above, only those classes in which six or more students responded to a given item are included in this analysis.

From the correlations of the general evaluative items with the selected non-evaluative variables, one can see that the highest values are with the variable "When enrolling, was this a course you wanted to take?" This is particularly impressive given the fact the mean for this variable was 2.73, thus revealing a high degree of positive response. (These data evidence that by and large students at the UW want to take the classes in which they are enrolled.) These correlations are higher with the items relating to the course and its content than those relating to the instructor and his decisions.

The average expected grade for a class is also positively correlated with all of the common items. The highest correlation is with "Evaluative and grading techniques" ($r = .43$), thus indicating that the higher the average expected grade, the better liked the grading technique.

The average class size and course level are generally correlated positively, but small with all items except for Use of class time, which is slightly negative. The largest correlation for both variables is with "Relevance and usefulness of course content." Apparently higher level courses tend to be considered more relevant by students. Genrally, the correlations with "content" items are greater than with "instructor" items. Class size, on the other hand, generally shows a small negative correlation, with the correlations yielded by the square root transformation being slightly larger in magnitude. The strongest relationship is with the item, "Instructor interest in whether students learned." "Evaluative and grading techniques" is only slightly smaller.

Finally, each of the three publication questions is correlated positively with all items, although the magnitudes are consistently small, indicating that chairmen, deans, and students are <u>not</u> receiving a highly biased set of evaluations.

Table 9

Non-evaluative Variables

Variable	Response categories and code
Wanted to take course	Yes = 3 No = 1 Neutral = 2
Course level	100 level = 1 200 level = 2 300 level = 3 400 level = 4 500 level = 5
Class	Freshman = 1 Sophomore = 2 Junior = 3 Senior = 4 Graduate = 5
Expected grade	A = 4 B = 3 C = 2 D = 1 E = 0 Fail = 0
Class size	Actual number of questionnaires
Square root class size	Square root of above
Chairman copy	Yes = 1 No or omit = 0
Dean copy	Yes = 1 No or omit = 0
Student report	Yes = 1 No or omit = 0

Table 10

Correlations between Common Items and Selected Non-evaluative Variables across All Forms*

Variable	Item 1	2	3	4	16	17	18	19	20	21	22
Wanted to take course	.42	.44	.29	.29	.29	.26	.41	.45	.26	.26	.24
Course level	.09	.13	.02	.01	-.05	.05	.07	.18	.06	-.01	-.03
Class	.10	.14	.03	.03	-.04	.07	.07	.17	.07	.01	-.02
Expected grade	.34	.30	.25	.28	.17	.34	.31	.32	.43	.35	.31
Class size	-.09	-.08	-.09	-.07	-.03	-.17	-.11	-.11	-.16	-.11	-.06
Square root class size	-.15	-.13	-.14	-.12	-.07	-.23	-.16	-.17	-.21	-.15	-.09
Chairman copy	.05	.05	.06	.06	.06	.06	.07	.06	.06	.03	.05
Dean copy	.11	.11	.10	.11	.08	.08	.11	.12	.09	.07	.07
Student report	.09	.09	.11	.12	.11	.10	.11	.08	.10	.10	.12

*See Table 7 or 8 for item wordings.

Discussion

There are many potential areas of discussion embedded in the analyses which have been presented. I have chosen to focus on three such areas, which are: the reliability of items, the extent to which items are diagnostic, and biases or factors outside the course and instructor which might affect ratings.

The reliability of items. As mentioned earlier, inter-rater reliability is a very necessary attribute of a successful instrument. Indeed, individual items should be evaluated in this regard and sub-standard items discarded. Reliability is not sufficient, of course, since one can reliably measure something which has no relationship to the purpose of the measurements.

Based on the data presented, two assertions seem warranted. First, there are no "bad" items in this regard, i.e., every item seems to have adequate reliability. Even the least reliable item reaches .64 with only 10 raters and .78 with 20 raters. Secondly, as results from smaller and smaller classes are considered, more items become of questionable reliability, and interpretation becomes more suspect.

It is also worth mentioning the tendency for the items from Form C to have higher reliabilities than those of the other forms, because this might help dispel a myth extant within higher education. The myth is that seminar type classes are universally easy to teach and liked by students. These data strongly suggest that students are able to consistently discriminate good seminars, from their point of view, from less good seminars. In fact, students are able to better make this discrimination than they can for other types of courses.

Are Items Diagnostic?

Section 2 items were designed to be diagnostic. By this I do not necessarily mean that these items can reveal specific instructional problems, but that these items can reveal areas of problems which can then be looked at more closely. This corresponds to what Smock and Crooks (1974) called level II items. These items can be contrasted with the general items (Section 1) which are indicative of overall quality, but give no hint of where the problems may lie.

We hoped that the directions given to students for the items in Section 2 would elicit more critical information than yielded by items in Section 1. It seemed reasonable to suspect that one could criticize someone more easily in order to help him improve than if it would be information mostly useful in determining a person's promotion or termination. As mentioned in the results section, the closeness in magnitude of item means between the sections indicates that this attempt was probably not successful.

However, the items within Section 2 still could be achieving a diagnostic function. In such a case, we would expect an instructor to be rated favorably on some of the items in Section 2 and not so favorably on others. The mix would depend upon the extent of his instructional problems and the specific items on a form. What we would not expect is all items rated at a level roughly equal to the rating of the general item for a given instructor. Equivalently, we would not expect high inter-item correlations within this section. But, as reported earlier, we do get fairly high inter-item correlations. These correlations can be described as a "halo" effect, which is "...the tendency, in making an estimate or rating of one characteristic of a person, to be influenced by another characteristic or by one's general impression of that person" (English & English, 1958, p. 236). Insofar as there is a halo effect operating, then the items can not be diagnostic.

The diagnosis versus halo question cannot be fully resolved from the data at hand. Even though the inter-item correlations are high, there is still specific variance within each item, that is, variance attributable neither to that which is in common with other variables nor to that which is measurement error. Furthermore, items with similar content correlate more highly than items with dissimilar content. To illustrate, on Form A item 7 (Explanations by instructor were) and item 8 (Instructor's ability to present alternative explanations when needed) correlated at .92. But these two items correlated with item 13 (Encouragement given students to express themselves) at .69 and .72 respectively. Thus, there is some differential responding by students.

Another point which can be made is that when thinking in terms of a halo effect, it is easy to lose sight of the possibility that one who

does well in his teaching in one area also tends to do well in other areas, and vice versa. In other words, the halo may be, in fact, an accurate perception. We choose to accept this view, but also allow for some exceptions. If a given instructor may be strong in most areas, but weak in a few, or vice versa, then the pattern of correlations obtained is just what we would expect.

It seems too early to give up the notion that student rating information can be diagnostic. On the other hand, there is reason for pessimism in this regard. Certainly diagnostic clues may be provided by open-ended comments from students. If these are followed by more detailed and precise closed-ended questions, more successful diagnosis may result. Further research is needed to determine the diagnostic value of student ratings items both within Section 2 and in general.

Biases. One hopes that ratings given a course are reflective of the content and teaching of that course, and not influenced greatly by non-instructional factors. We have isolated four variables which seem to relate to ratings to a non-trivial extent: whether, when registering, students wanted to take the class, expected grades, the evaluation form used, and class size. These are ordered in terms of apparent importance. Showing a relationship is one thing, however, and understanding it causally is quite another.

The question, "When registering, was this a course you wanted to take?" seems to be the potentially most important. Not only does it account for more variance than any of the others (almost 20% for some items), it is also information which could be collected at the beginning of the course, and later ratings could then be appropriately adjusted. The causation can only go one way. The only reservation about this variable is the extent to which the reputation of the course or instructor influences whether or not students wnat to take it or him. If the correlation were so explained, then clearly adjustments in end-of-course ratings are not appropriate.

The relation between ratings and expected grade is an explosive issue. The argument goes that the way to get high ratings is to promise students high grades. However, one could also argue that if students like the course they will work harder and get better grades, or a well

taught course will result in both more learning, and hence higher grades, and high ratings. For any of the above, a positive correlation would result.

The fact that instructors using Form E tend to get somewhat higher ratings is equally fuzzy in interpretation. It is possible that there is something inherent in skill-acquisition type courses which students like better. Also, it is equally plausable that instructors tend to work a bit harder to put this type of course together, or perhaps tend to be more student oriented, which is reflected in their willingness to teach this kind of course. If one chooses to make any of these arguments, he must also be willing to make the same argument in a negative sense concerning problem solving courses, since they come out at the bottom of the heap. One thing is clear. It is the course which is important, not the form, since it is the common items on which Form E users are higher, not the diagnostic items. Thus, the act of choosing Form E alone does not help anybody's ratings.

Finally, I mention class size. In actuality, it is not a major influence on ratings, possibly much less of an influence than people think. Perhaps the most interesting aspect is that large classes do not automatically lead to low ratings.

Reference Notes.

1. Gillmore, G. M. A brief description of the University of Washington Instructional Assessment System (EAC Report 276). Seattle: University of Washington, Educational Assessment Center, 1974.
2. Hodgson, T. F. A systematic approach to instructional assessment by Students. Paper presented at the Washington Educational Research Association, Seattle, May, 1974.
3. Kane, M. T., Gillmore, G. M., & Crooks, T. J. The application of generalizability theory to course evaluation questionnaires. Paper presented at the National Council of Measurement in Education, Chicago, April, 1974.

References

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

English, H. B., & English, A. C. A comprehensive dictionary of psychological and psychoanalytical terms. New York: Longmans, Green & Co., 1958.

Smock, H. R., & Crooks, T. J. A plan for the comprehensive evaluation of college teaching. Journal of Higher Education, 1973, 44, 577-586.