DOCUMENT RESUME

ED 179 263 JC 790 635

AUTHOR Morsch, Joseph E.: And Others
TITLE Student Achievement as a Measure of Instructor Effectiveness.
INSTITUTION Air Force Personnel and Training Research Center, Lackland AFB, Tex.
PUB DATE 14 Mar 55
NOTE 28p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Achievement Gains: *Effective Teaching: *Evaluation Criteria: Evaluation Methods: Military Personnel: Military Training: Peer Evaluation: Performance Based Education: Performance Tests: *Predictive Validity: Predictor Variables: Student Evaluation of Teacher Performance: Teacher Characteristics: *Teacher Evaluation: Teacher Qualifications: Technical Education: Vocational Education Teachers

ABSTRACT

Using an eight-day hydraulics maintenance course taught by 121 instructors on a regular basis to classes of about 14 students using the same classrooms and materials, the Air Force conducted a study designed to determine instructor measures which correlate with, and are therefore predictive of, student achievement. The study report first discusses the problem of measuring student achievement, including the "ceiling effect" induced by inadequate precourse tests, the contamination of test results and rating measures, and the frequent lack of adequate samples of teachers imparting the same subject matter to similar students in the same learning environment. A description of the study methods and instruments follows, by which student gains criteria (as measured by a pre-course exam, a written post-test, and a performance examination) are correlated with instructor variables determined by peer and supervisor evaluation and by tests used to measure instructors' subject knowledge and general intelligence. The report then outlines the statistical criteria interrelations and validities and summarizes the study conclusions. Major findings include the result that students' ratings of instructors' subject knowledge produced significant correlation with instructors' proficiency tests while peer and supervisor ratings did not significantly correlate with student gains criteria. (JP)

STUDENT ACHIEVEMENT AS A MEASURE OF INSTRUCTOR EFFECTIVENESS

By Joseph E. Morsch, George G. Burgess,

and Paul N. Smith

Personnel Research Laboratory
AIR FORCE PERSONNEL AND TRAINING RESEARCH CENTER
Air Research and Development Command
Lackland Air Force Base, Texas

Project No. 7950
Task No. 77243

Approved by:
Lloyd G. Humphreys, Director
Personnel Research Laboratory

ACKNOWLEDGMENTS

The authors wish to express their gratitude to Col. Harold F. Layhee, then Commander of the 3750th Technical Training Group, to Mr. Herbert E. Johanson, Director, and to Mr. William A. Harriman, Proficiency Measurements Officer of the Training Analysis Division, to Capt. Lorne A. Davis, Mr. C. L. May, Mr. L. P. Gray, and to the supervisors and instructors of the Hydraulics Branch of the Aircraft Mechanics General Course at Sheppard Air Force Base. The genuine interest and wholehearted cooperation of these persons in making student airmen and facilities available made the study possible.

M. Sgt. John E. Burks, Jr., M. Sgt. Joseph M. Donze, S. Sgt. James F. Pierce, S. Sgt. Valentine W. Schreiber, S. Sgt. Richard E. Shaw, A/1C Robert G. Breckenridge, Jr., A/1C Joseph J. Leslie, and A/1C Richard H. Perkins developed and administered tests, observed instructor performance, and assisted with the statistical analysis.

IMPLICATIONS

The Air Force has a continuing need for measures to be used in the evaluation of instructors' on-the-job performance. Such standards of success in teaching are required so that the most effective instructors may be promoted, upgraded, or given supervisory responsibility while the less effective instructors may be given remedial training or reassignment.

This Research Report describes an experimental study aimed at developing a reliable criterion of instructor effectiveness. Although the work was done with instructors in an aircraft mechanics course at Sheppard Air Force Base, it was anticipated that the results of the study would find application in other courses at other bases.

Because it provides the most objective criterion so far discovered, the effectiveness measure used was the actual subject-matter achievement of an instructor's students, often called student gains. Since a precise comparison of instructors in terms of student achievement is usually impractical in an operational situation, an attempt was made in the study to find other characteristics of the instructor, besides imparting subject matter, which might be related to gains and could thus be used to predict student achievement.

In planning the study it was necessary to locate from among Air Force technical schools a course in which an adequate sample of instructors could be compared on the basis of their students' achievement. This required that the instructors should teach the same subject matter to students similarly selected, in similar classrooms, with the same training aids. The Hydraulics Phase of the Aircraft Mechanics Course at Sheppard Air Force Base was chosen because it provided 121 instructors and because this phase was taught in a single building arranged with special classrooms, each with training equipment for seven successive days of training (an eighth day being devoted to testing). Six new classes of about 14 students each, in three shifts, entered each first day's training. Each class was assigned to an instructor who moved along with his class through seven days. The Hydraulics Phase was sufficiently short to permit each instructor to teach two classes a month apart, which allowed the gains of his first class to be compared with gains of his second class.

In order to obtain student gains, a pretest was used. On the basis of this pretest and of grades that students achieved in three previous phases (fundamentals, structures, and electrical) the scores students would make on a hydraulics posttest were predicted. An index of instructor effectiveness, the measure of his students' gains, was determined in terms of whether or not his classes exceeded expectation. Besides the gains of their students, several other criteria of instructor effectiveness were used in the study. These included forced-choice ratings and rankings by supervisors, rankings by fellow instructors, and several kinds of student ratings. Other instructor measures included ratings on verbal facility and scores on tests of hydraulics proficiency and intelligence.

The chief results of the investigation are the findings that student gains can be reliably measured and that students' ratings of their instructors' teaching effectiveness and supervisors' ratings of instructors' verbal facility are correlated significantly with student gains.

Technical school students appear to know when they are well taught. Their ratings of their instructors offer promise as a technique for instructor evaluation. Teaching which emphasizes acquisition of subject matter may induce a favorable attitude of students toward their instructors.

The high relationship found between ratings and rankings by fellow instructors and supervisors, together with the fact that these measures appear unrelated to student gains, suggests that fellow instructors and supervisors judge instructor effectiveness on the basis of factors other than what students learn. One of these factors appears to be the instructor's knowledge of subject matter. To obtain a completely adequate evaluation of an instructor it may be that a multiple criterion composed of supervisor ratings, student ratings, and student gains should be used. The study also suggests that further more detailed investigation might be made of speech factors as related to instructor effectiveness.

This report will be of interest to instructors and instructor-supervisor personnel in technical training schools. The report will also be of interest to military psychologists and to anyone confronted with the problem of instructor evaluation.

Hq, AFPTRC
Lackland Air Force Base
San Antonio, Texas
27 May 1955

Arthur W. Melton
Technical Director

Herbert N. Cowles
Col., USAF
Commander

TABLE OF CONTENTS

LIST OF TABLES

STUDENT ACHIEVEMENT AS A MEASURE OF INSTRUCTOR EFFECTIVENESS*

INTRODUCTION

The present investigation was undertaken to determine whether or not a reliable measure of instructor effectiveness in terms of student achievement could be developed. Having found such a criterion an attempt was to be made to identify instructor measures that would correlate with and hence could be used to predict the criterion or could be used in place of it.

The accomplishment of the mission of the United States Air Force depends upon training large numbers of men in difficult and complex technical skills, and their adequate training requires a competent, effective instructor staff. Since the Air Force constantly must replenish its instructor corps, it is faced with three major problems, i.e., the selection of potential instructor personnel, their subsequent training, and the evaluation of their on-the-job performance.

In selecting potential instructors from a large group of men valid and reliable selection measures should be used (or devised if they are not available) because they will help to insure that the right men are assigned to an instructor training program, i.e., those men who will become effective instructors.

Standards for evaluating the instructor training programs must be found to bring assurance that such programs are the best that can be developed for teaching the men those skills they will need to become effective instructors.

Finally some standards of success in teaching are required so that the most effective instructors may be promoted, upgraded, and given supervisory jobs, and so that the poorer, less effective instructors may be identified. In addition, the identification of instructors who are most effective will make it possible to study their characteristics and teaching procedures. This information will in turn assist the Air Force in arriving at some solution of the problems involved in selecting, training, and evaluating instructors. The present study was designed to investigate this last problem.

CRITERIA OF INSTRUCTOR EFFECTIVENESS

The identification of successful instructors presents problems which are neither new nor unique to the Air Force. The importance of these problems has been recognized and attempts made to solve them in several hundred studies over the past fifty years. The criterion of instructor effectiveness

---

* Manuscript received 14 March 1955

most widely used in these investigations has been some form of rating. Usually supervisors' ratings have been used; but peer ratings, student ratings, and self ratings have also been tried out as criteria.

While the studies appear to show that instructors can be reliably rated, ratings still are not completely satisfactory as criteria of instructor effectiveness. Ratings are merely one method of estimating an instructor's effectiveness, and there has been no way of knowing whether or not the ratings include all (or any) of the important aspects of that behavior. Furthermore, ratings are subject to many faults, such as the halo effect, and the persistent attempts made to improve and refine them have not been very successful.

Such problems have led to a search for other criteria that would avoid some of these deficiencies. It is generally agreed that student growth or gains (the amount of subject matter that the students learn with respect to their ability) is one of the most if not the most important criteria of instructor effectiveness. There is also general agreement that it is one of the most difficult criteria to measure because of the many controls required to make effectiveness comparisons meaningful. The fundamental value of student achievement as a criterion of teaching efficiency has long been recognized. As early as 1921, Courtis (3) pointed out the value of student gains for comparing teachers with respect to effectiveness, as well as the importance of holding the variable of class ability constant. Except for four studies during the 1930's however, experimental work with the gains criterion made us of relatively simple statistical controls of initial student ability such as subtracting pretest from posttest raw scores, or of gain corrected for one or two extraneous factors such as student intelligence or previous training. The fact that initial student abilities must be controlled statistically was apparently neglected until 1933 when Betts (1) used student gain in reading as a criterion of teacher effectiveness after partialling out student age, initial score, and class variance. Bolton (2) used the ratio of mean pupil achievement to its standard deviation as a measure of teaching effectiveness. Seyfert and Tyndal (16) determined student achievement in terms of the mental age necessary in order that a student of the less able of two teachers might achieve the same score level as a corresponding student of a better teacher. Residual pupil gain, that is, the difference between actual gain and gain predicted on the basis of pupil factors such as initial score or intelligence quotient has been used by Gotham (4), Jayne (9), Jones (9) LaDuke (10), Lins (11), Remmers et al (12), Riesch (13), Rolfe (14), Rostker (15) and Von Haden (18). A modification of this type of gains criterion, to be discussed in detail later, was used in the present investigation.

The gains criterion provides a more objective and probably a more logically defensible criterion than those provided by ratings. The reported results of gains studies, however, have shown considerable inconsistency and thus far have failed to demonstrate the practical utility of student gains as a criterion of instructor effectiveness.

Any determination of gain is dependent upon the availability of valid and reliable instruments for measuring such growth. The measures used to show student gain, especially in the case of some of the earlier investigations were quite inadequate. Some of these studies were based on essay type tests. Other studies were concerned with more or less inconsequential aspects of student achievement. The reliability of the measures used usually was not reported.

The "ceiling" effects of tests used appear to have been largely ignored. A maximum or perfect score on a precourse test limits the amount of gain possible. Thus, a very effective instructor whose students obtained high precourse scores might show up poorly under a gains measure because his students would have to "gain" very little to achieve a maximum score on the postcourse test, i.e., they would have little opportunity to exhibit gain.

In certain instances proper precautions were not taken to eliminate the possibility of contamination of test results or of the rating measure used. In cases in which gains measures and rating scales were compared the person who judged the teacher's effectiveness might also administer and score tests to determine the achievement of the teachers' students.

Most investigators of student gains as a measure of instructor effectiveness have been confronted with sampling difficulties. In general, student gain in one subject matter area cannot be compared with gains in different students in other areas, and often it has not been possible to secure adequate samples of teachers all imparting the same subject matter to the same kind of student in the same type of school situation. Consequently, in some studies the numbers of teachers and of pupils have been so small that any findings must be regarded as highly tentative.

If student gains are to be used as a measure of instructor effectiveness it is necessary to hold constant insofar as possible, all relevant variables other than the effects of the teaching itself. The instructor is only one of many factors operating to produce changes in the students. Variables such as interest, motivation, and aptitude, which affect student learning, often have not been properly controlled. In addition to variables which can be experimentally controlled, at least in part, there are other variables which are very hard to control. Changes in student behavior in a particular phase of instruction may be due only in part to the instructor presently teaching that phase. Certain aspects of the student change may reflect the influence of previous instructors, some of whom may have had a continuing influence upon their students. In other cases personal and environmental factors not measured may outweigh the effects of the instructor's performance to such an extent that even the most refined of statistical techniques will fail to isolate the elements of student change for which the instructor is responsible.

Too, the instructor is called upon to accomplish many changes in his students which are not measurable in terms of subject-matter achievement.

Any measure of student gains, therefore, represents only a part of the instructor's total effectiveness. This particular limitation of the gains criterion is not as applicable to the Air Force situation in which the instructor's chief concern is the imparting of course materials of a technical nature as it is in a civilian school where attitudes and other objectives which are difficult to measure are emphasized.

If it can be determined that students of one instructor make greater gains than do those of another instructor, the problem still remains to determine what behaviors, traits, or characteristics of the successful instructor are responsible for the changes produced in the students. Since the operational use of the gains criterion is in many cases impractical, more easily obtained correlates of this criterion may be determined and substituted for measures of student gains. Such simple measures may then be used for evaluative or predictive purposes.

DESIGN OF THE RESEARCH

In planning this study particular attention was given to selecting from among all Air Force Technical Schools the course or courses in which an adequate sample of instructors could be compared on the basis of their students' achievement. This required that the instructors should teach the same subject matter to students similarly selected, in similar classrooms, with the same training aids. Courses in several of the larger technical schools were canvassed. From these the Aircraft Mechanics Course at Sheppard Air Force Base was selected, and from this course the Hydraulics Phase was chosen as best fitting the requirements of the study. This phase was taught in a single building in which two rows of classrooms were arranged on either side of a central corridor. On each side, the first classroom had training equipment for the first day's training, the second classroom had equipment for the second day's training and so on for each of the seven days of training in the phase. Every week day six new classes of about fourteen students each entered the first day of training, two classes on each of three shifts. Each class was assigned an instructor who moved along with his class through the seven days of training. The eighth day was devoted to written and performance testing over the subject matter of the seven previous days. The instructor did not participate in the testing. Since the school was using two instructors per class in the hydraulics phase, 121 instructors were available. During the present investigation, however, only one instructor per class was used so that the effects of his teaching would be isolated.

Because the hydraulics training phase lasted only eight days (including the testing day), it was feasible to duplicate the experiment, thus obtaining data based on two classes for each instructor. The correlation between the gains of the instructors' first class and of their second class (taught approximately a month later) was used in determining the reliability of this measure of instructor effectiveness.

DESCRIPTION OF VARIABLES USED

Student Variables

The most important of the student measures was the posttest. A special posttest had to be developed for the experiment because the content of written tests previously used in the Hydraulics Phase was too likely to be known to the instructors, and the many easy items in these tests tended to introduce ceiling effects.

In developing the written hydraulics posttest to be used with the experimental student groups six 50-item tentative posttests were constructed. Each was administered to approximately 84 airmen in order to determine which items fell within a predetermined difficulty range of 30% to 70% passing. (The scores of these tentative tests were used by the Hydraulics Branch in determining the written portion of the phase grade thus insuring high student motivation.) In constructing these tests 205 discrete four-choice items were used. Some 95 items were revised and retested on the basis of results of sequential item difficulty analysis. In the course of developing items two members of the research team sat in as students in the hydraulics classes. All items were subjected to rigorous scrutiny by supervisors in the Hydraulics Branch.

A final test composed of 75 items which satisfied the conditions of the experiment with respect to difficulty, adequacy of incorrect alternatives, and internal consistency was assembled. Four forms of this test containing identifical items but with varied order of items and/or alternatives were prepared in order to reduce the possibility of compromise of the key. That the compromise was successfully eliminated was shown by the fact that mean class scores of subsequent classes did not increase beyong chance expectancy throughout the course of the experiment. This 75-item multiple choice hydraulics posttest was administered to all students in the experimental group as their final written test in the Hydraulics Branch. All pre- and posttests were administered and scored by research personnel and were not seen by any hydraulics instructor.

The regular hydraulics performance examination, a nonwritten job-sample type test, used by the Branch was used in this study. It included checking units for internal leakage, identifying units and fittings, and making operational checks of various hydraulics systems. Each student was required to complete 10 of the 28 performance items available. The time varied from 10 to 30 minutes per item, the total practical test time being about four hours. Hydraulics instructors trained in the administration of the performance test were used as examiners. In no case, however, did a class instructor test his own class. Each item was graded from one to five points, examiners being provided with a grading guide which listed the number of points allotted for each step completed satisfactorily. In determining final performance grades, total raw scores of the performance test were converted to T-scores.

If the instructors were to be evaluated on the basis of their students' achievement, then the initial ability of all instructors' classes had to be considered. In the Air Force situation where the Mechanical Aptitude Indexes of students vary from 4 to 9 it cannot be certain that classes taught by different instructors are of equal ability. Since matching of students to insure classes of approximately equal ability presented almost insurmountable difficulties, it was decided to try to make adjustments statistically rather than experimentally. For this purpose a special pretest of student knowledge about hydraulics was developed and administered to all students before they entered the Hydraulics Branch. Since the period of instruction in hydraulics was only eight days, the use of identical or even similar pre- and posttests to measure student gain might have introduced considerable memorization and practice effects. Therefore, the pretest was composed of items appropriate for students who had had some experience with the subject matter to be taught but had not been exposed to the specific course subject matter. It contained items pertaining to the background knowledge and theory required for learning the phase content as contrasted with the posttest which consisted of items covering specific subject matter information expected of students who had just completed the phase. The content of the pretest was selected to correlate maximally with the posttest, but was not equivalent to it.

In constructing the pretest 211 four-choice items were tried out, many of these being revised and readministered. In the course of making the preliminary item difficulty analysis by sequential procedures, 25 different tentative 20- to 25-item pretests were given to a total of 650 students before they had entered the Hydraulics Branch. The items were chosen from examinations that had previously been used in the Branch or were suggested by knowledge of hydraulics, physics, or mechanics that students might have learned before entering the Aircraft Mechanics Course. On the basis of the item difficulty analysis a 75-item pretest was assembled. After statistical determination of internal consistency and item validity as determined by correlation with the posttest scores this was reduced to 65 items, and four forms were constructed in a manner similar to that of the posttest. The pretest was then administered to all students in the experimental group during the last day of the Electrical Branch immediately prior to their entry into the Hydraulics Branch. Students used in the item analysis of the pre- and posttests took no further part in the experiment.

Other student variables used included three final grades, one each from the Fundamentals, Structures, and Electrical Branches, the three previous phases of instruction in the Aircraft Mechanics School. Aptitude indexes, the standard composite classification scores described by Gragg and Gordon (1950) were obtained from the students' record card. However, preliminary analysis showed that these indexes did not increase the multiple correlation between the other predictor student variables and the posttests. Accordingly aptitude indexes were not used in the final analysis.

Instructor Variables

Supervisors' ratings of the hydraulics instructors were obtained by means of the Instructor Description Form-C, (7) the form officially designated by the Air Training Command to be used in rating AF technical school instructors. The "forced choice" part of this form consists of 29 four-item blocks, the rater being required to choose in each block two statements which best describe the instructor being rated. A graphic scale is also included on the Form-C. This consists of nine behavioral categories on each of which the supervisor rates the instructor as compared with other instructors (1) generally weak, (2) somewhat lacking compared to others, (3) compares favorably with others, (4) considerably better than most, (5) one of the best. The graphic scale was scored by assigning weights of zero to four to the five categories and then summing over the nine areas rated to yield a single score per instructor. The scoring of Form-C is described by Highland and Berkshire (7). In addition, supervisors were required to indicate the area in which the instructor was strongest, second strongest, and third strongest and the area in which he was weakest, second weakest, and third weakest. Shift supervisors also were asked to rank-order on a roster the approximately 40 instructors on each of three shifts in terms of their "general effectiveness." Each instructor received a T-score based on rank and group size.

Hydraulics instructors were asked to rank-order their colleagues on the same shift in terms of their effectiveness as instructors, omitting those whom they did not know well enough to rank. The rankings were done anonymously. Mean ranks were obtained and converted to T-scores.

Just before taking their final phase examination students rated their instructors as outstanding, very good, good, or poor or unsatisfactory, on each of four qualities of behavior: (a) knowledge of subject, (b) teaching methods, (c) understanding of students, and (d) as a personal friend. They then rankd these four qualities in the order of their instructor's relative strength in them. Students also were required to make an "overall" rating of their instructor by indicating whether or not "they felt they were fortunate in having had this instructor." In subsequent portions of the report, the student rating of instructors is referred to as the "student graphic rating," and the ranking of qualities is called the "student forced choice rating." The instructors were scored separately on each of the four qualities by averaging over students of the two classes to produce a mean score on each quality.

The Wonderlic Personnel Test was used as a measure of instructor general intelligence. This test consists of 50 items and has a 12-minute time limit. The score was the number of items right.

Subject matter knowledge was measured in terms of the Basic Knowledge Proficiency Examination, Airplane Hydraulic Mechanic, Experimental Form - 104X-2 as devised (March, 1951) by the Proficiency Analysis Division of the

Technical Training Research Laboratory. This test consists of 125 multiple choice items arranged to be used with an IBM answer sheet. Of these items 122 were five-choice and 3 were four-choice items. The instructors were allowed all the time they wished in answering this test. The score was the number of correct answers.

Verbal facility ratings were obtained by having six supervisors (no members of the Hydraulics Branch) rank-order (in groups of six) individual instructors on their organization and presentation of special material which they had been allowed to prepare for 15 to 20 minutes. Simultaneous tape recordings were made of the presentations for further analysis. Verbal facility scores were determined by averaging the ranks assigned by supervisors and converting to T-scores.

Three instructor personality measures and three observation check lists were also used. The evaluation of these measures requires extensive item analysis which has not yet been completed. Results obtained from these measures will be given in a later report.

RESULTS

In addition to the other instructor variables three scores which will be referred to as instructor effectiveness scores were determined. These scores are based upon the average student posttest scores adjusted for differences in the average pretest and aptitude scores.

The statistical procedure used to estimate the adjustment to be made to posttest scores to compensate for original aptitude and ability is described at length in the Appendix. Briefly, the adjustment was based upon a multiple regression equation computed from within instructor and between classes variances and covariances. Two different effectiveness scores were computed for each instructor: one based upon the written posttest and one upon the performance posttest. Since these scores have variances of similar magnitude and lack a basis for arbitrary weighting a combined effectiveness score was obtained by adding these scores without weighting. The reliabilities of instructor effectiveness scores were estimated from the correlation between scores based upon the achievement of the first class and those based upon the achievement of the second class. These are presented in Table 1.

Criterion Intercorrelations

The correlation coefficients shown in Table 2 are based on 106 instructors. With this sample size a coefficient of .253 is significantly different from zero at the .01 level of confidence while .192 is significant at the .05 level.

Table 1

Reliability Estimates of Student Gains

(N = 106)

	r	r^a
Written gains -- First class vs. Second class	.34	.51
Performance gains -- First class vs. Second class	.32	.48
Combined written and performance gains -- First class vs. Second class	.38	.55

---

[a] r adjusted by Spearman-Brown formula for double length; thus, represents estimate of reliablity when two classes are used to obtain an effectiveness score.

Table 2

Intercorrelations of Criterion and Predictor Variables[a]

Variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
1 Student gains -- Total		88	88	40	46	47	28	09	18	-22	00	-36	13	11	12	16	-04	00
2 Student gains -- Performance			55	39	41	42	26	00	22	-19	08	-22	14	12	12	07	-16	-12
3 Student gains -- Written				32	41	41	24	16	11	-20	-07	-40	09	07	09	20	08	11
4 Students' over-all rating					86	64	76	24	53	-34	61	-53	18	18	18	07	-05	01
5 Students' rating-- Teaching ability (G)[b]						72	77	22	56	-35	55	-59	19	21	18	03	-07	01
6 Students' rating-- Teaching ability (FC)[c]							40	-18	35	-35	22	-47	19	15	17	01	-06	-05
7 Students' rating-- Understanding of students (G)								59	44	-44	69	-53	09	14	07	09	-15	-06
8 Students' rating--Understanding of students (FC)									14	-53	34	-27	-22	-14	-24	15	-27	-06
9 Students' rating--Knowledge of subject (G)										33	29	-58	31	30	27	03	19	02
10 Students' rating--Knowledge of subject (FC)											-47	-18	26	24	24	-08	43	12
11 Students' rating--as a friend (G)												-08	02	04	01	11	-22	-11
12 Students' rating--as a friend (FC)													-25	-27	-19	-09	-16	-03
13 Peer ranking														77	52	11	31	03
14 Supervisors' ranking--Form C															67	11	23	05
15 Supervisors' (FC) rating--Form C																10	25	15
16 Verbal facility rating																	03	23
17 Hydraulics Proficiency Test																		44
18 Wonderlic Personnel Test																		

As shown in Table 2, three student gains criteria of teaching effectiveness were used. These were based on the written posttest, the performance posttest and a combination of these tests. The written and performance posttests were correlated .55 and each of these when correlated with the total measure yielded and r of .88.

Criterion Validities

Student Ratings

Students' over-all rating of instructors was correlated significantly with all three of the gains criteria (written r = .32, performance r = .39 and total gains r = .40). Slightly higher correlations were found for students' graphic and forced choice ratings of instructors' teaching ability, the r's being .41, .41, and .46, respectively, for the graphic and .41, .42, and .47 for the forced choice with the three gains criteria.

Students' graphic ratings of instructors' understanding of students were correlated significantly with the three gains criteria (r's = .24, .26, and .28) but students' forced choice ratings of instructors' understanding failed to correlate significantly with any gains measure.

Students' graphic ratings of instructors' knowledge of subject were correlated significantly only with the performance gains criterion (r = .22). Students' forced choice ratings of instructors' knowledge of subject showed a significant negative correlation with written gains (r = -.20) and with the total gains criterion (r = -.22).

Students' graphic rating of the instructor as a friend was not significantly correlated with any of the three gains criteria, while their forced choice rating of the instructor as a friend was correlated negatively with all three gains criteria (r's = -.40, -.22, and -.36, respectively). The negative correlations found in the case of the forced choice ratings probably represent an artifact due to the requirements of forced choice rating. The sum of the correlation coefficients of the four forced choice student ratings with an outside variable should approximate zero because a high score on one trait is compensated for by a low score on another trait. As pointed out by Thomson (17, p. 304) and others, among n variates the limit of the average correlation in the negative direction is $-1/(n-1)$. The magnitude of the negative correlation to be expected here under the simplest assumption possible would, therefore, be of the order of r = -.33. The negative correlations observed thus seem logically attributable to artifact.

Since these results were encouraging, the reliablity of student ratings was estimated by computing the correlation between instructor ratings made by the first class and by the second class. These results are shown in Table 3.

Table 3

Reliability of Student Ratings

Student Rating	r^a	r^b
1. Fortunate in having this instructor	.26	.41
2. Teaching ability (Graphic)	.34	.51
3. Teaching ability (Forced choice)	.32	.49
4. Understanding of students (Graphic)	.26	.41
5. Understanding of students (Forced Choice)	.42	.60
6. Knowledge of subject (Graphic)	.33	.49
7. Knowledge of subject (Forced choice)	.32	.49
8. As a friend (Graphic)	.18	.31
9. As a friend (Forced choice)	.21	.34

[a]Correlations between mean ratings by first class and second class

[b]r adjusted by Spearman-Brown formula for double length

However, the high correlations between student gains and student ratings could be accounted for by factors specific to the student or to the class situation with little reliability over time. To test this possibility the correlations between first class student ratings and second class gains and vice versa were determined. Results are shown in Table 4.

As can be seen from Table 4, these cross correlations are smaller than those shown in Table 2. However, the correlations between second class rating and first class gains for the more valid measures continued significant. Unfortunately, the other set of correlations fails to show this consistency. Possibly the true value of teh correlations lies somewhere between the two. If so, some real validity can be attributed to the student ratings.

Peer Ratings and Supervisor Ratings and Rankings

Neither peer rankings nor supervisor ratings nor rankings were correlated significantly with any of the three student gains criteria.

Table 4

Correlations of Student Ratings of Instructors with Combined Written and Performance Mean Gains

Student Rating of Instructor	Graphic r	Forced Choice r
1st Class Student Ratings vs. 2nd Class Gains		
1. Teaching ability	.08	.25
2. Understanding of students	-.05	-.04
3. Knowledge of subject	.09	-.09
4. As a friend	-.12	-.13
5. Over-all student rating	.09	
2nd Class Student Rating vs. 1st Class Gains		
6. Teaching ability	.34	.25
7. Understanding of students	.26	.25
8. Knowledge of subject	.14	-.21
9. As a friend	.15	-.08
10. Over-all student rating	.29	

Other measures

Verbal facility ratings correlated significantly with the written gains measure ($r = .20$) but with neither of the other gains criteria. Neither the Hydraulics Proficiency Test nor the Wonderlic Personnel Test were correlated significantly with any of the student gains criteria. We see, therefore, that student ratings of their instructors were the only instructor measures which seemed to predict the student gains criterion.

Intercorrelation of Student Ratings

All student rating variables tended to be highly intercorrelated with the students' forced choice rating of the instructor as a friend and their forced choice rating of instructor's knowledge of subject matter tending to give significantly negative correlations. As explained above these negative correlations were probably artifacts of the forced choice rating technique.

Intercorrelations of Peer Rankings and Supervisor Ratings and Rankings

Besides the student ratings, three other instructor rating or ranking measures were used. These were supervisors' Form C rating, supervisors' ranking, and peer ranking. The two supervisor measures when correlated yielded a coefficient of .67. Supervisor Form C ratings were correlated with peer ranking ($r = .52$) while supervisor ranking and peer ranking gave rise to a correlation coefficient of .77. This high correlation found be-between supervisor and peer rankings coupled with the fact that neither of these measures correlated highly with the student gains criterion suggests that supervisors and peers judge instructors on the basis of factors other than teaching effectiveness as measured by the gains criteria. As will appear in the next section, one of these factors seems to be instructors' knowledge of subject matter. We thus see rather close agreement between peer and supervisor opinion but fellow instructors and supervisors agree only slightly with student opinion. This latter finding has been reported by Guthrie (6) and other investigators.

Intercorrelations of Verbal Facility, Subject Matter, Knowledge, and Intelligence Measures

Verbal facility was correlated positively with scores on the Wonderlic Personnel Test ($r = .23$) but with no other instructor measure. There was a correlation of .44 between Wonderlic Personnel Test scores and scores on the Hydraulics Proficiency Examination. A correlation coefficient of .43 was found between scores on the proficiency test and students' forced choice rating of instructors' knowledge of subject matter and an r of .19 between the proficiency test and students' graphic rating of instructors' knowledge of the subject. There was a negative correlation ($r = -.22$) between the proficiency test and students' graphic rating of the instructor as a friend. The correlation between the proficiency test and students' forced choice rating of the instructor as a friend was not significant. Students' forced choice rating of the instructors' understanding of students also correlated negatively ($r = -.27$) with instructors' proficiency test scores. Here again, these negative correlation coefficients are due to the forced choice techniqe used.

The variables may be thought of as falling into several "clusters": a gains cluster, an intelligence-verbal facility cluster, a peer-supervisor ranking cluster, a knowledge of subject matter cluster, and a student rating cluster. The latter clusters appear to share more variables with each other than they do with the peer-supervisor ranking cluster. Intercorrelations among the variables, however, suggest that some variables seem to be involved in more than one cluster. For example both peer rankings and supervisor ratings and rankings are related to students' forced choice and graphic ratings of instructors' knowledge of subject matter. Peer rankings and supervisor ratings and rankings are also correlated with instructors' scores on the hydraulics proficiency test.

SUMMARY AND CONCLUSIONS

1. Student gains can be reliably measured. It has been demonstrated that the gains measure has some reliability. The reliability obtained, however, was not comparable to the reliability of such pyschometric instruments as the conventional intelligence tests.

2. Little relationship between student gains and instructor test scores has been found.

3. Little relationship between supervisor or peer estimates of instructor effectiveness and student gains was shown.

4. Students appear to know when they are well taught.

5. If reasonable precautions can be taken to preclude student evaluation of instructors from becoming a popularity contest or mere guesses at the instructors' ability, student ratings offer promise as a technique for instructor evaluation.

6. Teaching which maximizes acquisition of subject matter may induce favorable attitudes of students toward their instructors.

7. The high correlations found between peer and supervisor rankings plus the fact that neither of these measures correlates highly with the student gains criterion suggests that peers and supervisors judge instructors on the basis of factors other than teaching effectiveness as measured by student gains. One of these factors appears to be subject matter knowledge.

8. Students' ratings of instructors' subject matter knowledge produces significant correlations with instructors' proficiency test scores.

9. The correlation between gains on the written and practical tests indicates that they have about 25% common variance. These measures show a logical patterning of correlations with other measures. This rather definitely implies that gains scores are specific to the type of posttest used.

10. The correlation between verbal facility rating of the instructor and the student written gains criterion while not high was significant (at the .05 level) and suggest further investigation of speech factors as related to instructor effectiveness.

REFERENCES

1. BETTS, G. L. The education of teachers evaluated through measurement of teaching ability. In National survey of the education of teachers. U. S. Off. Educ. Bull., 1933, No. 10 (5), 87-153.

2. BOLTON, F. B. Evaluating teaching effectiveness through the use of scores on achievement tests. J. educ. Res., 1945, 38, 691-696.

3. COURTIS, S. A. Standards of teaching ability. Educ. Rev., 1921, 62, 183-186.

4. GOTHAM, R. E. Personality and teaching effectiveness. J. exp. Educ., 1945, 14, 157-165.

5. GRAGG, D. B., and GORDON, MARY AGNES. Validity of the Airman Classification Battery AC-1. San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, December 1950. (Research Bulletin 50-3.)

6. GUTHRIE, E. R. The evaluation of teaching. TA&D Informational Bull., 4 (No. 3), 1953, 199-206.

7. HIGHLAND, R. W., and BERKSHIRE, J. R. A methodological study of forced-choice performance rating. San Antonio, Tex.: Human Resources Research Center, Lackland Air Force Base, May 1951. (Research Bulletin 51-9.)

8. JAYME, C. D. A study of the relationship between teaching procedures and educational outcomes. J. exp. Educ., 1945, 14, 100-134.

9. JONES, R. DeV. The prediction of teaching efficiency from objective measures. J. exp. Educ., 1946, 15, 85-99.

10. LaDUKE, C. V. The measurement of teaching ability. Study number three. J. exp. Educ., 1945, 14, 75-100.

11. LINS, L. J. The prediction of teaching efficiency. J. exp. Educ. 1946, 15, 2-60.

12. REMMERS, H. H., MARTIN, F. D., and ELLIOTT, D. N. Are students' ratings of instructors related to their grades? In H. H. Remmers (Ed.) Student achievement and instructor evaluation in chemistry. Purdue Univ. Stud. higher Educ., 1949, No. 66, 17-26.

13. RIESCH, K. P. A study of some factors in pupil growth. J. exp. Educ., 1949, 18, 31-55.

14. ROLFE, J. F. The measurement of teaching ability. Study number two. J. exp. Educ., 1945, 14, 52-74.

15. ROSTKER, L. E. The measurement of teaching ability. Study number one. J. exp. Educ., 1945, 14, 6-51.

16. SEYFERT, W. C. and TYNDAL, B. S. An evaluation of differences in teaching ability. J. educ. Res., 1934, 28, 10-15.

17. THOMSON, G. H. The factorial analysis of human ability. New York: Houghton Mifflin, 1948.

18. VON HADEN, H. I. An evaluation of certain types of personal data employed in the prediction of teaching efficiency. J. exp. Educ., 1946, 15, 61-84.

APPENDIX

The estimation of the contribution made by an instructor to the achievement test scores of his students becomes more accurate as the other factors affecting posttest scores are better controlled. Controls used in this study may be roughly grouped into three categories: experimental controls, sampling controls, and statistical adjustment. Experimental controls have been described earlier and will not be discussed here. (See the section Design of the Research.) Control by sampling in relation to student variables may be achieved by any of the following procedures: (1) the students of one instructor are matched one by one with the students of another instructor, the error being inversely proportional to the degree of relevance of the matching variable; (2) students taught by each instructor are a random sample of all students and a large number of students are taught by each instructor, error variance being inversely proportional to the number of students; or (3) each instructor teaches a random sample of classes with error variance inversely proportional to the number of classes. Of these possibilities, the only one that was feasible in this study was the third. Students in different classes could not be matched because of administrative difficulties no could they be logically assumed to be a random sample of all students. As a check, a rough statistical comparison of all experimental classes showed that the differences between classes were greater than could be accounted for by chance. However, each instructor taught two classes which were presumed to be a random sample of all classes: Inspection of class means showed neither a steady nor cyclical change over time; the two classes taught were spaced about a month apart; the correlation between class pretest mean scores for the first and second class to be taught be the same instructor did not differ from zero more than would be expected by chance. Thus, the two classes taught by an instructor were presumed to be a random sample of all classes.

The third type of control used in this study was that of statistical adjustment. In this procedure a weighted estimate of the initial ability of the student as determined by a composite of various available scores was subtracted from the posttest score. Since there was available a variety of scores from which initial student ability and aptitude were estimated, since the predictors did not account for all factors other than instructor differences and since the units of measurement for predictor scores were not necessarily comparable to that of the posttest, it was decided to determine the weights to be assigned to the initial student score by a multiple regression equation. However, two considerations prevented direct computation of the regression equation. One was that posttest scores were partially determined by the differences in instructor effectiveness. In fitting a regression equation to these scores, chance variability in the predictor scores, not due to the instructor, would be fitted to variability in the posttest scores that was due to instructor difference thus suppressing the factor to be measured. A second consideration was that although variance was greater between classes than within classes for both previous phase grades and posttests (as would be expected since members of a class had had common experiences), correlations between

phase grades and posttests were higher when computed between students than when computed over class means. This could be explained by the fact that each of these measures was obtained for all members of a class under nearly identical situations. Since situational factors vary more between classes than within classes these factors attenuate the between class correlation. Instructor scores are based on total class performance; therefore an estimate of correlation based on within class variances and covariance would be an overestimate which when applied to interclass difference would overcorrect for initial differences in ability.

In light of these considerations it was decided to compute the regression equation using between class but within instructor variances and covariances. Since each instructor taught two classes the number of degrees of freedom available for computing these values was equal to the number of instructors.

If the mean scores for predictor A are represented by A_1 for the first class and A_2 for the second class, the correlation between A and B, any other variable, can be estimated for within instructor between class variances and covariance as follows:

$$r_{AB} = \frac{\Sigma x_{A_1} x_{B_1} + \Sigma x_{A_2} x_{B_2} - \Sigma x_{A_1} x_{B_2} - \Sigma x_{A_2} x_{B_1}}{\sqrt{\Sigma x^2_{A_1} + \Sigma x^2_{A_2} - 2\Sigma x_{A_1} x_{A_2}} \sqrt{\Sigma x^2_{B_1} + \Sigma x^2_{B_2} - 2\Sigma x_{B_1} x_{B_2}}}$$

This procedure was followed for each pairing of six variables, namely, the three previous phase grades, the written pretest, and the written and performance posttests. The obtained correlations are shown in Table 5. From this table, two regression equations were computed: one to predict the written posttest and one to predict the performance posttest. From these equations predicted posttest scores were computed for each of the 212 classes. The differences between the predicted and actual posttest scores for the two classes taught by an instructor were summed and this sum was used as the instructor effectiveness score. The procedure used to determine the correlations upon which the multiple was based capitualizes on chance relationships to increase the reliability obtained. However, as explained earlier, other procedures would result in distortion of the desired score.

Figure 1 is a schematic diagram which shows the breakdown of posttest variance and suggests the procedures used to isolate estimates of instructor effectiveness. No attempt has been made to construct the diagram in proportion to the actual findings but rather to include components essential to the planning of the analysis.

In Figure 1 Block A shows posttest variance separated into common variance (posttest variance shared with the student predictor variables),

specific reliable variance and error variance. Block B represents a breakdown of posttests variance by sampling units namely: within class, between classes but within instructors, and between instructors. Block C represents a combined breakdown obtained by superimposing A upon B. In this diagram we can also place a component section representing the situational bias in the within class situation. Also in Block C the posttest variance that is

Table 5

Intercorrelations of Student Variables and Regression Weights

	1	2	3	4
1. Fundamentals, phase grade				
2. Structures, phase grade	.478			
3. Electrical, phase grade	.366	.342		
4. Written pretest	.492	.558	.315	
5. Written posttest	.431	.507	.349	.540
6. Performance posttest	.459	.343	.291	.415
B weights for written posttest	.388	.136	.229	.218
B weights for performance posttest	.132	.264	.244	.109

actually attributed to instructor differences as shown. It includes all posttest variance indistinguishable from the true instructor effects by the procedures of the analysis. As would be expected, it is determined from between instructor variance. As shown in Block D it includes predictor error, true specific posttest variance, posttest error, as well as variance due to true instructor effect. The first three, however, are presumably independent of the instructor effect and uncorrelated between classes. Thus, an increase in the number of classes taught by each instructor would lead to a greater saturation of the estimated instructor effectiveness with variance due to true instructor effect.

APPENDIX (Cont.)

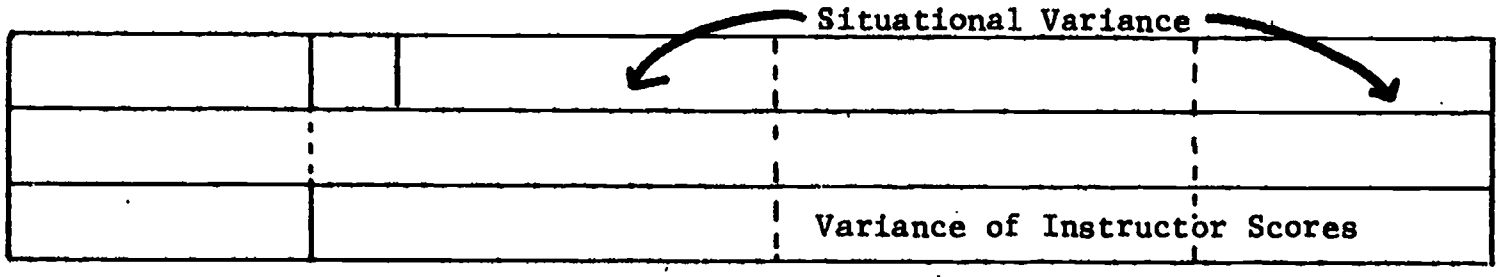
A. Posttest Factors

Predictable Factors	Predictor Error	Specific and Instructor Effect	Posttest Errors

B. Posttest Variance by Sampling Unit

Within Classes
Between Classes Same Instructors
Between Instructors

C. Posttest Factors by Sampling Unit



D. Variance of Instructor Scores

Predictor Error	True Instructor Effect	True Specific Posttest Variance	Posttest Error

Fig. Components of posttest variance. ("C" is "A" superimposed on "B".