ANALYSIS OF VARIABILITY OF GRADERS AND STUDENTS

by

WILLIS NEWTON MCKEEL, JR.

A THESIS

submitted to

OREGON STATE COLLEGE

in partial fulfillment of
the requirements for the
degree of

MASTER OF SCIENCE

June 1951

APPROVED:

[redacted]

Professor of Educational Psychology

In Charge of Major

[redacted]

Dean of Education and
Chairman of School Graduate Committee

[redacted]

Dean of Graduate School

Date thesis is presented May 12, 1951

Typed by Helen Pfeifle

ACKNOWLEDGEMENTS

Grateful appreciation for his aid in carrying out this research and his helpful suggestions in bringing this paper to presentable form is hereby given to Dr. H. R. Laslett of the Oregon State College School of Education. Helpful assistance in grading these papers was rendered by Dean F. R. Zeran, Dr. Laslett, Dr. Frank Parks, Professor W. R. Crooks, and Professor C. L. Hagen.

Many thanks are extended to Dr. J. C. R. Li of the Oregon State College Mathematics Department for his expert help in setting up the proper statistical design for analysis in this type of experiment and for aiding in the clarification of the explanation of the method which the author hopes has been achieved.

TABLE OF CONTENTS

INDEX OF TABLES

ANALYSIS OF VARIABILITY OF GRADERS AND STUDENTS

CHAPTER I

INTRODUCTION

The fairness or appropriateness of classroom grades in representing student knowledge and accomplishment is one of those things quite often discussed in both student and staff circles. Many things are said, but comparatively little is offered in the way of something better to take the place of marks or grades or to make their use more efficient. It has often been said, and occasionally proven, that the results which graders show when grading the same papers differ to a reprehensible extent. What is needed (27, p.24) in educational measurement is not the utterance by onlookers of criticisms and suggestions with which those people actually at work with measurements are as familiar as they are with their own names, but expert assistance in overcoming the weakness. This experiment was made to try to help eliminate some of the shortcomings resulting from the discrepancies of graders in judgment and in consistency from one paper to another.

In preparing this thesis, the writer first secured a random sample of twenty answers to two ordinary essay questions, one general and the other specific. These were graded by five graders, each grading an original typewritten copy. The students were in a class in Educational Psychology in the autumn quarter of 1950-1951.

The technique used for this experiment involved the finding of the sample variances or unbiased estimates of the population variances from which the samples were drawn. These unbiased estimates of variance were found for all the possible sources of variation, and the F-tests, or ratios of the appropriate variance estimates (18, p.196-200) were made to test the hypotheses resulting from the purposes stated below.

This experiment was designed to demonstrate that these five grader-sample means did or did not differ to a statistically significant degree in the marking of this set of twenty papers and how they should be grouped. A second purpose was that of showing whether the student-sample means could be called significantly different by the written information and how they are grouped, or whether they all should be considered in the same group on the basis of this set of answers. This was done by the accepted method of analysis of variance (4, p.47-49; 8, Chapter 10; 9, Chapters 10-11; 10, Chapters 7-8; 13, Chapters 10-11; 17, Chapter 5; 19, Chapters 13-14; 20, p.472-476; 26, Chapters 10-11).

This application of the analysis of variance to demonstrate the acceptance or rejection of the hypotheses concerning these points serves as a good illustration of one of the uses of the inferential type of statistics available today. By this method, such issues as those above can be determined with a known probability of error based on the significance level chosen for the experiment. Any method having a known probability of error is

better than a haphazard estimate such as ordinary grading seems to be today and has been since time immemorial, and this use of the analysis of variance is suggested as one possible method of placing grading on a more concrete and scientific basis.

This thesis is presented for the perusal of others who are in search of more accurate methods of evaluating the results and the meaning of grading. This method is believed to be feasible enough to be worthy of consideration. Many advantages are offered by the minimization of work when proper and applicable inferential statistical methods are used, as much time and money can be saved by correct experimental design. The mistake is often made of attempting to use descriptive statistics, which are really based on samples, as a basis for inferences about other samples. This can only lead to errors and confusion because descriptive statistics are those dealing with a known population and its parameters, such as the mean and variance, while inferential statistics deal with samples and the estimates of the population parameters which are derived from them. To clarify these statements, the two main functions of statistics should be stated. The descriptive function is that of finding population parameters and compiling data into tables, charts, and similar systems in order to simplify the presentation and understanding of such data. In this case, the population is merely being described. The inferential function of statistics involves the making of inferences about the characteristics of an unobtainable population on the basis of the "statistics" or characteristics of a

feasibly obtainable sample or set of samples. Experimental or inferential statistics deals with the latter of sample "statistics." In this case, the statistic is to the sample as the parameter is to a population. There are two types of inferential statistics, one being that of estimation of the probable range of population parameters and the other in the testing of hypotheses. This experiment deals with the latter type of statistical inference. Inferential and descriptive statistics are types developed for different uses, and indiscriminate use of the one in the place of the other is like the mixing of gasoline and water and bound to be confusing.

In the following chapters, a background is offered for the suggested uses of the inferential type of statistics in grading and student analysis. The use of such methods is dependent on several factors, a few of which involve the availability of equipment, trained personnel, and acceptability in local situations. This method, like anything else, can become more generally accepted only when it is more generally understood, and it is believed that this thesis will help to further such understanding. A discussion of some of the ramifications of the problem of grading is undertaken in the following chapter to emphasize that teachers need to become more familiar with available scientific tools and methods if they are to do the type of job that the teaching profession should demand.

CHAPTER II

SOME OF THE PROBLEMS OF VARIABILITY IN GRADING

The experiment undertaken as a foundation for this thesis called for the grading of two essay questions. No specific instructions were given to the five graders who cooperated except that they were to grade each of the questions on a fifty-point basis and state their reasons for grading as they did. A partial list of the most prevalently stated reasons follows, in the order of the frequency in which they were found from the most to the least often mentioned.

1. Originality, techniques, and ideas presented.
2. Statements vague, superficial, useless, short-sighted, or too general.
3. Poor spelling.
4. Understanding, diagnosis, depth of thinking, knowledge of what information is needed and how it might be obtained.
5. Organization, language mechanics and wording.
6. Unrealistic, lazy approach, or too little said.
7. Insufficient answers, and neglect of important factors such as family, home, school, background, health, abilities, and so on.
8. Jumping to conclusions.
9. The use and understanding of tests.

10. Poor or defeatist attitude, and the avoiding of responsibility.

11. Incorrect use of underscoring.

These were the main apparent sources of the differences in the grades given by the five graders to the answers in this experiment. Many other sources of variability were probably present also, but were not stated by the graders on the papers. This is not unique, but it is the expected result of individual differences, one phase of which might be the extent of the thoroughness that the grades represent. As these graders did not know the students who wrote the papers, this experiment is similar to cases in which instructors have such large classes that they seldom are able to become acquainted with most of the students in their classes. The only means by which teachers can judge student achievement and knowledge in many instances is by the material represented on paper. Students do not ordinarily exhibit all their knowledge pertaining to a particular subject or specific question when they are under test tension. One reason for this might be that their organization of material (if they have any) can be easily disrupted to a certain extent under such circumstances. Many students have had the experience of remembering something that was asked on a test immediately upon leaving the testing situation or when they get back to their room or the place where they study before they hear the answer or see any pertinent notes or references. Psychologically, this is not uncommon. Things are remembered best under the circumstances and in the way that they were learned. The

studying situation is usually one of relaxed atmosphere and surroundings. Most students have a desk or a table at which they study. In most instances, examinations are given under conditions of tension. Small desk chairs or boards which certainly are not conducive to relaxation are quite common classroom equipment and so on. These things constitute situations and surroundings that are quite at odds with the conditions under which most learning presumably has been done, and it is quite understandable that many students feel that such tests and examinations are not fair trials or examples of their ability (without even realizing that this is a possible psychological reason why they feel as they do). If test situations must continue, and it appears they must, it seems wise to use evaluation methods that take into account the variability of results arising from these and other sources. A method suggesting elimination of one source of variation when grading is done by several graders is noted below, but other sources have to be considered too, or their effect should be at least minimized as much as possible by proper analysis methods such as the one used in the experiment presented in this thesis.

An experiment attempting to eliminate grader-variability by the use of scoring rules was successful in that the variation among the readers was held to be minimized, or at least reduced. Some of the possible causes of differences were brought out as

> disagreement as to what is being measured (content, organization, English, neatness, etc.); disagreement as to the combination of the various elements to get the total and the weighting of the questions;

> differences in standards of grading and the subjectivity of the materials being marked (24, p.20).

No mention is made in this experiment of the further problem which is considerably more basic, that is, are the student differences "agreed upon" by the graders valid and discriminatory or are they just estimates with no known probability of error? Such questions will need to be answered if grading is to be placed on a better scientific basis and really become more worthy of the esteem it demands unconditionally at the present time. A few other statements concerning the variability of teachers' marks are:

> The Starch and Eliott investigations in the unreliability of teachers' marks showed that the same final examinations in English would have marks assigned from fifty to ninety-eight by different teachers. A complete measurement of, say, a composition might include the exact definition of its spelling, its usage of words or word forms, its wit, its good sense, etc.; and each of these might again be subdivided into a score or more of component elements so that every measurement represents a highly partial and abstract treatment of the product (6, p.86-87).

> It has been repeatedly demonstrated that letter symbol grades as assigned in the typical secondary school are unreliable. Teachers tend to place widely varying judgments upon the same piece of work. They vary in their evaluation too if they attempt to evaluate a single unit of work at different times. These facts have been brought out again and again in elementary and secondary schools. The situation is not much different at the college level, but higher education has not made many investigations of the reliability of grades (22, p.21).

> Early studies showed that marks assigned by one teacher did not agree closely with marks assigned by another to the same series of examinations. The results of such studies were often incorrectly interpreted as meaning that marks were 'unreliable.'

> In most cases, however, different teachers were appraising different outcomes and agreement could not be expected. If the expected outcomes of teaching are accurately defined, then teachers can in most instances agree rather closely on the marks to be assigned to examination papers, and this source of the 'unreliability' of them can be largely eliminated (28, p.369).

USES AND PURPOSES OF GRADING

Grading is considered as one of the necessary factors in the present educational system and will probably remain so for some time. A statement concerning the importance and purposes of recording grades for students mentioned the use of grades as a basis for reaching understanding of individuals so that effective guidance can be given, making information transferable for later guidance, reporting to the home, and giving evidence of readiness for succeeding experiences. The purposes of recording for teachers are to stimulate their consideration of and decisions concerning their objectives, the relative importance of their aims, the results of their work and the progress of their students toward these objectives, and to help the teachers gain a wider vision and more constructive influence (25, p.465-466). Since grades are and evidently will continue to be used as the assumedly best possible index of the things mentioned here, it is the responsibility of the whole educational system and everyone who uses grades to make certain that they are on a basis that is as scientifically accurate and correct as can be discovered and used.

Other discussions along this line deal with the basis and

results of grading in a similar fashion:

> The main purpose of recording marks in a central office in any college is to provide a record of the student's accomplishment. When the record shows that the student has achieved certain objectives, then his accomplishments are recognized by the granting of a degree. Centrally recorded marks usually form part or all of the criterion used in the award of a degree at the graduate level, and degrees will be correctly granted in so far as marks represent a valid measure of success. While teachers may disagree on the outcomes that should be measured in appraising the products of a student's labor, there is also another important source of variation in the meaning of marks. Although many instructors assign them on the basis of the extent to which the objectives of the course are achieved, there are some who mark on the basis of factors which have little correlation with final achievement. If such is the case, then a central record of marks will inevitably have rather limited value as a criterion for awarding degrees, and this record may be largely unpredictable because its composition is unpredictable (28, p.369-370).

A survey made at the University of Washington (3, p.122-123) found grades in lower division courses mostly determined by examination alone, while upper division course grades were based on a combination of equal amounts of written work and examinations. Final grades frequently contained a weighted examination score. Some other discussions along this line show additional objectives or reasons for measuring and giving grades:

> From birth to death almost every aspect of our daily lives is touched by measurement in its numerous forms. These common experiences are characteristic of the emphasis placed on measurement in the modern world. In fact, if all our various measuring devices were suddenly destroyed, contemporary civilization would collapse like a house of cards (28, p.1).

> An objective of a marking system to which most persons would subscribe is to recognize both achievement and

> attitude in the process of appraising pupil progress. The average school system would do well to provide a uniform report form which would list the most significant factors of appraisal applicable to most of the school subjects, such as (1) achievement on tests, (2) quality of recitation, (3) quality of completed assignments, (4) promptness in completing work, (5) persistance for mastery, (6) self-reliance in work, (7) application during study, and (8) attention to class activities. Factors which are not applicable to certain courses would not have to be checked (1, p.19).
>
> The task of determining the 'rightness' of pupils and education is a task for the process of evaluation. Pupils' attributes and opportunities must be ascertained. Pupil behavior must be evaluated at all stages of the interaction between pupil and education to determine the fitness of one for the other (21, p.1).

"A tabulated analysis, where educational test scores are compared with achievement grades can give valuable information for counseling students and indicating weaknesses in faculty grading, student attitudes and administration policies" (16, p.322). This demonstrates a method somewhat improved over ordinary systems in that recognition is given to the fact that grades vary, and analysis is made with that recognition taken into consideration.

ASSUMPTIONS AND FACTORS MEASURED

> We assume that tests as given by different teachers and at different times have called forth equal or approximately equal effort; we assume a sufficient sensory and motor equipment; we assume that the sampling as drawn out by the test questions constitutes a fair and sufficient sampling of ability. If we cannot avoid making these assumptions, we can at least pause long enough to steep our souls in the conviction that they are present and obscure our findings. We have assumed test scores may with entire propriety be added, subtracted, multiplied,

> and divided. They seldom can. Test devisers have apparently been quite successful in obtaining test-score units which are substantially equal and can be added and subtracted, but they have failed quite signally in determining reasonable zero points, so that the product or quotient technique rests upon shifting ground (14, p.16-17).

Teachers believe that the best indication of achievement is the student's ability to use facts and principles in new situations and to act consistently with valid conclusions, and they attempt this type of classroom instruction (22, p.21). In an opinion study made at the University of Washington, most of the instructors thought their examinations were generally well constructed, while some thought that examinations were too comprehensive for the time allowed (3, p.123). Considering the points ventured above, if application is the best indication of achievement and even expert test devisers make some errors in setting up their scales, it seems incongruent that teachers should try to devise and evaluate their own tests without expert help. It would be wise in this instance to incorporate standardized tests as much as possible and leave the final evaluation and grading to more competent personnel because most teachers are not expert test designers and they do not have the time to do this job nearly as well as experts who design and standardize tests. Factors of importance could be checked and brought to the teachers' attention with the intention of aiding their teaching. Teachers would have more time for creating an atmosphere conducive to better student achievement, development, application, and classroom instruction if they were not required to do so much evaluating.

If the objectives are known, the teachers can help the students to develop the right abilities, knowledge, attitudes, and interests necessary to gain those objectives (6, p.115). Unless such objectives are recognized, teaching facts, testing, drills, and so on cannot be of either immediate or permanent usefulness.

> "A noteworthy attempt to study the newer procedures in evaluating the results of secondary education was undertaken as a part of the Eight Year Study sponsored by the Progressive Education Association some years ago. In connection with this study, a committee worked to develop instruments of evaluation for interests and aptitudes, work habits and study skills, abilities involved in interpreting data, awareness of significant problems in our society, and abilities involved in applying facts and principles to specific real life situations. The unique feature of this research was the attempt to evaluate teaching in terms of the achievement of accepted educational purposes" (22, p.21).

One of the major problems in achievement testing is the necessity of being careful as to what is being tested. Students are achieving under many teachers, different methods of instruction, and varied curriculum requirements. It is still difficult to define achievement, as it may be considered as something immediate; as a prerequisite for later work; as a mental discipline; as a special application, knowledge or skill; or as a measure for prediction (6, p.95). Evaluation is much more useful if it is done as part of the learning situation and not solely as a measure of learning afterward. Testing should be done for the benefit of the students and not to their detriment as when one or two tests determine a course grade. Much of the test-taking tension would probably disappear if

tests were used in this manner. Achievement cannot be expressed merely in grades without ignoring the fact that evaluation has many functional uses in addition to the descriptive phases, such as aids in spotting deficiencies or pointing out mechanical difficulties and so on (6, p.124). This points out the fact that descriptive grades are used for prognosis only by taking the risk of making dire mistakes. Teachers must learn what type of behavior typifies the desired outcomes and when this behavior may be correctly interpreted as fulfilling these outcomes. Schools must accept the responsibility of such measurement and make a real attempt to do something about it (2, p.321).

> "Because of native differences in capacity for the different subjects or because of an earlier differentiation in interest and effort which has persisted with the years, we discover, perhaps for the first time in the middle or late school years, a genuine difference in relative accomplishment which is, however, more than merely that, for it is a prophecy of differences in capacity to achieve in the future along various related lines" (14, p.111).

The above quotation brings out the necessity for the recognition and understanding of individual differences. This concept is looked upon as prerequisite for good teaching and seems to be closely allied with the ideas of variability. The need for proper analysis of the variability of the results in any measurement of achievement in order to be more meaningful to the students and the teachers using that measurement seems to be an inherent part of considering individual differences. All persons differ in characteristics and abilities, and any method whereby decisions can be

made as to when differences are really present and how great a difference constitutes a real difference in the component or components under comparison will be a boon to the understanding and judgment of people.

GRADES AND THEIR BASES

Local grading and reporting systems should be: in harmony with the local philosophy of education; designed for the benefit of the students; cooperatively developed and accepted by students, parents, teachers, and administration; informative and meaningful to all concerned; in line with the objectives of the courses; reported with a frequency determined by relative value; economical and sure of reaching their destination; useful in computing final marks on a basis other than the outmoded one of competition; and continuously evaluated and modified cooperatively (1, p.16-24).

Grade and age scores are misleading as they imply that those measured meet at or near the point called the mean or average, and percentile scores avoid some of the misinterpretations but give no satisfactory picture of progress from year to year (6, p.107). A common fault made by those who use percentile scores is that they they assume that the distribution is a rectangular one. "Percentile scores are still widely used due to the force of tradition and in spite of the fact that this device has long been discredited (partly due to the faulty assumption of a rectangular distribution) in educational circles. Grades are not absolute and cannot honestly be

made to appear as if they were merely by representation as so many fractional parts short of perfection" (30, p.300).

> "In practice percentile grades are relatives based upon their own average, rather than upon 100% perfect accomplishment. One cannot fail everybody without soon being forced to change his occupation. An array of grades distributed within the acceptable range must be produced. This is so whether the teacher realizes what he is doing or not. Unfortunately, too often he does not know what is taking place, or if he does, is not frank about it. However, tacit admission of the state of affairs is seen when references are made to certain members of the faculty as 'hard markers' or 'easy markers'" (30, p.300).

A literal system of grading, using A, B, C or analogous letters, commonly gives an order of merit within the group, with no precise quantitative relationships stated or understood (15, p.488). Single marks cannot reveal teachers' estimates of the pupils' comparative accomplishments and attitudes, and much variation in importance and meaning is attached by graders to these various factors. Such a system is inadequate, unreliable, and generally not informative enough to be as completely useful as it should be for effective guidance (1, p.18). Grades are unreliable, and can too easily be wrongly used for discipline or punishment. Any representation that absolute standards prevail in grading is false and an educational fraud, and this kind of deceit does not seem conducive to a wholesome way of thinking by the superintendents, teachers, and principals who are the leaders in education (30, p.301). This might possibly be the result of permitting or requiring teachers to use a tool (grades) with which they are not truly versed in understanding and about the criteria for which they seldom reach

agreement. It seems that more of the fundamentals of mathematics such as number theory and inferential statistics should be required in their training if teachers are to use grades, as they evidently must. Figures or grades do not "lie" when properly used and when the underlying assumptions requisite for their proper use are understood and fulfilled.

> "Who causes the greater sorrow, the physician who wrongly diagnoses thirty in one hundred ailments or the school principal who wrongly judges intellect and effort and gives unsound advice as to training and vocation to some thirty in one hundred of his graduating class? The onus is great in either case and but little relieved by pointing with pride to the seventy correct diagnoses" (14, p.20).

The following chapter deals with the experimental phase of this research and is a suggested way by which student grouping methods can be made more scientific, and possibly better understood.

CHAPTER III

THE PRESENT STUDY

THE EXPERIMENT

Two essay questions of the usual type were presented to a class of seventy-five students as part of a regular examination program. These students were practically all juniors in Education taking Educational Psychology at Oregon State College. Numbers were assigned to all of the students in the original class, and a random sample of twenty of their papers was drawn by using a random number table. (A random number table is made up of a set of numbers arranged such that a random succession of numbers may be selected according to any procedure, subject to the sole restriction that the selection of a number from the set be influenced only by its location in the table (12, p.294). Such tables may be entered at any place, row, or column, and any direction may be taken in the table to obtain a set of random numbers.) These twenty papers were typed in five copies in exact reproductions of the original papers, even to duplicating all of the errors as the students had made them. Five graders of extensive experience in grading, two professors, one associate professor, and two assistant professors, who were willing to cooperate were found, and each one graded a set of the test copies on the basis of fifty points for each question, listing his corrections and reasons for marking the papers as he did. This resulted in a total of two hundred observations. The two questions were on the same subject.

The first question was one of a more general type, while the second question was more specific about the type of discussion desired. A detailed description of a problem case concerning an eleven-year-old boy named Albert was presented. The two questions asked following this description were: (1) What would you want to know and try to find out about Albert before giving advice on this case? (2) What advice would you give to: (a) Albert; (b) his parents; (c) his home-room teacher; (d) his physical education instructor; (e) his other teachers; (f) the school nurse; (g) the principal; and (h) the boy's advisor? The results of the grading of the students' answers to the two questions are shown in Table I and Table II. Table I shows the grades given by each grader for each student on the answer to the first question, and Table II shows the relative grades for the second question. The numbers in these two tables would be analogous to percentages if they were doubled and the tables or the two questions were considered separately.

For simplification, Q-1 and Q-2 are the questions, G-1 through G-5 are the graders, and S-1 through S-20 are the students. These letters will be used in the following discussion.

EXPLANATION OF TABLE I

This table shows that G-3 graded S-14's answer 37, G-4 graded S-17's answer 47, and so on. The total grade that each student received on his answer to Question One is found in the right hand column labeled "Totals", e.g., S-6 made a grade of 84 while S-12 was

graded 171. Similarly, the total of all the grades given by each grader on Question One appears in the bottom row labeled "Totals", e.g., G-1 gave 560 points while G-4 gave 808 points. The total number of points given for Question One on all of the papers by all of the graders appears in the lower right hand corner, and is 2971.

TABLE I

GRADES FOR QUESTION ONE

Graders / Students	G-1	G-2	G-3	G-4	G-5	Totals
S-1	20	20	30	30	25	125
S-2	30	20	11	40	10	111
S-3	30	25	26	35	25	141
S-4	10	10	30	30	5	85
S-5	40	45	26	45	20	176
S-6	20	20	4	30	10	84
S-7	30	10	26	40	30	136
S-8	30	35	37	45	35	182
S-9	30	40	15	48	35	168
S-10	40	30	30	42	40	182
S-11	30	30	19	45	40	164
S-12	20	45	26	50	30	171
S-13	30	40	15	40	35	160
S-14	30	45	37	40	40	192
S-15	10	15	7	30	10	72
S-16	30	30	15	40	20	135
S-17	30	40	33	47	40	190
S-18	50	50	37	46	40	223
S-19	20	20	7	40	20	107
S-20	30	30	22	45	40	167
Totals	560	600	453	808	550	2971

EXPLANATION OF TABLE II

In this table, the grades given by the graders to all of the students for their answers to Question Two are shown. The grade given by G-4 to the answer made by S-19 was 45, the grade G-2 gave S-7's answer was 20, and so on. The total grade given to each student on the answer he made to this question appears in the right hand column labeled "Totals", e.g., S-4 received a total grade of 162. The total grade given by each grader appears in the bottom row labeled "Totals", e.g., G-5 gave a total of 475 points. The total number of points given for Question Two on all of the papers by all of the graders appears in the lower right hand corner and is 3014.

TABLE II

GRADES FOR QUESTION TWO

Graders / Students	G-1	G-2	G-3	G-4	G-5	Totals
S-1	50	40	22	40	25	177
S-2	20	35	18	40	25	138
S-3	40	25	21	45	20	151
S-4	40	30	27	45	20	162
S-5	40	40	17	45	25	167
S-6	20	35	18	40	20	133
S-7	40	20	14	45	20	139
S-8	30	40	20	45	20	155
S-9	30	40	26	47	25	168
S-10	40	35	21	40	40	176
S-11	40	35	21	47	25	168
S-12	10	30	5	25	15	85
S-13	30	35	17	40	20	142
S-14	50	35	25	42	35	187
S-15	30	10	14	30	10	94
S-16	40	40	15	45	20	160
S-17	20	25	26	49	35	155
S-18	50	45	30	47	35	207
S-19	20	25	22	45	20	132
S-20	20	20	15	43	20	118
Totals	660	640	394	845	475	3014

STATISTICAL ANALYSIS AND INTERPRETATION

Table I and Table II, shown just previously, are very similar. The only difference between the two tables is that the grades arranged in them were obtained from answers to two different questions. Each grader graded all twenty of the papers, or each paper was graded by all five of the graders. The experiment was set up this way in order that the type of design called a split-plot experiment with sub-unit treatments in strips would be applicable. The students and the graders are the two variables or sub-unit treatments under study here, while the two questions are the two replications (repetitions of the experiment) that are the main parts of the split-plot design. The problems posed are: (a) to determine whether or not any differences found among the five graders; and (b) whether or not any differences found among the students are statistically significant. In the event that the decision reached is that the students or the graders do differ significantly, then the problem arises as to how the students and graders are to be grouped. For solution, the analysis of variance with two-way classification and single observation was used along with a method of comparing individual means in the analysis of variance.

The main hypotheses tested by this analysis are that: (a) there is no difference between the mean grades of the graders and (b) there is no difference between the mean grades of the students. Two types of error are possible in the testing of all hypotheses. A

"Type One" error is made when a correct hypothesis is rejected and a "Type Two" error is made when an incorrect hypothesis is accepted. The significance level of a given test is the probability of making a Type One error.

Separate estimates of error were obtained for S, G, and S x G (4, p.234). These are designated as error (a), error (b), and error (c) in Table IV, and they are the differences or variations in the tabulated grades due to the interaction of Q x S, Q x G, and Q x S x G respectively. Their appropriate degrees of freedom are 19, 4, and 76 respectively (4, p.232). Interaction means that some of the students will be marked higher by one of the graders and lower by another. The number of degrees of freedom is the rank of the quadratic form or matrix on which the analysis of variance is based. It can be found in this instance by noting the number of observations in the sample and subtracting one from that number, e.g., there were five graders and therefore the number of degrees of freedom of the grader sum of squares is four, and so on. The number of degrees of freedom of the interaction sum of squares is the product of the degrees of freedom of the parts of the interaction, e.g., for the S x G interaction the number of degrees of freedom is nineteen times four or seventy-six. It is necessary to have the degrees of freedom of the numerator and denominator in making the F-test of significance so that the critical region can be found. If the value of a statistic falls within a certain range of the distribution being used, the hypothesis being tested is rejected.

The term applied to this range or these ranges is the "critical region". The term "significant", as used in statistics, merely means that the statistic falls in the critical region for the given test (11, p.13-14).

It is assumed that the original observations are random samples drawn from normal distributions with equal variances. It follows that the sample variance ratios formed have the F-distribution as a sampling distribution (5, p.97). The test of significance used in the analysis of variance is Snedecor's F-test. "F" is defined as the ratio of the sample estimates of two variances. The distribution of "F" is determined by the degrees of freedom of both numerator and denominator. "F-values" for various significance levels are tabulated (26, Table 10.7). It is also assumed that the treatment effects or effects from the source of variance and the error effects are additive.

The level of significance or magnitude of the "Type One" error used for this experiment was 5%. The critical regions for rejection of the hypotheses are larger than 6.39 for G, and 2.16 for S because these are the 5% points in the distributions of F-values for the degrees of freedom of 19 and 19, and 4 and 4. The computations (Table IV, p.29) show F-values of 16.66, and 2.73 with 4 and 4 and 19 and 19 degrees of freedom respectively. This leads to the conclusions that the hypotheses are rejected on the basis of these results being significant at this level. This means that the difference among student means is too great for all of the students to be considered

as belonging to the same group or population, and the difference among the grader means is too great for them to be considered as belonging to one population. In order to group the students or the graders, further tests are necessary after the F-test. As the graders graded the papers under the same set of directions and the students took the test under the same conditions; and, as the numbers of observations are equal for all the student or grader samples, a suggested practical method of grouping the respective means becomes applicable. First, apply the gap test to break up the means into one or more broad groups. Second, apply the straggler test within these groups to further break off stragglers within groups. Third, apply the F-test to these new sub-groups (if there are three or more means in the group) to detect excess variability (29, p.102). The gap test referred to above is the least significant difference (L.S.D.) This is defined for the case of equal numbers of observations in the samples as $L.S.D. = t_a\sqrt{2s^2/N}$. The level of significance for which the t-value with the same degree of freedom as s^2 is to be looked up (5, Table 5, p.12) is decided upon as "a". The level of significance used in this experiment was 5%. The error mean square for the source of variation being tested is s^2, while N is the number of observations that go to make up each of the means. The straggler test is the u-test:

$$u = \frac{\dfrac{|\bar{x}_1 - \bar{\bar{x}}|}{\sqrt{s^2/N}} - \frac{6}{5}\log_{10}k}{3(1/4 + 1/n)} \quad \text{for k greater than 3,}$$

where $\bar{x}_1$ is the extreme mean of the group being used,

$\bar{\bar{x}}$ is the general mean of that group,

s^2 is the error mean square of that group in the analysis of variance,

N is the number of observations that go to make up each mean,

k is the number of means, and

n is the number of degrees of freedom of the error sum of squares in the analysis of variance.

For the case where k equals three, replace $(6/5) \log_{10} k$ by 1/2. The third step, or application of the F-test amounts to making a new analysis of variance of the sub-group means, using the original error mean square to test the hypothesis that all of the means within the groups belong to the same population. The hypothesis tested by the L.S.D. is that the two sample means being tested are equal. The hypothesis tested by the u-test is that the extreme mean of the group belongs to the group being tested. If the u-value is found to be greater than 1.645 it is significant at the 5% point and the hypothesis is rejected.

The L.S.D. for the graders was found to be equal to 7.24. The grader L.S.D. separated the G-4 mean from the rest of the group. The u-value of the G-3 mean in the group of the four remaining means was equal to 1.80 and, therefore, the G-3 mean was separated from this group. The u-value for the G-5 mean in the remaining group of three means was 0.934 and therefore not significant. The F-value for the group of the three grader means Graders (B) was found to be 2.60 and as this was below the 5% point which was 6.94 for this test, the

hypothesis that these three grader-means are samples drawn from the same population is accepted. The interpretation of this is that G-4 graded higher than the rest of the graders, G-3 graded lower than the other three graders in the lower-grading group, and the other three graders grade the same.

The L.S.D. for the students was found to be equal to 10.93, and this did not separate off any of the twenty student-means. The u-value for the S-15 mean gave a value of 2.25 and, therefore, the S-15 mean was separated off. The u-value for the S-18 mean in the remaining group of nineteen means was found to be 2.56, so the S-18 mean was separated off. The u-value for the S-6 mean in the remaining eighteen means was 0.795 and therefore it was not separated off. The F-test for the group of eighteen remaining student-means Students (C) gave a value of 1.55 and, as the 5% point was 2.21 in this case, the hypothesis that all eighteen of the remaining means were drawn from the same population was accepted. The interpretation of this is that S-15 received a lower grade than the rest of the students, S-18 received a higher grade than the remaining eighteen students, who all received the same grade.

Additional information obtainable shows the F-value for Q as .07 with 1 and 19 degrees of freedom, and for S x G as 1.81 with 76 and 76 degrees of freedom. At this level of significance, the critical region for Q and S x G would be F-values greater than 4.38 and 1.46 respectively. This indicates that there is no significant difference between the questions, but that there is significant

interaction between S and G.

Table III is not ordinarily given when reporting an analysis, as it shows the preliminary calculations only. It is believed, however, that its inclusion here will help to clarify the method used in the analysis of variance.

EXPLANATION OF TABLE III

In this table, the respective columns contain the following information:

(1) A list of the sources of variation. Correction (A) refers to the square of the sum of all the observations divided by the number of observations. It is included because the method of computation to find the sum of the squares of the differences about the mean in all the samples uses the mathematical equivalent of subtracting this correction from the sum of the squares of the totals found in each source of variation, divided by the number of observations in each total. The questions, graders (A), and students (A) are the main sources of variation. The SG, SQ, and QG subclasses are the Tables of Totals V, VI, and VII respectively. Correction (B) refers to the square of the sum of all the observations that go to make up the sub-group of grader-means G-1, G-2, and G-5 divided by the number of observations. The graders (B) are G-1, G-2, and G-5. Correction (C) refers to the square of the sum of all the observations that go to make up the sub-group of student-means which includes all means except S-15 and S-18. The students (C) are all of the students

excepting two, S-15 and S-18. The individual observations are also a source of variation (and usually this is the only type of variation recognized and considered in grading).

(2) The total of squares refers to the sum of the squares of the main parts that are found in the sources of variation.

(3) The number of items squared tells us the number of main parts there are in each source of variation.

(4) The number of observations per squared item is the number of primary observations that go into making up each main part in (2) and (3) above.

(5) The total of squares per observation is the sum of the squares of the main parts in the sources of variation divided by the number of observations in each part.

(6) The sum of squares is the sum of the squared differences of the main parts about the general mean.

TABLE III

ANALYSIS OF VARIANCE CALCULATIONS

Experiment: Split-Plot Experiment on Grading with Sub-Unit Treatments in Strips.

Preliminary Calculations

(1)	(2)	(3)	(4)	(5)	(6)
Source of Variation	Total of Squares	No. of items Squared	Observations per Squared Item	Total of Squares per Observation (2) ÷ (4)	Sum of Squares (5) - correction
Correction (A)	35,820,225	1	200	179,101.13	0
Questions	17,911,037	2	100	179,110.37	9.24
Graders (A)	7,526,443	5	40	188,161.08	9,059.95
Students (A)	1,861,927	20	10	186,192.70	7,091.57
QG Subclass	3,774,259	10	20	188,712.95	9,611.82
SQ Subclass	944,011	40	5	188,802.20	9,701.07
SG Subclass	399,363	100	2	199,681.50	20,580.37
Ind. Obs.	205,279	200	1	205,279.00	26,177.87
Correction (B)	12,147,225	1	120	101,210.21	0
Graders (B)	4,076,625	3	40	101,915.62	705.41
Correction (C)	29,041,321	1	180	161,340.67	0
Students (C)	1,649,471	18	10	164,947.10	3,606.43

Table IV actually demonstrates the analysis of variance and is ordinarily the only part of the computation given in reporting an analysis of variance experiment.

TABLE IV

ANALYSIS OF VARIANCE CALCULATIONS

Experiment: Split-Plot Experiment on Grading with Sub-Unit Treatments in Strips.

Analysis of Variance

Variation Due to:	Sum of Squares	Degrees of Freedom	Mean Square	F	Remarks 5%
Questions	9.24	1	9.24	.07	-
Students (C)	7,091.57	19	373.24	2.73	Significant
QxS [Error (a)]	2,600.26	19	136.86		
Graders (B)	9,059.95	4	2,264.99	16.66	Significant
QxC [Error (b)]	542.63	4	135.66		
SxG	4,428.85	76	58.27	1.81	Significant
QxSxG [Error (c)]	2,445.37	76	32.18		
Total	26,177.87	199	131.55		
Graders (B)	705.41	2	352.70	2.60	-
QxG [Error (b)]	542.63	4	135.66		
Students (C)	3606.43	17	212.14	1.55	-
QxS [Error (a)]	2600.26	19	136.86		

EXPLANATION OF TABLE IV

In this table the columns contain the following information:

(1) The sources of variation are shown separated for testing into first, second, and third order interactions.

(2) The first order and total sums of squares are taken directly from Table III, column (6). The error (a), or interaction Q x S sum of squares is found by subtracting the sum of squares of S and Q from the sum of squares of the SQ subclass. Similarly, error (b), or interaction Q x G is found by subtracting the sum of squares of Q and G from the sum of squares of the QG subclass. This is done because the total sum of squares of these subclasses and the sum of squares of each part of the subclass, other than the error, can be found. As the total is made up of the sum of squares of all of its parts, the error sum of squares is equal to the total sum of squares minus the sum of squares of the known sources of variation. Error (c) or the third order interaction Q x S x G is found by subtracting all the other sums of squares from the total sum of squares for the same reason.

(3) The degrees of freedom are found as the product of the degrees of freedom of the parts in each source of variation, and must add up to the total degrees of freedom. The number of main parts in each source of variation minus one is the number of degrees of freedom.

(4) The main square column contains the various statistics or

estimates of the various population variances.

(5) The F column indicates the ratios of the variance estimates of the source of variation being tested over the error associated with that source.

(6) The remarks column contains only the statement "significant". The common practice is to draw a line or dash if the F-value is not in the critical range or is not "significant".

Table V shows the 100 SG totals used in finding the total of squares for the SG subclass. Tables V-VIII are not ordinarily contained in statistical reports of the analysis of variance, but they are given here for the purpose of clarity.

EXPLANATION OF TABLE V

This table contains the total number of points given by each G to each S without separating the points given for each Q. The Totals column shows the total points for each S, and the Totals row shows the total number of points given by each G. The box where the two totals meet shows the total number of all of the points given.

The bottom row shows the mean grade given by each G on each Q or G-mean. The right hand column shows the mean grade received by each S on each Q.

Table VI shows the 40 SQ totals used in finding the total of squares for the SQ subclass.

TABLE V

SG TOTALS

Graders / Students	G-1	G-2	G-3	G-4	G-5	Totals	S-Means
S-1	70	60	52	70	50	302	30.2
S-2	50	55	29	80	35	249	24.9
S-3	70	50	47	80	45	292	29.2
S-4	50	40	57	75	25	247	24.7
S-5	80	85	43	90	45	343	34.3
S-6	40	55	22	70	30	217	21.7
S-7	70	30	40	85	50	275	27.5
S-8	60	75	57	90	55	337	33.7
S-9	60	80	41	95	60	336	33.6
S-10	80	65	51	82	80	358	35.8
S-11	70	65	40	92	65	332	33.2
S-12	30	75	31	75	45	256	25.6
S-13	60	75	32	80	55	302	30.2
S-14	80	80	62	82	75	379	37.9
S-15	40	25	21	60	20	166	16.6
S-16	70	70	30	85	40	295	29.5
S-17	50	65	59	96	75	345	34.5
S-18	100	95	67	93	75	430	43.0
S-19	40	45	29	85	40	239	23.9
S-20	50	50	37	88	60	285	28.5
Totals	1220	1240	847	1653	1025	5985	
G-Means	30.5	31.0	21.2	41.3	25.5		

EXPLANATION OF TABLE VI

This table contains the total number of points given on each Q to each S without separating the points given by each G. The right hand column shows the total points for each S, and the bottom row shows the total points given on each Q. The lower right hand section shows the total number of all of the points given.

TABLE VI

SQ TOTALS

Students	Question 1	Question 2	Totals
S-1	125	177	302
S-2	111	138	249
S-3	141	151	292
S-4	85	162	247
S-5	176	167	343
S-6	84	133	217
S-7	136	139	275
S-8	182	155	337
S-9	168	168	336
S-10	182	176	358
S-11	164	168	332
S-12	171	85	256
S-13	160	142	302
S-14	192	187	379
S-15	72	94	166
S-16	135	160	295
S-17	190	155	345
S-18	223	207	430
S-19	107	132	239
S-20	167	118	285
Totals	2971	3014	5985

Table VII shows the 10 QG totals used in finding the total of squares for the QG subclass.

TABLE VII

QG TOTALS

Graders	Question 1	Question 2	Totals
G-1	560	660	1220
G-2	600	640	1240
G-3	453	394	847
G-4	808	845	1653
G-5	550	475	1025
Totals	2971	3014	5985

EXPLANATION OF TABLE VII

This table contains the total number of points given on each Q by each G without separating the points given to each S. The right hand column shows the total points given by each G, and the bottom row shows the total points given on each Q. The lower right hand section shows the total of all of the points given.

It was found desirable to separate the grader-means and the student-means into their respective groups so that is done in the next section. The type of calculations appended here is usually omitted in making statistical reports, but it is believed that its inclusion will help to clarify the process.

ADDITIONAL CALCULATIONS

The analysis of variance indicate that there is excess variability among the graders and also among the students, so the means have to be more closely examined.

For the graders, the grader-means are arranged in order from the lowest to the highest:

Graders	3	5	1	2	4
Means	21.2	25.5	30.5	31.0	41.3

The least significant difference between any two grader-means is determined:

$$\text{L.S.D.} = 2.78\sqrt{2(135.66)/40} = 7.24.$$

The L.S.D. separates the G-4 mean away and the mean of the remaining group of four means is found to be $\bar{\bar{x}} = 27.5$. Next the u-value for the extreme mean (G-3) in this group was obtained:

$$u = \frac{\dfrac{|21.2 - 27.5|}{\sqrt{135.66/40}} - \dfrac{6}{5}\log 4}{3(1/4 + 1/4)} = 1.80.$$

Since this value is significant, the G-3 mean is separated away from the group. The remaining group of three means has a mean value of $\bar{\bar{x}} = 29.0$. The u-value for the test of the extreme mean (G-5) in this group was obtained:

$$u = \frac{\dfrac{|25.5 - 29.0|}{\sqrt{135.66/40}} - \dfrac{1}{2}}{3(1/4 + 1/4)} = 0.934.$$

Since this value was not significant, this group of three means remains intact. The analysis of variance for this group of grader-means (GRADERS B) was appended to Table III and Table IV for convenience. The F-test for this group gave a value of $F = 2.60$ with 2 and 4 degrees of freedom. This was not significant and therefore this group of means does not have excess variation and so it remains intact. These calculations show that the grader-means are grouped as follows:

$$G\text{-}3 < G\text{-}5 = G\text{-}1 = G\text{-}2 < G\text{-}4.$$

For the students, the student-means are arranged in order from the lowest to the highest:

Students	15	6	19	4	2	12	7	20	3	16
Means	16.6	21.7	23.9	24.7	24.9	25.6	27.5	28.5	29.2	29.5

Students	1	13	11	9	8	5	17	10	14	18
Means	30.2	30.2	33.2	33.6	33.7	34.3	34.5	35.8	37.9	43.0

The least significant difference between any two grader-means is determined:

$$\text{L.S.D.} = 2.09\sqrt{2(136.86)/10} = 10.93.$$

The L.S.D. does not separate the means of the group. The mean of the group of twenty student-means is $\bar{\bar{x}} = 29.92$. The u-value for the extreme mean (S-15) in this group was obtained:

$$u = \frac{\dfrac{|16.6 - 29.92|}{\sqrt{136.86/10}} - \dfrac{6}{5}\log 20}{3(1/4 + 1/19)} = 2.25.$$

Since this value is significant, the S-15 mean is separated away from the group. The remaining group of nineteen means has a mean value of $\bar{\bar{x}} = 30.626$. The u-value for the test of the extreme mean (S-18) in this group was obtained:

$$u = \frac{\frac{|43.0 - 30.626|}{\sqrt{136.86/10}} - \frac{6}{5}\log 19}{3(1/4 + 1/19)} = 2.56.$$

Since this value is significant, the S-18 mean is separated away from this group. The remaining group of eighteen means has a mean value of $\bar{\bar{x}} = 29.94$. The u-value for the test of the extreme mean (S-6) in this group was obtained:

$$u = \frac{\frac{|21.7 - 29.94|}{\sqrt{136.86/10}} - \frac{6}{5}\log 18}{3(1/4 + 1/19)} = 2.56.$$

Since this value is not significant this group of eighteen means remains intact. The analysis of variance for this group of student means (STUDENTS C) was also appended to Table III and Table IV for convenience. The F-test for this group gave a value of $F = 1.55$ with 17 and 19 degrees of freedom. This F-value is not significant, and therefore this group of means does not have excess variation, and so it remains intact. These calculations show that the student-means are grouped as follows:

$$S\text{-}15 < S\text{-}6 = \ldots = S\text{-}14 < S\text{-}18.$$

STATISTICAL SUMMARY

This experiment was set up as a split-plot because the only

differences between the replications were due to two types of essay questions. The sub-unit treatments were considered "strips" because each of the five instructors graded the same papers.

The method of the analysis of variance was used. The conclusions arrived at are that the grader-means and the student-means differ respectively to an extent that is statistically significant at the 5% significance level. Further calculation showed that the G-3 mean is lower and the G-4 mean is higher than the rest of the grader-means which are in a group that does not have excess variation. It was shown, also, that the S-15 is lower and the S-18 mean is higher than the rest of the student-means which are in a group that does not have excess variation. This means that there are three groups of graders and three groups of students and no more for this particular experiment. This means that if we were giving grades to this group of students on the basis of their written answers to these questions, only a three-grade system would be applicable. This shows that one grader graded much higher than the others, and one graded somewhat lower. This fulfills the purpose of the author as the result that this research study was intended to show was that (a) the grader-means or the student-means did or did not differ to a statistically significant degree and (b) the ways in which the student-means and the grader-means, respectively, should be grouped. Additional information derived as a consequence of the use of this method is that (a) there is no significant difference in this case between the specific and the general types of essay

questions used here, and (b) that interaction existed between students and graders in this experiment.

This experimental method could be used as one way of arriving at more scientific conclusions, thus being better than methods not having such a basis. A scientific method is far better than a guess - in the same way that almost any measurement is better than none. Similar analysis could be carried out on other sets of grades of classes as marked on different tests by one grader at different times. This would show the way in which the students should be grouped and the number of grades which are statistically applicable in specific instances. Expert help in setting up such an experiment is advised, as the comparatively new field of experimental design is involved, and the requirements call for expert understanding and assistance. Further discussion and recommendations appear in the next chapter, with some of the reasons why a statistical analysis should be considered whenever grades are used.

CHAPTER IV

DISCUSSION AND RECOMMENDATIONS

Education has been defined as the changing of human beings for the better. These changes are made known to us by comparing the degrees of development during progress from one step to another, such as ideas understood, words spoken, acts performed, and traits shown. To measure any of these degrees of development, definitions of the developments and of their amounts are called for in order that anyone can grasp the meanings of the differences better than could be done without measurement (27, p.17). If the changes taking place are to lead to progress, one must be able to evaluate the extent to which teachers influence student development. As teachers are in key positions to help the students profit from evaluation, they must know how to evaluate achievement and how to interpret that evaluation (6, p.113). Teachers are supposed to help instill habits of accuracy in students. If this is desirable, as it must be, then the teachers should set good examples at least insofar as the grades given are concerned. It should be evident that teachers will not be able to recognize the degree of accuracy in their grades if they are not versed in the variability of the numbers which are translated into grades and are not able to interpret them scientifically. A teacher can too easily overlook the scientific method of thought in grading members of his classes, and often does. Under such circumstances, students will not develop attitudes that are as scientific as they

would be if the grading were more scientific, thorough, and consistent. Education is done by example as well as by words (30, p.301)

> General education, as well as the teaching of science aims at the worthy goal of encouraging clear thinking. Therefore, teachers face the responsibility of all men who want to know the truth. To be sincere about some things and not about others is not acceptable. In another profession a practitioner might be called a quack for no greater offense (30, p.302).

Research should be undertaken to improve the many evaluation methods now employed. Dependable inferences concerning a child's understanding can be made only under procedures having a known degree of significance. The present trend indicates that textbooks in "Educational Measurement" are improving and that teachers will meet their responsibilities and become better able to obtain understanding of methods of measurement and grading. This will improve teaching, and assure sounder and more worthwhile learning on the part of the children or young people whom they are teaching (2, p.328-330). Teaching and testing are two sides of the same medal, and it is probable that teaching will improve as the understanding of the uses and evaluation methods of tests are developed and made more a part of the training of teachers. For this reason it is strongly recommended that more test training be required of teachers.

Tools and techniques of measurement have by no means been developed to a state of perfection. Sources of error may stem from the measuring instruments themselves or the ways in which they are used. Three ways of controlling errors in measurement are suggested: (1) the improvement of existing measuring instruments, (2) the

devising of adequate methods that estimate or allow for errors, or (3) the development of skill in applying instruments of measurement and interpreting results in terms of estimated errors (23, p.13).

It seems evident that teachers should be required to obtain a thorough foundation in the techniques and statistics of achievement evaluation, or leave evaluation of achievement and its interpretation to experts in those fields. If teachers know what achievements they really wish to measure and what standardized tests actually measure the desired characteristics, they can quite easily obtain and give such tests. Care must be taken to use norms that reflect results in the light of the reasons for making the measurements. Results should be compared to local as well as national norms in order to obtain a more realistic comparison. A few other reasons for using statistics - which requires statistical training - and seeking expert statistical advice for more successful analysis are given in the following quotations:

> The difficulty in using Educational evaluations is that no worth-while, reliable results can be obtained unless the appraisal becomes a controlled scientific experiment, with significant variables held constant and with statistical tests of significance applied to the results. The complexity and refinement of the procedures and precautions that are necessary may be appreciated from a perusal of recent textbooks in advanced statistics and educational research. Principals, superintendents, and others who make appraisals must either acquire a mastery of these scientific methods for themselves, or turn the problems over to experts, or cease to use evaluations for the appraisal of educational instrumentalities. The appraisal function can be realized only in the form of controlled, statistically analyzed experimentation. (21, p.12).

> Seek advice on experimental design and associated statistical treatment. Any study which is to utilize statistical treatment should be carefully planned before the data are collected. Finding an appropriate statistical treatment for data collected in a haphazard fashion is difficult and never very satisfactory. Presented with a statement of the hypothesis to be tested, a competent statistician can devise an appropriate design for collection of data and a statistical treatment suitable to the questions raised. Some personnel workers are statistically sophisticated to the point of doing this themselves; most are not (7, p.337).

These suggestions were followed in making the experiment on which this thesis is based, and hope is expressed that both the method and reasons for its use will be helpful to others who might be venturing in this direction.

> Creative teaching is only one of many challenging occupations, but among scholars and administrators, among scientists and laborers, in all walks of life, there is an all too common contempt for teaching as a daily activity. This attitude may be related to the conflict between good teaching and percentile grading. Petty measurements have a part in making our profession into a stronghold of taskmasters, bereft of inspiration (30, p.303).

It is of great importance that teachers who interpret test scores and classify students know the errors of their techniques (14, p.174). If teachers are to continue toward the objective of improving teaching, it is believed necessary to eliminate guess-work as rapidly as feasible in terms of the required training. Decisions on whether or not students are achieving at their optimum rate and in the direction of the greatest benefit to themselves and the society in which they live require a minimizing of guess-work. This thesis is presented as a possible step toward that end.

BIBLIOGRAPHY

1. Bolmeier, E. C. Principles pertaining to marking and reporting pupil progress. The school review 59:15-24. Jan. 1951.

2. Brownell, William A., et al. The measurement of understanding. Forty-fifth yearbook of the national society for the study of education, Part 1 45:321-330. 1946.

3. Cavanaugh, Joseph A. A survey of opinion on examinations. Educational research bulletin 29:120-125, 139-140. 1950.

4. Cochran, William G. and Gertrude M. Cox. Experimental designs. New York, Wiley, 1950. 454p.

5. Dixon, Wilfred Joseph and Frank Jones Massey, Introduction to statistical analysis. Eugene, University of Oregon press, 1949. 220p.

6. Donahue, Wilma T., Clyde H. Coombs and Robert M. W. Travers (eds.). The measurement of student adjustment and achievement. Ann Arbor, University of Michigan press, 1949. 256p.

7. Dressel, Paul L. Personnel services in high school and college. A discussion of evaluation. Occupations, the vocational guidance journal 29:331-340, 1951.

8. Edwards, Allen L. Experimental design in psychological research. New York, Rinehart, 1950. 446p.

9. _____. Statistical analysis for students in psychology and education. New York, Rinehart, 1946. 360p.

10. Fisher, Ronald A. Statistical methods for research workers. 11th ed. New York, Hafner, 1950. 354p.

11. _____. The design of experiments. 5th ed. New York, Hafner, 1949. 242p.

12. James, Glenn and Robert C. James (eds.), et al. Mathematics dictionary. New York, Van Nostrand, 1949. 432p.

13. Johnson, Palmer O. Statistical methods in research. New York, Prentice-Hall, 1949. 377p.

14. Kelley, Truman Lee. Interpretation of educational measurements. New York, World book, 1927. 363p.

15. _____. The use of literal grades. The journal of educational psychology 41:488-492. 1950.

16. Krathwohl, W. C. A 3 by 3 analysis of the predictive value of test scores. Journal of applied psychology 28:318-322. 1944.

17. Lindquist E. F. Statistical analysis in educational research. Boston, Houghton, 1940. 266p.

18. Lewis, Donald. Quantitative methods in psychology. Iowa City, The bookshop, 1950. 290p.

19. McNemar, Quinn. Psychological statistics. New York, Wiley, 1949. 364p.

20. Micheels, William J. and M. Ray Karnes. Measuring educational achievement. New York, GcGraw-Hill, 1950. 496p.

21. Remmers, H. H. and W. L. Gage. Educational measurement and evaluation. New York, Harper, 1943. 580p.

22. Rogers, Virgil M. Improved methods of reporting pupil progress. The school executive 60:19-22. Oct. 1950. (Our schools No. 71)

23. Ross, C. C. Measurement in today's schools. 2nd ed. New York, Prentice-Hall, 1948. 551p.

24. Sims, Verner Martin. Improving the measuring qualities of an essay examination. Journal of educational research 27:20-31. 1934.

25. Smith, Eugene R., Ralph W. Tyler, et al. Appraising and recording student progress. Adventure in American education 3:463-469. 1942.

26. Snedecor, George W. Statistical methods. 4th ed. Ames, The Iowa state college press, 1946. 485p.

27. Thorndike, Edward L. The nature, purposes, and general methods of measurements of educational products. Seventeenth yearbook of the national society for the study of education, Part 2 17:16-24. 1918.

28. Travers, Robert M. W. and Norman E. Gronlund. The meaning of marks. The journal of higher education 21:369-374. 1950.

29. Tukey, John W. Comparing individual means in the analysis of variance. Biometrics 5:99-114. 1949.

30. Walter, John T. Grades, fact or fraud. American association of university professors bulletin 36:300-303. 1950.