Columbia University

Contributions to Education

Teachers College Series

TEACHERS' MARKS

THEIR VARIABILITY AND STANDARDIZATION

BY

FREDERICK JAMES KELLY, PH.D.

TEACHERS COLLEGE, COLUMBIA UNIVERSITY
CONTRIBUTIONS TO EDUCATION, No. 66

PUBLISHED BY
Teachers College, Columbia University
NEW YORK CITY
1914

ACKNOWLEDGMENTS

Any study such as this is possible only as the result of the coöperation of scores of persons. It is hoped that each one whose time and labor have entered into the data here presented has served ungrudgingly in the interest of our profession. In appropriate places in the text I have expressed my thanks to those who have contributed most generously to the separate phases of the study. My appreciation of the work of all others who have helped is keenly felt. I must forego the mention of other names, however, except Professors Strayer, Thorndike, and Hillegas, whose counsel and assistance have been so freely given throughout the year.

F. J. K.

CONTENTS

INTRODUCTION 1

The Problem
The Material and Method

STANDARDS OF MARKING IN ELEMENTARY SCHOOLS 5

STANDARDS OF MARKING IN HIGH SCHOOLS 11

STANDARDS OF MARKING IN COLLEGES 48

THE MARKING OF EXAMINATION PAPERS 51

STANDARD TESTS AND SCALES AS AIDS IN STANDARDIZATION 85

The Courtis Tests in Arithmetic
The Thorndike Drawing Scale
The Thorndike Handwriting Scale
The Hillegas Composition Scale

CONCLUSIONS 133

BIBLIOGRAPHY 135

TEACHERS' MARKS

INTRODUCTION

Two distinct questions are involved in the problem of assigning marks to pupils. The first concerns the average standard of achievement which should be expected of normal children of a given age and grade, and the second concerns the distribution of the ability within the normal group around that standard. The first is a question for the school administrator primarily, while the second must be solved by the psychologist. The interrelation between the two questions is obvious and suggests the need for psychology in school administration and the need for keen insight into the problems of administration on the part of the educational psychologist. However, I shall not examine the second question beyond what its relation with the first demands.

The answer to the first question lies in our discovery of some method of defining merit in school work. We have long depended upon the examination paper and still do depend upon it almost universally. Out of a growing recognition of the inadequacy of the examination as at present administered, there has developed in this country a disposition to depend more and more upon the individual teacher's notion of what is proper to expect in the way of student achievement. This is resulting in wide differences in demands because the standards of teachers are far from uniform. The standardizing influence of examinations is being removed without anything being put in its place. For example, the custom has come to be quite universal throughout the middle and western states of having admission to college based, not upon examinations as was the custom not so many years ago, but upon high school accreditment, which is an arrangement between the high school and college whereby the graduates of the high school are admitted to college without examination, provided certain standards in equipment, instruction, course of study, etc., are maintained by the high school. To be sure, there is usually a representative of the college designated as high school inspector whose duty it is to keep the high

schools up to a fairly uniform standard of scholarship, but anyone familiar with the work of inspectors in general will not claim that marked success attends their efforts. Standardization of equipment, teachers' salaries and experience, number of recitations, etc., does little more than begin to standardize the requirement for a passing achievement in a given subject of study. Work which satisfies a teacher of chemistry in one accredited school would be considered far from satisfactory by another teacher of chemistry in another accredited school. This absence of uniformity of requirement seems fairly spread over the whole field of education. With the increasing emphasis we are placing in this country upon the teacher's individuality, the situation is likely to grow worse unless some measures for standardization can be put into operation.

It is with this question of standards among teachers that the present study is concerned. The effort is here made to point out the extent of variability among teachers in rating work of equal merit. From this the need for practical definitions of standard achievements may be appreciated. When we appreciate the need for defining achievement we are ready to consider some of the tests and scales which have been devised for the purpose of making possible these definitions.

The Problem

The problem undertaken in this study, then, is twofold: first, to set forth the situation as it exists with respect to teachers' marks, and, second, to examine certain standard tests and scales to determine their effectiveness in improving this situation.

The Material and Method

In the first part of the study which undertakes to set forth the existing conditions with respect to the variability of standards among teachers, my main task has been to summarize and evaluate the work of former students of the subject. I have found it necessary in many cases to give the results of earlier studies in the form of tables which the authors have used to summarize their findings.

In the second part of the study I have undertaken to try out certain recently devised tests and scales as instruments for re-

ducing the variability among standards of marking which the present situation reveals. Comparison between the variabilities, found without the scale and with it, forms the basis of the study. There is a definite limitation to be kept in mind throughout this comparison. In every case the results recorded are obtained from persons thoroughly experienced in the use of the common systems of marking and completely unpracticed in the use of the derived scales. Until some future study reveals the effects of practice in the use of the derived scales, no final judgment can be made as to the ultimate service of the scales in establishing standards. The present study contains a few evidences tending to show that the practice effect is rapid and great.

The two most significant studies in this field support the generally accepted notion that marks mean very different things to different teachers. The investigation by Ralph E. Carter[1] in Milwaukee, Wis., in 1911, where a uniform system of marking prevails throughout the elementary schools, revealed some striking facts. He considered only classes which completed the eighth grade in 1907, thereby assuring uniform instructions about grading and uniform curricula among the several schools. The following variation was found in the marks given by three schools: Of all the marks given in arithmetic,

In School A, $\frac{1}{3}$ were below 79, and $\frac{1}{3}$ above 84.

In School B, $\frac{1}{3}$ were below 71, and $\frac{1}{3}$ above 78.

In School C, $\frac{1}{3}$ were below 82, and $\frac{1}{3}$ above 88.

Two thirds of School B fall within the range of the lowest third of School A, while two thirds of School C fall within the range of the highest third of School B with a margin of four points to spare.

As an indication of how much real difference in ability these differences in marks indicated the records of the members from these schools were traced in the high schools. It was necessary to determine first what proportion of the members from the poorer and better sections of the three classes entered high school. It was found that the school grading lowest had sent a larger proportion of its poorer members to high school than the school grading highest. Nevertheless, when all the algebra marks of the members from the three schools were ranked together, it was found that "a greater percentage of School B excelled in maintaining their original rank or increasing it. In fact there was a complete reversal of things from what the absolute marks alone might indicate."

In Iowa City, Iowa, Walter R. Miles[2] made a similar study of the marks of pupils entering the high school from the elementary schools of that city. Using the cases of all pupils whose scholastic

[1] Ralph E. Carter, Correlation of Elementary Schools and High Schools, *Elementary School Teacher*, 12:109–118.

[2] Walter R. Miles, Comparison of Elementary and High School Grades, *Univ. of Iowa, Studies in Education*, Vol. I, No. 1.

records were complete for the last four years of their elementary school and at least two years of their high school careers, he covered a period of twelve years, and obtained 106 cases. To obtain a pupil's rank, all of his elementary school marks were averaged for his elementary school rank and all of his high school marks were averaged for his high school rank. By this means the inequalities of rating pointed out by Carter in Milwaukee were largely balanced from year to year and subject to subject since, in every case, the rank of a pupil represented the combined judgment of several teachers. The average of the elementary school marks thus determined was found to be 89.15 while the average of the high school marks was 82.49. By subjects, the averages varied as follows: In elementary school, from 87 to 91.33; in high school, from 79.94 to 86.92.

A list of coefficients is given representing the correlation between the marks given in one department or school and another. The average of the fifteen coefficients of correlation between one elementary school subject and another is .567; the average of the ten between one high school subject and another is .618; the average of the eighteen between elementary school subjects and high school subjects is .446. The highest coefficient of the list is, naturally, that between the average of all elementary school grades, and the average of all high school grades. It is .71. When we remember that the marks used in these calculations were the average of several teachers' ratings in every case, the coefficients do not seem very high. It appears that the greater the number of marks which enter into the averages, the higher the correlation. We shall consider this point a little later.

There are three other studies, the first by W. F. Dearborn[1] at Madison, Wis., the second by H. I. Miller[2] at Kansas City, Kan., and the third by F. W. Johnson[3] at Chicago, which point to the same absence of standards among teachers in elementary schools. The data which they furnish, however, may be accounted for in large part by other factors than variations in standards among

[1] W. F. Dearborn, School and University Grades, *Univ. of Wisconsin Bulletin*, No. 368, 1910.

[2] H. I. Miller, A Comparative Study of Grades of Pupils from Different Elementary Schools in Subjects of the First-Year High School, *Elementary School Teacher*, 11:161–175.

[3] F. W. Johnson, A Comparative Study of Grades of Pupils from Different Elementary Schools, in Subjects of the First-Year High School, *Elementary School Teacher*, 11:63–68.

teachers, and so it seems best to omit here any detailed statement regarding them.

To supplement these rather meager data pointing to the unreliability of the marks given by elementary teachers, I made in December, 1913, a study of the cumulative record cards for the Hackensack, N. J., schools. Several considerations prompted me to use these schools for this investigation. Besides the effective way in which the records are kept, and the cordial spirit with which Superintendent Stark and his corps of teachers welcome an investigator, the fact that departmental teaching is done in the seventh and eighth grades seemed very significant for my purposes. There are four ward schools which send their pupils at the completion of the sixth grade to this common seventh grade. It seemed to me important to determine how far the pupils from each school maintained their relative positions in the common seventh grade classes. If there should be found a difference in the amount of increase or decrease in marks from sixth to seventh grade among the four school groups as wholes, it would be possible to measure with some degree of security the difference in standard between a given mark in one sixth grade and the same mark in another sixth grade.

One fact must be taken into account in estimating the worth of such a measure. The seventh grade pupils are classified in three "courses": academic, commercial, and manual arts. The work is not identical in these courses, and, in part, the subjects—language for example—are not taught by the same teacher in all the courses. Nor do the representatives from the several sixth grades distribute themselves similarly among the three courses. Hence if different standards are held by the different seventh grade teachers, it may affect the results in some degree. Since it is impossible to calculate the amount of this influence, however, I have disregarded it in the figures, and have assumed that the seventh grade marks are a common standard by which to measure the variation among the four sixth grades.

The marks recorded on the cards in Hackensack are letters as follows: "E," "G," "F," and "P," for excellent, good, fair, and poor, respectively. The plus or minus after each letter is used, thus making twelve steps from the poorest to the best. For purposes of this study I have called "P−," 1; "P," 2; "P+," 3; "F−," 4; and so on to "E+," as 12. The smallest difference

recognized is one of these steps and must be carefully distinguished from those differences recognized when the common basis of 100 is used.

The data were gathered for all the pupils who made the two promotions in succession as follows: first group, from 6A in June, 1912, and from 7B in January, 1913; second group from 6A in January, 1913, and from 7B in June, 1913. Only the term's grade appears on the card, one mark for each subject; hence we have a composite mark in each subject derived from whatever sources, daily recitations, tests, etc., the teachers thought fit to determine the pupil's standing at promotion time. The marks for the six subjects, language, penmanship, history, geography, arithmetic, and spelling were used.

The following simple plan was used for arranging the data:

SCHOOL A

Pupils' Names	Language				Penmanship			
	6A	*7B*	*Gain*	*Loss*	*6A*	*7B*	*Gain*	*Loss*
Adams	9	7		2	5	6	1	
Brown	4	6	2		7	7	0	0

From the tabulations thus made the average gain or loss of the pupils from each of the four schools was determined by simply dividing the algebraic sum of the gains and losses by the number of the pupils from the school. The median grades or marks given in both the sixth and seventh grades were also determined. These two groups of data appear in the following tables, numbered 1, 2, 3, and 4.

TABLE 1

The Median Marks Received by the Pupils Promoted from the Four Sixth A Grades in June, 1912, and the Median Marks Received by the Same Four Groups of Children in January, 1913, When They Were Promoted from the Seventh B Grade

School	No. of Pupils	Lang.		Penman.		Hist.		Geog.		Arith.		Spell.		Medians of Totals		Gain or Loss in Totals
		6A	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	
A.....	19	8	5	7	5	9	5	8	8	6	5	8	5	45	33	−12
B.....	29	9	8	8	8	8	7	7	8	8	6	9	10	49	45	− 4
C.....	20	8	5	8	5	9	5	8	8	8	5	9	8	49	35.5	−13.5
D.....	20	6.5	8	7	7	7.5	6.5	8	8	8	6	9.5	11	43	42	− 1

TABLE 2

AVERAGE GAINS OR LOSSES OF PUPILS, BY SCHOOLS, IN THE VARIOUS SUBJECTS BETWEEN THE MARKS GIVEN IN JUNE, 1912, AND TO THE SAME CHILDREN IN JANUARY, 1913. *G* STANDS FOR GAIN, *L* FOR LOSS

SCHOOL	NO. OF PUPILS	LANG.	PENMAN.	HIST.	GEOG.	ARITH.	SPELL.	AVERAGE OF TOTAL GAINS AND LOSSES
A.....	19	*L* 2.32	*L* 3.06	*L* 2.73	*L* .10	*L* 1.89	*L* 1.05	*L* 1.86
B.....	29	*L* .86	*L* .88	*L* 1.39	*L* .04	*L* 1.48	*G* .79	*L* .64
C.....	20	*L* 2.15	*L* 3.53	*L* 4.16	*L* 1.40	*L* 1.30	*L* .70	*L* 2.21
D.....	20	*G* 1.10	*L* .26	*L* .80	*L* .05	*L* 1.30	*G* .35	*L* .16

TABLE 3

THE MEDIAN MARKS RECEIVED BY THE PUPILS PROMOTED FROM THE FOUR SIXTH A GRADES IN JANUARY, 1913, AND THE MEDIAN MARKS RECEIVED BY THE SAME FOUR GROUPS OF CHILDREN IN JUNE, 1913, WHEN THEY WERE PROMOTED FROM THE SEVENTH B GRADE

SCHOOL	NO. OF PUPILS	LANG.		PENMAN.		HIST.		GEOG.		ARITH.		SPELL.		MEDIANS OF TOTALS		GAIN OR LOSS IN TOTALS
		6A	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	*6A*	*7B*	
A.....	21	7	7	6	5	7	7	7	6	6	6	9	9	45	38	7
B.....	17	6	6	none		6	6	6	7	5	5	9	10	32	32	0
C.....	6	7	5.5	7.5	5	8	6	8	4	6.5	5.5	8	6.5	42	31.5	−10.5
D.....	20	8	8	6	5	7	8	7	7	6.5	6	10	11	47	44.5	− 2.5

TABLE 4

AVERAGE GAINS OR LOSSES OF PUPILS, BY SCHOOLS, IN THE VARIOUS SUBJECTS BETWEEN THE MARKS GIVEN IN JANUARY, 1913, AND TO THE SAME CHILDREN IN JUNE, 1913. *G* STANDS FOR GAIN, *L* FOR LOSS

SCHOOL	NO. OF PUPILS	LANG.	PENMAN.	HIST.	GEOG.	ARITH.	SPELL.	AVERAGE OF TOTAL GAINS AND LOSSES
A.....	21	*L* .57	*L* 2.09	*L* .38	*L* .86	*L* .48	*G* .33	*L* .68
B.....	17	*L* .06	None	*G* .29	*G* .06	*G* .68	*G* .46	*G* .24
C.....	6	*L* 1.16	*L* 2.00	*L* 2.00	*L* 2.33	*L* 1.16	*L* .83	*L* 1.58
D.....	20	00	*L* .80	*G* .50	*L* .25	*L* .30	*G* .35	*L* .08

From the foregoing tables it is evident that there is no uniformity among the standards used by the several teachers in giving marks to pupils. On the whole, the marks are considerably reduced from 6A to 7B. With few exceptions, the median marks given by the teacher of 6A in School C are higher than

the median marks given by the teacher of 6A in School D, for example, although when these two groups of children are marked at the close of the following semester in 7B, it is found that the pupils from School D are almost uniformly higher than those of School C. Considering the average gains and losses in some of the more extreme cases, we find many striking variations. In the June, 1912–January, 1913 group, the difference in standards as represented by the common 7B marks the succeeding semester, amounts to 3.25 steps in language between schools C and D, 3.27 steps in penmanship between the same schools, 3.35 steps in history between the same schools, and an average difference of 2.05 steps between the same schools. This means that for work which the teacher in School C would give a mark of "G" in language, penmanship, or history, the teacher in School D would give less than a mark of "F." And that a pupil whose monthly report card in School C had been a "G" card, on the whole, would be dismayed by receiving an "F+" when he moved to School D. It appears that School A marked somewhat lower than School C while School B marked higher than School D.

A nearer approach to uniformity seems to prevail in the January, 1913–June, 1913 group, although wide variations appear there also. This greater uniformity may be partly accounted for by the fact that between June, 1912, and January, 1913, three of the sixth grade teachers who had given the marks recorded in the first group were transferred and their places filled by teachers who had formerly been grammar grade teachers.

STANDARDS OF MARKING IN HIGH SCHOOLS

The first important study of this subject was made by F. W. Johnson,[1] principal of the University High School of the University of Chicago. He investigated the marks given by the various departments in his school for the years 1907–08, and 1908–09 to determine the variation among them. His data are deserving of careful study. The plan of marking used in the University High School is as follows: F for failure, and D, C, B, A, for the successive ranks above failure. The percentages of the different letters given by the several departments for 1908–09 are given in Table 5.

TABLE 5

GIVING THE DISTRIBUTIONS OF THE MARKS OF THE SEVERAL DEPARTMENTS OF THE UNIVERSITY OF CHICAGO HIGH SCHOOL
(From Johnson)

DEPARTMENT	TOTAL NO. OF MARKS	% OF F	% OF D	% OF C	% OF B	% OF A
Greek and Latin	886	10.6	16.1	31.8	23.5	17.9
German	416	8.4	19.5	26.4	28.6	17.1
French	475	10.9	18.7	33.0	28.0	9.3
English	1514	15.5	21.7	32.8	23.4	6.5
Mathematics	1466	14.5	25.2	27.6	21.1	11.5
History	825	8.1	15.9	31.2	30.0	14.7
Science	672	8.3	16.8	27.7	32.6	14.6
Domestic Science	176	5.7	2.3	27.3	51.7	13.1
Average	7297	11.5	18.9	30.6	27.0	12.0

One cannot fail to notice from the table that the failures in English and mathematics far outnumber the failures in either history, science or German, while the A's are nearly three times as frequent in Greek and Latin as in English.

When the marks of individual teachers are separated from these department groups, still wider variation appears. For example, the marks given by two different teachers in the same department are as follows:

	% OF F	% OF D	% OF C	% OF B	% OF A
First Teacher	8.0	16.0	47.5	22.0	7.5
Second Teacher	4.5	6.0	24.0	30.5	36.0

[1] F. W. Johnson, A Study of High School Grades, *School Review*, 19: 13–24.

The comparison of the marks of two teachers in different departments reveals even more striking variations:

	% of F	% of D	% of C	% of B	% of A
First Teacher	26.5	42.5	25.5	4.5	1.5
Second Teacher	4.5	6.0	24.0	30.5	36.0

It is conceivable that a set of conditions might prevail in which the above variations would be justified, at least in part. Johnson offers them, however, as examples of variation which have no justification. They are simply due to different standards held by different teachers.

Franklin O. Smith[1] at the University of Iowa used a different method for discovering the variability of standards of marking in use in the high schools of Iowa. He compared the high school marks and the college marks of 120 Liberal Arts students who graduated from the University of Iowa in 1910. The average of all high school marks was used as the student's high school standing, and the average of all university marks as his university standing. The correlation between the high school and university standings of these 120 students is represented by a Pearson coefficient of .53. This seems surprisingly low in view of Smith's use of the average. Much lower correlation appears, however, when the separate subjects are compared with one another, or even with the same subject in the two schools. If the marks of individual teachers are fairly reliable, we should expect to find the correlations rather high between, say, mathematics in high school and mathematics in university. The following portion of his list of coefficients is illuminating:

English, high school and university	.34
Mathematics, high school and university	.29
History, high school and university	.18
Ancient Language, high school and university	.43
Modern Language, high school and university	.28
Science, high school and university	.34
Average	.31

Pettit[2] also found that the Pearson coefficient of correlation between average high school marks and average freshman college marks to be .63, but found the average of the coefficients

[1] Franklin O. Smith, A Rational Basis for Determining Fitness for College Entrance, *Univ. of Iowa, Studies in Education*, Vol. 1, No. 3.

[2] W. W. Pettit, A Comparative Study of New York High School and Columbia College Grades, Master's essay, Teachers College, 1912.

when calculated by departments, high school English with college English, mathematics with mathematics, etc., to be .49.

It must be borne in mind that even these rather low coefficients are derived from marks which are for the most part averages from several teachers' ratings. Teachers' marks in a single high school subject with those in the university would probably show even less relation.

This plan of using the rank of a student in the next higher school as a guide for determining the correctness of rating in the lower school possesses such great possibilities for forcing us to derive standards that it seems worth while to give two of Smith's tables indicating quintile changes and retentions from school to school. The averages are used to determine rank either high school or university, and then each fifth of the high school group is traced through the university, thus giving a simple indication of how consistently a given rank is maintained. These two tables follow as Table 6 and Table 7.

TABLE 6

DISTRIBUTION BY QUINTILES IN THE UNIVERSITY RANKINGS OF EACH QUINTILE OF THE HIGH SCHOOL RANKINGS. GENERAL AVERAGES OF MARKS USED IN EACH SCHOOL DETERMINE RANK

HIGH SCHOOL	DISTRIBUTION IN UNIVERSITY BY PER CENTS				
	1st Q.	*2nd Q.*	*3d Q.*	*4th Q.*	*5th Q.*
1st Quintile	54.0	16.6	16.6	4.0	8.0
2nd Quintile	25.0	29.0	16.6	12.5	16.6
3rd Quintile	16.6	25.0	21.0	21.0	16.6
4th Quintile	0.	25.0	25.0	33.3	16.6
5th Quintile	4.0	4.0	21.0	29.0	42.0

This table (from F. O. Smith's study, page 142) reads as follows: Of the lowest one fifth in high school rank, 54 per cent are found in the lowest one fifth in college rank; 16.6 per cent are found in the next fifth, and so on.

TABLE 7

Same as Table 6 except that instead of the general average of all university grades, the senior grades alone are used. (F. O. Smith, p. 145.)

HIGH SCHOOL	DISTRIBUTION IN UNIVERSITY BY PER CENTS				
	1st Q.	*2nd Q.*	*3rd Q.*	*4th Q.*	*5th Q.*
1st Quintile	25.0	25.0	21.0	12.5	16.6
2nd Quintile	29.0	25.0	25.0	16.6	4.0
3rd Quintile	21.0	12.5	16.6	33.3	16.6
4th Quintile	21.0	16.6	16.6	25.0	21.0
5th Quintile	4.0	16.6	21.0	16.6	42.0

If there were no tendency for students to maintain their previous rank when they went on to a higher school, all the per

cents in the tables reproduced would be just 20. Whether the amount of the tendency indicated is sufficient to satisfy us regarding the reliability of teachers' marks, each reader must judge. The average of the ten figures representing the retention in the same quintile is 31.3. A chance distribution accounts for 64% of the retention of quintile rank.

From these tables as well as from the coefficients of correlation (they are lower than those found by Dearborn, Miles, or Pettit) a fair inference can be made that the fact of absence of standards is a very large factor in producing this change of rank from one school to the next. Smith is using representatives from a large number of small schools instead of a small number of large schools. Perhaps one or two students from a school is the rule rather than fifty or more, and there is less chance for these isolated schools and teachers to approach uniformity of standards than there is in the cases of the large schools. Smith indicates in the last sentence of his study his appreciation of the need for standardization: "But when this is done (meaning the adoption of a rational method of the distribution of marks), there still remains the problem of standardizing the teacher's judgment."

One of the most striking illustrations of how largely a matter of tradition the passing standard is, is afforded by the figures in Table 8, page 15, which were given to me by the principal of one of the New York City high schools. The difference between the percentage of pupils allowed to pass in the various subjects during the year previous to his becoming principal and the first year of his service was due almost wholly to the determination on his part to break up the tradition that a large percentage of each class ought to fail.

Consider that in the large high school where this change took place this meant a reduction of the number of failures by nearly if not quite 500 a year. All this depended primarily upon the notion of one man. In the same high school during the last three years the average time of attendance of students to win graduation has been reduced by more than one year. There are undoubtedly many other factors entering into the remarkable changes, but the largest factor, certainly, is that the present principal and the former principal happen to have radically different standards.

TABLE 8

REPRESENTING THE CHANGE IN PERCENTAGE OF PUPILS PASSED IN THE DEPARTMENTS OF A NEW YORK HIGH SCHOOL FROM ONE YEAR TO THE NEXT

DEPARTMENT	TERM	PERCENTAGE OF PUPILS PASSED	
		1910	*1911*
Biology	1st	69	81
	2nd	79	83
Algebra	1st	48	75
	2nd	61	80
English	1st	77	86
	2nd	84	90
French	1st	68	83
	2nd	66	85
German	1st	65	83
	2nd	65	86
Latin	1st	64	78
	2nd	72	82
Average		68.2	82.7

In an unpublished report of a study made in 1912 by Carter H. Alexander, at the time Professor of School Administration at the University of Missouri, some interesting facts concerning standards in the high schools of Missouri are brought out. The amount of variability among the standards employed in the thirty-one schools whose records were studied is best shown in the following table which I copy from the report: *M* stands for median, 1*Q* for 25 percentile or that point below which 25 per cent of the cases fall, 3*Q* for 75 percentile.

TABLE 9

PERCENTAGES BY SCHOOLS OF ALL GRADES ISSUED BY EACH SCHOOL IN THE VARIOUS SUBJECTS WHICH ARE BELOW PASSING. THE MEDIANS AND LIMITS OF THE MIDDLE 50 PER CENT OF THE DISTRIBUTIONS FOR 31 MISSOURI HIGH SCHOOLS, 1911–12, ACCREDITED TO THE UNIVERSITY (Alexander)

	1ST YEAR			2ND YEAR			3RD YEAR			4TH YEAR		
	M	*1Q*	*3Q*	*M*	*1Q*	*3Q*	*M*	*1Q*	*3Q*	*M*	*1Q*	*3Q*
English	8.7	4.3	17.6	4.0	0	12.5	2.8	0	11.0	0	0	5.0
History	12.4	8.0	23.5	10.0	0	16.9	5.5	0	20.0	0	0	2.4
Mathematics	13.2	6.5	22.0	12.9	4.6	23.9	11.0	0	16.3	0	0	0
Latin	17.8	8.4	34.5	11.8	2.1	18.1	0	0	8.1	0	0	0
German	16.0	11.0	20.0	10.0	0	11.4						

This table reads as follows: In first year English, half the schools give more than 8.7 per cent of grades below passing, and half give less; one fourth of the schools give less than 4.5 per cent of grades below passing, and one fourth give more than 17.6 per cent below passing.

Every item in this table is significant. Surely no one can maintain that such wide variation regarding the number not passed in the several schools corresponds to a similar variation in student merit from school to school. Look, for example, at the third year column. One fourth of the schools fail none of their students in English, history, mathematics, or Latin, while another fourth of the schools fail more than 8 to 20 per cent of their students in the same branches. In spite of such differences, Dearborn[1] argues from a close correlation of rankings between averages in high school and averages in the university, that the plan of accrediting high schools forms a successful way of selecting students.

Mention may be made here of the variation among the averages of marks given by the several departments, and in the percentages failed by departments in a representative high school for which figures are available. In Iowa City, Miles reports in the study referred to above, the following averages by departments, and failure marks:

	Average of Marks	Per cent Failed
Science	79.94	13
Foreign Language	81.53	14
Mathematics	80.51	19
English	83.39	9
History	84.20	7
Drawing	86.92	Not given

In a certain large Illinois high school the principal reports percentages of failures as follows:

Commercial	28
Mathematics	23
Modern Languages	22
Ancient Languages	18
History	16.5
English	16
Science	13.5

Notice that twice as many are failed in commercial subjects as in science. The enrollment of the two departments concerned is 850 and 620 respectively, so we see that about ninety more pupils each year fail in commercial subjects than in science.

In the two studies referred to above by W. F. Dearborn, at that time a member of the faculty of the University of Wis-

[1] W. F. Dearborn, Relative Standing of Pupils in the High School and in the University, *Univ. of Wisconsin Bulletin*, No. 312.

consin, we find not only a mine of information on the subject of grading but we find also the source of inspiration for three other most painstaking investigations in the same field. These three are a Master's essay written at Teachers College in 1911 by W. W. Pettit,[1] entitled "A Comparative Study of New York High School and Columbia College Grades." A Doctor's dissertation written at the University of Chicago in 1912, by John A. Clement,[2] entitled "Standardization of the Schools of Kansas"; and the third a recent number of the Educational Psychology Monographs, prepared at the University of Chicago by Clarence Truman Gray[3] and entitled, "Variations in the Grades of High School Pupils." The latter two studies were written under the direction of Dearborn at Chicago, but differ from Dearborn's study in one essential particular which will be described later. Pettit's study follows the same plan as Dearborn's and his conclusions are supposed to support the conclusions reached by Dearborn, whose chief purpose was to establish the superior merit of the plan of admission to college by accreditment over the plan of admission by examination. As supporting this purpose, the method of Dearborn contains two fallacies, it seems to me, which should be pointed out. I shall, therefore, defer consideration of the material in the later studies until after a criticism of the method of Dearborn and Pettit. While the latter author did not have the same purpose in view he used the same method as Dearborn to determine the " relative standing of pupils in the high school and college," and one of the chief points of significance which attaches to such information is its bearing upon the question of method of admission to college. Pettit cannot be held guilty, however, of the fallacies which are present in the Dearborn method. These two fallacies are, first, the use of *averages* to determine rank in both the high school and the university, when the results are to be applied to the question of admission to college by accreditment, and second, the failure to take account of differences in standards of rating *by schools* among the group of schools studied.

The pointing out of these two fallacies is not a mere academic matter. If facts establish the conclusion that accreditment is

[1] Unpublished.
[2] University of Chicago Press.
[3] Warwick and York, Baltimore, Md.

"a successful means of selecting students for college," the corollary must follow that the standards of marking among the teachers in the high schools concerned are satisfactorily uniform, and there is not the urgent need for standardization in marking which is being claimed at present. Dearborn closes his study with the comment that his results "are in sharp contrast to those secured by the test of the entrance examinations at Columbia College." (Professor Thorndike's study is meant.) While I do not regard entrance examinations at all satisfactory as a means of selecting students for college, I do believe it can be shown that the fallacies above referred to are responsible for the "sharp contrast" which Dearborn establishes in favor of accreditment.

In Dearborn's study, "The Relative Standing of Pupils in the High School and the University," the high school marks of all the representatives from the six cities which furnished the largest number of students in the College of Letters and Science at the University of Wisconsin from 1900 to 1905 were secured, as well as all the marks received by these same students in all their undergraduate classes in the University. This made a group of 472 students in all. Only three subgroups were considered: Madison High School furnished 238; the three high schools of Milwaukee together furnished 139; four smaller high schools in the state furnished the remainder, 92. Because no closer differentiation into single high school groups is made, the second fallacy indicated above, namely, the disregard for difference in standards *by schools*, is not so clearly evident although demonstrable. On this account I shall use Pettit's data for the first and more complete illustration.

Pettit studied the ratings of all the individuals who entered Columbia College from 1900 to 1910 from three high schools which we shall call A, B, and C. All the high school marks received in English, history, mathematics, science, Greek, Latin, and modern languages by each boy were averaged together to make his rank in the total high school group. Similarly the college ranks were determined by averaging all marks received in a similar group of departments in college. Of the total group of 218 boys, School A furnished 53, School B, 88, and School C, 77.

Pettit followed the method used by Dearborn except in one particular. Where Dearborn used the quartile division of the

group, that is, divided the whole range of ranks into four equal parts, Pettit used the quintile division, dividing the whole group into five equal parts. The individuals in each quartile or quintile were then traced through the higher school by quartile or quintile. To illustrate I here reproduce one of Pettit's tables indicating the location in the sophomore rankings of the members of each quintile in the high school group of 218 boys:

High School Quintiles	College Sophomore Quintiles				
	I	II	III	IV	V
I	60%	30%	7%	3%	0%
II	23	20	23	17	17
III	6	19	31	22	22
IV	0	17	11	37	34
V	0	10	31	24	34

This table reads as follows: of the fifth ranking highest in high school, 60 per cent are found in the highest fifth in college sophomore class; 30 per cent are found in the next to the highest fifth in rank, and so on.

The two criticisms which I wish to make of the two studies are, then, specifically these: first, that data such as these give but slight ground for the conclusion that the system of accredited schools in vogue in the West is a successful method of selecting students for college, and second, that the amount of quartile and quintile change from high school to college, which is due directly to different standards prevailing in the several high schools composing the group, is far from negligible.

In support of my first contention I hold that unless admission to college by accreditment means that all the high school grades of applicants for admission to college are averaged and admission granted on the basis of this average, then the data submitted do not bear directly upon the success of the scheme of accreditment. So far as I am aware, no college secures its students that way. Instead, the common way is for the high school to satisfy the college that its work is up to the standard. In return the high school gets the assurance that any student whom it graduates will be admitted to college *provided* he *has been passed* by the school in a certain list of subjects prescribed by the college. The *passing* in these subjects is determined, as a rule, by the standard of a single teacher in each subject. The variability of these standards from subject to subject, and school to school, has been abundantly shown, especially in Iowa and Missouri. In order successfully to maintain that a close correlation between high

school and college rank is a fair indication of the reliability of a teacher's mark in a particular subject, it would have to be shown that the ranking of a group from several schools in a single subject, say, physics, correlates closely with the rankings of the same group in a similar subject in the university. This, neither Dearborn nor Pettit has shown. In fact there is every reason to believe from the other studies in this field that the nearer we approach the marks of individual teachers the lower will be found the correlations between the rank in one group with the rank in the next. It will be recalled, for example, how much lower was the correlation between the marks of a department in high school with the marks of the same department in the college found by Smith in Iowa, than was the correlation between the average of all high school grades with the average of all college grades. This is very significant for our purposes. We should expect the average of the estimates of a dozen or more teachers to come pretty close to the correct ranking of young people. We should expect the average estimates of a dozen teachers in the higher school to come pretty close to the same ranking. Nearly everything of importance about marks, however, attaches to a particular teacher's mark in a particular subject. Even admission to college by the accreditment plan depends finally upon it.

To indicate clearly how children's averages from one term to the next correlate more closely than do the marks which enter into the averages, I calculated Pearson coefficients to designate the correlation between one teacher's marks in a given subject and the succeeding teacher's marks in the same subject in the case of forty-two separate pairs of classes. I then averaged the marks given the same child in his six different subjects, and correlated these averages in the case of the seven different groups of children who constituted the forty-two separate pairs of classes.

The data gathered at Hackensack for the study of the variation of teachers' marks in elementary schools were conveniently arranged for calculating the above coefficients. While it would have been desirable to use high school marks for this purpose, the same principle should hold throughout.

If the teacher's mark is a reliable index of the child's ability, then the marks of two successive teachers in the same subject should show a close correlation, closer in fact than the average

of many marks extending over a long time and including a variety of subjects. On the other hand, even if the teacher's mark is a very poor index of the child's ability, the average of thirty or forty such estimates is bound to approach fairly near that "general ability" which will be approached again by thirty or forty more estimates made in the higher school. Thus the rank obtained by the method of averages will be likely to hold fairly consistently from school to school, regardless of how inconsistently the teachers may mark from term to term. It is the teacher's mark which determines passing or failing, and it is passing or failing which determines college entrance by the accreditment plan.

In Table 10, the correlations are indicated by Pearson coefficients. Great accuracy cannot be claimed for any individual figures indicating relationships where the number of the cases in the distribution so correlated is so small, but the Pearson coefficients seem as accurate as any figure. It becomes significant, however, when a large number of such coefficients are secured, and their averages used as a measure of the correlation between one fact and another. It will be observed that in every case, the average of the coefficients obtained from the marks of the separate subjects is decidedly less than the coefficient obtained from the *averages* of the marks of the same

TABLE 10

PEARSON COEFFICIENTS OF CORRELATION BETWEEN THE MARKS GIVEN IN 6A, JUNE, 1912, AND THE MARKS GIVEN TO THE SAME CHILDREN IN THE SAME SUBJECT IN 7B, JANUARY, 1913, AND THE SAME BELOW FOR THE JANUARY, 1913–JUNE, 1913 GROUP

SCHOOL	LANG.	PEN.	HIST.	GEOG.	ARITH.	SPELL.	AVG. OF THE 6 COEFFICIENTS	COEFFICIENT OF THE AVGS. OF THE 6 MARKS	GAIN BY METHOD OF AVG.
A.....	.20	.05	.21	.52	.63	.39	.334	.48	.146
B.....	.51	.62	.59	.37	.20	.63	.488	.84	.352
C.....	—.10	.00	.13	.18	.03	.75	.166	.35	.184
D.....	.52	.11	.71	.162	—.16	.64	.39	.61	.22
A.....	.70	—.07	.69	.62	.64	.54	.519	.60	.081
B.....	.74	none	.48	.24	.39	.64	.506	.70	.194
D.....	.74	.54	.33	—.04	.51	.37	.409	.49	.081
Averages	.471	.211	.449	.364	.321	.566	.401	.581	.18

Note: All of the above coefficients are plus except where minus is indicated.
In the lower group, School C is omitted because the numbers in the class were too small to make valuable a figure indicating relationship between successive marks.

pupils. If now these averages were averaged with the averages from a half dozen other semesters, the coefficients would become rather high, even though the individual marks seem to be but poor indices of ability in the several subjects.

From this table it is seen that the coefficient is increased .18 by taking the average for a single semester, instead of individual teachers' marks by subjects. How much more would it be increased if the average for several years and many teachers were used? I submit then that the coefficients of correlation given by Smith, Miles, Dearborn and Pettit (.55, .71, .80, and .63, respectively) are poor evidence of the reliability of teachers' marks individually, and it is those marks, not averages, that count for accreditment.

My second contention is that a far from negligible part of the changes in rank from high school to college is due directly to different standards of marking in use in the several schools making up the group. On this point Dearborn says, "as the average of the marks of pupils entering from one high school was often 1 or 2 per cent higher than that of another high school it was the practice at first to weight the marks of all pupils to a common average of all the high schools included in the group. It was found, however, from actual trial, that such weighting did not affect the general comparison sufficiently to be worth while. In some cases at least the differences in averages of the high schools may represent real differences in the efficiency of the pupils concerned. But however that may be, the weighting of marks did not affect appreciably the large units of comparison employed in this study, and has not for that reason been used in the final results."

Pettit makes no mention of the point. He ranks his individuals on the supposition that "80" in one school means the same as "80" in the others.

The method employed to determine just how much of the change in rank from high school to college was due to the error in the above supposition was as follows: I calculated on the basis of the numbers from each school just what proportion of the changes should be attributed to each school group. I then determined by count how many quintile losses and quintile gains were actually made by each school group. If the losses were found to be proportionately too high, I assumed that the

standard of rating in the high school concerned had been lower than in the other schools. That is, an "80" had meant less in that school than in all the others. First, however, I compared the whole range of marks given each group in the high school and in the Freshman class in college. The data for this comparison in the cases of the three high schools studied by Pettit are given below in Table 11:

TABLE 11

GIVING THE MEDIAN MARK AND QUINTILE DIVISION POINTS IN DISTRIBUTIONS OF MARKS BY HIGH SCHOOL GROUPS

		QUINTILE DIVISION POINTS, *High School Marks*			
	Median	*1st*	*2nd*	*3rd*	*4th*
School B.....	78	66.75	74.25	79.6	84.8
School A.....	80	72	77.5	80.75	84
School C.....	74	67	71.75	76	82.6

GIVING THE MEDIAN MARK AND THE QUINTILE DIVISION POINTS IN DISTRIBUTIONS OF COLLEGE FRESHMAN MARKS BY HIGH SCHOOL GROUPS

	QUINTILE DIVISION POINTS, *College Freshman Marks*				
	Median	*1st*	*2nd*	*3rd*	*4th*
School B.....	81	73	79.1	84.2	88.5
School A.....	75.5	71	74	78	83
School C.....	81	74.5	79	83	88

From the high school marks School A boys seem the strongest students. When the marks of the first college year are taken, a

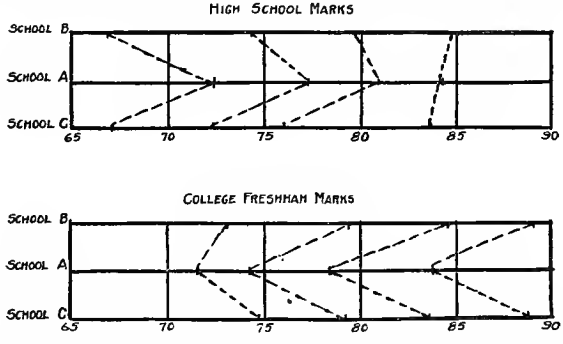

FIG. 1. (Data from Pettit.) Division points between quintiles in the distribution according to high school marks, and again in the distribution according to college freshman marks for the same pupils.

change appears. The median for School A drops 4.5 points while the median for School B rises 3 points and the median for School C rises 7 points. Similar reversal of the situation occurs all along the distribution. This is represented in the diagram, Fig. 1.

Consider now the fate of the lowest and highest fifths of each high school group. Table 12 is prepared to indicate clearly the change which appears in the representation in the highest and lowest quintiles, by schools, between high school and freshman college ranks.

TABLE 12

SHOWING THE NUMBER FROM EACH SCHOOL WHICH MAKE UP THE LOWEST AND HIGHEST FIFTHS OF THE HIGH SCHOOL AND FRESHMAN COLLEGE GROUPS, AND THE PER CENT THIS NUMBER IS OF THE NUMBER WHICH EACH SCHOOL WOULD HAVE IF REPRESENTATION IN THESE QUINTILES WERE PROPORTIONATE TO THE WHOLE NUMBER FROM THAT SCHOOL

School	Lowest Quintile H. S. Group		Lowest Quintile College Group		Highest Quintile H. S. Group		Highest Quintile College Group	
	No.	*% of quota*	*No.*	*% of quota*	*No.*	*% of quota*	*No.*	*% of quota*
B	19	108	16	91	22	125	20	114
A	3	28	18	170	13	122	7	66
C	22	143	10	65	9	58	17	111

From Table 12 we notice that while School A furnishes only 28 per cent of its quota to the lowest quintile, according to the ratings of high school teachers, it furnishes 170 per cent of its quota to the lowest quintile when the boys are rated in college. On the other hand, School C furnishes 143 per cent of its proportion to the lowest quintile in the high school ranking, but only 65 per cent of its share in the college ranking. Similar reversals appear in the highest quintile individuals, except that the shifting is in the opposite direction. School B, on the other hand, seems to have used a standard more nearly like the college, and midway between School A and School C.

In view of these obvious differences in standards among the three high schools it seemed worth while to calculate accurately just what portion of the shifting of position in the ranks between high school and college was due to these differences in standards. In doing this it seemed wise to use a method in the calculations a little more exact than that used in the study. When a distribution, say, of fifty marks, is divided into quintiles, the tenth mark needs to change but one rank in order to fall in the next quintile, and thus register as one quintile change. The first individual in the distribution, on the other hand, has to change by as much as ten ranks in order to register one quintile change.

The method used here was devised to obviate that feature in counting quintile changes, because in using the expression, "dropped from the first to the second fifth of the class," we convey the idea of having shifted position by as much as one fifth of the number in the class.

To explain the method most simply let us consider the cases of the fifty-three School A boys. If we record in the left hand column of the accompanying table the ranks of the boys in their own high school group, and in the second column their ranking when separated from the freshman college group, we may count the quintile gains or losses by subtracting each rank from the corresponding rank in the other series. If this difference equals one fifth of the total number of ranks in the series, it will register as one quintile change. If it equals two fifths of the number of ranks in the series, it will register as two quintiles change, etc. For example, in the table given herewith, from fourth to eighteenth rank is a change of fourteen places and we register a loss of one quintile. From tenth to forty-ninth place is a drop of three quintiles, etc.

TABLE 12A

To Illustrate Method of Computing Quintile Gains and Losses

H. S. Ranks	Freshman Ranks of Same Boys	Quintile Gains or Losses
1	3	0
2	5	0
3	2	0
4	18	−1
5	8	0
6	19	−1
7	7	0
8	14	0
9	9	0
10	49	−3
11	17	0
12	6	0
..	..	..
..	..	..
..	..	..
42	38	0
43	13	+3
44	22	+2
45	25	+2
46	41	0
47	42	0
48	23	+2
49	34	+1
50	48	0
51	53	0
52	32	+2
53	52	0

The same method exactly was followed when calculating the quintile gains and losses of the whole 218 group, with the addition of a mark attached to each number in the first column to indicate from what high school the boy came so that the quintile gains and losses could be credited to the proper school.

By this method of calculation the following Table, No. 13, was constructed, showing the number and percentage of individuals from each school who maintained their quintile position, who gained rank by one-fifth, two-fifths, three-fifths, or four-fifths of the number in the group, and who lost rank by one-fifth, two-fifths, three-fifths, or four-fifths of the number in the group.

TABLE 13

SHOWING THE NUMBERS RETAINING SAME RANK, AND NUMBERS CHANGING RANK, FROM HIGH SCHOOL TO FRESHMAN COLLEGE, IN ENTIRE GROUP OF 218 BOYS (Compiled from Pettit)

	Retain Same Quintile Position	Gain 1 Quintile	Gain 2	Gain 3	Gain 4	Lose 1	Lose 2	Lose 3	Lose 4
School B									
1st Quintile ..	19					3			
2nd Quintile ..	10	1				4	2		
3rd Quintile...	11	2	1				2		
4th Quintile...	12					2			
5th Quintile...	10	6	3						
Total	62	9	4			9	4		
Per cent.....	71	10	4			10	4		
School A									
1st Quintile...	7					3	2	1	
2nd Quintile ..	8	1				1	4		
3rd Quintile...	1					7	4		
4th Quintile...	10	1							
5th Quintils...	3								
Total	29	2				11	10	1	
Per cent.....	55	4				21	19	2	
School C									
1st Quintile...	8					1			
2nd Quintile ..	12	1							
3rd Quintile...	7	4	1			3			
4th Quintile...	9	6	2			2			
5th Quintils...	12	4	4	1	1				
Total	48	15	7	1	1	6			
Per cent.....	62	19	9	1	1	8			

If now we consider the gain of two quintile ranks by one boy the equivalent of two quintile gains, and so on for gains and losses of three, or four quintiles, we may summarize the gains and losses by schools as follows:

Schools	Quintile Gains	Quintile Losses	Excess of Gain or Loss
B.....	17	17	0
A.....	2	34	Loss 32
C.....	36	6	Gain 30
Totals.....	55	57	62

From this we see that of the total changes (55 plus 57), sixty-two or fifty-five per cent were due to the sliding up or down the scale of the group from a particular school *in mass*. Surely so much of the transfers should not be considered negligible.

The above fact can be verified from the less exact tables given by Pettit himself. By taking the difference between the numbers from each school found in each quintile of the high school group and in the freshman group, we get a measure of the same fact. Table 14 gives those data:

TABLE 14

Membership from Each High School in Each Quintile in High School and Freshman Distributions

School B	1st	2nd	3rd	4th	5th	total	Total Differences
High School Group.....	22	17	16	14	19	88	
Freshman Group.....	20	21	15	16	16	88	
Differences.....	2	4	1	2	3		12
School A							
High School Group.....	13	14	12	11	3	53	
Freshman Group.....	7	7	9	12	18	53	
Differences.....	6	7	3	1	15		32
School C							
High School Group.....	9	13	16	17	22	77	
Freshman Group.....	17	16	20	14	10	77	
Differences.....	8	3	4	3	12		30
Total Differences.....	16	14	8	6	30		74

According to Pettit's Chart 1, showing individual transfers from quintile to quintile, there were in all 126 quintile changes from high school ranks to freshman ranks. By the above table it appears that seventy-four of them, or fifty-eight per cent, can

be accounted for by the different standards of rating in the three high schools.

Now for Dearborn's claim that to weight the marks to a common average of all the schools would not alter the results. I have thus far shown only that if every boy had held exactly his high school rank among his own schoolmates when he did his freshman college work, there would still have been more than half as many changes in rank as Pettit found, and all because of the different standards in the three schools. If now there shall be

TABLE 15

SHOWING THE NUMBERS RETAINING SAME RANK AND THE NUMBERS CHANGING RANK FROM HIGH SCHOOL TO FRESHMAN COLLEGE, IN EACH SCHOOL GROUP CONSIDERED SEPARATELY (Compiled from Pettit)

	Retain Same Quintile Position	Gain 1 Quintile	Gain 2	Gain 3	Gain 4	Lose 1	Lose 2	Lose 3	Lose 4
School B									
1st Quintile ..	14					3			
2nd Quintile ..	13	1				2	2		
3rd Quintile...	11	2	1			3	1		
4th Quintile...	16		2						
5th Quintile...	10	6	1						
Total	64	9	4			8	3		
Per cent.....	73	10	4			9	3		
School A									
1st Quintile ...	7					2		1	
2nd Quintile ..	6	2				2	1		
3rd Quintile...	7	1	1				2		
4th Quintile...	7	2		1		1			
5th Quintile...	5	1	4						
Total	32	6	5	1		5	3	1	
Per cent.....	60	11	10	2		10	5	1	
School C									
1st Quintile ...	12					3			
2nd Quintile .	10	1				2	3		
3rd Quintile...	10	2				4			
4th Quintile...	9	4				3			
5th Quintile...	9	2	3		1				
Total	50	9	3		1	12	3		
Per cent	64	11	4		1	15	4		
Totals 3 schools	146	24	12	1	1	25	9	3	

found to be as great shifting of position within each school group as was found in the entire group, so that the same amount of shifting will be found with weighting marks to a common measure as without weighting them, as Dearborn claims, that will be mere accident. Certainly it will be a different fact when determined for each school separately than when determined for the group as a whole.

To enable me to make the comparison suggested above, it was necessary to calculate the quintile changes for each school group separately. These data are given in Table 15.

From this table, No. 15, it will be observed that the total quintile changes, when calculated as indicated for Table 13, are 107. It will be recalled that this is about the same number as was found to represent the quintile changes in the whole group of 218 taken together (that number being 112), but it represents a very different fact. In the first we had a measure of the change in rank due to two causes combined, namely, the sliding up or down of entire school groups, and the shifting of position within the entire group; in the latter figure we have a measure of the shifting of positions by members within their own school group. The fact that the two measures are so nearly the same is an indication that the difference in standards from school to school is about the same as the difference in standards among teachers of the same school, and leads to the suspicion that standards are a mixture of about equal parts of tradition, which influences a school group, and individual notions of teachers.

Turning now to the data furnished by Dearborn, I shall stop only long enough to point out that the differences in standards in Wisconsin high schools are not less than those in New York. His three groups of schools are, (1) Madison, (2) Milwaukee, three high schools together, and (3) four smaller high schools. Constructing a table for them similar to Table 12, for the New York high schools showing how the makeup of the highest and lowest quartile groups change from high school to freshman college, we have the following table, No. 16:

TABLE 16

SHOWING THE NUMBER FROM EACH SCHOOL GROUP WHICH MAKE UP THE LOWEST AND HIGHEST FOURTHS OF THE TOTAL HIGH SCHOOL AND FRESHMAN COLLEGE GROUPS, AND THE PER CENT THIS NUMBER IS OF THE NUMBER WHICH EACH SCHOOL SHOULD HAVE TO MAKE THE REPRESENTATION PROPORTIONATE TO THE WHOLE NUMBER FROM THAT SCHOOL (Compiled from Dearborn)

	Lowest Quartile H. S. Group		Lowest Quartile College Group		Highest Quartile H. S. Group		Highest Quartile College Group	
	No.	*% of Quota*	*No.*	*% of Quota*	*No.*	*% of Quota*	*No.*	*% of Quota*
Madison	85	144	64	108	41	69	49	83
Wilwaukee	25	72	35	100	45	129	48	137
Small High Schools	8	33	19	80	32	133	21	87

Without using any more exact calculation than Table 16 affords, we may see what share of the quartile changes is due to differences among the groups as wholes. In the lowest quartile we note that

Madison loses 21
Milwaukee gains 10
Small high schools gain 11

while in the highest quartile

Madison gains 8
Milwaukee gains 3
Small high schools lose 11

A grand total of sixty-four quartile changes in first and fourth quartiles, due to shifting of whole school groups.

From Dearborn's table (page 14), we note that there are retained

In the first quartile 76
In the fourth quartile 54
A total of 130 retained in these two quartiles.

The number in each quartile is 118, thus making 236 the number in both. If 130 retain their position, there are 106 as the total number of changes in the two quartiles. Of these there are sixty-four, or sixty per cent, which may be accounted for by differences in standards among the schools, *as groups*. It is scarcely to be expected that a calculation carried through

the two middle quartiles would change this percentage greatly. The diagram, Fig. 2, indicates that the reversal of position is about equally pronounced all along the line.

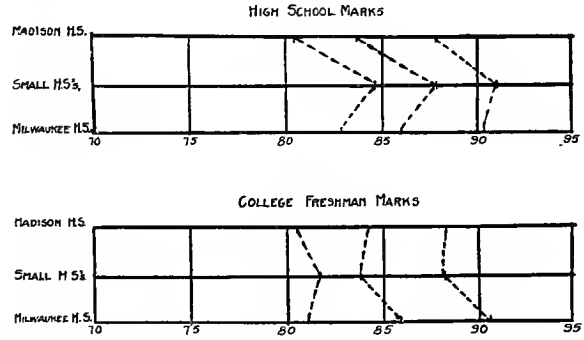

Fig. 2. (Data from Dearborn.) Division points between quartiles in the distribution according to high school marks and again in the distribution according to college freshman marks of the same pupils.

Summarizing, it seems to me that in a study supporting the plan of entrance to college by accreditment, a plan which regards each high school as the final judge of the fitness of its students for college, a factor which is great enough to account for from fifty-five per cent to sixty per cent of the changes in rank (which changes are the bases of the study), should not have been disregarded. Furthermore, the plan of using averages of a number of teachers' marks to indicate student strength hides the most serious defect of the plan of accreditment, namely, the wide differences in standards among the teachers of any subject in the various accredited high schools.

Whether or not the conclusion reached by Dearborn is sound, we cannot overlook the fact that in his carefully tabulated data is a fund of information which must be appraised. There is significance in the determination of how many pupils retain their quartile position from high school to college even when average of four years' work makes the high school rank in every case. As the easiest way to bring together the facts which we wish to evaluate, I have averaged the quartile retentions for each of the groups compared in the study, and have listed the averages in the following table, No. 17. To make clear the derivation of these "average quartile retentions," let me take the case of the

first one in the table, 45.2. On page 14 of Dearborn's study is the table indicating the quartile retention in university freshman class of the 472 students from eight high schools. Of the lowest quartile in high school (that is, the lowest 118 pupils) 64.4 per cent are in the poorest quartile in freshman college work. Of the second quartile, in high school, 39.8 per cent are found in the second quartile in college freshman work. Of the third quartile, 31.4 per cent, and of the fourth quartile, 45.8 per cent, are in the corresponding quartile in college freshman work. The average of these four per cents is 45.2. The table will now be clear.

TABLE 17

DATA CONCERNING QUARTILE RETENTION IN COLLEGE OF THE REPRESENTATIVES OF EIGHT HIGH SCHOOLS IN WISCONSIN (Tabulated from Dearborn)

GROUPS COMPARED			
High School Group	*College Group*	*No. of Pupils*	*Avg. Quartile Retention*
8 H. S.'s, Genl. Avg.	Freshman, Genl. Avg.	472	45.2
" "	Sophomore "	357	43.6
" "	Freshman " *	180	40.7
" "	Sophomore " *	180	47.2
" "	Junior " *	180	41.0
" "	Senior " *	180	41.5
Madison "	Freshman "	238	43.7
" "	Sophomore "	188	45.7
Milwaukee "	Freshman "	139	42.7
" "	Sophomore "	99	42.7
4 small H. S.'s "	Freshman "	92	35.7
" "	Sophomore "	82	40.5
8 H. S.'s English	Freshman English	255	36.5
" Mathematics	" Mathematics	216	33.7
" German	" German	189	42.2
" History	" History	219	45.2
Madison, Genl. Avg.	" Genl. Avg. *	115	41.0
" "	Sophomore " *	115	47.5
" "	Junior " *	115	33.0
" "	Senior " *	115	44.5
Madison Mathematics	Freshman Mathematics	181	51.2
" English	" English	244	36.0
" German	" German	126	42.5
" History	" History	97	37.5
" Latin	" Latin	39	46.0
" Mathematics	" Mathematics	69	34.5
Milwaukee English	" English	92	46.2
" History	" History	64	48.5
4 small H. S.'s English	" English	53	33.5
" History	" History	42	33.2
Average (counting all of equal weight)			41.43

* Those who completed the college course.

SUMMARIZING AND AVERAGING GROUPS FROM THE ABOVE TABLE

8 H. S.'s Genl. Avg.	Freshman Genl. Avg.		45.2
Madison "	" "	Average	40.7
Milwaukee "	" "		
4 small H. S.'s Genl. Avg.	" "		
8 H. S.'s Genl. Avg.	Sophomore Genl. Avg.		43.6
Madison "	" "	Average	42.9
Milwaukee "	" "		
4 Small H. S.'s Genl. Avg.	" "		
(For groups finishing college)			
8 H. S.'s Genl. Avg. and Fresh., Soph., Junior, and Senior			42.6
Madison " " " " " " "			41.4
(Individual subjects in H. S. and College)			
8 H. S.'s English	Freshman English		36.5
Madison "	" "	Average	38.6
Milwaukee "	" "		
4 small H. S.'s English	" "		
8 H. S.'s History	Freshman History		45.2
Madison "	" "	Average	39.7
Milwaukee "	" "		
4 small H. S.'s History	" "		

From the above table it will be observed that the average quartile retentions range from 51.2 to 33.0. As a central tendency for all these retentions the rough average was calculated by simply giving each figure for retention its face value. The average thus determined is 41.43.

The question which these data present to us is this: How satisfactory is an average quartile retention of 41.43 per cent? To be sure, no definite answer can be given, but it is possible to consider the question and get a clearer idea of it than appears on the surface.

It will be noted first that in a quartile arrangement an absolutely random redistribution would result in a quartile retention of 25 per cent. In our retention of 41.43 per cent we have evidence that 16.43 per cent more than a random redistribution would make, retain the same quartile position in college classes that they had in the average of high school. In other words, we have such a quartile retention that 60.3 per cent is accounted for by a random redistribution. I state it in this way so as to make it comparable with results secured in Clement's and Gray's studies to be considered later.

In the second place, a glance at the summary at the bottom of the table will indicate another feature of this retention. It will be observed that whenever the general averages of freshman marks in all subjects are considered, the retention is greater for the group of eight high schools taken together than for each high school group taken separately. In this connection I must repeat what was said in criticism of using correlation of averages to support the accreditment plan of college admission, that the farther the groups being compared are removed from the ranks or marks given by individual teachers the closer is the correlation between them. If, however, the marks of individual teachers were a close approximation of student ability in the subject, then the nearer the groups would approach to individual teachers' marks in allied subjects, the closer would be the correlation. When we turn to the retention indicated for separate subjects, we find that while they are in practically every case lower than the general averages, those in which the eight high schools are grouped together are on the whole a little higher than those for the individual school groups. In this case we have an indication of different standards of marking in the different high schools composing each group, so great that it cannot be counterbalanced without the use of averages of several subjects.

We are now ready to consider the evidence bearing upon standards of rating pupils in the high school which is given in the two most recent studies of the subject, the one by Clement, and the other by Gray. In both of these the plan of comparing marks of pupils with marks given the same pupils in later years is used, but there is a minimum of averaging, and little combination of several schools into one group. Thus the fallacies pointed out in Dearborn's work are avoided in these, and we have the task of evaluating the wealth of material which these two studies provide.

Clement used the records of nearly 5,000 high school graduates, mostly in Kansas schools. Twenty-three high schools of representative sorts were included. The records of as many of these 5,000 high school pupils as possible were traced back into the grammar school and forward into college. Of course relatively few could thus be traced, but nevertheless the long list of comparisons which he was able to make affords the richest mine of information concerning marks that we have.

His method was to compare each group with itself in some later class, indicating in the comparison the per cent who retained their original tertile position. To make this plain I have here reproduced one of the tables of comparison which he uses. From this it will be clear that of the thirty-seven pupils who were in the first tertile in seventh grade history, 22 were in the first tertile in eighth grade history, eleven in the second tertile, and four in the third tertile. The retention, then, is twenty-two out of thirty-seven, or 59.45 per cent. The total or average retention of the whole class is seen to be 51.78 per cent. It is this average tertile retention which represents the most significant fact for our purposes, and we shall, therefore, assemble into one series of tables for easy study the figures representing tertile retention from group to group which are scattered through Clement's study.

History, School 5, Eighth Grade

		1	*2*	*3*	*% Tertile Retention*
History, School 5, Seventh Grade	1	22	11	4	59.45
	2	11	15	12	39.49
	3	4	12	21	56.75
Total Retention.....	.....	.....	.....	.....	51.78

While tabulating the tertile retentions we shall also tabulate tertile division points, those points below which one third of the group fall, and above which another third of the group fall, so as to make easy the comparison of range of ratings between school and school, or group and group.

Clement uses a second method of indicating retention of position, a method which he calls the "modified median method." For this index he calculates the percentage of the lowest third who remain below the median in the next grouping, and the percentage of the highest third who remain above the median in the next grouping, and then averages the two percentages. Wherever this method was used by Clement, I have copied his figures in the column headed "Median Retention."

These data are all given in the following tables, numbered 18, 19, 20, and 21:

TABLE 18

DATA FOR MARKS GIVEN SUCCESSIVE CLASSES IN THE GRAMMAR SCHOOLS

(Tabulated from Clement)

SCHOOL	PUPILS	CLASSES COMPARED	DIVISION POINTS BETWEEN	DIVISION POINTS TERTILES	AVG. TERTILE RETENTION	MEDIAN RETENTION
No. 5	112	7th history	88.1	94.1		
		8th history	89.1	94.1	51.76	79.76
No. 5	112	7th English	85.8	91.0		
		8th English	86.4	90.1	54.46	77.05
No. 5	112	7th arithmetic	87.4	92.5		
		8th arithmetic	90.5	94.9	45.53	67.56
				Average	50.58	74.79
No. 5	112	7th English				
		7th history			51.78	
No. 5	112	7th English				
		7th arithmetic			49.10	
No. 5	112	8th English				
		8th history			53.57	
No. 5	112	8th English				
		8th arithmetic			52.67	
				Average	51.78	
No. 7	78	7th English	86.2	93.2		
		8th English	85.3	91.2	51.28	76.73
No. 7	78	7th arithmetic	83.9	92.6		
		8th arithmetic	86.6	92.9	56.41	74.98
				Average	53.84	75.85

TABLE 19

DATA FOR MARKS GIVEN TO SUCCESSIVE CLASSES IN HIGH SCHOOLS
(Tabulated from Clement)

SCHOOL	PUPILS	CLASSES COMPARED	DIVISION POINTS BETWEEN TERTILES		AVG. TERTILE RETENTION	MEDIAN RETENTION
No. 8	126	Fresh. English	84.4	90.4		
		Soph. English	83.4	90.8	59.52	88.09
No. 8	126	Soph. English	83.4	90.8		
		Jun. English	80.3	90.4	73.01	90.47
No. 8	114	Fresh. Latin	85.2	92.1		
		Soph. Latin	80.5	90.6	65.78	86.83
No. 8	125	Fresh. math.	84.1	90.7		
		Soph. math.	80.6	90.0	48.00	71.44
				Average	61.58	84.20
No. 9	160	Fresh. English	82.9	88.4		
		Soph. English	81.8	87.4	52.50	72.63
No. 9	160	Soph. English	81.8	87.4		
		Jun. English	80.4	85.5	60.00	75.45
No. 9	93	Fresh. Latin	83.6	89.9		
		Soph. Latin	79.4	84.6	54.83	70.96
No. 9	117	Fresh. math.	83.0	90.5		
		Soph. math.	78.9	85.5	58.11	87.17
				Average	56.36	76.55
No. 5	212	Fresh. English	81.6	87.9		
		Soph. English	81.6	86.5	49.52	76.05
No. 5	212	Soph. English	81.6	86.5		
		Jun. English	80.4	86.5	55.66	80.84
No. 5	217	Fresh. Latin	85.0	90.4		
		Soph. Latin	78.5	86.3	52.99	81.24
No. 5	212	Fresh. math.	85.6	90.3		
		Soph. math.	81.6	89.7	58.49	76.80
				Average	54.16	78.73
Nos. 8, 9 and 5	633	Fresh. English	82.9	88.7		
		Soph. English	82.0	88.2	54.50	
Nos. 8, 9 and 5	633	Soph. English	82.0	88.2		
		Jun. English	80.5	86.5	52.76	
Nos. 8, 9 and 5	467	Fresh. Latin	84.6	90.8		
		Soph. Latin	79.6	87.6	58.68	
Nos. 8, 9 and 5	589	Fresh. math.	84.8	90.6		
		Soph. math.	81.4	89.4	55.68	
				Average	55.40	

TABLE 20

DATA FOR MARKS GIVEN TO CLASSES IN GRAMMAR SCHOOL AND HIGH SCHOOL

(Tabulated from Clement)

SCHOOL	PUPILS	CLASSES COMPARED	DIVISION POINTS BETWEEN TERTILES		AVG. TERTILE RETENTION	MEDIAN RETENTION
No. 5	212	8th English Fresh. English	86.4 81.6	90.5 87.9	46.17	66.18
No. 5	212	8th English Soph. English	86.4 81.6	90.5 86.5	44.33	64.78
No. 5	212	8th English Jun. English	86.4 80.4	90.5 85.5	53.17	71.83
No. 5	212	8th English Sen. English	86.4 80.9	90.5 86.4	43.39	67.60
No. 5	212	8th arithmetic Fresh. math.	88.7 85.6	93.8 90.4	44.81	61.96
No. 5	212	8th arithmetic Soph. math.	88.7 81.5	93.8 89.7	41.50	68.30
No. 5	212	8th history Soph. history	89.2 82.4	93.8 88.0	48.11	73.93
No. 5	181	8th English Fresh. Latin	86.4 83.0	90.9 89.1	50.82	71.66
No. 8	126	8th English Avg. 3 yrs. H. S. English	88.3 82.5	93.8 90.1	46.30	71.43
No. 10	150	8th English Avg. 3 yrs. H. S. English	82.9 88.2	89.7 91.3	48.00	73.00
No. 7	270	8th English Soph. English	84.3 86.0	90.4 92.6	45.74	69.94
No. 7	270	8th English Fresh. English	84.3 85.2	90.4 92.6	43.70	67.21
No. 7	270	8th arithmetic Fresh. math.	83.8 86.2	91.4 92.7	48.51	69.99
No. 6	338	8th arithmetic Avg. Fresh. and Soph. math.	(Coarse grouping of marks makes these division points uncertain)		46.15	69.02
No. 6	302	8th English Avg. Fresh. and Soph. Latin			53.31	74.75
No. 6	338	8th English Avg. Fresh. and Soph. English			56.21	77.87

TABLE 20—*Continued*

DATA FOR MARKS GIVEN TO CLASSES IN GRAMMAR SCHOOL AND HIGH SCHOOL

(Tabulated from Clement)

SCHOOL	PUPILS	CLASSES COMPARED	DIVISION BETWEEN	POINTS TERTILES	AVG. TERTILE RETENTION	MEDIAN RETENTION
No. 2	97	7th arithmetic Avg. Fresh. and Soph. math.			53.60	
No. 2	72	7th English Fresh. Latin	(Schools 2, 3, and 4 in the same city complete grammar school with 7th grade		46.05	
No. 2	97	7th English Fresh. English			50.51	
No. 2	97	7th English Avg. Fresh. and Soph. English			56.70	
No. 3	93	7th arithmetic Avg. Fresh. and Soph. math.			56.98	
No. 3	78	7th English Fresh. Latin			55.12	
No. 3	93	7th English Fresh. English			59.13	
No. 3	93	7th English Avg. Fresh. and Soph. English			46.23	
No. 4	73	7th English Avg. Fresh. and Soph. Latin			43.97	
No. 4	73	7th English Fresh. English			47.94	
No. 4	73	7th English Avg. Fresh. and Soph. English			46.57	
No. 4	73	7th arithmetic Avg. Fresh. and Soph. math.			43.14	
Nos. 2, 3 and 4	299	7th English Avg. Fresh. and Soph. English	80.5 79.0	88.8 86.0	42.47	66.00
Nos. 2, 3 and 4	299	7th arithmetic Avg. Fresh. and Soph. math.	81.7 75.1	89.0 83.0	43.00	65.50
Nos. 2, 3 and 4	166	7th English Avg. Fresh. and Soph. Latin	83.5 76.0	89.8 85.3	46.67	68.27
		Average (giving equal weight to each group regardless of size)			48.36	69.43

TABLE 21

DATA FOR MARKS GIVEN TO CLASSES IN HIGH SCHOOL AND COLLEGE

(Tabulated from Clement)

SCHOOL	PUPILS	CLASSES COMPARED	DIVISION POINTS BETWEEN TERTILES		AVG. TERTILE RETENTION
H. S. No. 1	266	H. S. Fresh. English	91.2	95.0	
Col. No. 1		Col. Fresh. English	84.0	92.0	50.00
H. S. No. 1	266	Avg. 3 yrs. H. S. English	89.0	94.0	
Col. No. 1		Col. Fresh. English	84.0	92.0	59.02
H. S. No. 1	86	Avg. 3 yrs. H. S. English	91.5	96.0	
Col. No. 1		Avg. 4 yrs. Col. English	84.0	92.0	60.46
H. S. No. 5	81	H. S. Fresh. English	82.6	87.9	
Col. No. 2		Col. Fresh. English	80.4	88.5	35.80
H. S. No. 5	81	H. S. Soph. English	82.6	87.4	
Col. No. 2		Col. Fresh. English	80.4	88.5	53.08
H. S. No. 5	81	Avg. 4 yrs. H. S. English	82.6	86.7	
Col. No. 2		Col. Fresh. English	80.4	88.5	45.67
H. S. No. 5	60	Avg. Fresh. and Soph. H. S. Math.	85.5	90.5	
Col. No. 2		Col. Fresh. Math.	76.5	84.3	40.00
H. S. No. 7	84	H. S. Fresh. English			53.57
Col. No. 1		Col. Fresh. English	These schools used only three marks, and the tertile retention was calculated on basis of these marks.		
H. S. No. 6	184	H. S. Fresh. math.			60.32
Col. No. 3		Col. Fresh. Math.			
H. S. No. 6	165	H. S. Fresh. English			43.00
Col. No. 3		Col. Fresh. English			
		Average (giving equal weight to all groups)			50.09

SOME AVERAGES FROM THE ABOVE TABLES TO INDICATE INFLUENCE OF DIFFERENCE OF STANDARDS AMONG THE SCHOOLS MAKING UP A GROUP, UPON TERTILE RETENTION

First, from Table 19:

The average of all retentions by single subjects from year to year in Schools 8, 9, and 5 taken separately	57.4
The average of all retentions by single subjects from year to year in Schools 8, 9, and 5 taken together	55.0

Second, from Table 20:

The average of all retentions from seventh arithmetic to the average of freshman and sophomore mathematics in Schools 2, 3, and 4 taken separately	51.2
The average of all retentions from seventh arithmetic to the average of freshman and sophomore mathematics in Schools 2, 3, and 4 taken together	43.0
The average of all retentions from seventh English to the average of freshman and sophomore English in Schools 2, 3, and 4 taken separately	49.8
The average of all retentions from seventh English to the average of freshman and sophomore English in Schools 2, 3, and 4 taken together	42.5

In connection with the foregoing tables we must ask the same question as was asked concerning the data in Dearborn's study: How satisfactory is the retention here indicated? Between successive classes in grammar school there is a tertile retention of slightly more than 50 per cent. It is no greater on the average than the retention between different subjects taken during the same year, however, which indicates that a pupil who is good in history, say, this year is as likely to be found among the good pupils in arithmetic this year as he is to be found among the good pupils in history next year.

Among successive classes in high school we find a little higher tertile retention. This may be accounted for in part perhaps by the fact that most of the work in high schools is done departmentally so that the class in freshman Latin, say, this year will be taught sophomore Latin next year by the same teacher. In that case the personal equation would weigh in the same direction in successive years, and work to increase tertile retention. At any rate, there seems to be a retention averaging about 57 per cent between the same subjects in successive years in high schools.

Turning to the retention in high school of grammar school ranks, we find much lower figures. For all the schools considered the average retention is a little below 50 per cent. It may be urged that this is probably caused by the abrupt change

in both subject matter and method between grammar school and high school. While this no doubt has some effect it must be remembered that the retention between successive years in grammar school where the subject matter is very closely similar from year to year, was but little higher. The children change teachers from seventh grade to eighth grade, just as they do from eighth grade to freshman high school, and that is probably the greatest reason for the lower retention found between these two groups than occurs between successive years in high school where teachers do not as a rule change.

Between high school and college the retention is, on the average, about 50 per cent. It will be observed that the highest figure in the list is that for the average of all high school with the average of all college English, an evidence again of the oft-noted fact that averaging marks for several years or several subjects tends to cover up the most serious fault of our present marking system.

It appears, then, that the tertile retention for all classes in all schools and between the various schools is a little above 50 per cent. Now how satisfactory is a tertile retention of 50 per cent? Bearing in mind that a perfectly random redistribution at each successive marking would produce a tertile retention of 33.3 per cent, we have in this retention of 50 per cent such a retention that a random redistribution accounts for 66.7 per cent of it. If we use the figures given for "modified median retention," we note that on the whole they run a little less than 75 per cent. By this method of calculating retention, a random redistribution would produce 50 per cent retention. Here again, then, we have evidence that chance accounts for 66.7 per cent of the retention. It seems fair to make a comparison on this basis, with Dearborn's data for certain Wisconsin high schools. It was found that chance accounted for but 60.3 per cent of the retention there, although comparisons were made only between high school and college marks.

Before leaving Clement's study attention may be called to the list of tertile division points. It will be seen that not only are there some rather marked differences in the standards of marks used among the various departments of the same school as well as among like subjects in various schools, but there is a most consistent tendency to reduce the marks from one school

to a higher school. That is to be expected, however, when we consider that the poorer members of the group drop out before they enter the next higher school, and of course a normal distribution made for the members of each successive group would tend to distribute ever lower and lower those members of the original group who persisted to the end. I call attention to this fact here because it has a very definite bearing upon the question as to whether we should plan for a more and more skewed distribution as we advance in the grades where elimination takes place.

In Gray's study we have the most significant type of data yet gathered bearing upon the subject of marks. He considers the individual records of pupils from ten different high schools, mostly in Indiana. He does not tell[1] how many records enter into his conclusions, but it is fair to assume that he used enough to make his figures valid. Neither is it stated that only high school graduates were used, but that fact is implied throughout, and I shall act upon that assumption.

The method used by Gray was that of calculating the number of *points change* which occurred from one mark of a pupil to the next in the same subject in the high school. For example, if a mark of 80 was received in freshman history and 85 in sophomore history, a change of 5 points was recorded for that promotion. Similarly all changes were recorded and averages struck for each school in each department of study.

Incidental to this main purpose Gray pointed out many irregularities and variations in standards of grading as well as forms of distributions of marks which were to be found in the several schools and departments. These furnish most impressive evidence of the need of some method of standardization in high school work, but I shall not undertake detailed comment upon them. I shall rather confine my attention to his main

[1] Upon direct inquiry from Mr. Gray I learn that the following numbers of pupils from each school entered into his records and that all had completed high school except 75 in School 2:

School 1	140	School 6	25
School 2	100	School 7	25
School 3	135	School 8	30
School 4	35	School 9	25
School 5	25	School 10	30

tables in which he summarizes the variations in marks above referred to. From his tables we assemble the data in Table 22.

TABLE 22

AVERAGE VARIATION IN POINTS AT EACH PROMOTION FOR EACH PUPIL IN EACH SUBJECT (Tabulated from Gray)

	No. of Schools	Avg. Var. in Points	Schools 1 and 3	Other Schools
English	10	4.0	3.95	4.04
History	9	3.8	3.85	3.77
Mathematics	10	4.3	4.65	4.22
Latin	9	3.5	3.55	3.53
Mod. Lang.	9	3.2	3.70	3.02
Science	10	4.7[1]	4.30	4.87
Averages		3.92	4.00	3.91

To discover whether the larger numbers of pupils studied in Schools 1 and 3 than in the other schools makes the results from them differ widely from the other schools, I separated them from the rest and compiled column 2 of the table.

In trying to answer the question, "How satisfactory is a variation of 3.92 points?" we find it difficult to get a satisfactory basis of comparison with the results given by Clement and Dearborn. In the effort to make such comparison we have worked upon an assumption which is not capable of absolute proof. For those who may wish to discount the assumption, however, this statement will form a basis of comparison which will be more helpful than no basis.

In the following discussion I shall try to answer the question, "What per cent is 3.92 points variation, of the variation which would occur by a perfectly random redistribution at each instance of remarking by the teachers?" To do this I have to make an assumption of the typical range of marks which these classes would fall into, and also the form of distribution into which they would fall. Several elements enter to guide this assumption. First, if the list is confined to those who continue four years in high school, the distributions of the early high school classes will be bunched pretty high on the scale, and therefore not make a wide distribution. In the second place, the large number of graphs given by Gray represent in almost every instance no cases below seventy-five, and I therefore

[1] Gray's text errs in making this figure 3.7.

assume that the schools had seventy-five as a passing mark. In the third place, an examination of the graphs published by Gray reveals that from 80 per cent to 90 per cent of the cases there represented do actually fall, in the majority of instances, within a range of fifteen marks. On these grounds I have assumed that the groups from which Gray's cases were drawn distribute themselves midway between the two forms recommended as most appropriate by J. McK. Cattell,[1] and Max Meyer,[2] in which each of the five steps of the scale is considered as spread over five points as follows:

	73 to 77	78 to 82	83 to 87	88 to 92	93 to 97
Cattell	10%	20%	40%	20%	10%
Meyer	3%	22%	50%	22%	3%

For the sake of simplicity, consider a class of 100 pupils distributed according to the Cattell distribution. Upon a random redistribution at promotion time, or rather the time of remarking, they would be found in the following arrangement:

	73 to 77	78 to 82	83 to 87	88 to 92	93 to 97
Lowest 10	1	2	4	2	1
Next 20	2	4	8	4	2
Middle 40	4	8	16	8	4
Next 20	2	4	8	4	2
Highest 10	1	2	4	2	1

To obtain these positions, the following number of points changes or variations would be sustained:

Lowest 10	0	10	40	30	20	
Next 20	10	0	40	40	30	
Middle 40	40	40	0	40	40	
Next 20	30	40	40	0	10	
Highest 10	20	30	40	10	0	
Totals	100	120	160	120	100	600

It is evident, then, that these 100 pupils if distributed according to the Cattell scheme would make a total of 600 points changes with a random redistribution, or an average of six points per pupil.

[1] J. McK. Cattell, Examinations, Grades and Credits, *Pop. Sci. Monthly*, 66:367–378.

[2] Max Meyer, The Grading of Students, *Science* (n. s.), 28:243–252.

Consider now the Meyer distribution in the same way. A random redistribution of the 100 pupils would produce the following arrangement of the members of each group:

Lowest 3	.09	.66	1.5	.66	.09
Next 22	.66	4.84	11.0	4.84	.66
Middle 50	1.50	11.00	25.0	11.00	1.50
Next 22	.66	4.84	11.0	4.84	.66
Highest 3	.09	.66	1.5	.66	.09

To accomplish the above arrangement, the following amount of changes in points would be involved:

Lowest 3	0	3.3	15.0	9.9	1.8	
Next 22	3.3	0	55.0	48.4	9.9	
Middle 50	15.0	55.0	0	55.0	15.0	
Next 22	9.9	48.4	55.0	0	3.3	
Highest 3	1.8	9.9	15.0	3.3	0	
Totals	30.0	116.6	140.0	116.6	30.0	433.2

From this table it appears that a random redistribution of 100 pupils arranged after the Meyer plan produces 433 points changes, or an average of 4.33 points per pupil.

It seems fair to take the average of the two figures obtained from these two calculations, 5.17 (that is, 6 plus 4.33, divided by 2) and consider it the number of points change which would accompany a chance arrangement of grades at each remarking of Gray's people. Now we have a basis of comparing the retention of position in these schools taken pupil by pupil, subject by subject, with the retention found by Clement, who combined several classes to make his groups, and with Dearborn, who used the averages of several years' marks. The average number of points change actually found is 3.92 per pupil. A chance redistribution would make 5.17 points change per pupil. We have, then, but 25 per cent improvement over a chance redistribution.

In the case of Dearborn's data, we were able to make the statement that the retention was such that chance accounted for 60.3 per cent of it. With Clement's data, chance accounted for 66.7 per cent of the retention. While we cannot make a similar statement for Gray's data we can get a fairly clear idea of how it compares in respect to retention by saying that the points changes per pupil are 75.8 per cent as great as they would be by chance. I recognize that it is dangerous to press this

comparison far. It cannot be demonstrated that it is at all a sound basis of comparison, but is the only way that we may think a relationship among them. Unfortunately the data are not so arranged that a coefficient of correlation can be calculated for each resemblance. By means of this basis of comparison we can at least see that the close correlation which we should expect between marks in successive years in the same subject does not exist, but that on the contrary, the more estimates we average to get a pupil's rank, the closer the correlation. In other words, just as Clement found in his few classes, teachers' marks in any subject are an index of general ability quite as much as they are an index of special ability in the given subject. This surely points to a sort of dead level of student interest in high school.

But turning now to the actual retention found, are we ready to accept the standard which Clement says we may judge our schools by, namely, a tertile retention of approximately 50 per cent? If we can come no nearer than that in ranking our children for general ability, we cannot hope to command much respect as a teaching profession. Rather should the revelations made by these studies open our eyes to the real need for some more effectual method for establishing standards whereby both teachers and pupils may measure progress. No more striking illustration of the far-reaching effect of having no definite standards could be found than just what these studies reveal: teachers do not draw out special abilities from their high school pupils. No more fruitful source of discouragement and of elimination exists to-day than just that failure to find and develop the special interests of the pupils.

STANDARDS OF MARKING IN COLLEGES

The non-uniformity of standard of marking among the instructors in colleges was first brought forcibly to public attention by Professor Max Meyer[1] in the University of Missouri. He collected all the marks for a period of five years of forty instructors, mostly in the College of Arts and Sciences, all but two of whom had the rank of professor or assistant professor. The marks were all in terms of the uniform system, A, B, C, D, and E. D meant failure with the privilege of another examination, and E meant failure without such privilege. Meyer combined the D's and E's, using the letter F for the combined group. He then tabulated for each instructor the number of classes he had taught during the five years, the total number of marks he had given, and the per cent which he had used of each letter, A, B, C, and F. In addition to these facts, he calculated also the coefficient of variability in the giving of each letter from class to class. This coefficient is derived by dividing the average variation from the average percentage which any professor assigns a given mark, by the average percentage assigned that mark. For example, the philosophy professor listed in the table gave 55 per cent of his people A, on an average, but he varied from class to class by an average variation of 11 per cent. Therefore, the coefficient of variability is 11/55, or .2. These data for the half dozen instructors at either extreme of the list are reproduced in Table 23.

The need for the reform in marking, which was effected in the University of Missouri shortly after Meyer's investigation, is evident from the above table. It is not to be supposed, however, that Missouri was exceptional in this absence of uniformity. There have been enough similar investigations in other institutions to prove that just such variation is the rule among college instructors.

We are not much surprised by facts brought out by Meyer. In fact there still persists a very general feeling that college instructors should be allowed practically absolute freedom to

[1] Max Meyer, The Grading of Students, *Science*, 28:243–252.

TABLE 23

SHOWING THE VARIABILITY OF MARKING BY INSTRUCTORS IN THE UNIVERSITY OF MISSOURI (From Meyer)

INSTRUCTOR IN	%A	%B	%C	%F	TOTAL MARKS	NO. OF CLASSES	COEFFICIENTS OF VARIABILITY			
							A	*B*	*C*	*F*
Philosophy	55	33	10	2	623	29	.2	.3	.8	1.2
Latin I	52	42	6	0	130	9	.3	.3	1.2	—
Sociology.....	52	30	13	5	958	47	.3	.5	.9	.9
Mathematics I	40	31	16	13	208	19	.6	.6	.8	.9
Economics....	39	37	19	5	461	28	.4	.4	.7	.9
Greek.....	39	26	24	11	287	30	.4	.4	.5	.9
Average.....	46	33	15	6						
Engineering I .	13	36	42	9	813	39	.6	.3	.2	1.0
Mech. Drawing	18	29	41	12	558	28	.4	.4	.3	.9
Mechanics.....	18	26	42	14	495	12	1.1	.3	.3	.4
Engineering II	16	26	46	12	826	?	.3	.3	.3	.9
English II.....	9	28	35	28	1098	44	.8	.3	.3	.4
Chemistry III .	1	11	60	28	1903	12	1.0	.6	.1	.3
Average.....	12.5	26	44	17						

conduct their classes in any way they see fit, and so we rather expect to see individual standards manifested in the marks given. It should be kept in mind, however, that the adoption of some method whereby a given mark may signify more nearly the same merit in the several departments, is not a restraint upon that cherished independence.

Since all the studies made in this field point to the same variation, it seems unnecessary to do more than indicate the institutions where such investigations have been made. This will suffice to establish the claim that standardization is as much needed in college as in high school or elementary school.

In the appendix to Dearborn's "School and University Grades" is a series of tables setting forth in great detail the distributions of marks given at the University of Wisconsin. William T. Foster[1] worked out with similar care the marks given at Harvard. His graphical representations tell a very plain story of the situation there. While we scarcely need a proof of the contention that low standing in a course is not prophetic of failure in one's career, yet the table indicating the undergraduate marks received by men of honor standing in the professional schools shows pretty plainly where the relationship does hold. Foster

[1] William T. Foster, Scientific *vs.* Personal Distribution of College Credits, *Pop. Sc. Mo.*, 78:378–408.

gives also the significant facts concerning the variation among instructors' marks at the University of California.

In the 1910–11 report of the President of the University of Chicago, pages 91 to 94, we have a table indicating the wide variations among marks given by the different instructors in that institution.

Edwin E. Slosson[1] after examining the situation at Amherst College records his conviction that the marking there tells more about the instructors than about the students.

In 1905 Cattell[2] tabulated a few of the markings of Columbia University instructors as a basis for his recommendation concerning a proper type of distribution. A far more exhaustive study of the marking system of Columbia was made in 1906, however, by Miss Mary T. Whitley.[3] From her report, which appeared in the form of a Master's essay, it is clear that Columbia stood at that time high in the list of institutions giving to instructors a maximum of individual liberty.

By all these studies the significance of President Foster's question is emphasized: "Can the personal equation as the chief factor in the awarding of college marks be supplanted by scientific guidance?" A partial answer to this question is what is attempted in a later section of this discussion. First, however, we must evaluate our present common means of standardization, the examination paper.

---

[1] E. E. Slosson, A Study of Amherst Grades, *Independent*, April 20, 1911.

[2] J. McK. Cattell, Examinations, Grades and Credits, *Popular Science Monthly*, 66; 367–378.

[3] Mary T. Whitley, Statistical Study of College Marks, Master's essay, Teachers College, 1906.

THE MARKING OF EXAMINATION PAPERS

The use of examination papers as a means of measuring knowledge, or efficiency, or mental ability, or whatever name may be given to that which is supposed to indicate one's fitness for a particular grade of work is almost a universal custom in our schools. It is being extended more and more each year to civil service and industrial positions. In spite of this the few studies which have been made reveal a very wide difference of rating upon the same paper among supposedly competent judges. We shall not in this section attempt to analyze the situation to determine the causes of variation among judges. We shall merely indicate how reliable examinations in actual practice are, in order to have some basis for our expectations concerning the use of standard tests or scales for evaluating papers.

F. Y. Edgeworth, professor of Political Economy at the University of Oxford, was among the first to call wide attention to this variation. The care with which his first experiment was conducted justifies a full statement of it here. In 1889 he inserted a specimen of Latin prose composition in the English *Journal of Education* accompanied by a request that competent persons rate the paper. Quoting from his article: "I propose, through the medium of the *Journal of Education*, to invite any competent person to assign a mark to the subjoined piece of Latin prose, upon the supposition that he is marking the work of a candidate for the India Civil Service. Let it be distinctly understood that in giving his mark the examiner is not to look to, or wish to illustrate, his own ideal of classical elegance nor yet the degree of proficiency which may be current in the school or other institution with which he may be connected. Let him imagine that he has been appointed examiner in Latin for the India Civil Service, and let him give his mark, having regard only to what may be expected from a candidate for that prize. Let 100 be the maximum attainable by any candidate.

"To avoid accidental divergence as much as possible, to perform the experiment under the most favorable conditions, I

would suggest that the examiners should consist of a pretty homogeneous class—of much the same class as those who actually conduct our public examinations. To be more definite I would invite to take part in this experiment only those who have taken high honors in classics at one of the universities, or classical masters of the sixth form in a public school. All such are earnestly invited to examine the accompanying piece with as much care as if they really were exercising the function of public examiner; and send to the editor their verdict, guaranteed by their name and status, which, it need hardly be added, it is not intended to publish. It is desirable that the examiners should assign their respective marks independently, and without mutual conference."

In response to this appeal, "twenty-eight highly competent examiners were so kind as to mark this piece of Latin prose."

The twenty-eight marks distributed themselves as follows: 45, 59, 67, 67.5, 70, 70, 72.5, 75, 75, 75, 75, 75, 75, 77, 80, 80, 80, 80, 80, 82, 82, 85, 85, 87.5, 88, 90, 100, 100.

While two examiners thought the paper met the requirements perfectly, four others marked it less than 70.

Upon discovering so much divergence among these "highly competent examiners," Edgeworth entered into a very careful study of examinations, giving especial attention to the Civil Service papers. A full account of his work appears in the 1890 report of the Royal Statistical Society. It seems unnecessary to quote from his tables since his reputation as a statistician and economist insures us against any overstatement in his conclusion. His most significant conclusion he states thus: "I find the element of chance in these public examinations (India Civil Service, Army, and Home Civil Service clerkships of the second order) to be such that only a fraction—from a third to two thirds—of the successful candidates can be regarded as quite safe, above the danger of coming out unsuccessful if a different set of equally competent judges had happened to be appointed."

We surely need no other justification for studying further the soundness of our examination system.

In 1911 Allen Mead Ruggles conducted an experiment in marking papers, the results of which are reported in a Master's essay submitted at Columbia University. He had twenty sixth-grade geography papers rated by eleven graduate students

in Teachers College. To indicate the range of marks for each paper, the following table, No. 24, is quoted:

TABLE 24

SHOWING THE MARKS BY EACH OF ELEVEN JUDGES, DESIGNATED BY LETTERS, UPON EACH OF TWENTY GEOGRAPHY PAPERS

(From Ruggles)

PAPERS	X	E	S	B	W	H	A	P	K	C	O	MEDIAN	A. D. FROM MEDIAN
1	40	70	53	37	77	53	63	65	37	57	65	54	11.0
2	21	40	15	10	29	32	30	23	30	60	25	29	8.9
3	33	30	55	13	53	29	40	50	47	35	55	40	10.9
4	27	50	60	28	59	60	48	40	90	72	15	50	18.5
5	24	15	10	14	40	26	26	25	30	34	60	27	10.7
6	63	50	85	45	58	56	40	60	25	35	55	55	12.3
7	29	30	45	21	40	47	30	25	20	20	65	37	12.0
8	59	75	85	38	72	74	55	55	75	45	60	59	11.8
9	27	25	10	53	20	48	35	60	25	30	70	34	14.8
10	36	35	15	25	17	65	31	53	25	48	65	39	14.1
11	39	35	75	40	49	57	35	52	100	44	40	47	13.4
12	58	45	65	47	56	43	42	50	50	50	60	49	5.9
13	28	22	25	30	0	58	50	50	25	59	40	39	15.1
14	49	50	55	53	44	77	59	45	40	69	50	53	7.6
15	45	40	78	41	46	74	47	55	25	67	90	49	15.4
16	57	12	60	20	22	35	46	35	60	26	20	27	15.1
17	53	50	90	54	93	63	46	60	100	39	100	59	18.6
18	67	55	90	50	65	65	58	80	65	48	50	60	10.1
19	43	25	70	40	38	54	44	40	15	43	65	45	11.4
20	53	35	90	47	56	60	60	53	58	51	45	54	8.6
Medians	41.5	37.5	60.0	39.0	47.5	56.5	45.0	51.0	38.5	46.5	57.5	48.0	12.15
A.D. from M.	12.1	13.1	22.7	11.1	17.1	11.1	8.9	10.5	21.1	11.4	14.8		

Median of the A. D.'s on the bottom row 12.1

Note: The A. D.'s are my own calculations.

Rather surprising variations are revealed in this table. Paper 18 has the highest average, 60, and the other papers range down to 27, the mark assigned to papers 5 and 16. However, judge S considers the entire set of papers worth 60 on the average while judge E considers them worth only 37.5. In fact, the median of the average deviations from the median among the marks assigned the same paper by the several judges is just as large as the median of the deviations among the marks of each judge upon the several papers. In other words, there is as much variation among the several judges as to the value of each paper as there is variation among the several papers in the estimation of each judge. And the set of papers are of widely different values too.

Another brief experiment was performed at Columbia by H. Jacoby.[1] He asked six professors of astronomy to mark a set of eleven astronomy papers. The rating was to be done on the

[1] H. Jacoby, The Marking System in the Astronomical Course at Columbia College, 1909–1910, *Science*, 31: 819.

scale of 10, with 7 as the passing mark. One judge misunderstood directions, and so the marks of only five are significant. These marks are reproduced in Table 25.

TABLE 25

THE MARKS OF FIVE JUDGES ON ELEVEN ASTRONOMY PAPERS (Jacoby)

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	
Judge A....	9	7	9	10	7	10	6	9	8	10	9	
Judge B....	9	6.6	9	9.4	6.2	9.8	5.8	9.3	5.7	8.5	9	
Judge C....	8.5	7	8.8	9.9	6.7	9.6	6.3	9.7	9	9.1	9.5	
Judge E....	9	6	8	10	7	10	7	9	10	9	8	
Judge F ...	7.3	6.5	8	9.2	5.9	9.5	5.4	8.8	8.7	9	9	
Maximum Divergence	1.7	1.5	1	.8	1.1	.5	1.6	.9	4.3	1.5	1.5	Avg. 1.5

Thus there appears an average difference of 1.5 on a scale of 10 between the highest and lowest mark on a paper. In the cases of four papers out of eleven some judges would pass the student while other judges would fail the student.

Perhaps the most striking demonstrations of the divergence of marks of teachers upon individual papers have been afforded by the recent experiments by Starch and Elliott[1] at the University of Wisconsin. They used two English papers which were written by two pupils at the end of the first year of high school English, and a geometry paper handed in as an exercise in one of the largest high schools in the state. A facsimile reproduction was made of each paper and printed upon exactly the same sort of paper as that on which the pupil had written it. These sheets, along with a set of the questions to which they had been given as answers, were sent out to the high schools constituting the North Central Association of Colleges and Secondary Schools with a request that the papers be rated by the ones in each high school best qualified to pass upon them, presumably the heads of the departments of English and mathematics, respectively. The replies were treated in two separate groups, those from high schools having 75 as a passing mark, and those having 70. The English papers were then rated by a class in

[1] Starch and Elliott, Reliability of Grading High School Work in English, *School Review*, Sept., 1912. Also, Reliability of Grading Work in Mathematics, *School Review*, April, 1913.

the Teaching of English, in the University of Wisconsin, and by a Summer School class of teachers, in the University of Chicago. The distributions for all these groups of judges are given in Table 25a.

TABLE 25a

DISTRIBUTION OF MARKS UPON THREE PAPERS—"A" AND "B" ARE ENGLISH AND "C" IS GEOMETRY. THE MARKS ARE ALL COMBINED IN CASE OF "C" BY WEIGHTING THE MARKS FROM SCHOOLS HAVING 70 AS PASSING MARK, BY ADDING 3 TO EACH MARK (Data from Starch and Elliott)

	Paper A, Pass. 75	Paper A, Pass. 70	Paper A, U. of Wis. Class Pass. 70	Paper A, U. of Ch. Class Pass. 70	Paper B, Pass. 75	Paper B, Pass. 70	Paper B, U. of Wis. Pass. 70	Paper B, U. of Ch. Pass. 70	Paper C, Pass 75
28									1
39									1
41									1
44									1
48									2
50 to 54									6
55 to 59					1	1			8
60 to 64	2			1	1	2		1	17
65 to 69	1	1			6	6	2	10	19
70 to 74	2	1	1	5	5	11	7	14	13
75 to 79	5	6	1	4	24	9	7	20	27
80 to 84	18	7	6	24	27	13	27	21	11
85 to 89	24	17	16	31	19	5	24	23	7
90 to 94	30	15	40	25	7	3	18	9	2
95 to 100	9	4	22	7	1	1	1		
Total	91	51	86	97	91	51	86	98	116
Medians	88.3	87.2	92.4	86.7	80.4	78.8	84.5	80.5	70.0
Med. Dev.	4.5	4.2	3.0	4.3	4.4	5.8	4.2	5.8	7.5

Note: The median deviations are my own calculations.

These tables may be allowed to speak for themselves. We need to point out only two features: There is a difference of more than five between the median mark given by the high school teachers and the class in the Teaching of English in the case of either English paper; and the chances are about even that, in the case of any group of judges, the paper will be changed five points or more when given from one teacher to the next for rating. I wish to call attention to these two facts because of their similarity with those revealed in the study of the New York Regents Examinations to be reported a little later.

I shall report upon but one other experiment with marking. That experiment is described at the close of Gray's study to

which attention was directed earlier. Gray secured sets of examination answer papers in mathematics and in English from an Indiana high school, and asked five other competent persons besides the class teacher to rate them. These five ratings along with the rating of the class teachers who had furnished the papers are given in Gray's book. I shall give in Table 26 only the average mark for each paper and the average of the variations from the average, and the average of the marks which each judge gave to the entire set. The judges were experienced teachers, and the passing mark was understood to be 70 in each case.

TABLE 26

THE AVERAGE AND AVERAGE DEVIATION AMONG SIX JUDGES OF A SET OF MATHEMATICS AND A SET OF ENGLISH PAPERS, AND THE AVERAGE OF ALL MARKS GIVEN BY EACH JUDGE (Gray)

MATHEMATICS			ENGLISH		
Papers	*Avg.*	*A. D.*	*Papers*	*Avg.*	*A. D.*
1	66.5	8.0	1	59.3	9.6
2	77.3	4.2	2	82.5	9.3
3	86.3	3.1	3	82.7	7.5
4	28.5	9.2	4	77.8	10.8
5	45.5	14.7	5	75.6	6.7
6	57.0	3.0	6	66.8	11.8
7	76.6	7.2	7	71.2	10.8
8	83.5	7.3	8	79.3	7.6
9	67.5	10.2	9	79.3	7.6
10	78.8	4.2	10	77.1	5.3
11	44.5	7.0	11	87.0	12.0
			12	85.3	6.0
Average.....		7.1	13	68.3	11.2
			14	76.5	8.2
			15	76.1	10.5
			16	74.0	8.3
			17	65.3	9.5
			18	62.8	16.8
			19	77.8	8.5
			20	79.8	5.4
			Average.....		9.2

Average of Each Judge's Several Marks	Average of Each Judge's Several Marks
Judge A.....78.7 (original teacher)	Judge A .....80.3 (original teacher)
Judge B.....74.0	Judge B 83.7
Judge C.....61.4	Judge C 78.5
Judge D 65.5	Judge D 54.0
Judge E.....75.5	Judge E 70.0
Judge F.....58.0	Judge F 79.5

Note: The A. D.'s are my own calculations.

In this table we have a greater variation among marks than was found by Starch and Elliott. In the marking of the mathematics papers the judges varied about as much as with the geometry paper in the above study, but in marking the English papers the variation was nearly twice as great as Starch and Elliott found. There is a difference of 20.7 points on the average between the marks of judges A and F of the mathematics set, and a difference of 29.7 points between the averages of judges B and D of the English set. In fact, judge D failed all but one of the papers, while judge B passed all but one, in the English set.

In all of the above studies we see very serious lack of standards among teachers. It is true that in all these cases the judges were selected from an area where no especial effort had been made to standardize the judgments. On this account I undertook to measure the variation between the marks of the teachers in New York state on the one hand and the Regents on the other. Here is a place where through several decades examinations have been given regularly throughout the state, and where there has been not only the opportunity but the necessity of standardizing the judgments of the many teachers as far as the present type of examinations accomplishes such standardization. It is in such a situation that the greatest care is exercised by the teachers because they recognize that they are themselves judged somewhat by the correlation between their own and the regents' marking.

Before giving the results of this study I wish to indicate something of the extent of this system of examinations and some of its tendencies. For the series of years from 1889 to 1895 inclusive, Thomas O. Baker[1] has tabulated the data found in Table 27, page 58, set opposite those years, and the reports of the Department of Education of the State of New York furnished the data for the years 1911, 1912, and 1913.

From this table the extent of the system is apparent. The two tendencies to which attention is called are the constantly increasing per cent of papers which the regents have passed, and, at the same time, the constantly increasing per cent of papers rejected by the regents of those passed by the teachers.

[1] Thomas O. Baker, An Analysis of the Regents' Examinations in Relation to Secondary Schools, Doctor's essay, New York University, New York City, 1896.

TABLE 27

DATA CONCERNING NEW YORK STATE REGENTS' EXAMINATIONS

YEAR	SCHOOLS TAKING EXAMS.	PAPERS WRITTEN	PAPERS CLAIMED BY THE SCHOOLS AS PASSING	PAPERS ALLOWED BY REGENTS AS PASSING	PAPERS REJECTED BY THE SCHOOLS	ADDITIONAL PAPERS REJECTED BY THE REGENTS	PER CENT OF PAPERS PASSED BY REGENTS	PER CENT REJECTED BY TEACHERS	PER CENT REJECTED BY REGENTS OF THOSE PASSED BY THE SCHOOLS
1889...	304	193,197	107,149	99,079	86,048	8,070	51	44.5	7.5
1890...	311	201,488	117,267	107,915	84,231	9,342	53	41.3	7.9
1891...	358	244,979	153,788	146,565	91,191	7,223	59	37.2	4.7
1892...	357	278,907	176,516	155,869	102,391	20,647	56	36.9	11.7
1893...	393	302,471	185,677	165,676	116,794	20,001	55	38.3	10.8
1894...	410	357,908	238,319	211,533	119,589	26,786	59	33.4	11.2
1895...	468	388,945	259,932	231,231	129,013	28,701	59	33.0	11.0
1911...		452,703	363,708	309,608	88,995	54.100	68.3	19.6	14.9
1912...		327,043	273,624	235,768	53,419	39,856	71.5	16.3	14.6
1913...		392,252	319,582	279,035	72,670	40,547	71.1	18.5	12.7

There is, of course, a corresponding decrease in per cent of those rejected by the teachers in the schools. These two tendencies seem to me significant. While an ever-increasing number of pupils in the high schools of the state are able to meet the requirements of the examiners, the difference in standards of judging papers by teachers and examiners grows ever greater. While the requirements for high school teachers are constantly being increased, their judgment of the value of examination papers is being more and more rejected. At the same time that this tendency is present, the custom of accepting without reëxamination the ratings of certain well known teachers is growing among the regents' examiners. This latter custom is used to such an extent, in fact, that in the report of 1913 above, if only the papers were counted which the regents reëxamined, only 60.1 instead of 71.1 per cent would be found to be passed by the regents. In short, we seem driven by the facts here revealed to the conclusion that as the work in the high school becomes richer, the examination paper becomes a less satisfactory means of determining promotion, and we feel more and more the need of objective standards which are capable of consistent interpretation by all good teachers, as a means of measuring progress.

This situation seemed to call for still further investigation. I therefore computed Table 28, from data contained in the 1913 report, State Department of Education, pages 826 to 834:

TABLE 28

THE NEW YORK STATE REGENTS' EXAMINATIONS IN HIGH SCHOOL SUBJECTS, JANUARY, 1912, AND JUNE, 1912

SUBJECTS	PAPERS WRITTEN	% PASSED BY TEACHERS	% REJECTED BY TEACHERS	% PASSED BY REGENTS	% REJECTED BY REGENTS OF THOSE PASSED BY TEACHERS	REGENTS' RATINGS (Of those which the regents reëxamined)			
						Below 60 Failed	*60 to 74*	*75 to 89*	*90 to 100*
English	71,902	86.8	13.2	80.5	7.3	22.2%	45.5	28.0%	4.1%
German	22,459	77.7	22.3	62.9	19.0	38.2	39.7	19.8	2.3
French	9,689	80.6	19.4	67.7	16.0	31.0	46.8	30.6	1.6
Latin	32,522	78.7	21.3	63.9	20.1	35.3	42.7	19.7	2.3
Mathematics	79,786	74.2	25.8	65.1	25.7	37.2	29.9	21.6	11.3
Science	61,989	85.8	14.2	76.9	9.2	27.5	36.2	30.5	5.8
Hist. and Soc. Sci.	46,344	83.0	17.0	71.8	13.5	28.2	47.2	21.3	3.3
Commercial	33,517	77.7	22.3	61.6	20.9	44.9	29.6	21.7	3.8
Drawing	29,848	86.5	13.5	75.1	13.2	24.9	39.4	32.2	3.5
Music	2,485	81.7	18.3	80.1	1.9	19.9	22.8	35.6	21.7
Other Subjects	1,812	89.0	11.0	74.6	16.2				
Total	392,252	81.6	18.5	71.1	12.8	39.6	34.7	21.5	4.2

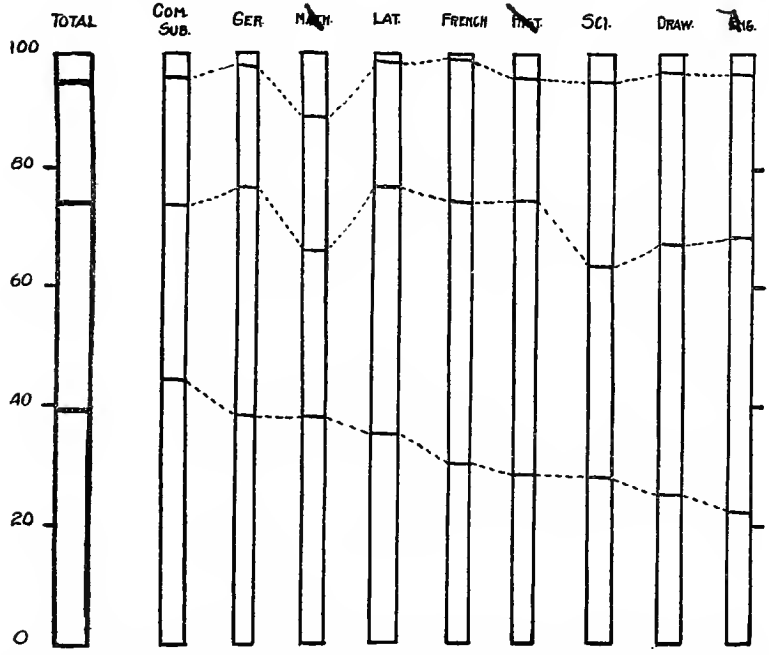

FIG. 3. Representing the per cents of papers in the various departments of study, marked as follows by the regents in 1912: Lowest section, failed; next section, 60 to 74; next section, 75 to 89; top section, 90 to 100. The apparent discrepancy between the total and the several subjects is due to the fact that certain subjects were more liberally excused from reëxamination by the regents than others.

Examination of Table 28 reveals, first of all, a wide variation in the percentages of papers passed in the various subjects, the failures in commercial subjects, for example, being practically double the percentage of those in English. The distribution of the various ratings, four groups being designated, (1) below 60, (2) 60 to 74, (3) 75 to 89, and (4) 90 to 100, shows no special similarity among the various subjects. These data are represented graphically in Figure 3, and there we see at a glance how unequal these examinations must be considered as tests of student ability. If the contention held so generally by students of education to-day has any validity, namely, that ability as represented by marks in school should be distributed in any large normal group of pupils approximately according to the probability surface of frequency, then these examinations as marked at present either by teachers or regents cannot be held to test at all adequately the abilities of the pupils in the several subjects of study.

Table 28 discloses another fact of at least equal importance. Consider for a moment the two columns, "per cent rejected by the teachers" and "per cent rejected by the regents of those passed by the teachers." We find the columns running thus:

English	13.2	7.3	History	17.0	13.5
German	22.3	19.0	Commerce	22.3	20.9
French	19.4	16.0	Drawing	13.5	13.2
Latin	21.3	20.1	Music	18.3	1.9
Mathematics	25.8	25.7	Other Subjects	11	16.2
Science	14.2	9.2			

This phenomenon seems hard to understand. At first thought, one would suppose that the greater the per cent failed by the teachers, the fewer additional papers would be failed by the regents. We find, on the contrary, that with remarkable consistency, the greater the per cent failed by the teachers, the greater the additional per cent failed by the regents. The rule is not even violated in the case of mathematics, which by all tradition offers the greatest possibility of exactness in marking papers. If great care in speech is not demanded, we may say that in nearly all the subjects, the regents' examiners reject the judgment of the teachers to just about the same degree that the teachers reject the judgment of the pupils.

In the graphical representation of these data given in Figure 4 we see how closely the two areas correspond not only in extent, but in shape as well. The only explanation which occurs

to me for this is the absence of all harmonious standards among the examiners of the various subjects. When the questions are prepared by the examiners, a certain standard of excellence in high school work is set up in each subject. The questions are an attempt to measure ability by these several standards. When the thousands of teachers over the state get the questions with the answer papers from their respective classes, if the questions seem easy as measured by the small number of their pupils whom

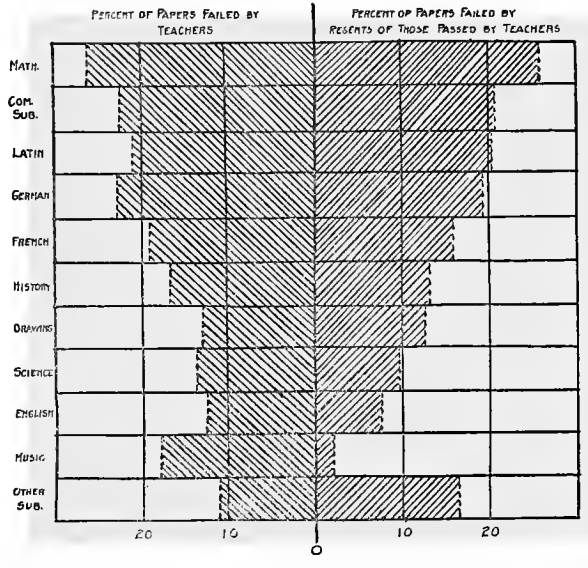

FIG. 4. Showing the percentages, by departments, of all papers written in the regents examinations in January and June 1912, which were failed by the teachers, and also the percentages failed by the regents of those passed by the teachers.

they feel compelled to fail, it is an evidence that the standard held by the examiner in that subject is not high as compared with the standard of the teachers of the subject throughout the state. If, on the other hand, the questions seem very hard to the teachers, and they must fail a larger per cent of the pupils, it is evidence that the standard of the examiner in that subject is higher than that held by the teachers. Consequently, when the teachers mark the papers by their own standards, the examiner whose standard is lower than theirs finds fewer papers to reject among

those passed by the teachers than does the examiner whose standard is higher than that of the teachers.

Whether or not this explanation is the true one, it seems certain that if the examinations were as valuable an instrument for standardizing the work of the schools of New York as its advocates claim, this situation would not exist.

As an interesting sidelight upon the variation among these standards as they exist in the various high schools of the state, Table 29 is presented. This was prepared from the 1913 report of the Department of Education from the table beginning on page 830. Not all the schools were used to make this distribution, but the first 393 were taken with no omissions, and they form thus a sufficiently large random selection.

TABLE 29

DISTRIBUTION BY SCHOOLS, OF THE PERCENTAGES OF PAPERS PASSED BY THE REGENTS OF THOSE MARKED PASSED BY THE TEACHERS. ALL ACADEMIC EXAMINATIONS IN 1912. FIRST 393 SCHOOLS IN THE ALPHABETICAL LIST

PER CENT OF PAPERS PASSED BY REGENTS OF THOSE PASSED BY TEACHERS	NUMBER OF SCHOOLS	PER CENT OF SCHOOLS
40 to 49.9	2	.51
50 to 59.9	10	2.55
60 to 61.9	7	1.79
62 to 63.9	6	1.53
64 to 65.9	8	2.04
66 to 67.9	9	2.29
68 to 69.9	8	2.04
70 to 71.9	20	5.09
72 to 73.9	14	3.56
74 to 75.9	12	3.06
76 to 77.9	24	6.11
78 to 79.9	11	2.80
80 to 81.9	25	6.36
82 to 83.9	32	8.14
84 to 85.9	24	8.65
86 to 87.9	35	8.91
88 to 89.9	33	8.40
90 to 91.9	40	10.20
92 to 93.9	33	8.40
94 to 95.9	20	5.09
96 to 97.9	9	2.29
98 to 100	11	2.80
Average		82.74
Middle 50%		76.3 to 91.0

This table should be read as follows: .51 per cent of the schools of New York state have fewer than 50 per cent of the papers passed by the regents which were passed by the teachers; 2.55 per cent of the schools have 50 to 59.9 per cent passed by the regents; one fourth of the schools have less than 76.3 per cent of their papers accepted which they had passed, while another one-fourth of the schools have more than 91 per cent accepted.

The significance of this table is not great. The schools are so different in size that any distribution of schools as units must be interpreted guardedly. However, in conjunction with the table giving percentages of papers failed, it is a certain indication that there is as little agreement among the teachers of the state concerning standards hoped for by the regents themselves in the examinations, as there is among the examiners of the various subjects.

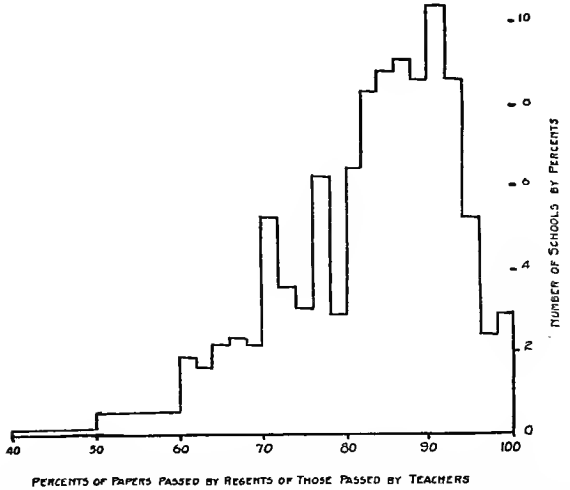

FIG. 5. Surface of frequency of schools having the various per cents of their claimed papers passed by the regents. (Data, Table 29.)

The data of Table 29 are represented graphically in Figure 5. If the spread here indicated persists at this time after so many decades of service of the examination system, it cannot be maintained that such a system is a very effective means of standardizing work either among the schools or the teachers.

Turning to the study of the marks themselves which are on file in the examinations division of the Department of Education at Albany, I must first express my appreciation of the courtesy extended to me while examining the records. Not only was free access to all the records given, but also every facility for most readily transcribing the data was afforded.

The study was confined to four questions:

(1) What is the distribution of differences between the marks given by teachers and examiners on the same papers?
(2) What is the distribution of marks of teachers on papers marked a certain figure, say 75, by the examiners?
(3) What is the distribution of marks of teachers on papers failed by the examiners? and the related question, What are the examiners' marks on papers marked near the failing point by the teachers?
(4) What differences if any exist between the standards maintained by the small high schools and the large ones?

A detailed description of the system of examinations carried out in New York State seems unnecessary. Questions are supplied to all the high schools every half year on practically each year of work in each subject taught in the high school. The papers are first graded by the teachers in the schools, and then those marked 60 or above are sent to Albany to be reëxamined by the examiners. This is, indeed, a task for a small army of readers. With the development of the department certain customs have become quite fixed. Significant among these the following may be mentioned as bearing most directly upon the findings of this study:

(1) Before the readers start the rating of any set of papers, all those who are to help with any given set go over several papers together so as to gain as great uniformity as possible before beginning to mark.
(2) Any paper marked failed by the reader which was passed by the teacher, must be read by another reader before it is finally failed.
(3) Where the difference between the examiner's mark and the teacher's mark is 3 or less, the examiner gives the same mark to the paper that the teacher had given, except in cases where the examiner's mark is below 60. In those cases, the examiner holds to his own mark, thus failing the paper.
(4) The ratings of certain teachers, and afterwards certain schools, come to be accepted, and the papers rarely if ever reëxamined.

These traditions must be kept in mind in connection with all phases of the study.

The ratings of the June, 1913, examinations were chosen since they were the most recent as well as the most accessible. Among the subjects the following were selected as perhaps the most rep-

resentative: English Grammar, Latin II, Elementary Algebra, American History and Civics, Physics, and Elementary Representation. The bases for the selection of schools were as follows:

(1) The schools were taken in alphabetical order beginning at the first.
(2) All "Union Schools" were thrown out.
(3) All large schools (those the record for whose ratings required more than one book) were thrown out.
(4) All schools having fewer than three English Grammar papers were thrown out.
(5) All schools whose ratings in English Grammar were accepted without reëxamination by the regents were thrown out.

When the list of schools meeting these requirements totaled 36, no more were added, but without exception, all the data in these thirty-six were used.

For the five large high schools to use for the brief comparative study, the five double books, which came first to hand, were taken. I have not been asked to withhold the names of these schools, but it seems only courteous to do so.

The following distribution, Table 30, furnishes an answer to the first question above: What is the distribution of differences between marks given by the teachers and the examiners to the same papers? The facts are represented graphically in Figures 6 and 7.

TABLE 30

THE DISTRIBUTION OF DIFFERENCES BETWEEN TEACHERS' MARKS AND REGENTS' MARKS ON THE SAME PAPERS (36 Schools)

Range of Differences between Marks on Same Papers		English Grammar		Latin II		Elementary Algebra		American History and Civics		Physics		Elementary Representation		Total		Theoretical Distribution of Totals	Range of Differences.
		No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*%*	
	15 or more	0		0		1	.2	3	.8	0		0		4	.1	.04	17
	10 to 14	7	1.8			8	1.2	13	3.4	1	.4	2	.6	31	1.3	.04	16
	9	1	.3			5	.8	4	1.0	1	.4	1	.3	12	.49	.08	15
																.12	14
	8	1	.3			13	2.0	4	1.0	1	.4	1	.3	20	.81	.2	13
	7	8	2.1	1	.2	10	1.5	10	2.6	0		0		29	1.18	.25	12
	6	7	1.8	3	.7	21	3.2	5	1.3	4	1.5	0		40	1.63	.33	11
Regents' Mark More	5	8	2.1	4	1.0	18	2.7	19	4.9	3	1.1	0		52	2.11	.37	10
	4	16	4.3	1	.2	25	3.8	18	4.6	5	1.9	3	.9	68	2.76	.49	9
	3	5	1.3	1	.2	5	.8	3	.8	3	1.1	2	.6	19	0.77		
	2	2	.6	3	.7	8	1.2	3	.8	2	.8	1	.3	19	0.77		
	1	4	1.0	2	.5	2	.3	11	2.8	2	.8	1	.3	22	0.89		
	0	122	32.5	106	24.7	385	58.5	148	38.1	67	25.5	73	20.9	901	36.60	7.48	0
	1	22	5.86	54	12.6	30	4.6	6	1.6	5	1.9	13	3.7	130	5.28	1.00	1
	2	10	2.66	34	7.9	9	1.4	7	1.8	5	1.9	11	3.2	76	3.08	9.96	2
	3	15	3.99	18	4.3	11	1.7	6	1.6	5	1.9	5	1.4	60	2.43	7.48	3
	4	27	7.19	37	8.6	12	1.8	18	4.6	17	6.5	16	4.6	127	5.16	5.16	4
Regents' Mark Less	5	15	3.99	29	6.8	23	3.5	16	4.1	18	6.8	23	6.6	124	5.04	5.04	5
	6	14	3.73	22	5.2	16	2.4	27	7.0	24	9.1	26	7.5	129	5.24	5.24	6
	7	27	7.19	20	4.7	5	.8	18	4.6	19	7.2	15	4.3	104	4.23	4.23	7
	8	11	2.93	17	4.0	11	1.7	13	3.4	16	6.1	13	3.7	81	3.30	3.30	8
	9	12	3.19	14	3.3	8	1.2	5	1.3	16	6.1	20	5.7	75	3.05	3.05	9
	10 to 14..	28	7.45	48	11.3	19	3.0	17	4.4	34	13.0	65	18.6	211	8.57	2.64	10
	15 to 19..	7	1.86	14	3.3	8	1.2	10	2.6	13	4.9	25	7.2	77	3.13	2.16	11
	20 to 24 .	3	.8	1	.2	2	.3	4	1.0	0		19	5.6	29	1.17	1.71	12
	25.....	4	1.0	0		3	.5	0		2	.8	14	4.0	23	.93	1.26	13
																.81	14
	Totals....	376		429		658		388		263		349		2463		.73	15
																.69	16
																.65	17
																.57	18
																.49	19
																.41	20
																.28	21
																.20	22
																.16	23
																.13	24
																.12	25
																.12	26
																.12	27
																.12	28
																.12	29
																.12	30

This table reads as follows: In English Grammar there were 7 papers, or 1.8 per cent of the papers on which the regents' mark was from 10 to 14 higher than the teacher's mark on the same paper. The column at the right gives a theoretical distribution of the totals, distributing the coarser groupings at the upper and lower ends of the original distribution, and dividing the large number at zero among the three numbers on either side of it as nearly as possible as they would have been found had not the custom prevailed of changing no mark unless it differed three or more.

Before commenting upon this table, another one whose data bear upon the meaning of these distributions must be given. This is the table of distributions of the teachers' marks on papers

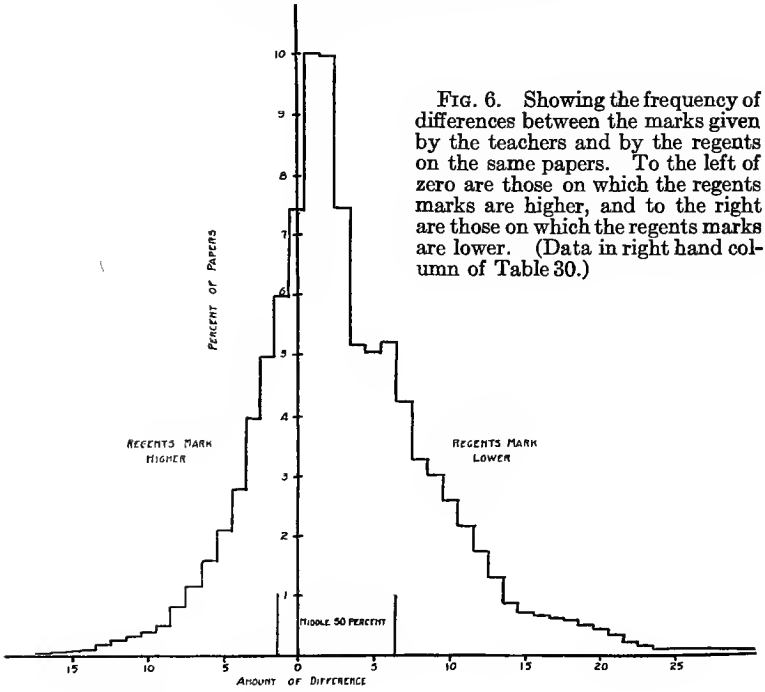

FIG. 6. Showing the frequency of differences between the marks given by the teachers and by the regents on the same papers. To the left of zero are those on which the regents marks are higher, and to the right are those on which the regents marks are lower. (Data in right hand column of Table 30.)

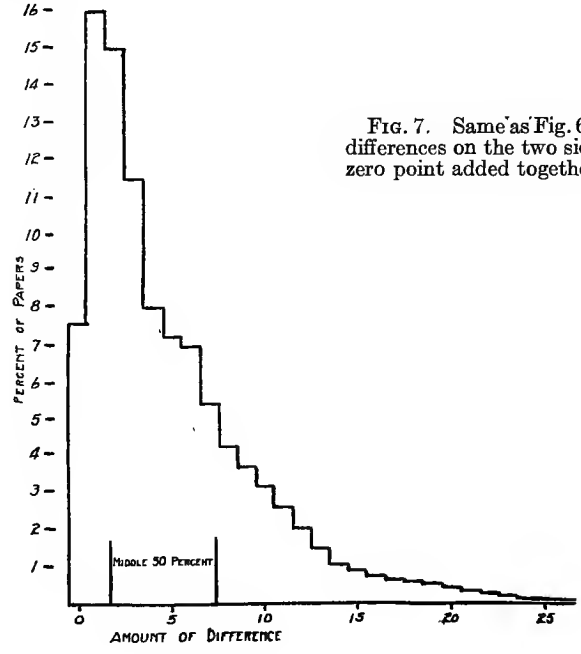

FIG. 7. Same as Fig. 6, with the differences on the two sides of the zero point added together.

which the regents marked failed. This information is given in Table 31, and is represented graphically in Figure 8.

TABLE 31

DISTRIBUTION OF TEACHERS' MARKS ON PAPERS WHICH THE REGENTS, UPON REËXAMINATION, MARKED FAILED

TEACHERS' MARK	ENGLISH GRAMMAR		LATIN II		ELEM. ALGEBRA		PHYSICS		AM. HIST. AND CIVICS		ELEM. REPRESENTATION		TOTAL	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
60	15	41.7	47	41.6	11	45.9	4	5.8	5	22.8	8	6.3	90	23.0
61	1	2.8	20	17.7	1	4.2	2	2.9	3	13.7	7	5.5	34	8.7
62	3	8.3	6	5.3	2	8.3	2	2.9	2	9.1	3	2.4	18	4.6
63	4	11.1	8	7.1	3	12.5	7	10.2	2	9.1	4	3.1	28	7.1
64	1	2.8	6	5.3	1	4.2	4	5.8	0	0	3	2.4	15	3.8
65	0	0	6	5.3	1	4.2	11	18.1	1	4.6	6	4.7	25	6.4
66	1	2.8	3	2.7	2	8.3	9	13.1	1	4.6	6	4.7	22	5.6
67	0	0	4	3.5	0	0	7	10.2	1	4.6	5	3.9	17	4.3
68	2	5.6	0	0	0	0	5	7.3	0	0	11	8.6	18	4.6
69–73	3	8.3	11	9.7	1	4.2	12	17.5	2	9.1	38	29.7	67	17.1
74–78	3	8.3	1	.9	1	4.2	5	7.3	3	13.7	20	15.7	33	8.4
79–83	2	5.6	1	.9	0	0	0	0	2	9.1	11	8.6	16	4.1
84	1	2.8	0	0	1	4.2	1	1.5	0	0	6	4.7	9	2.3
Totals	36		113		24		69		22		128		392	

This table reads as follows: Of the thirty-six papers marked failed by the regents in English Grammar, the teachers had marked fifteen or 41.7 per cent of them at 60, etc.

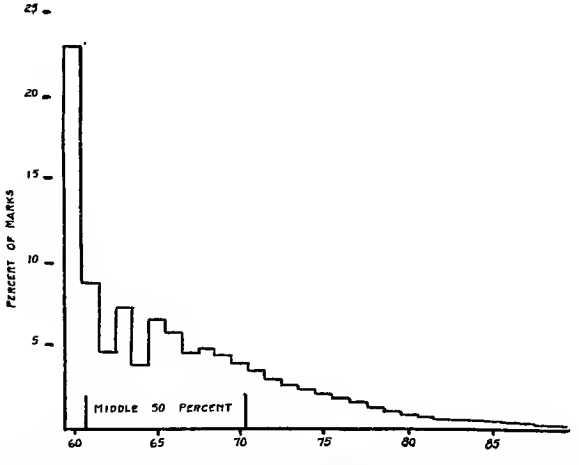

FIG. 8. The surface of frequency of teachers' marks on papers failed by the regents. (Data Table 31.)

It is necessary to discuss these two tables, Nos. 30 and 31, together. The marks on papers which the regents failed enter into the composition of both tables. However, since no marks were put upon them by the regents except the failure mark, we cannot tell how much below 60 they would have been reduced had we known the mark. As it was, we arbitrarily called all papers which were marked failed, 59. Thus, when a teacher had marked a paper 60, the difference between the teacher's mark and the regents' mark is called *one* if the regents fail the paper. In the same way, the difference is called *eight* on a paper marked 67 by the teacher and failed by the regents. This will tend to a reduction from the real differences, and the tables of differences understate the facts somewhat. Since 392 of the total 2,463 papers used were failed by the regents, this influence is considerable. Since the median reduction of marks on all failed papers is 6 points even when all failed papers are rated at 59, it is fair to assume that at least half of those differences which are in the 1, or 2, or 3, column due to failure, represent really differences as great as 5. It will be observed, also, that of the 130 papers reduced 1 point by the regents, 90 of them were due to failures; of the 76 reduced by 2, 34 were due to failures; of the 60 reduced by 3, 18 were due to failures.

With these facts in mind we may now briefly examine the tables separately. It should be said in the first place that the theoretical distribution made to allow the proper spread of the extremes and the six middle steps, three on either side of the zero point, was not constructed with mathematical precision. There was not enough of the distribution given to allow of precise determination of its form. Without this, the undistributed portions could be spread by approximations only. It is, however, sufficiently accurate for all practical purposes, and if it errs at all, the error is in favor of making the differences smaller than they really are.

Considering the distribution first which takes account of the differences on both sides of the zero point, we find the median difference is a reduction by the regents of 1.3 points, with the middle 50 per cent of the cases lying between an increase of .8 points and a decrease of 7 points. It thus appears that there is one chance in four that a paper marked, say, 70 by the teacher, will be marked 71 or higher, another chance in four that it will

be marked 62 or lower, and one chance in two that it will be marked between 62 and 71, with an even chance that it will be changed at least 3 points. When we consider this in connection with the rather narrow range within which the marks lie (less than 25 per cent being above 74 on all the papers marked in 1912), it seems a serious situation. The median mark given by the regents in 1912 to the 319,582 papers examined was probably 65, although the method of reporting the figures does not allow of absolute determination of the median. Thus the median paper of the group has less than three chances in four of being passed by the regents. Furthermore, it will be noted by Table 31 that 25 per cent of the papers which were failed had been marked 70 or more by the teachers.

Marked variation among the subjects exists in several particulars. Notice first the percentages of papers on which the difference between the teacher's mark and the regents' mark is three or less; then notice the percentages of papers reduced by 10 or more; finally notice the percentages of failures in each subject. These data are summarized in the following table, No. 32.

TABLE 32

SIGNIFICANT VARIATIONS AMONG THE VARIOUS SUBJECTS

	PERCENTAGES OF PAPERS HAVING A DIFFERENCE OF 3 POINTS OR LESS BETWEEN TEACHER'S MARK AND REGENTS' MARK	PERCENTAGES OF PAPERS REDUCED 10 POINTS OR MORE BY THE REGENTS	PERCENTAGES OF PAPERS FAILED BY THE REGENTS AFTER BEING PASSED BY THE TEACHERS
English Grammar	47.5	11.1	9.58
Latin II.....	50.9	14.8	26.35
Algebra.....	68.4	5.0	3.65
Am. Hist. and Civics ...	47.5	8.0	17.80
Physics.....	33.9	18.7	8.37
Elem. Representation...	30.4	35.4	36.70
Totals of all subjects..	49.82	13.80	16.30

In Table 32, compare, for example, the figures given for two standard subjects such as algebra and physics. Algebra has more than twice as many papers where differences cluster around zero, and less than a third as many where differences of ten or more exist, and less than half as many failures. Judged by these figures, the grading of the teachers of algebra is more than twice as reliable in the eyes of the regents as that of the physics. In elementary representation the situation is even worse than in

physics. In fact, the teachers must have had little notion of the standard to be applied to the papers when more than a third of them were reduced by 10 or more points, and 9.6 per cent of them reduced by 20 or more points.

The percentages of failures in these subjects are seen to vary much more than the percentages recorded in Table 28 for mathematics, English, Latin, science, etc. The variations there seemed very large, but if the variations are due to varying standards among the judges, it is only natural that in the individual subjects we should find this increased as the last table reveals. These extreme differences in standards of individuals are largely concealed in the group of subjects taken together to make, say, English. The uncertainty which exists in the mind of the teacher as to the outcome of the visit of her papers to Albany, however, is determined by the hazard of the individual subject, and she has no way of knowing the outcome. She may lose them all. Indeed, several schools were encountered in this study which had every paper in certain subjects rejected by the regents.

In this connection I may say that in collecting the data for this entire study, the distributions were first made for each school separately. Many most interesting things were revealed thereby but the tables become so very long it seems scarcely wise to publish them. One illustration of striking nature may be noted. A certain school had sent in seventeen English grammar papers. On these papers the regents raised two marks by 7 points, raised one by 5, left twelve unchanged, lowered one by 5, and lowered the other by 6. None were failed. This made all round the best record of any school in English grammar so far as grading was concerned. When we came to the same school in Latin II we found sixty-one papers. On these papers no marks were raised, one was left unchanged, two reduced by 1, one reduced by 2, one by 3, three by 4, five by 5, three by 6, three by 7, five by 8, one by 9, twenty-six by from 10 to 14, and the other ten by from 15 to 19 points. Twenty-eight papers were failed. This made all round the worst record of any school in Latin II.

Another illustration seems worthy of note. Among the algebra papers, which proved on the whole subject to the least variation of any, one school sent in twenty-two and had all but six of them reduced by 10 points or more, with five of them reduced by 20 or more.

Scores of other anomalies in grading can be seen at a glance from these separate distributions by schools, but the one impression left by them all is that judgments as to the worth of examination papers such as are written to-day are too variable to permit of substantial justice in making awards of things of such supreme worth to students, by means of such judgments.

The second question which this study was devised to answer, namely, "what is the distribution of teachers' marks on papers rated at 75 by the regents?" is in a sense a detail of the more inclusive study of differences above. However, it seems a little more definite, and certainly reveals some few additional facts. Then, too, a reduction from 100 to 75 as a mark on a paper seems a greater reduction than from 85 to 60. It seemed desirable also to use this more exact form of differences in comparing large with small schools.

The following distributions, Table 33, need but little comment. The total of the thirty-six small schools is given first, then the total of the five large schools. All the papers in English, mathematics, Latin, and science in both groups of schools were used, and the German papers were added to the group of five large high schools to make the numbers in the two groups more nearly alike. It must be noted that the regents did not reëxamine all papers from the large schools. Approximately, the following omissions are correct:

Four fifths of science papers from School 1.
Four fifths of mathematics papers from Schools 3 and 5.
Two thirds of German papers from Schools 3, 4, and 5.
One third of Latin papers from School 3.

It must be noted also that in making the theoretical distributions of the large groups at 75 to take into account the custom of changing few or no marks less than 3 points, the group at 75 in the case of the 5 schools was not large enough to smooth out the surface at 76 and 77 alone, hence no changes were made from the actual figures in the other steps.

TABLE 33

DISTRIBUTION OF TEACHERS' MARKS ON ALL PAPERS IN ENGLISH, LATIN, MATHEMATICS, AND SCIENCE, WHICH WERE MARKED 75 BY THE REGENTS; THIRTY-SIX SMALL SCHOOLS, AND FIVE LARGE SCHOOLS

TEACHERS' MARKS	36 SCHOOLS		5 SCHOOLS		Theoretical distributions of the groups between 72 and 78 to allow for the custom of not changing any grade unless the difference is at least 3	
	No.	*%*	*No.*	*%*	36 schools	5 schools
					%	%
60	1	.24	3	.95		
61	0	0	0	0		
62	2	.49	1	.32		
63	2	.49	2	.63		
64	0	0	0	0		
65	2	.49	1	.32		
66	7	1.71	1	.32		
67	2	.49	4	1.26		
68	10	2.45	5	1.58		
69	9	2.20	2	.63		
70	10	2.45	0	0		
71	10	2.45	5	1.58		
72	14	3.43	10	3.16	3.50	3.16
73	22	5.38	16	5.05	5.75	5.05
74	21	5.14	17	5.36	7.75	5.36
75	102	24.92	32	10.10	8.50	6.10
76	6	1.47	4	1.26	8.00	3.26
77	10	2.45	4	1.26	7.50	3.26
78	15	3.67	17	5.36	6.50	5.36
79	15	3.67	13	4.11		
80	32	7.82	29	9.15		
81	13	3.18	15	4.73		
82	12	2.93	23	7.26		
83	15	3.67	10	3.16		
84	9	2.20	16	5.05		
85	17	4.16	21	6.63		
86	10	2.45	12	3.79		
87	3	.73	12	3.79		
88	7	1.71	8	2.53		
89	4	.98	5	1.58		
90	14	3.43	13	4.10		
91	4	.98	5	1.58		
92	4	.98	4	1.26		
93	0	0	5	1.58		
94	2	.49	1	.32		
95	0	0	0	0		
96	1	.24	1	.32		
97	0	0	0	0		
98	1	.24				
99	0	0				
100	1	.24				
Totals	409		317			
Median marks.		75		80	77	80
Middle 50%		74–82		75–85		

Examining first the distribution and corresponding surface of frequency, Table 33 and Figure 9, of the marks for the thirty-six schools, we note that the median mark is at 75 in the actual distribution, and at 77 in the theoretical distribution. The middle 50 per cent lie between 74 and 82. Thus it appears that 25 per cent of the papers marked 75 by the regents had been rated at 82 or above by the teachers. One paper had been rated at 100, another at 60, while enough were rated at 80, 85 and 90 to show modal tendencies for those points.

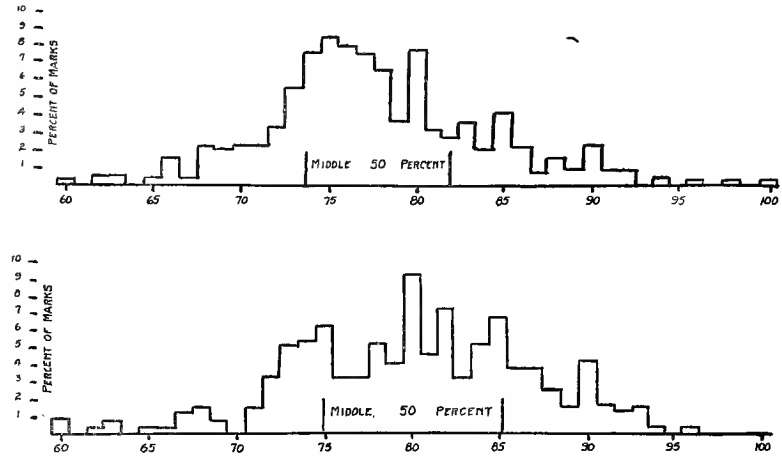

FIG. 9. Surfaces of frequency of teachers' marks on papers marked 75 by the regents. For 36 schools above; for 5 large schools below.

In the case of the five large schools we find the median mark at 80, with the middle 50 per cent extending from 75 to 85. Thus one fourth of the papers were reduced by 10 points or more, while the median paper was reduced by 5. If it were not for the modal points at 75, 80, 85, and 90, the marks would seem spread with remarkable indifference over a space of nearly 20 points. On the whole, the large schools present a decidedly less close grouping than do the small schools. When we consider that these large schools probably employ only teachers of special training and successful experience, we are forced to either of two conclusions, that training and experience do not lead to familiarity with the standards of the regents, or that "familiarity breeds contempt."

In the light of the results shown here, we feel surprise that the regents should have found cause for not reëxamining so many sets of papers, only in part, from these schools. If these schools are reaching approximately satisfactory standards in marking papers, then we must conclude that the regents do not find fault with a practically rectangular distribution of marks from 72 to 90 on papers of equal merit as judged by their own examiners. Or is it possible that the extent of the differences between their own marks and those of the teachers has not come to their attention? In any case, the situation does not seem very satisfactory.

To anticipate the possible criticism that these figures are not very reliable on account of the smallness of the numbers of papers which enter into the computations, I have determined mathematically, from formulæ in common use,[1] the extent of unreliability of the medians and median deviations from the medians, in both distributions, the one for thirty-six schools, and the one for five schools. In the case of the distributions for thirty-six schools the formula gives a mean square deviation of the divergence of the true median from the obtained median (in this case 77) of .348. By this we are assured that if marks from all the schools of the state of which these thirty-six are typical had been secured, the chances are more than two to one that the median of the total distribution would not differ from the median of the present distribution by more than .348 points. The chances are more than two hundred to one that the true median is not more than 78, nor less than 76.

Again in the case of the thirty-six schools, it is found that the mean square deviation of the divergence of the true median deviation from the obtained median deviation (in this case 3.8) is .22. We know from this that if all the small high schools had been used, the middle 50 per cent of cases would not in more than one chance in three have been increased or decreased by more than .44, not in one chance in ten thousand would it be found to equal that of the five schools.

Applying the same formula for unreliability to the measures found for the five schools, we find that the mean square deviation of the divergence of the true median from the obtained median

[1] Thorndike, Theory of Mental and Social Measurements, page 195.

(in this case 80) is .52. Thus there is less than one chance in three that the median mark here found is less or greater than the true median for all such schools by as much as .52 of a step. The chances are, indeed, less than one in fifty thousand that if all such schools had been used the median mark would have been found as low as 77. By the formula for the unreliability of a difference between two central tendencies[1] it can be shown that it is out of the range of possibility that the medians for the two distributions would be changed enough by the use of larger numbers of schools to become equal.

Similarly determined, it can be stated that there is only one chance in three that the middle 50 per cent of the complete distribution for the five large schools would be found to include less than 9.34 or more than 10.66 steps.

Thus the measures found are seen to be quite reliable so far as the number of cases used is concerned. There may be some question still as to whether these five schools (a number too small to be considered a fair random selection) may not be an unfair representation of the large schools of the state by the very reason that they are still among those whose papers are reëxamined by the regents. This cannot be answered with certainty, since we cannot compare the differences of ratings of these schools with those which the regents do not reëxamine. There are two considerations to offer in this connection, however. One is that since I was entirely strange to the conditions in all the high schools chosen as well as all other New York high schools, practically, no prejudice could have entered into the case if I had possessed any. The other and far weightier consideration is that in the case of four of the five schools certain sets of papers were looked over only in part, the inference being that the rating was being found satisfactory, and, therefore, the teachers' marks could be accepted for the remainders of the sets of papers. Thus it appears that these schools share in the confidence of the regents and are surely not a blacklisted lot which I happened to select. Furthermore it develops upon inspection that the one of these schools which had none of its papers exempted from reëxamination made the best showing in the distributions of differences.

[1] Thorndike, Theory of Mental and Social Measurements, page 193.

Notice the following table, No. 34, of medians and limits of middle 50 per cents.

TABLE 34

TEACHERS' MARKS ON ALL PAPERS IN ENGLISH, LATIN, GERMAN, MATHEMATICS, AND SCIENCE MARKED 75 BY THE REGENTS. FROM THE DISTRIBUTIONS FROM FIVE LARGE HIGH SCHOOLS TAKEN SEPARATELY

	MEDIAN MARK	LIMITS OF MIDDLE 50%
School I	80	75 to 84
School II	75	73 to 81
School III	82	79 to 87
School IV	80	75 to 85
School V	82	76 to 86

If grading approximating that of the regents is the thing desired, it seems singular that School II should be the one school of these five to have no papers exempted. School III, though its record is the worst of the group, so far as the regents' marks set the standard, had papers exempted from reëxamination in Latin, German, and mathematics. Presumably, too, the decision not to reëxamine the remainder of the sets of papers from this school was reached by the reëxamination of those which go to make up the major part of the record here made.

It seems, then, that so far as our meager evidence goes, these schools are typical of all the large schools, including those which are even more largely exempted.

One further type of comparison seems worth while. If the large schools have advantages surpassing the small schools in any subjects, they are the sciences, perhaps, which require for their proper study expensive equipment and laboratories which are seldom furnished in small schools. Since we had made a distribution of differences between teachers' marks and regents' in physics for the small schools, it seemed fitting to make a similar distribution of the same subject for the large schools. Since School V was largely exempted in this subject, we used only schools I, II, III, and IV. In Table 35 the distribution from the four large schools is placed side by side with that from the thirty-six small schools as it appeared in Table 30. The distributions of failed papers with the marks which the teachers had given them are also given, Table 36, page 79.

These tables need little comment. In the distribution of differences it will be seen that the large schools make the poorer record so far as tallying with the regents is concerned. Of the

TABLE 35

DISTRIBUTION OF DIFFERENCES BETWEEN TEACHERS' MARKS AND REGENTS' MARKS ON THE SAME PAPERS IN PHYSICS; THIRTY-SIX SMALL SCHOOLS AND FOUR LARGE SCHOOLS LISTED SEPARATELY

AMOUNTS OF DIFFERENCE		36 SMALL SCHOOLS	4 LARGE SCHOOLS	36 SMALL SCHOOLS	4 LARGE SCHOOLS
		No. Papers	*No. Papers*	*% of Papers*	*% of Papers*
Regents' Mark *Greater*	10 to 14	1		.4	
	9	1	2	.4	1.2
	8	1		.4	
	7				
	6	4		1.5	
	5	3	2	1.1	1.2
	4	5		1.9	
	3	3		1.1	
	2	2		.8	
	1	2		.8	
	0	67	26	25.5	16.3
Regents' Mark *Less*	1	5	8	1.9	5.0
	2	5	2	1.9	1.2
	3	5	1	1.9	6.6
	4	17	19	6.5	11.9
	5	18	11	6.8	6.9
	6	24	15	9.1	9.4
	7	19	12	7.2	7.5
	8	16	8	6.1	5.0
	9	16	18	6.1	11.3
	10 to 14	34	24	13.0	15.0
	15 to 19	13	11	4.9	6.9
	20 to 24				
	25 or more	2	1	.8	.6
Total papers		263	160		
Medians		Teachers' mark reduced 5		Teachers' mark reduced 6	
Middle 50%		0 to 8		2.5 to 9	

263 physics papers from the small schools 69 or 26.3 per cent were marked failed by the regents. Of the 160 papers from the large schools 68 or 42.5 per cent were marked failed by the regents. And on the papers which were marked failed the teachers had given about the same range of marks, the small schools faring a little the worse. On the whole, physics does not serve to strengthen the case of the large schools.

Only one other question remains of those which the study was undertaken to answer: What is the fate of papers marked near the failing point by the teachers when they come into the hands of the regents? To answer this question the following method was used: All papers which were marked 60, 61, 62, 63, 64, or 65

TABLE 36

DISTRIBUTION OF TEACHERS' MARKS ON PAPERS MARKED FAILED BY THE REGENTS IN PHYSICS; THIRTY-SIX SMALL SCHOOLS AND FOUR LARGE SCHOOLS TAKEN SEPARATELY

TEACHERS' MARK	36 SMALL SCHOOLS	4 LARGE SCHOOLS	36 SMALL SCHOOLS	4 LARGE SCHOOLS
	No. Papers	*No. of Papers*	*% Papers*	*% Papers*
60	4	7	5.8	10.3
61	2	2	2.9	3.0
62	2	1	2.9	1.5
63	7	6	10.2	8.8
64	4	6	5.8	8.8
65	11	10	16.1	14.8
66	9	5	13.1	7.4
67	7	7	10.2	10.3
68	5	9	7.3	13.3
69 to 73	12	9	17.5	13.3
74 to 78	5	6	7.3	8.8
79 to 83				
84 or more	1		1.5	
Totals of papers	69	68		
Median marks	66	66		
Middle 50%	64 to 69	64 to 68		

by the teachers were taken. The departments of English, mathematics, Latin and science were used, no papers in them being omitted. The same thirty-six schools constituted the list. The table of distributions follows in Table 37, page 80, and the facts are represented in Figure 10.

It will be observed that of these low papers, the regents fail 41.3 per cent. In the case of Latin they fail more than half, while in science they fail more than three fifths. Only 5.64 per cent of the marks are raised above 65, while more than one half of those left within the original range of marks, 60 to 65, are found at 60.

The chief interest in this table, No. 37, lies in the report common among high school teachers that they push up the grade on doubtful papers to "take a chance" on their passing. From this table one would judge that the report is true. Being true it is indicative of an attitude on the part of the teachers toward the regents' examinations which is not altogether good. If the teachers felt the confidence in the examinations which should be felt, they would give them in the same spirit in which teachers elsewhere give examinations of their own making. If weakness to the extent of that indicated by a grade of below 60 were

TABLE 37

DISTRIBUTION OF REGENTS' MARKS ON PAPERS WHICH THE TEACHERS HAD MARKED AT 60, 61, 62, 63, 64, AND 65, ALL LUMPED TOGETHER. FROM THIRTY-SIX SCHOOLS

REGENTS' MARKS	ENGLISH		LATIN		MATHEMATICS		SCIENCE		TOTAL	
	No.	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*	*No.*	*%*
Failure	176	32.7	151	54.1	114	33.6	119	60.7	560	41.3
60.....	181	33.6	82	29.3	97	28.7	14	7.1	374	27.6
61.....	27	5.0	9	3.2	14	4.1	5	2.6	55	4.0
62.....	25	4.6	9	3.2	23	6.8	5	2.6	62	4.6
63.....	27	5.0	7	2.5	16	4.7	11	5.6	61	4.5
64.....	36	6.7	4	1.5	20	5.9	13	6.6	73	5.4
65.....	30	5.6	15	5.4	15	4.4	17	8.7	77	5.7
66.....	6	1.1			6	1.7	2	1.0	14	1.0
67.....	5	.9			3	.9			8	.6
68.....	6	1.1			4	1.2			10	.7
69.....	3	.6	2	.7	7	2.0	5	2.4	17	1.3
70.....	10	1.9	1	.4	18	5.3	2	1.0	31	2.3
71.....					1	.3	1	.5	2	.1
72.....										
73.....							1	.5	1	.07
74										
75.....	5	.9			1	.3	1	.5	7	.5
76.....	1	.2							1	.07
Totals	538		280		339		196		1353	

This table reads as follows: Of the 538 English papers which the teachers had marked from 60 to 65 inclusive, the regents marked 176 as failed; 181 at 60; 27 at 61, etc.

revealed in the examination, the teacher would be glad to take that information at its face value, and fail the paper. But where the spirit grows up that makes the aim of the teachers to get as many students "through the regents'" as possible, then the examinations have lost their chief value. The battle then becomes one between the regents and the teachers, the one taking every possible precaution that no student "gets through" who does not deserve to, and the other using every device to enable the students to pull through. Instead of a device welcomed by the teachers to measure their work by, the examinations have become the goal, and the passing of them, the victory sought by the students.

In order to discover whether the failing of low papers was confined to a few schools, or whether it was well distributed over the whole lot, Table 38, page 82, indicating failures by schools, first in various subjects and then in totals, was compiled.

From this it is plain that the disposition to send low papers having generous ratings to the regents to "save as many as

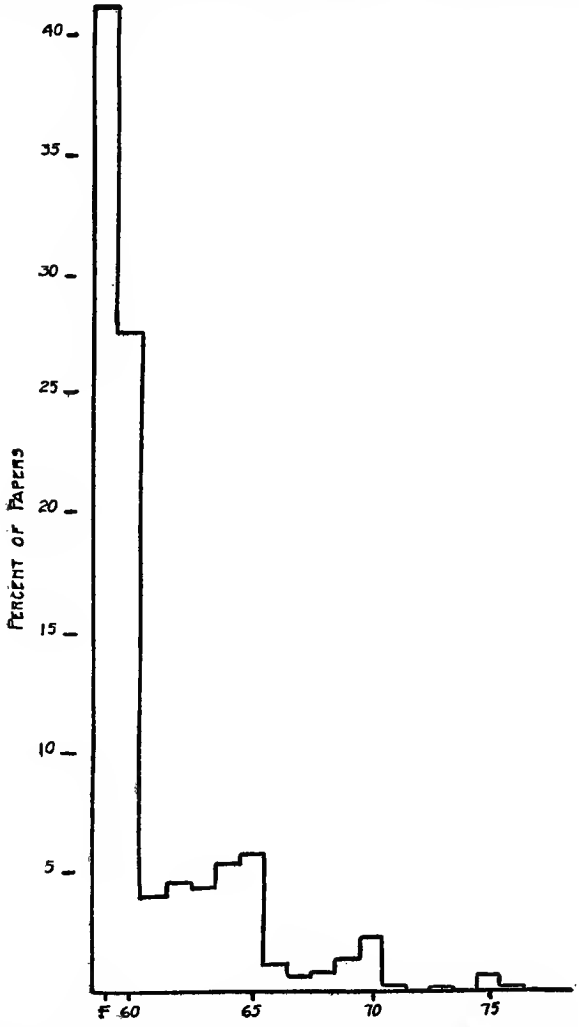

FIG. 10. Surface of frequency of regents' marks on papers marked 60 to 65 inclusive by the teachers. "F" means failed.

possible" is pretty well distributed among the schools. The lowest percentage saved by any school was 22, that for School 8, while the highest percentage saved was 75, that for School 29.

TABLE 38

TOTALS BY SCHOOLS OF PAPERS MARKED BETWEEN 60 AND 65 INCLUSIVE BY TEACHERS AND THE NUMBERS OF THOSE PAPERS FAILED BY THE REGENTS

SCHOOLS	ENGLISH		LATIN		MATHEMATICS		SCIENCE		TOTALS		%
	T	*F*	*T*	*F*	*T*	*F*	*T*	*F*	*T*	*F*	*Failed*
1	39	10	19	18	21	11	2	2	71	41	57
2	14	4	9	6	13	3	4	2	40	15	37
3	13	2	7	4	14	5	13	4	47	15	30
4	16	0	30	10	30	10	16	9	92	29	31
5	7	3	5	1	4	1	8	7	24	12	50
6	8	1	2	2	2	1	7	4	17	8	47
7	7	1	6	3	3	0	8	7	24	11	46
8	2	2	2	2	4	2	1	1	9	7	78
9	15	2	2	2	3	0	3	2	23	6	26
10	9	1	0	0	10	0	10	9	29	19	65
11	10	6	4	4	7	2	8	6	29	18	62
12	12	3	18	5	10	3	6	4	46	15	32
13	13	3	2	0	22	10	1	1	38	13	34
14	7	2	4	1	3	1	1	1	15	5	33
15	23	8	13	11	4	2	9	8	59	29	49
16	17	6	6	1	9	7	2	2	34	16	47
17	5	4	3	2	6	2	6	3	20	11	55
18	5	0	10	4	14	7	15	3	44	14	32
19	11	5	8	5	5	2	6	2	30	14	46
20	50	21	17	17	34	15	5	2	106	55	52
21	11	3	2	2	4	0	7	5	24	10	42
22	18	7	6	6	4	0	0	0	23	13	56
23	20	7	9	7	7	0	3	2	39	16	41
24	4	2	8	2	1	0	1	0	14	4	28
25	13	1	5	4	1)	4	1	1	29	10	34
26	30	10	15	12	8	0	15	6	68	28	41
27	10	4	6	4	4	0	2	1	22	9	41
28	27	10	3	0	3	0	4	4	37	14	38
29	11	3	0	0	16	3	5	2	32	8	25
30	18	5	10	3	8	2	2	2	38	12	32
31	6	2	0	0	6	5	5	3	17	10	59
32	22	13	5	2	11	6	6	3	44	24	55
33	40	11	28	6	21	5	9	9	98	31	32
34	6	3	2	1	6	3	1	0	15	7	46
35	1	0	6	2	3	0	2	2	12	4	33
36	23	11	8	2	5	2	2	0	38	15	39
Totals	538	176	280	151	339	114	196	119	1353	660	41.
Per cent Failed	32.7		53.9		33.6		60.7		41.3		

This table reads as follows: Of the 39 English papers from School 1 which the teachers had marked from 60 to 65 inclusive, the regents failed 10; and of the total 71 low papers which the teachers sent in from School 1 the regents failed 41, or 57 per cent.

The average saved to the schools from among these low papers was found to be 58.7 per cent.

But few features of this table call for comment. In the case of School 20, all of the seventeen Latin papers were failed, a record but little poorer than that made by School 1. School 10 lost none of the ten mathematics and lost all but one of the ten science papers. The relatively high record made by the English teachers in grading to suit the regents is a little surprising, since the impression prevails that there is less chance of establishing uniform standards understandable by all the teachers in English than in most other subjects. The result in that

particular here coincides with the result in the totals of the 1912 examinations reported several pages back.

By all these findings concerning the New York State system of examinations, we are compelled to conclude that the type of examination now in common use is not a successful means of standardizing school achievement.

Through the interest and coöperation of Superintendent Muir and his teachers at Orange, N. J., I was able to get the data for the following brief experiment in marking. There were two aims in mind in gathering the data: first, to find the extent of differences in rating elementary school papers by the teachers in the grades, and second, to determine the extent of reduction of these differences which would be accomplished by having the several teachers follow a uniform standard of values for the different parts of each question. To this end, Superintendent Muir had all of his fifth grade teachers give a uniform arithmetic test to their pupils, rate the papers, and send them to him without any marks upon them. When the papers were thus assembled he asked one of the teachers who is unusually systematic in arithmetic work to make out an appropriate scheme for the marking of the papers, which should be simple and yet should take account of the various processes involved in the several problems. When this was done, a substitute was provided in this teacher's room, and she was asked to rate all the papers by the scheme she had provided. Afterwards, the teachers of the several rooms were asked to rate their own papers again by using this teacher's scheme of marking. Thus each paper was rated three times. The questions which I desired to investigate could be answered by comparing each teacher's ratings made by her own method and by the systematic method with the ratings of the special teacher (called judge hereafter). These comparisons are given in Table 38a, page 84, where distributions of differences between the teacher's mark and the judge's mark on the same papers are given for each of the six teachers who did both ratings.

From Table 38a it will be observed that there is a very considerable range of differences when teachers use their own standards of marking, there being one fourth of the cases where the judge's mark is greater by 3 points or more, and another fourth of the cases where the teacher's mark is greater by 5 points or more.

TABLE 38a

THE DISTRIBUTIONS OF DIFFERENCES BETWEEN TWO TEACHERS' MARKS ON SETS OF FIFTH GRADE ARITHMETIC PAPERS, FIRST WITHOUT ANY EFFORT TO UNIFY THE METHODS USED, AND SECOND BY A COMMON STANDARD

		WITHOUT STANDARD							WITH STANDARD						
		A	*B*	*C*	*D*	*E*	*F*	*Total*	*A*	*B*	*C*	*D*	*E*	*F*	*Total*
	RANGE OF DIFFERENCES														
Judge's Mark *Greater*	21 or more					2		2							
	16 to 20	1				1	1	3							
	15						2	2							
	14						1	1							
	13					1	2	3							
	12		1			1		2							
	11			1		1	2	4			1				1
	10						1	1		1					1
	9		1			2	1	4							
	8				1	3	1	5							
	7	1	1		1	1	1	5				1			1
	6		2			1	1	4							
	5		1	2	1	1	2	7							
	4	2	2	2	1	1	2	10	1					1	2
	3		4	2	1	2	2	11			1	1	1		3
	2	2	2	1	1	1	1	8	4	1	1	3	7	1	17
	1		5	4	3	2	4	18	2	3	4	5	1	1	16
	0	1	4	4	1	1	1	12	22	30	16	16	29	26	139
	1	2	5	2	2	2	1	14	5		2	2	1	3	13
	2	6	1	3	2	3	1	16	1	1	3				5
	3	9		2		2		13		2	2	1		1	6
	4	5	1	4	1	5	1	17		2	3	3			8
	5	2	3	2	2	1		10		1	1	2			4
Judge's Mark *Less*	6	1	1		3	2		7			1	1			2
	7		1	1	6	1		9							
	8		2	1	2		1	6							
	9	1		1	2			4							
	10	1			1		1	3							
	11		1	1				2							
	12	1			1		1	3							
	13		1			1	1	3							
	14						1	1							
	15			1	1			2				1			1
	16 to 20		2					2							
	21 or more			1	3	1		5							
	Totals	35	41	35	36	39	33	219	35	41	35	36	39	33	219
	Medians	+3	0	+1	+6	−1	−4	+1							

When the same standard of rating is used by both teacher and judge the range of differences is very much reduced, considerably more than half the cases being 0. Individual differences among teachers appear plainly in the medians at the bottom of the columns. While teacher D made a median mark 6 points higher than the judge, teacher F made a median mark 4 points lower than the judge. As measured by the standard of this judge, the teachers differed by 10 points as to the value of equivalent papers.

From this brief experiment we may draw one lesson: If the superintendent expects to place much significance upon the uniform tests which he gives he must either have the marking done by a single judge, or else must make out a scale for the rating of the papers by which the variations of the several teachers may be greatly reduced.

STANDARD TESTS AND SCALES AS AIDS IN STANDARDIZATION

As illustrations of the means being advocated during recent years for overcoming in part this variability of standards among teachers we may mention the following: for arithmetic, Stone[1] and Courtis[2]; for handwriting, Ayres[3] and Thorndike[4]; for composition, Hillegas[5]; for drawing, Thorndike[6]; for reading, writing and composition together, Courtis[7]; and for spelling, Buckingham.[8] It has been impossible for me to examine all of these. In fact, it has been impossible for me to examine any of them exhaustively. I shall, however, submit some data concerning Courtis's arithmetic tests, Thorndike's drawing and writing scales, and Hillegas's composition scale. These data will be presented as far as possible in such a way that they will be of service to anyone who wishes to carry the study further.

These standard measures are of two distinct types. The first, illustrated by the Courtis arithmetic tests, is a special test so devised that the rating of the results is wholly objective, and practically all variability among markers is, therefore, eliminated. The other type is designed to define merit in the ordinary productions of the pupils. By their use it is expected that the same paper will be given more nearly the same mark

[1] C. W. Stone, Arithmetical Abilities and Some Factors Determining Them, Teachers College Contributions to Education, No. 19.

[2] S. A. Courtis, Standard Tests in Arithmetic, 82 Eliot St., Detroit, Mich.

[3] L. P. Ayres, Scale for the Measuring of Quality of Handwriting in Children, Russell Sage Foundation, Publication No. 113.

[4] E. L. Thorndike, Handwriting, *Teachers College Record*, March, 1910.

[5] M. B. Hillegas, Standard for Measuring the Quality of English Composition by Young People, *Teachers College Record*, Sept., 1912.

[6] E. L. Thorndike, The Measurement of Achievement in Drawing, *Teachers College Record*, Nov., 1913.

[7] S. A. Courtis, Standard Tests in Reading, Writing and Composition, 82 Eliot St., Detroit, Mich.

[8] B. R. Buckingham, Spelling Ability, Its Measurement and Distribution, Teachers College Contributions to Education, No. 59.

I shall omit any extended account either of the nature or origin of the tests and scales to be discussed in this section because I assume that anyone interested in the discussion will be familiar with them. It is difficult to do justice to them with any brief description. All of them are available in their complete form as indicated by the above addresses of publishers.

by several judges than would be the case without the scales. The Hillegas and Thorndike scales are examples of this type.

Each type has its advantages and its shortcomings. In the case of the special test there is greater definiteness, and less variation among the judges, but it is narrower in scope and involves a great amount of care and labor in its preparation and administration. In addition, there is doubtful value in the continued use of the same test with the same children. In the case of the standard scales the results are less precise because more subjective but can be applied to the specimens of the regular work of the children. Also they increase in helpfulness with time and repeated use.

In the following discussion of the standard tests or scales for measurement, it must be kept in mind that our chief interest in this study is the establishment in the minds of the teachers of a uniformity of standards such that the injustices which surely follow from the variability pointed out in the preceding sections may be materially reduced. The data will be available for further study of other phases of the tests and scales, but we shall be primarily concerned with their serviceableness as instruments for the establishing of uniform standards in the minds of teachers by which variability of rating a given degree of merit can be reduced.

I. The Courtis Tests in Arithmetic

The above limitation of my purpose makes unnecessary anything but the briefest sketch of my findings in regard to the Courtis arithmetic tests because by them the rating becomes a mechanical process subject to almost no variation. Upon only one basis can we properly inquire into the effectiveness of the Courtis tests with reference to their soundness as a means of measuring merit, and that is the basis of the material which is selected as an index of the ability which the author seeks to measure. This I shall consider very briefly.

Two fellow students, Mr. P. P. Brainard and Mr. R. L. McLaughlin, were associated with me in a study of the Courtis tests in their application to the schools of Hackensack, N. J. Under the direction of Superintendent Stark we assisted in the administering of the tests, and we did practically all of the calculations by which the results were made of service to the

superintendent and his teachers. By means of this test it was possible to make very definite statements regarding each of the eight sorts of arithmetical work called for in the test. Since the superintendent had given the same test four months earlier, statements of progress as well as condition were possible by comparisons with previous records of individuals, rooms, buildings, and the system as a whole. It seems to me beyond question that such information is of great value to the school system of Hackensack. The question which is related to my purpose is whether the abilities upon which the test enabled us to report, are the abilities which we are trying to measure by means of the tests. To be specific, is the ability to do single combinations in addition, subtraction, multiplication and division a good indication that the person can do well the long processes in the same fundamentals? If not, then by establishing standards in the single combinations we are using a false index of ability, because of course the ability which we wish developed is that by which success in the long processes is achieved.

To determine whether there is a close correlation between facility with single combinations and with long processes in the fundamental operations of arithmetic, I calculated Pearson coefficients to indicate this correlation between the sum of a pupil's scores in tests 1 to 4 (single combinations tests) and his score in "rights" of test 7 (the abstract examples involving all of the four fundamental operations). I used six groups of children from different grades, selecting approximately fifty papers at random from each larger group. (The means which I used to assure random selecting was to take the papers just as they came in the pile.) The coefficients were found to be as follows:

	R's Between Results of Courtis Tests 1 to 4, and Test 7 Rights
4B grade	.028
5B grade	.20
6B grade	.10
7B and 7A (Academic Course)	.34
7B and 7A (Commercial Course)	−.015
8B (Commercial Course)	.41
Average	.177

It thus appears that facility in long processes is not dependent primarily upon facility in single combinations. From this we

may conclude that it is ill-advised to try to standardize the work in single combinations, especially in the upper grades. Its significance has yet to be established even in the lower grades.

In this connection I wish to call attention to what seems to me a significant fallacy in an article by Courtis in the March, 1913, *Elementary School Teacher*. He there states that a very high correlation exists between tests 1 to 4 and test 7 attempts ("Pearson coefficient of correlation of .98"), and a "slightly lower" correlation between tests 1 to 4 and test 7 rights. He uses score sheets from 55,200 children as a basis for his computation and naturally his statement carries great weight. The obvious corollary to it is that the teacher who gets the best results in the single combinations is producing the greatest facility in practical processes in the fundamentals. It is, therefore, a matter of consequence.

In Courtis's table he divides his 55,200 children into 45 groups, and records with each group its average in tests 1 to 4, and its average in test 7, both attempts and rights, in separate columns. These columns of averages are the bases of his correlation. His groups, although he does not tell us their origin, are presumably class groups, the lowest being the third grades of some city, the next being the group of the first step higher, and so on up to the best twelfth grade at the other end of the series. It is, then, not surprising that the lowest group should be lowest in both tests 1 to 4, and test 7, nor that the highest group should be highest in both tests. That is just what we should expect whether there is any correlation between the two abilities tested or not. Both things are taught in school, and as children advance in years, they become more efficient in both processes, on the average. Even if there is no correlation between the two abilities in individual children, we should expect to see them improve, on the average side by side, just as improvement in either one would correlate with physical growth. I contend, then, that Courtis's discovery of a high correlation is no indication of correspondence between the two abilities in individuals, and therefore constitutes a mistaken doctrine which it is injurious to advance.

The above error accompanies, if it does not originate in, an attempt to devise a test suitable for all grades alike. The use of tests 1 to 4 in the upper grades, anywhere above the fifth, cannot be justified. The same attempt has led to another

mistake, it seems to me, which I shall mention. The use of test 8, the two step reasoning test, in grades below the fifth, or perhaps the sixth, is hard to defend. According to Courtis's published averages, the achievement for many thousands of children indicates the following:

	Rights, Test 8
Third grade,	.6
Fourth grade,	.8
Fifth grade,	1.2

For securing even as high figures as these the provision in his method of calculation whereby each child getting none correct is credited with .5 of one, and each child getting one, is credited with 1.5 and so on, is largely responsible. Thus in the third grade, out of a hundred children probably 90 get none right. It is absurd to suppose that these ninety averaged a half one right. In fact, probably the majority can make absolutely nothing out of the jumble of words which constitute the problem. The situation is little better in the fourth grade, and it is surely vain to try to standardize such processes where achievement is so low. Rather let us abandon the notion of a uniform test for all grades and adapt the test which we do give to the age of the pupils who are to take it. It is probably something of the same thought which has prompted Courtis to publish recently his separate sheets for testing fundamentals.

II. The Thorndike Drawing Scale

In our examination of the scales for measurement of regular school products, we shall consider first the Thorndike Drawing Scale because tradition has as yet done less to fix a standard of any sort for drawing than for most other school subjects. The value of the scale can be the more readily pointed out on that account.

As a basis for the study of the drawing scale a set of thirteen drawings were rated by from twenty-five to thirty-five teachers by both methods, the ordinary percentage method, and with the scale. Professor Hillegas very kindly permitted the rating to be done by his advanced class in Current Problems in Elementary Education during one of his class hours. The samples of drawing were those which Professor Thorndike has had printed on heavy paper for purposes of experimentation and perfection of the scale. It was possible thus to have a copy of the drawings

in the hands of all the teachers at one time. The percentage rating was made first. They were instructed to rate the drawings first as they would if they were fourth grade teachers, and the drawings had been done by children in their class. They then repeated the rating supposing the drawings to be done by sixth grade children. Next they were to consider them as eighth grade productions, then as tenth grade, and finally as senior high school productions. Thus every teacher who finished the task made sixty-five judgments in all.

After this was finished, the records were taken up and the scale for measuring drawing, consisting of fourteen drawings of stated values as determined statistically, was handed to each teacher and the request made that the thirteen drawings be again rated, this time by giving to each one the value which was assigned on the scale to the drawing most nearly equal to it in merit. Thus we secured the judgment of each of from thirty-four to thirty-six teachers on these thirteen drawings by means of the scale. Our question concerns itself mainly with a comparison of the two groups of data thus secured.

The distributions of the judgments by the percentage scale are given in the five parts of Table 39, and the distributions of judgments by the Thorndike scale are given in Table 40. With each division of the tables are given the average of all the judgments made and the average deviation of the judgments from that average. At the lower right hand corner of each table is given the average of the thirteen average deviations found for that table. The drawings are numbered, the numbers being the same as designate the drawings on the sheet prepared by Professor Thorndike for experimentation.

The wisdom of using the average instead of the median as the central tendency in these distributions may be questioned. The reason which seemed to me to justify it is that the distributions are so wide, and often so dispersed in the middle, that the median would be shifted considerably away from the average, even though the distribution was fairly symmetrical. Of course the undistributed extremes of the distributions point to the proper use of the median, but, on the other hand, for purposes such as these tables are compiled, full weight should be given to extreme measures which are far from the central tendency. At any rate, probably either measure answers the purpose with sufficient accuracy for our use.

TABLE 39

DISTRIBUTION OF MARKS ASSIGNED TO SAMPLES OF DRAWINGS BY TEACHERS. THE NUMBERS CORRESPOND TO THOSE USED ON THE SHEET OF DRAWINGS REPRODUCED BY PROFESSOR THORNDIKE

I. When Considered as Fourth Grade Productions

MARKS	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0 to 2			2	8										
3 to 7			1	3					1		1			
8 to 12				5			1	1	1					
13 to 17														
18 to 22				3	1				1	1				
23 to 27	1		2		2				1					
28 to 32		1	3						2				1	
33 to 37														
38 to 42				2			1		2				2	
43 to 47							1							
48 to 52	1	1	2	2	1	2	2	1	3	1			1	
53 to 57										1				
58 to 62			1	3	1		1	1		1	1		1	
63 to 67			3			1	1							
68 to 72		1	7	2	1				4	1	2	1	1	
73 to 77	1		2	1	3		4		3		1		3	
78 to 82		4	3		3		5	4	3	3	3		4	
83 to 87	3	1	1	1	2	1	1	4	1	4	4		3	
88 to 92	9	5	2		12	6	5	3	4	7	3	2	5	
93 to 97	7	7		1	2	8	3	5		4	8	8	4	
98 to 100	9	11	2		3	12	4	10	3	6	6	18	4	
Totals	31	31	31	31	31	30	29	29	29	29	29	29	29	
Average	90	91	60	29	79	91	77	87	63	84	86	97	80	
A. D.	8.5	10.0	22.5	27.0	15.0	9.0	16.5	13.0	23.5	12.5	12.0	4.0	13.5	Avg. 14.4

Note: All the A. D.'s in this table were computed not from the true average but from the nearest mid-point of a step.

II. When Considered as Sixth Grade Productions

MARKS	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0 to 2			4	16				1	5		1		1	
3 to 7			1	2			1							
8 to 12			3		1				1	1			1	
13 to 17	1		1	2					1					
18 to 22				1	1					1			1	
23 to 27	1	1	1	1		1			1					
28 to 32							2		2					
33 to 37		1											1	
38 to 42	1		5	4	2	1	1		1	2	1		1	
43 to 47							1		2				1	
48 to 52		1	1	1			3	2	1		1		1	
53 to 57			3		1	1	1						1	
58 to 62		2	4	2	3				4		2	1	3	
63 to 67					1		1		1	2	1	1	1	
68 to 72	1	1	3	1	2		3	1	2	3			3	
73 to 77	3	1	1		3		3	2	1	2	4	1	2	
78 to 82	2	2	1	1	6	6	6	5	4	4	5		6	
83 to 87	9	5			3	3	4	4		3	2	1		
88 to 92	7	6	2		4	8		4	2	5	6	5	2	
93 to 97	3	6				5	2	6	1	2	3	11	2	
98 to 100	3	5	1		2	5	1	4		4	2	8	2	
Totals	31	31	31	31	31	30	29	29	29	29	28	28	28	
Averages	81	82	44	19	71	85	68	82	49	76	78	92	67	
A. D.	13.0	14.0	25.0	21.5	15.5	11.0	17.0	13.5	26.5	17.0	13.5	7.5	20.0	Avg. 16.55

III. When Considered as Eighth Grade Productions

Marks	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0 to 2	1		10	20	1		2	1	7	1	1		5	
3 to 7									2	1			1	
8 to 12			1	1			1		1		1			
13 to 17					2	1			1					
18 to 22	2	2	3	1			1		1	1				
23 to 27				2	1		1				1			
28 to 32		1	4		1	1	1	1		1			3	
33 to 37			1		1									
38 to 42		1	1	2		1	1	3	5		1	2	2	
43 to 47							2						1	
48 to 52	1	1	4	3	3				1	3	1	1	2	
53 to 57		1												
58 to 62	3	1	2		7	1	4	2	5	5	5	2	3	
63 to 67	1						2			1				
68 to 72	3	2	2		4	3	6	4	1	3	2		4	
73 to 77	8	4		1	1	1	1		1		3	1	1	
78 to 82	4	7	2		3	7	4	5		3	4			
83 to 87	2	1			3	5		2	2	4	2	1	1	
88 to 92	2	4			1	4	2	7	1	2	3	9	1	
93 to 97	2	3			1	3	1	1		2	2	4	2	
98 to 100	1	2				2		2		2	1	7	1	
Totals	30	30	30	30	29	29	29	28	28	29	27	27	27	
Averages	70	73	29	13	59	77	58	73	36	66	68	80	49	
A. D.	15.0	16.0	23.0	18.0	18.0	13.5	22.0	17.5	27.0	20.5	18.5	14.0	27.0	Avg. 19.25

IV. When Considered as Tenth Grade (Second Year High School) Productions

Marks	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0 to 2	2		13	21	1		4	2	11	2	4	1	7	
3 to 7									1					
8 to 12		3	3	1	1	1	1		2		1	1	1	
13 to 17	1		2	1	2		1							
18 to 22	1	1	1	1	3	1	2	2		3	1		1	
23 to 27			1		1					1		1		
28 to 32			2	2			1	1	1				1	
33 to 37						1	1		3	1			1	
38 to 42	2	3	2	2	3		1	1	1	2		1	4	
43 to 47	2				2		1			1				
48 to 52	1		1		3	2	2		3	1	6		1	
53 to 57							1			1	1		1	
58 to 62	9	3	3		2	1	4	1	1	3	2	1	2	
63 to 67					2	1	1	2			1		1	
68 to 72	2	5	1	1	2	3	3	1	2	1	2	1	1	
73 to 77	2	3			1	2		1		3	2			
78 to 82	2	4			4	8	2	5	1	1	2	2	2	
83 to 87	2	1				2	1	5	1	4		4		
88 to 92	2	3			1	3		4		3	4	4	2	
93 to 97	1	1				2						5		
98 to 100		2				1		2		1		4		
Totals	29	29	29	29	28	28	26	27	27	28	26	25	25	
Averages	58	65	19	9	49	72	43	68	27	58	53	78	37	
A. D.	18.0	21.0	19.5	13.0	21.0	16.5	23.5	22.0	25.5	24.0	23.5	17.5	25.5	Avg. 20.8

To make comparisons easy, and to indicate the extent of reduction in marks from fourth to sixth grade, and from sixth to eighth and so on, I have assembled the lists of averages in a

V. When Considered as Twelfth Grade (Senior High School) Productions

Marks	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0 to 2	4	2	16	22	4	1	6	2	14	3	5	2	8	
3 to 7		2	2	1	2	1	1			1	1			
8 to 12	2	1	1	1	2	1	1	3	1	1			1	
13 to 17					1					1			1	
18 to 22	1		2	1	1		1	1	2	2		1	1	
23 to 27	1	1	1						1		1			
28 to 32		1	1	2	2	1	3		2	2	1	1	1	
33 to 37	2	1							2				2	
38 to 42	1	1	2		3	2	2		1	1	1		2	
43 to 47		1	1								3			
48 to 52	7	1	2		4		4		1	3	3		2	
53 to 57						1						1		
58 to 62		4			2	2	2	2	1	3	1		4	
63 to 67	2			1		2					2	1		
68 to 72	2	5			3	3	3	3		1	2		1	
73 to 77	1				1	5	2	1		1	1	3		
78 to 82	3				3	4		6	2	5	1	2		
83 to 87		3						4		2	2	3	1	
88 to 92	2	3				2		1		1	1	5	5	
93 to 97							2		1			3		
98 to 100		2						1				3		
Totals	28	28	28	28	28	27	25	25	27	27	25	25	25	
Averages	46	56	12	6	39	63	35	62	19	48	45	74	32	
A. D.	24.0	26.5	14.5	8.5	24.0	19.0	24.0	27.0	20.5	26.5	25.0	21.0	25.0	Avg. 21.95

TABLE 40

Distribution of Marks Assigned to Samples of Drawings by Teachers Using the Thorndike Scale. The Numbers Correspond to Those Used on the Sheet of Drawings Reproduced by Professor Thorndike

Marks	No. 118	No. 121	No. 123	No. 124	No. 125	No. 129	No. 130	No. 135	No. 139	No. 140	No. 145	No. 146	No. 153	
0				9										
2.4			3	15				1					2	
3.9			2	4			1		1				1	
5.7			2	1	1		1		1					
6.5			8	7			1		3				1	
7.8		1	12			1	7		14	3	1			
8.6			8		2	1	2		10	1	2		6	
10.5			1		4	2	10	1	5	2	4		5	
11.8	11	3			11	2	3	5		4	3	2	3	
12.6	10	7			11	2	2	3	1	9	7		3	
13.5	7	15			4	3	5	7		1	5		1	
14.4	4	5			3	9	2	7		5	5	2	8	
16.0	3	4				13		9		9	6	11	2	
17.0	1	1				3		2				19	1	
Totals	36	36	36	36	36	36	34	35	35	34	34	34	34	
Averages	13.03	13.58	7.08	3.15	12.13	14.07	10.58	13.89	8.17	13.00	12.94	16.47	11.08	
A. D. (In *steps* of scale)	.99	.83	1.18	.96	1.09	1.38	1.72	1.23	.89	1.71	1.49	.88	2.47	Avg. 1.29

separate table, No. 41, page 94. In this table the drawings are listed in the order of excellence as determined by the average of the five averages of ratings made upon them by the percentage method. The average by the scale method is listed also.

TABLE 41

GIVING THE AVERAGES OF THE RATINGS MADE UPON THIRTEEN DRAWINGS BY TEACHERS WHEN THESE DRAWINGS WERE CONSIDERED IN TURN, FOURTH, SIXTH, EIGHTH, TENTH, AND TWELFTH GRADE PRODUCTIONS, USING THE CUSTOMARY PERCENTAGE METHOD, AND THE RATINGS BY THE THORNDIKE SCALE

No. of Drawing	As 4th	As 6th	As 8th	As 10th	As 12th	Avg.	Thorndike Scale
124	29	19	13	9	6	15.2	3.15
123	60	44	29	19	12	32.8	7.08
139	63	49	36	27	19	38.8	8.17
153	80	67	49	37	32	53.0	11.8
130	77	68	58	43	35	56.2	10.58
125	79	71	59	49	39	59.4	12.13
145	86	78	68	53	45	66.0	12.94
140	84	76	66	58	48	66.4	13.0
118	90	81	70	58	46	69.0	13.03
121	91	82	73	65	56	73.4	13.58
135	87	82	73	68	62	74.4	13.89
129	91	85	77	72	63	77.6	14.07
146	97	92	80	78	74	84.2	16.47
Average	78.1	68.8	57.8	48.9	41.3		

From Table 41 it will be observed that this group of teachers consider that a drawing is worth about 10 points more for fourth grade than for sixth, and about 10 points more for sixth than for eighth, and so on. It will be observed further that drawing 139 is considered about as good for fourth grade as drawing 153 is for sixth, and this in turn as good as drawing 140 for eighth, and 121 for tenth, and 129 for twelfth. The average value of these five drawings by the scale is seen by the right hand column of the table to be, in order, 8.17, 11.8, 13.0, 13.58, 14.07. These values indicate steps of difference between what is expected from the successive grades, as follows:

From 4th to 6th grade, a gain of 2.63 units of the scale.
" 6th " 8th " " " 1.2 " " " "
" 8th " 10th " " " .58 " " " "
" 10th " 12th " " " .49 " " " "

Comparing this rapid decrease in the amount of gain expected from grade to grade as we advance in the grades, with the fact pointed out above that the drop in percentage from grade to

grade is about constant, being about 9 or 10 points for each two-year change of grade, we have revealed one of the most conspicuous defects in our present system of marking. While each two years are expected to add 10 points value by the percentage scale, they are in fact expected to add ever-decreasing amounts of actual value. These points on the percentage scale do not have a fixed value, and they vary in the different portions of the scale, and under different circumstances. There being no absolute standard fixed, each judge attaches his own value to the scale. As seen in this case, there is no consistency about it. There is for any drawing, on the average, about a 10 point higher standard required from fourth to sixth to eighth, and so on, but the papers valued practically of equal merit for the successive groups differ by very unequal amounts. Even the averages of the percentage markings on the five drawings which come nearest to a rating of 65 in the five successive groups, stand at very unequal intervals. These averages are, respectively, 38.8, 53, 66.4, 73.4, and 77.6, indicating increases of 14.2, 13.4, 7.0, and 4.2, respectively.

Before turning to the comparison of variabilities accompanying the two methods of rating, I wish to point out the evidence of the great diversity of standards held by the teachers. In order most clearly to point this out, I computed the difference between each teacher's judgment on each paper and the average of all the judgments on the same paper. For example, the average judgment of all teachers upon drawing 124 as a fourth grade production is seen to be 29. If a teacher rated it at 35 he was credited with a plus difference of 6. Similarly all the differences were calculated, and then the sum of all the plus differences and the sum of all the minus differences computed for each teacher separately. The same thing was then done for the judgments made by the Thorndike scale. These sums are tabulated in Table 42.

One very significant fact is revealed by this table. By the percentage method of rating there is a marked tendency for a teacher to be either much above, or much below the average on practically all papers. This is indicated by the wide difference between the plus and minus sums in the case of a great many teachers. The meaning is very plain. The teachers have as yet no uniform idea of how well a child in a certain grade should be

TABLE 42

SHOWING THE SUM OF THE DIVERGENCES, NEGATIVE AND POSITIVE, AND THE AVERAGE DIVERGENCE OF ALL MARKS GIVEN BY EACH JUDGE FROM THE AVERAGE OF THE MARKS OF ALL JUDGES ON THE SAME DRAWINGS

Thirteen drawings were rated five times by the percentage scale, considered in succession as fourth grade, sixth grade, eighth grade, tenth grade and twelfth grade drawings, and once by the Thorndike scale.

	PERCENTAGE SCALE				THORNDIKE SCALE				
	Divergences in Units of the Scale				*Divergences in Units of the Scale*				
Judges	No. of Judgments	Sum of Judgments Less Than Avg.	Sum of Judgments More Than Avg.	Average Divergence from Avg.	No. of Judgments	Sum of Judgments Less Than Avg.	Sum of Judgments More Than Avg.	Average Divergence from Avg.	Types of Experience
1	65	149	992	17.5	13	4.62	15.59	1.56	1
2	65	1406	88	22.9	13	17.04	4.75	1.67	2-3
3	65	133	545	10.4	13	8.60	6.81	1.18	5
4	65	372	750	17.2	13	10.81	12.72	1.81	3
5	65	1323	351	25.7	13	12.26	10.77	1.76	1-3
6	65	1053	759	27.8	13	27.67	8.08	2.59	3-4
7	65	740	114	13.1	13	7.89	10.46	1.41	4
8	65	58	692	11.5	13	9.84	8.65	1.42	1
9	65	118	1088	18.5	13	18.12	5.23	1.79	1
10	65	263	486	11.5	13	11.47	12.78	1.86	3-4
11	65	2982	29	46.5	13	4.45	2.58	.54	1
12	63	444	419	13.3	13	13.72	7.33	1.62	4
13	65	1002	245	19.0	13	18.59	5.40	1.84	1-2
14	65	2368	0	36.4	13	37.89	5.70	3.35	5
15	65	759	152	14.0	13	12.42	7.43	1.53	6
16	65	151	394	8.4	13	6.41	7.02	1.03	1
17	65	2	1684	25.9	13	7.48	16.19	1.82	1-2-3
18	65	49	952	15.4	13	7.28	8.10	1.18	3-4
19	65	807	306	17.1	13	13.57	11.58	1.93	1-2-3-4
20	65	33	925	14.7	13	4.94	9.45	1.11	2-3
21	65	1	1807	27.8	13	11.52	7.03	1.43	2-3-4
22	65	251	1102	20.8	13	15.86	11.87	2.13	3
23	65	1328	151	22.7	13	6.13	12.74	1.45	1-2
24	65	11	921	14.3	12	2.06	10.77	1.07	1
25	50	108	264	7.4	13	1.76	18.39	1.55	1
26	39	30	416	11.4	13	10.60	8.77	1.49	4
27	53	143	631	14.6	13	15.58	8.57	1.86	2
28	30	15	139	5.1	13	2.85	7.16	.77	5
29	53	402	206	11.4	13	3.26	18.63	1.68	2-3
30	39	126	418	13.9	13	6.61	6.20	.98	1-3
31	25	547	21	22.5	13	24.76	3.27	2.16	2
32					13	8.66	14.67	1.79	
33					13	15.23	7.44	1.74	
34					13	5.98	12.49	1.42	
35					9	1.76	7.78	1.06	
36					6	0	6.06	1.01	
	Sums	17074	17045			87.69	336.46		
	Averages			17.9				1.57	

expected to draw. A great improvement is effected in this respect by the use of the scale. As a measure of this improvement we may compare the ratio of the sum of the differences between the pairs in the plus and minus columns to the sum of the sums, for

the two methods of rating. For the percentage method the sum of the two columns headed "Sum of judgments less than average" and "Sum of judgments greater than average," is 34,119. The sum of the two corresponding columns for the Thorndike scale is 724.15. If now we calculate the differences between each pair constituting the two columns by each method, and then add these differences, we find that the sum in the case of the percentage method is 24,691, or 72 per cent as great as the sum of the two columns. The sum in the case of the scale method is 264.51 or 37 per cent of the sum of the two columns. There can be no question, then, but that the use of the scale tends decidedly to secure uniformity of standards of rating drawings.

I was interested to discover whether the variation of standards bore any relation to the type of experience which the various teachers had had who rated the drawings. Accordingly I asked the teachers to answer the question, "In what grades have you had teaching experience?" Opposite the numbers of the judges as they appear in Table 42, I tabulated their answers. I made only six classifications into which the experience would be placed, namely, (1) kindergarten or primary, or both; (2) intermediate grades; (3) upper grammar grades; (4) high school; (5) none; (6) not stated. This tabulation appears in the column at the extreme right of Table 42. By it we may see that no marked influence upon the standards is made by previous experience. There seems to be a slight tendency for primary teachers to demand more than do upper grade teachers, although this may be a mere chance indication for the few teachers here represented.

Before closing the discussion of Table 42 I wish to make plain what may at first sight seem to be an error of calculation. It will be observed that the sums of the two columns of plus and minus differences are not equal, and that the average of the column of average divergences is not the same as the average of the average deviations given in Table 40. Two facts account for these seeming errors. The steps of the Thorndike scale are unequal, the larger differences being found in the main at the lower end of the scale. The averages from which the divergences were all computed were derived by considering the steps all equal, that is by the short method of guessed averages corrected by plus and minus divergences. The average deviations in Table 40 were also in terms of steps. When, however, we came to calculate

the difference between each teacher's mark and the average mark for that paper, we subtracted the two figures, thus securing the difference not in terms of steps, but in terms of units. This makes the average found in Table 42 larger than that in Table 40, because one is in terms of units, and the other in terms of steps. The fact that the lower ranges of the scale contain steps of a larger number of units, makes the minus difference column of Table 42, larger than the plus difference column.

Returning now to the question of variability of judgments by the two methods of rating, we shall use Part III of Table 39 to compare with Table 40. This selection is made because it represents the median grade considered, and because the average deviations found are also midway between those of the grades below them and the grades above them.

We note by the tables that the average of the average deviations by the percentage scale is 19.25 points on the scale, while for the Thorndike scale method, the average deviation is 1.29 steps of the scale. Since the latter is calculated in steps, we shall have to think of the scale as consisting not of 17 units, but of 14 steps. To compare the deviations at all, it is necessary to reduce the value of the step on the Thorndike scale to units on the percentage scale. There is no absolutely correct way of doing this, but we may get an approximation which is near enough to justify our main conclusion. The range between the values of the poorest and best drawings on the one scale is about equal to the range between the poorest and best drawings on the other scale. If now we take the range between the average of the two poorest and the average of the two best in both cases, we shall come close enough to the relative size of steps on the two scales for our purposes.

By this calculation we get the following:

	Lower Limit	Upper Limit	Difference
Percentage scale	21	78.5	57.5
Thorndike scale	5.12	15.77	9.18

The derivation of the 9.18 as the difference on the Thorndike scale is as follows: 5.12 is .32 of a step below 5.7, the value next above it on the scale. Likewise, 15.77 is .86 of a step above 14.4, the value next below it. Between 5.7 and 14.4 there are 8 steps on the scale. Then between 5.12 and 15.77 there are 8 plus .32 plus .86 steps, or 9.18 steps.

The value of a step of the Thorndike scale in terms of the percentage scale thus calculated is 6.26 points. We may now reduce the average deviations in the Thorndike scale to their equivalents in points of the percentage scale. Multiplying 1.29 by 6.26 we get 8.08 points on the percentage scale representing the average deviation of the judgments by the Thorndike scale. Comparing this with 19.25, the average deviation by the percentage method, we see that the variability by the Thorndike scale is only 42 per cent as great as by the percentage method.

III. The Thorndike Handwriting Scale

A noteworthy experiment with the handwriting scales of both Ayres and Thorndike was conducted by Starch[1] at the University of Wisconsin during 1913. He had fifteen specimens of children's writing rated by ten business men, and ten teachers in each of three ways: The percentage scale, the Ayres scale, and the Thorndike scale. The order of papers was changed between each rating, and the order of methods of rating was changed from judge to judge. The average deviations were calculated for each group of judges, business men and teachers separately, upon each paper and the average of the 15 average deviations used as a basis of comparison of variability by the different methods. The instructions for the percentage scale ratings were that 100 was to be considered perfect writing, and 0 to be considered writing with no merit. To translate steps of the Thorndike scale into units of the percentage scale, 0 to 100 on percentage scale was considered equal to 0 to 18 on Thorndike scale. With the Ayres scale the problem was not so simple, but Starch adopted the method of equating the range from the poorest mark to the best mark on the two scales, and the range from the next to the poorest to the next to the best, and so on, and using the average of all these equations as the value of the Ayres step in terms of the percentage scale.

As a result of this calculation he found the A.D.'s to compare as follows:

	Thorndike Scale	Ayres Scale	Percentage Scale
Business men	6.32	6.04	10.04
Teachers	5.66	5.49	10.39

[1] Daniel Starch, The Measurement of Handwriting, *The Journal of Educational Psychology*, 4: 445.

From this we see that the variability by the Thorndike scale is, on the average for the two groups of judges, but 58.6 per cent as great as with the percentage scale.

These judges were all without practice in the use of the scale. Starch claims that with practice the variability can be reduced by nearly half. He does not give the basis of that opinion beyond the fact that his own judgments are only about half as variable from the average of all the judgments on the papers as the average of the unpracticed group. Of course that fact proves nothing about the effect of practice, but rather indicates how much better some people can use the scale than others. It would seem perfectly natural that practice should improve the efficiency of a judge in using the scale, but we have no proof as yet of the claim.

If the above study reveals the true gain in reliability of marking by means of the scale, it will prove a wonderful aid in standardization. A question arises in my mind as to the propriety of the instructions concerning the use of the percentage scale. In actual practice we do not think of the percentage scale as the distance between 0 merit and perfection. We always have a standard of some sort in mind, and use the 100 points to indicate the attainment of that standard. For example, if the teachers are rating a group of penmanship papers they always know what class of pupils wrote them, and they rate the papers on the basis of what they consider a proper standard of requirement for that grade of pupils. According to whether that standard is uniform or not in the several teachers who may be called upon to rate the paper, the ratings will be uniform or variable. In other words, it is possible that the concept of "standard work for seventh grade children" may be a much more uniform thing among teachers than the concept "perfect work." If so, then to indicate how serviceable the scale is for removing variability of marking we should have to have the papers rated on this basis, letting the teachers know what grade is being rated.

With this in mind I secured the ratings upon fifty papers, selected at random from fifth, sixth, seventh, and eighth grade papers, by sixteen teachers, using the regular system prevailing in the school, and then using the Thorndike scale.[1] The teachers

[1] These data were gathered in Providence, R. I., under the direction of R. L. McLaughlin, principal of Rochambeau Avenue Grammar School. My thanks are hereby expressed to him and his teachers.

were asked to rate each sample on the basis of what they considered to be the proper grammar school (eighth grade) achievement.

The system of marking which prevails in that city is that of letters, E, G, F, and P for excellent, good, fair, and poor, respectively. In tabulating the returns, these letters were changed to 9, 8, 7, and 6, respectively. The tabulations for the sixteen judges' marks by the letter method are found in Table 43, and

TABLE 43

A Set of Fifty Samples of Children's Handwriting Rated by the Common Letter Method, P, F, G, and E, by Teachers. These Letters Were Changed into Figures, 6, 7, 8, and 9, Respectively

Paper	Judges																Avg.	A. D.
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*		
1	9	6	7	6	6	6	6	6	7	6	7	7	7	6	7	6	6.50	.562
2	6	6	6	6	6	6	6	6	7	6	7	7	6	6	7	6	6.25	.375
3	6	6	6	6	6	6	6	6	7	6	7	7	6	6	7	6	6.25	.375
4	7	6	7	6	7	7	7	6	7	7	8	8	7	8	8	8	7.13	.547
5	7	7	7	7	7	7	7	6	7	7	7	7	7	7	7	7	6.94	.117
6	7	6	7	6	7	7	7	7	8	7	7	7	7	7	7	7	6.94	.234
7	7	6	7	6	7	7	7	6	7	7	7	7	7	7	7	7	6.81	.304
8	7	7	7	7	6	8	7	7	6	8	8	7	7	7	8	7	7.13	.437
9	7	6	7	6	7	7	7	7	8	8	7	8	7	7	8	7	7.13	.437
10	7	7	7	7	7	7	7	6	8	8	7	8	7	7	8	7	7.19	.406
11	7	7	6	7	7	7	7	7	7	7	8	7	7	7	7	7	7.00	.125
12	7	7	7	7	7	7	7	6	8	7	7	7	7	8	7	7	7.06	.234
13	7	7	7	7	7	7	7	7	8	8	7	8	6	7	8	8	7.25	.469
14	7	7	7	7	7	7	7	6	7	8	8	7	6	7	7	7	7.00	.250
15	7	7	7	7	7	8	7	7	8	7	8	7	7	7	7	7	7.19	.304
16	7	7	8	7	7	8	7	7	8	8	8	8	7	8	8	8	7.56	.492
17	7	7	8	8	8	8	7	7	8	8	8	8	7	8	8	8	7.69	.429
18	7	8	7	8	8	8	8	6	8	8	8	7	8	7	7	6	7.44	.639
19	7	7	7	7	7	7	7	6	7	7	8	7	7	7	7	7	7.00	.125
20	7	7	8	7	7	7	7	7	8	7	7	7	6	7	7	7	7.06	.234
21	7	7	7	8	7	7	7	7	7	8	8	8	6	7	8	7	7.25	.469
22	8	7	8	7	8	7	7	7	8	8	7	7	7	9	7	8	7.50	.562
23	7	8	7	7	7	7	7	6	8	7	7	7	7	7	7	6	7.00	.250
24	8	7	8	7	7	8	7	8	8	8	8	7	7	8	7	8	7.56	.492
25	8	7	8	7	7	7	7	8	8	8	8	7	6	8	7	8	7.44	.562
26	7	7	7	7	7	8	7	6	8	8	7	7	6	7	7	7	7.06	.350
27	7	7	7	8	7	8	7	7	7	8	7	8	7	7	8	8	7.37	.461
28	7	7	7	8	9	8	7	8	8	8	8	8	7	7	8	8	7.69	.516
29	8	7	8	8	7	7	7	7	8	7	7	7	7	7	7	6	7.19	.406
30	9	8	9	8	9	8	9	7	8	8	8	8	8	8	8	9	8.25	.469
31	7	7	7	8	7	7	7	7	7	7	7	7	6	7	7	6	6.94	.234
32	8	8	8	8	8	8	9	7	8	8	8	8	7	8	8	9	8.00	.250
33	9	8	8	8	9	8	8	7	8	8	8	7	7	8	7	9	7.94	.469
34	9	8	8	8	8	7	8	8	8	8	8	8	7	8	8	9	8.00	.250
35	8	8	8	8	8	8	8	7	8	8	8	7	7	8	7	8	7.75	.375
36	9	8	9	9	8	8	8	8	8	8	8	8	6	8	8	9	8.13	.438
37	8	7	7	8	9	8	8	7	8	7	8	7	7	7	7	8	7.56	.562
38	8	7	7	8	9	8	7	8	8	8	8	7	7	8	7	7	7.63	.547
39	8	8	8	8	8	8	8	8	8	8	8	7	7	8	7	8	7.81	.304
40	8	8	9	8	8	8	8	7	8	8	8	8	7	8	8	8	7.94	.234
41	8	8	8	8	8	8	8	8	8	8	8	8	7	8	8	7	7.87	.219
42	8	8	8	8	7	8	8	7	8	8	8	7	7	7	7	7	7.56	.492
43	7	7	7	8	7	7	7	7	7	7	7	7	6	7	7	6	6.94	.234
44	8	7	7	7	8	8	8	7	8	8	7	7	6	7	7	8	7.37	.547
45	8	8	8	8	8	9	8	8	8	8	8	7	7	8	7	9	7.94	.350
46	8	7	7	7	7	8	7	6	8	8	8	8	6	8	8	9	7.50	.688
47	6	6	6	6	6	7	7	6	7	7	7	6	6	6	6	6	6.31	.429
48	7	6	7	6	6	7	7	6	7	7	7	7	6	6	7	6	6.56	.492
49	7	8	7	7	7	7	8	7	8	7	7	7	7	7	7	7	7.19	.304
50	7	6	7	8	6	7	7	6	7	7	7	7	6	7	7	6	6.75	.469
Averages	7.44	7.08	7.34	7.28	7.30	7.42	7.28	6.84	7.64	7.52	7.54	7.30	6.72	7.30	7.32	7.54	7.30	.3899

TABLE 44

RATINGS BY THE THORNDIKE SCALE UPON A SET OF CHILDREN'S HANDWRITINGS BY TEACHERS. SAME PAPERS AND SAME JUDGES AS IN TABLE 43

PAPER	JUDGES																AVG.	A. D.
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*		
1	9	9	9	9	10	8	9	9	11	11	11	8	11	9	8	7	9.25	.97
2	8	8	9	8	14	9	9	8	14	8	8	11	11	8	9	9	9.44	1.53
3	9	7	9	7	11	9	8	9	9	9	9	9	9	8	8	8	8.56	.72
4	9	13	11	9	11	11	11	13	13	13	11	13	14	11	11	11	10.56	1.20
5	11	8	11	9	10	13	11	8	15	13	14	9	11	9	10	11	10.81	1.59
6	11	11	9	10	9	9	12	9	11	12	12	12	12	11	10	12	10.75	1.06
7	9	9	10	9	8	9	9	8	9	12	15	9	12	9	9	13	9.94	1.54
8	12	11	11	9	9	9	12	8	13	13	14	9	12	9	11	13	10.94	1.58
9	11	11	11	9	10	12	13	8	12	11	14	11	12	11	11	11	11.13	.92
10	9	13	11	10	11	13	16	8	16	11	12	12	13	9	10	12	11.56	1.56
11	11	14	11	10	9	14	11	14	13	14	9	14	13	9	10	11	11.68	1.77
12	9	13	10	9	11	11	11	9	13	11	13	11	11	11	11	11	10.94	.84
13	8	11	11	10	8	11	13	12	11	13	9	13	11	13	11	13	11.13	1.28
14	8	13	11	10	9	13	12	8	11	11	12	12	12	11	11	13	11.06	1.19
15	9	12	12	10	8	9	12	9	12	12	12	12	15	12	10	12	11.13	1.53
16	12	11	14	10	13	13	13	11	14	14	14	14	14	9	11	14	12.56	1.42
17	11	9	14	11	13	14	14	8	13	16	11	9	13	12	13	14	12.19	1.79
18	11	11	13	11	14	13	15	8	11	11	12	12	13	13	11	11	11.87	1.25
19	11	13	12	10	10	13	11	13	11	13	13	11	13	9	11	12	11.63	1.12
20	12	12	12	10	9	9	11	8	9	13	14	9	12	11	11	11	10.81	1.36
21	12	12	13	11	9	9	12	14	12	13	12	12	12	11	11	13	11.75	.97
22	15	15	15	13	11	13	13	13	13	17	13	13	13	14	13	13	13.56	1.02
23	11	13	11	11	11	11	13	8	11	11	16	13	11	13	11	9	11.50	1.31
24	14	15	17	11	11	11	13	8	13	13	14	13	13	13	13	13	12.81	1.28
25	14	11	16	11	12	13	11	11	12	13	13	13	12	13	13	14	12.63	1.05
26	13	15	15	10	9	9	12	8	9	16	16	12	12	12	13	13	12.13	2.02
27	13	12	15	11	9	12	15	9	12	12	13	12	12	12	13	15	12.44	1.23
28	15	14	16	11	9	13	15	11	12	12	12	13	13	13	13	13	12.81	1.23
29	13	11	16	12	13	13	13	11	13	13	13	13	11	16	14	11	12.87	1.05
30	17	16	18	13	15	13	13	13	15	14	13	14	13	13	14	15	14.31	1.30
31	12	13	14	10	14	9	12	8	13	11	16	11	12	11	12	11	11.81	1.46
32	15	11	17	13	15	14	15	9	16	14	11	13	13	14	16	14	13.75	1.56
33	17	16	17	13	12	14	15	9	13	13	13	13	13	13	13	15	13.69	1.48
34	16	13	16	13	13	14	16	11	16	14	14	14	14	14	14	14	14.13	.94
35	14	12	18	14	12	12	12	12	16	15	16	12	15	16	13	13	13.87	1.63
36	16	16	16	15	13	15	12	14	13	16	12	15	16	13	12	15	14.31	1.39
37	15	14	16	13	15	14	15	11	11	14	11	16	16	13	14	14	13.87	1.29
38	15	11	16	13	15	14	15	11	11	14	16	16	16	13	13	13	13.87	1.52
39	15	14	16	13	15	13	16	11	13	14	15	14	13	16	13	14	14.06	1.08
40	16	16	17	14	14	12	15	13	16	13	17	15	18	13	15	15	14.94	1.33
41	16	14	16	13	15	13	14	13	13	13	16	14	14	16	13	13	14.13	1.05
42	15	13	16	13	14	12	14	13	15	12	17	13	13	14	13	13	13.75	1.09
43	11	11	12	15	15	14	9	11	11	13	13	11	9	11	11	11	11.75	1.44
44	12	14	12	13	13	13	13	9	13	13	16	15	15	13	15	13	13.25	1.09
45	16	16	16	16	16	13	15	12	15	15	16	15	13	15	15	15	14.81	.80
46	14	12	15	14	11	11	15	13	17	17	17	11	15	12	15	15	14.00	1.75
47	10	8	9	9	11	9	9	8	9	8	9	11	8	9	9	8	9.00	.62
48	10	8	9	9	9	9	9	9	9	9	8	9	11	8	9	9	9.00	.47
49	11	14	12	12	14	14	13	9	11	11	9	12	13	11	11	11	11.75	1.25
50	11	9	11	12	10	12	12	8	13	12	12	12	11	9	9	12	10.94	1.21
Averages	11.8	12.3	13.0	10.9	11.3	12.5	12.6	9.3	12.7	12.8	13.0	12.4	12.7	11.9	11.5	12.7	12.11	1.222

the tabulations by the Thorndike scale are found in Table 44. A comparison of the variabilities there shown will now be made.

The task of translating the steps of the Thorndike scale into steps of the letter scale is difficult on account of the narrowness of range of the letter scale. However, the method used for making the change in the drawing scales will be employed, except

that the average of the lowest five and the average of the highest five will replace the averages of the lowest and highest two respectively. The low and high extremes by both scales are found by this method to be as follows:

	Low Extreme	High Extreme	Difference
Thorndike scale.....	9.05	14.50	5.45
Letter method.....	6.374	8.064	1.69

Equating these differences or ranges, we have one step of the letter scale equal to 3.22 steps of the Thorndike scale. By Table 43 the average of the average deviations by the letter scale is seen to be .3899. Converting this into steps of the Thorndike scale by multiplying by 3.22, we get 1.255. This is a trifle larger than the average of the variations found by the Thorndike scale, which is 1.222. The situation is reversed, however, if we make the correction of the A. D.'s for coarse grouping. The distributions spread over about six steps in the Thorndike scale, and over about three steps in the letter scale. It seems as nearly correct as we can estimate to subtract .04 steps from the A. D. of the letter scale, and .02 from the A. D. of the Thorndike scale.[1] That will leave the A. D. for the letter scale .3499 steps, or 1.117 in terms of units of the Thorndike scale. In like manner the correction will leave the A. D. for the Thorndike scale 1.202. It seems, then, that as far as this experiment goes, the Thorndike scale does not effect a reduction of variability among judges when the customary standard of the school is used instead of the unfamiliar standard of "zero merit" to "perfect writing." Of course, we must not forget that this lower variability is accomplished with long practice in using the standard of "Grammar Grade Achievement." What practice will accomplish with the standard scale is yet to be discovered.

Variability among the several judgments upon the same paper is not the only phase of variability which we wish to avoid by standardization. Some teachers grade all papers high, while other teachers grade all papers low by the common marking system. It is important to inquire whether that sort of variability is reduced by the use of the scale. To answer this inquiry I averaged the marks of each teacher upon the 50 papers by both

[1] E. L. Thorndike, Mental and Social Measurements, page 55.

methods of rating. These averages are found in the tables, but are reproduced below for purposes of comparison:

Judge	Avg. of Letter Judgments	Judge	Avg. of Scale Judgments
13	6.72	8	9.3
8	6.84	4	10.9
2	7.08	5	11.3
7	7.28	15	11.5
4	7.28	1	11.8
5	7.30	14	11.9
12	7.30	2	12.3
14	7.30	12	12.4
15	7.32	6	12.5
3	7.34	7	12.6
6	7.42	9	12.7
1	7.44	13	12.7
10	7.52	16	12.7
11	7.54	10	12.8
16	7.54	11	13.0
9	7.64	3	13.0
Average	7.30		12.1
A. D.	.167		.725

From these lists of averages we compute the average deviations from the averages in each column and find that the A. D. for the letter scale column is .167 while the A. D. for the scale column is .725. Converting .167 into units of the Thorndike scale by multiplying it by 3.22, we get .538 as the average deviation from the average among the sixteen teachers' average judgments of the fifty papers by the letter scale, as compared with .725 for the deviation among the judgments by the Thorndike scale. From this it seems that the handwriting scale is not entirely successful in leveling the varying standards among judges. Some judges rate all papers high by the scale and some others rate all papers low by the scale, on an average, more than they do by the letter method. In all probability, this would not be true with a group of teachers who had not had long experience in the grades to establish by practice a fairly uniform and fairly definite standard to assign letters by.

To determine whether this variation among judges by the scale was usual or exceptional, I secured the ratings made by six graduate students on each of two sets of papers, thirty-one seventh grade papers, and thirty-one graduate students' papers,[1] using the Thorndike scale. The distributions of these ratings

[1] These data were secured by Messrs. W. T. Bawden and J. Riley, to whom my thanks are hereby expressed for permission to use them.

are given in Table 45. The average of each judge's ratings on the thirty-one papers is given at the foot of the columns. The average deviation from the average among the six averages for seventh grade papers is .883, while for the other set of papers it is .633. If the average of these two marks be taken as typical of the rating of these six graduate students, all of whom have a vital interest in the problem of standardization, we find that it is larger than the deviation among the average judgments recorded for the teachers in Table 44. It seems then that we may expect from unpracticed judges about .75 of a step average deviation of their average judgment of a set of papers from that of any competent group of judges, and from 1 to 1.25 of a step average deviation among a group of judges on the same paper.

TABLE 45

THE JUDGMENTS UPON TWO SETS OF SAMPLES OF HANDWRITING BY EACH OF SIX JUDGES, GRADUATE STUDENTS IN TEACHERS COLLEGE, USING THE THORNDIKE SCALE FOR HANDWRITING

	SAMPLES OF SEVENTH GRADE WRITING								SAMPLES OF GRADUATE STUDENTS WRITING							
Paper	*J.1*	*J.2*	*J.3*	*J.4*	*J.5*	*J.6*	*Avg.*	*A. D.*	*Paper*	*J.1*	*J.2*	*J.3*	*J.4*	*J.5*	*J.6*	*Avg.*	*A. D.*
1	15	15	15	14	16	17	15.33	.78	1	15	12	12	13	13	17	13.67	1.55
2	15	14	14	13	15	16	14.50	.83	2	12	13	11	12	12	9	11.50	1.00
3	10	12	9	11	12	9	10.50	1.17	3	14	11	11	11	11	12	11.67	.89
4	14	9	11	12	14	12	12.00	1.33	4	14	13	12	13	14	11	12.83	.89
5	14	11	12	11	12	13	12.17	.89	5	8	9	8	9	9	8	8.50	.50
6	12	13	14	12	14	15	13.33	1.00	6	14	15	13	14	16	13	14.17	.89
7	15	15	15	14	16	17	15.33	.78	7	12	14	12	15	16	13	13.67	1.33
8	13	13	13	13	14	16	13.67	.89	8	14	13	12	15	14	15	13.83	.89
9	15	15	16	13	12	16	14.50	1.33	9	14	12	11	12	11	13	12.17	.89
10	15	15	13	14	11	16	14.00	1.33	10	12	11	11	12	10	12	11.33	.67
11	15	15	11	12	11	15	13.17	1.83	11	13	12	11	12	12	12	12.00	.33
12	14	13	11	13	11	14	12.67	1.11	12	13	11	12	12	11	14	12.17	.89
13	15	15	14	14	13	15	14.33	.67	13	14	11	11	13	12	11	12.00	1.00
14	11	11	13	12	12	13	12.00	.67	14	10	9	13	11	13	9	10.83	1.50
15	12	15	8	11	14	15	12.50	2.17	15	12	13	10	11	14	9	11.50	1.05
16	13	9	12	11	13	11	11.33	1.00	16	15	15	14	16	16	16	15.33	.67
17	16	11	15	12	14	13	13.50	1.50	17	9	9	8	8	9	8	8.50	.50
18	13	11	13	12	12	13	12.33	.67	18	16	14	12	11	13	16	13.67	1.67
19	15	12	16	13	17	15	14.67	1.45	19	15	13	8	11	10	14	11.83	2.17
20	14	15	12	14	16	15	14.33	1.00	20	11	11	10	11	11	12	11.00	.33
21	11	13	9	10	10	14	11.17	1.55	21	12	14	11	9	9	13	11.33	1.67
22	15	9	12	12	11	16	12.50	2.00	22	15	13	12	11	11	12	12.33	1.11
23	10	9	10	9	10	9	9.50	.50	23	14	12	13	11	14	13	12.83	.89
24	11	10	9	10	12	11	10.50	.83	24	11	12	11	13	15	13	12.50	1.17
25	15	10	14	13	13	16	13.50	1.50	25	14	10	12	14	16	15	13.50	1.67
26	15	9	15	14	11	16	13.33	2.22	26	13	9	10	11	13	13	11.33	1.67
27	12	11	14	13	10	15	12.50	1.50	27	15	12	13	15	15	14	14.00	1.00
28	15	8	8	11	9	12	10.50	2.17	28	12	12	12	12	13	13	12.33	.44
29	12	13	9	13	13	14	12.33	1.22	29	15	9	11	11	11	14	11.83	1.78
30	15	14	13	12	13	16	13.83	1.17	30	12	13	13	14	15	15	13.66	1.00
31	13	15	14	13	10	15	13.33	1.33	31	13	11	14	12	14	15	13.33	1.17
	14.9	12.7	12.9	12.6	12.3	14.9		1.17		13.4	12.0	11.6	11.9	12.9	13.0		1.00

There is one other significant inquiry relating to variability among judgments made by the scale. How does the amount of variation between two successive judgments by the same

judge compare with the variation between different judges on the same papers? For a tentative answer to this question I submit the data[1] which give the two successive judgments of four competent judges upon each of twenty-two specimens of handwriting, the judgments having been made several days apart in each case. These data are given in Table 46. To make the desired comparison it was necessary to calculate the difference between the judgment of each judge and that of each other judge on the same paper, and average those differences and then calculate the difference between the two successive judgments of each judge, and average those differences. The comparison of these two averages makes a fair answer to the question asked above.

TABLE 46

RATINGS OF FOUR GRADUATE STUDENTS OF TEACHERS COLLEGE UPON EACH OF TWENTY-TWO SPECIMENS OF HANDWRITING BY MEANS OF THE THORNDIKE SCALE. A SECOND SERIES OF RATINGS MADE BY THE SAME JUDGES SEVERAL DAYS LATER ARE RECORDED FOR PURPOSES OF COMPARISON

PAPERS	JUDGE I		JUDGE II		JUDGE III		JUDGE IV	
	1st	*2nd*	*1st*	*2nd*	*1st*	*2nd*	*1st*	*2nd*
1	9	9	10	11	11	10	9	8
2	14	14	14	14	13	13	13	13
3	11	10	13	9	12	12	10	10
4	14	14	13	14	14	13	12	13
5	9	9	14	13	10	9	12	11
6	13	11	13	12	11	12	11	12
7	11	10	10	11	9	10	11	14
8	11	10	12	11	12	11	9	11
9	13	12	11	10	11	13	10	11
10	9	9	10	11	9	9	8	10
11	9	9	9	11	10	11	8	10
12	11	10	11	10	12	10	9	10
13	11	11	10	9	10	10	9	9
14	10	9	10	10	8	9	8	9
15	9	9	9	9	10	11	9	10
16	11	10	11	14	14	13	11	11
17	13	13	12	14	14	14	12	13
18	11	11	11	11	11	11	11	11
19	10	10	11	11	11	10	8	10
20	10	10	8	11	11	10	8	9
21	12	12	11	11	13	12	9	11
22	12	10	12	12	14	14	10	12

The differences thus determined are tabulated for each of the twenty-two papers in Table 47. In the column next to the numbers of the papers, for example, are given the differences between the ratings of judges I and II on their first judgments

[1] These data were secured by R. O. Runnells, to whom my thanks are hereby expressed.

of the set, and in the first column under "Later Series" are given the differences between the ratings of the same two judges on their second judgments of the set. All of these twelve differences are averaged for the column "Avg. of 12 differences." In the next column to the right of this column of averages are given the extreme differences between the lowest and highest mark given to each paper in all of the eight judgments. Finally on the right of this column are given the differences between the two judgments of each judge.

TABLE 47

DIFFERENCES AMONG THE RATINGS RECORDED IN TABLE 46, IN TERMS OF STEPS ON THE THORNDIKE SCALE. EACH JUDGE COMPARED WITH EACH OTHER JUDGE AND EACH JUDGE COMPARED WITH HIMSELF

PAPERS	FIRST SERIES						LATER SERIES						AVG. OF 12	GREATEST DIFFERENCE AMONG THE 8 JUDGMENTS	EACH JUDGE WITH HIMSELF				
	Judges I and II	*I and III*	*I and IV*	*II and III*	*II and IV*	*III and IV*	*I and II*	*I and III*	*I and IV*	*II and III*	*II and IV*	*III and IV*			*I and I*	*II and II*	*III and III*	*IV and IV*	*Avg.*
1	1	2	0	1	1	2	2	1	1	1	3	2	1.42	3	0	1	1	1	.75
2	0	1	1	1	1	0	0	1	1	1	1	0	.67	1	0	0	0	0	.00
3	2	1	1	1	3	2	1	2	0	3	1	2	1.58	3	1	4	0	0	1.25
4	1	0	2	1	1	2	0	1	1	1	1	0	.92	2	0	1	1	1	.75
5	5	1	3	4	2	2	4	0	2	4	2	2	2.58	5	0	1	1	1	.75
6	0	2	2	2	2	0	1	1	1	0	0	0	.92	2	2	1	1	1	1.25
7	1	2	0	1	1	2	1	0	4	1	3	4	1.67	5	1	1	1	3	1.50
8	1	1	2	0	3	3	1	1	1	0	0	0	1.08	3	1	1	1	2	1.25
9	2	2	3	0	1	1	2	1	1	3	1	2	1.58	3	1	1	2	1	1.25
10	1	0	1	1	2	1	2	0	1	2	1	1	1.08	2	0	1	0	2	.75
11	0	1	1	1	1	2	2	2	1	0	1	1	1.08	3	0	2	1	2	1.25
12	0	1	2	1	2	3	0	0	0	0	0	0	.75	3	1	1	2	1	1.25
13	1	1	2	0	1	1	2	1	2	1	0	1	1.08	2	0	1	0	0	.25
14	0	2	2	2	2	0	1	0	0	1	1	0	.92	2	1	0	1	1	.75
15	0	1	0	1	0	1	0	2	1	2	1	1	.83	2	0	0	1	1	.50
16	0	3	0	3	0	3	4	3	1	1	3	2	1.92	4	1	3	1	0	1.25
17	1	1	1	2	0	2	1	1	0	0	1	1	.92	2	0	2	0	1	.75
18	0	0	0	0	0	0	0	0	0	0	0	0	0.00	0	0	0	0	0	0.00
19	1	1	2	0	3	3	1	0	0	1	1	0	1.08	3	0	0	1	2	.75
20	1	1	2	2	1	3	1	0	1	1	2	1	1.35	3	0	2	1	1	1.00
21	1	1	3	2	2	4	1	0	1	1	0	1	1.42	4	0	0	1	2	.75
22	0	2	2	2	2	4	2	4	2	2	0	2	2.00	4	2	0	0	2	1.00
Avg.	.86	1.23	1.45	1.27	1.41	1.86	1.32	.95	1.00	1.18	1.05	1.05	1.22	2.77	.50	1.05	.77	1.14	.864

By Table 47, we note that the average difference between one judge and another is 1.22 steps of the scale, while the average difference between the two judgments of the same judge is .864 steps of the scale. A judge is, then, less variable with himself than with other competent judges. A reasonable interpretation of this fact would seem to be that one factor in the production

of the variability in rating among competent judges is the varying standards of merit previously established in the judges' minds. Otherwise, if the variability were produced by chance inaccuracies in comparing the specimen with the scale, there should be found as great differences between successive judgments of the same judge, as between the judgments of different judges. Since this is found not to be the case, it is pretty good indication that when practice with the scale has established in the minds of judges the same standard of merit which the scale represents, the variability among judges will decrease markedly. Certainly for those teachers who begin their teaching with the scale standard as their guide, familiarity with any other standard will be unlikely to enter as a factor to produce variability. There is, therefore, strong probability that the variability found among unpracticed judges is much greater than will be found among those who make regular use of the scale, while the variability by the letter or per cent scales with which we have made comparison in this experiment would be found greater if we had teachers with less experience. In other words, the teachers have reduced the variability shown by the per cent method, by practice at the expense of children, while they have at the same time decreased their capacity for effective use of a standard scale.

It may be worth while to cite as further evidence on this point, the reduced amounts of difference shown in the second series of judgments over those of the first series. These differences are averaged at the foot of Table 47, and from the footings we may determine that the average difference for the first series is 1.35 steps, while the average difference for the second series is 1.09 steps. If this is a fair indication of the effect of practice, we may expect easily to overcome the major part of the variability found in these experiments by a practical use of the scale. Evidence pointing in the same direction may be obtained from Table 44 where the average of the A. D.'s for the first twenty-five papers rated by the judges is found to be 1.274, while the average of the A. D.'s for the last twenty-five papers is found to be 1.17.

IV. The Hillegas Composition Scale

There is less agreement among teachers as to what constitutes merit in composition than in any other subject, probably. This is seen in comparison with drawing by the fact that there are only

10 steps between 0 merit and a practically perfect composition by the Hillegas scale, while there are 17 steps between 0 merit and a practically perfect drawing by the Thorndike scale. In both these scales, a unit of difference is just that amount which is recognized by 75 per cent of the judges, and, therefore, the agreement as to what constitutes merit in drawing is far more general than in composition. This fact makes the production of a scale for general merit a most difficult thing, and also makes necessary the expectation of a high variability among unpracticed judges in rating compositions by the scale.

In order to indicate just how variable are the standards among teachers at present as to what constitutes merit in compositions, and as to how much merit should be expected from children of different grades, ratings by three teachers were secured upon a set of thirty-one compositions, ratings of four other teachers were secured upon a set of forty-two compositions, and ratings of four other teachers upon another set of thirty-seven compositions.[1] The teacher first rated the papers by the common letter method which is in use in the schools of that city, A, B, C, and D being used to designate the four steps from best to poorest. These data are given in Table 48 for the three groups of judges.

TABLE 48

GIVING THE COMMON LETTER RATING UPON THREE SETS OF COMPOSITIONS BY GROUPS OF TEACHERS CALLED JUDGES

	MARK *A*	MARK *B*	MARK *C*	MARK *D*	TOTAL
	First Set of Papers				
Judge I	4	6	7	14	31
Judge II	5	7	5	14	31
Judge III	5	10	6	10	31
	Second Set of Papers				
Judge I	2	12	15	13	42
Judge II	7	11	?	?	
Judge III	10	10	14	8	42
Judge IV	14	6	11	11	42
	Third Set of Papers				
Judge I	19	5	8	5	37
Judge II	9	14	9	5	37
Judge III	10	8	9	10	37
Judge IV	14	7	3	13	37

[1] These ratings were secured by W. H. Smith at East Orange, N. J., to whom my thanks are hereby expressed.

Table 48 calls for little comment. It is very evident that the teachers constituting each of the groups have no common idea of what is an A composition, or a B, C, or D. To discover some means of defining those marks is surely one of the greatest needs of educational administration. Afterwards the teachers were asked to arrange the set of compositions in order of merit from best to poorest. From this arrangement a rank number was given to each composition according to the position given it by each teacher. Needless to say that these positions were assigned by each teacher without the knowledge of the position assigned by previous teachers. The rank positions are listed in Table 49 for each set of papers as ranked by each teacher, called judge.

The only way to get an adequate notion of the difference between the positions of the papers as ranked by one judge and as ranked by another is to examine the tables. There is a little danger, however, of one's being misled by the fact that since the three sets contain unequal numbers of papers, a given difference in rank does not mean the same for the three sets. On this account, and also to make possible definite comparison with other variations in rating, I calculated coefficients of correlation between the relative positions assigned by each judge with the relative positions assigned the same papers by each other judge. Thus for set 1, the relationships between the positions given by judge I and judge II, those given by judge I and judge III, those given by judge II and judge III, were expressed by these coefficients of correlation. The same practice was followed for sets 2 and 3, there being six coefficients found for each set.

The method used for computing these coefficients was that of differences in relative positions or ranks, using the formula, $r = \text{sine}\ \frac{\pi}{2} R$, where $R = 1 - \frac{6\ \Sigma g}{n^2 - 1}$, Σg being the sum of the plus differences in rank, and n being the number of cases, and determining the r value by the use of Table 37 on page 169 of Thorndike's "Mental and Social Measurements" (ed. 1913).

TABLE 49

GIVING THE RANK POSITIONS OF COMPOSITIONS AS JUDGED BY DIFFERENT TEACHERS: THIRTY-ONE PAPERS IN GROUP ONE, RATED BY THREE TEACHERS; FORTY-TWO PAPERS IN GROUP TWO, RATED BY FOUR TEACHERS; THIRTY-SEVEN PAPERS IN GROUP THREE RATED BY FOUR TEACHERS. THE PAPER CONSIDERED *BEST* IS GIVEN RANK 1

SET 1				SET 2					SET 3				
PAPERS	JUDGES			PAPERS	JUDGES				PAPERS	JUDGES			
	I	*II*	*III*		*I*	*II*	*III*	*IV*		*I*	*II*	*III*	*IV*
1	10	11	13	1	31	36	35	39	1	7	9	4	14
2	4	10	14	2	13	13	20	31	2	20	23	16	26
3	3	4	3	3	11	6	11	16	3	13	8	1	12
4	27	27	27	4	8	17	14	2	4	21	16	37	35
5	6	6	4	5	6	2	12	18	5	34	30	35	33
6	12	7	12	6	41	40	38	40	6	17	5	19	15
7	14	12	11	7	3	8	13	1	7	28	34	28	28
8	18	20	19	8	42	43	36	41	8	2	2	21	11
9	25	28	21	9	29	41	24	21	9	8	13	22	13
10	9	16	10	10	2	3	25	13	10	27	26	23	29
11	20	19	22	11	28	22	29	15	11	18	3	14	4
12	1	2	1	12	12	7	7	17	12	10	15	17	8
13	7	24	8	13	9	25	26	26	13	6	4	3	5
14	19	14	18	14	19	20	1	10	14	5	6	5	3
15	16	8	7	15	5	9	2	4	15	26	22	24	27
16	31	18	25	16	27	33	27	27	16	36	12	30	31
17	23	21	26	17	18	14	18	23	17	1	21	13	23
18	30	31	29	18	12	4	3	11	18	23	29	29	24
19	5	3	5	19	26	24	23	30	19	31	11	25	17
20	2	1	2	20	25	11	15	14	20	35	33	31	36
21	13	13	6	21	33	30	37	42	21	37	36	34	32
22	22	17	30	22	34	34	39	37	22	24	28	15	2
23	21	25	24	23	35	26	19	6	23	25	35	36	30
24	29	9	17	24	15	19	17	7	24	14	19	18	7
25	17	5	15	25	36	35	40	34	25	15	20	2	20
26	24	26	31	26	40	42	42	33	26	19	10	6	16
27	8	15	9	27	21	16	21	20	27	11	17	8	18
28	11	22	16	28	16	15	6	9	28	32	14	11	22
29	26	23	23	29	37	32	30	25	29	16	1	7	9
30	15	29	20	30	24	39	31	22	30	12	18	9	10
31	28	30	28	31	14	12	5	3	31	22	24	12	21
				32	23	38	16	19	32	4	7	10	1
				33	7	1	4	12	33	29	27	27	19
				34	10	21	10	32	34	9	32	26	6
				35	30	27	32	38	35	33	25	32	37
				36	4	28	28	28	36	3	31	20	25
				37	17	4	9	8	37	30	37	33	34
				38	1	10	8	5					
				39	32	31	41	35					
				40	22	23	22	29					
				41	39	37	34	36					
				42	38	39	33	24					

These coefficients are given below:

SET 1

Between ranks given by	*Coefficients of Correlation*
Judges I and II	.70
Judges I and III	.88
Judges II and III	.78

SET 2

Judges I and II	.79
Judges I and III	.75
Judges I and IV	.65
Judges II and III	.77
Judges II and IV	.62
Judges III and IV	.78

SET 3

Judges I and II	.53
Judges I and III	.62
Judges I and IV	.72
Judges II and III	.62
Judges II and IV	.66
Judges III and IV	.62

In order to compare the coefficients derived by this method with those which would result from some other method, I used with set 2 the formula,

$r = 2 \sin \frac{\pi}{6} p$ where $p = 1 - \frac{\Sigma D^2}{\frac{n(n^2-1)}{6}}$ where D = differences in rank and n = the number of cases.

I also calculated the Pearson coefficient to show the relation between the positions assigned the papers of the first set by judges I and II. These coefficients are given below:

SET 2

Judges I and II	.80
Judges I and III	.76
Judges I and IV	.70
Judges II and III	.79
Judges II and IV	.71
Judges III and IV	.84

SET 1

Judges I and II	.655 (Pearson coefficient)

The average of all the coefficients listed above is a little less than .72. From this we may get some notion of the extent of the variation in standards among teachers regarding merit in composition work.

While we might expect considerable change of position among the compositions near the middle of the group, we are surprised to see, for example, the composition in set 3 which judge I considers the best, given rank 23 by judge IV. To determine the extent of agreement regarding the best compositions, I averaged the ranks given by the other judges of the group upon the five papers considered best by each judge. To illustrate, in set 1, judge I considered papers 12, 20, 3, 2, and 19 the five best. The average of his ranks on these five papers is, of course, 3. The average of the ranks assigned to the same five papers by the other two judges is 3.5. Carrying out a similar calculation for all the groups we have the following:

	The Papers Ranked 1, 2, 3, 4, and 5 by	Avg. Rank Among Other Judges	
Set 1	Judge I	3.5	
	Judge II	5.4	
	Judge III	3.3	
		Avg.	4.07
Set 2	Judge I	12.3	
	Judge II	10.6	
	Judge III	9.5	
	Judge IV	8.6	
		Avg.	10.25
Set 3	Judge I	13.6	
	Judge II	11.0	
	Judge III	9.8	
	Judge IV	10.4	
		Avg.	11.2

From this it appears that there is an average change of rank among the judgments on the best five papers in each group as follows:

For Set 1 ..... 4.07 minus 3, or 1.07
For Set 2 ..... 10.25 minus 3, or 7.25
For Set 3 ..... 11.2 minus 3, or 8.2

The average change of rank for all the papers for each set was found to be as follows:

For Set 1 ..... 4.28
For Set 2 ..... 6.66
For Set 3 ..... 6.94

From this it appears that the average change of rank among the best papers is much less than among the set as a whole in the case of set 1, but in the two other sets, the changes are greater among the positions of the best papers than among the papers

as a whole. This is significant because we are led by psychologists to believe that in any normal group we should find at the extremes of the distribution a small number whose ability should be easily distinguishable from the ability of the majority of the group. So far as these sets of papers go, either this assumption is not true, or else the teachers are not able to recognize this ability, as shown by the compositions making up sets 2 and 3.

This brief experiment should serve to emphasize two points: The need of standardization in composition work, and the great difficulty in the way of such standardization. We shall now proceed to our examination of the Hillegas scale.

In the following study of the Hillegas scale for English composition, it must not be forgotten that it is but one phase of its usefulness which is being investigated. Our entire thesis pertains to variability among standards of teachers as shown in the marking of students upon daily work and examinations. In the study of the Hillegas scale we confine ourselves to its availability as an objective measure by which variability of rating may be reduced, and to its responsiveness in locating the varying amounts of merit among the several papers to be marked by it. Its great value as a means of defining merit is not examined beyond these two points.

The amount of variability among the marks given by many judges to the same paper is always in terms of the steps used in the scale of marking, and to compare the variability of the marks given by two different methods, it is necessary to equate the steps of the two scales. This is undertaken with the first group of data which follows.

During the summer of 1913 under the direction of Professor Strayer, to whom I am very much indebted for permission to use the data, a set of twenty-eight seventh grade compositions written by the pupils in the schools of Baltimore County, Maryland, were first rated by about sixteen teachers of the county, using the common marking system, 0 to 100, with the step of 5 as the unit. Thus the marks ran 60, 65, 70, 75, and so on. The same papers were then marked by about sixteen teachers from the same group, using the Hillegas scale. The ratings of each teacher were put upon the reverse side of the sheet containing the composition, thus permitting each one to observe the ratings made by previous judges. The caution was given and without doubt

was conscientiously heeded, that each judge should have definitely made up his mind what value he attached to the composition before turning over the paper, and that he should then put that value down regardless of how it differed from marks of previous judges.

In the following table, No. 50, the distributions of marks given by the two methods of rating are given.

TABLE 50

DISTRIBUTION OF RATINGS UPON TWENTY-EIGHT SEVENTH GRADE COMPOSITIONS GIVEN BY ABOUT SIXTEEN TEACHERS IN BALTIMORE COUNTY, MARYLAND. THE COMMON PERCENTAGE METHOD WAS USED FOR THE FIRST RATING, AND THE HILLEGAS SCALE FOR THE SECOND

	PERCENTAGE SCALE RATINGS													HILLEGAS SCALE RATINGS								
Papers	*50*	*55*	*60*	*65*	*70*	*75*	*80*	*85*	*90*	*95*	*100*	*Avg.*	*A. D.*	*183*	*260*	*369*	*474*	*586*	*675*	*772*	*Avg.*	*A. D.*
1			1	2	7	4	2	1				72.0	4.77				3	8	6		6.01	.58
2				1	6	2	4	4	1			77.0	5.80			4	6	4	2		5.02	.81
3				2	2	2	3	5	2			79.1	6.80				5	9		1	6.63	.63
4						3	8	3	2			81.2	3.59				3	5	2	6	6.47	1.19
5				1	0	2	2	1	8	3		86.1	7.15			1	2	3	6	4	6.41	.97
6								1	8	8		92.1	2.75					1	14	1	6.75	.12
7				1		1	5	7	3			82.9	4.53			2	3	4	7		5.85	.87
8			4	3	3	2	2	1				69.4	6.95				8	8			5.29	.50
9	1		7	4	2	2	1					64.7	5.60		1		8	6	1		5.11	.62
10					3	5	5	1	2			78.2	4.85			6	2	5	3		5.08	1.06
11					2	1	4	3	6			83.2	5.85			2		7	9		6.10	.72
12		1		2	5	2	5					72.4	5.80			7	7	2			4.54	.71
13					12	2	2					71.80	2.81			2	4	6	3		5.49	.80
14									6	9	1	92.9	2.70						6	12	7.40	.44
15					3	5	3	3				77.15	4.60				1	4	11		6.42	.52
16				1	1	3		5	6	1		83.55	6.25					6	2	4	6.60	.83
17			1	2	8	1	4	1				72.35	5.40			2	2	4	6	2	6.07	1.00
18					1	6	4	3	1			79.0	4.40				7	7	1	1	5.57	.66
19			1	1	2	1	2	2	5			80.0	8.55				2	3	9	2	6.48	.68
20				1	8	7	1					72.35	3.08				12	3	3		5.29	.67
21	2		8	5	1							60.9	3.65	1	3	5	3	5			4.18	1.09
22		4	4	3	3	2						63.45	5.95		2	1	8	3	2		4.87	.79
23				3	6	5	3					72.35	4.25			3	12	1			4.61	.33
24				1	4	5	5	1	1			76.17	4.90			2	7	3	4		5.37	.88
25				1	6	3	4	1	1			75.3	5.35		2		5	6	4		5.40	.95
26			1	1	5	4	3	2				74.06	5.43			8	6	2			4.35	.62
27			1	1	2	3	2	4	3			78.75	7.65				2	6	9		6.22	.69
28								1	7	9		92.35	2.80				1	2	7	8	6.96	.69
Averages												76.45	5.08								5.69	.722

In calculating the averages and average deviations recorded in the table some difficulties were encountered with the Hillegas scale distributions. The successive steps are not equal, but it was necessary to indicate the deviations in steps. The size of the successive intervals in terms of Median Deviations (one M. D. being that difference in merit which exists between two compositions when just 75 per cent of competent judges recognize the difference) is found by subtracting the value of each sample

from the value of the one next above it, and dividing by 100. The successive steps of the scale are found to be,

1.83	Median	Deviations		.90	Median Deviations
.77	"	"		.97	" "
1.09	"	"		.66	" "
1.05	"	"		.97	" "
1.11	"	"			

In determining the average of the series of judgments assigned to a given paper, no account was taken of the difference in the sizes of steps except in the case of the step where the average fell. The method of deviations from the guessed average was used, and the steps of deviation all counted equal. However, when the location within the step was determined for the true average, the actual difference or size of step was used. For example, in the case of paper 1, the deviations above the guessed average were found to be 6, while the deviations below were 3, thus making a difference of 3. There were seventeen judgments, therefore the true average was calculated as 3/17 of the difference between 5.85 and 6.75, above 5.85, or at 6.01. In calculating the A. D., however, the 3/17 was counted simply as a fraction of a step, and the total steps of deviation determined by adding to 3 plus 6, the product of 8 plus 3 minus 6, and 3/17. Thus 9 15/17 divided by 17 gives the A. D. as .58 steps.

It cannot be claimed that this method of finding either the average or the average deviation is precisely correct. Since, however, the value of the study does not hinge upon the absolute accuracy of either measure, but upon the average of the 28 averages and the average of the 28 average deviations, it is believed that an error in one direction with one, will be balanced by an error in the opposite direction with another, and so in the end substantially the same results will be obtained as if a much more elaborate method had been used.

Turning now to the problem of equating a step of the Hillegas scale with steps of the other scale (called for want of a better designation, the percentage scale) several alternatives presented themselves. The whole number of places on the one could be called equal to the whole number of places on the other, thus making,

1 step, Hillegas scale, equal 10 steps, percentage scale.

This seemed unfair because the range on the percentage scale used by the teachers was near the top, while not a judgment by the Hillegas scale was at either 8.38 or 9.37. Similarly the custom of rarely grading papers below 50 shuts off the use of the lower half of the percentage scale.

Another alternative was to take the range between the highest mark and the lowest mark given to any paper in the group by both methods of rating, and equate those ranges. This alternative was not adopted because it seemed to give undue weight to extreme judgments.

The method which seemed fairest was to equate the ranges between the average of the five lowest papers and the average of the five highest papers found by both scales. This is the method used in the following calculations.

From Table 50 we find the five lowest papers as judged by the Hillegas scale are Nos. 21, 26, 12, 23 and 22. Their values by the two methods of rating constitute the first part of Table 51. The derivation of the other portions of the table will be apparent upon inspection.

TABLE 51

GIVING THE FIVE LOWEST AND FIVE HIGHEST PAPERS AS FOUND BY EACH METHOD OF RATING, AND THE CORRESPONDING VALUES GIVEN THE SAME PAPERS BY THE OTHER METHOD

LOWEST PAPERS BY HILLEGAS SCALE		CORRESPONDING VALUES BY PERCENTAGE SCALE	LOWEST PAPERS BY PERCENTAGE SCALE		CORRESPONDING VALUES BY HILLEGAS SCALE
No. 21,	4.18	60.9	No. 21,	60.9	4.18
No. 26,	4.35	74.06	No. 22,	63.45	4.87
No. 12,	4.54	72.4	No. 9,	64.7	5.11
No. 23,	4.61	72.35	No. 8,	69.4	5.29
No. 22,	4.87	63.45	No. 13,	71.8	5.49
Averages	4.51	68.63		66.05	4.99

HIGHEST PAPERS BY HILLEGAS SCALE		CORRESPONDING VALUES BY PERCENTAGE SCALE	HIGHEST PAPERS BY PERCENTAGE SCALE		CORRESPONDING VALUES BY HILLEGAS SCALE
No. 14,	7.4	92.9	No. 614,	92.9	7.4
No. 28,	6.96	92.35	No. 28,	92.35	6.96
No. 6,	6.75	92.1	No. 6,	92.1	6.75
No. 16,	6.6	83.55	No. 16,	83.55	6.6
No. 19,	6.48	80.0	No. 5,	86.1	6.41
Averages	6.84	88.18		89.40	6.82

From the above table the following ranges appear:

1st. Lowest five Hillegas scale to highest five same scale, 6.84 less 4.51, or 2.33 steps.

2nd. Corresponding values, percentage scale, 88.18 less 68.63, or 19.55.

3rd. Lowest five to highest five, percentage scale, 89.40 less 66.05, or 23.35 steps.

4th. Corresponding values, Hillegas scale, 6.82 less 4.99, or 1.83 steps.

From the three methods of equating possible from the above ranges, we get the following:

Equating the first and second, we have 1 step Hillegas scale equals 8.39 steps percentage scale.

Equating the first and third, we have 1 step Hillegas scale equals 10.02 steps percentage scale.

Equating the third and fourth, we have 1 step Hillegas scale equals 12.76 steps percentage scale.

The average of these three values, 8.39, 10.02, and 12.76 is 10.39, and this is taken to be a fair value of the step in the Hillegas scale in units of the percentage scale. If this be admitted as fair, we may now proceed to compare the variabilities existing with the two methods of rating.

The lists of average deviations for each paper with each method of rating is given in Table 50 under A. D. The averages of these two sets of A. D.'s are 5.08 steps for the percentage scale, and .722 steps for the Hillegas scale. Reducing the latter to its equivalent in units of the percentage scale by multiplying it by 10.39, we have the variability of the judgments by the Hillegas scale represented by an average deviation of 7.50 units of the percentage scale. This, it will be observed, is very much larger than the average deviation of the judgments given the same papers by the percentage method.

A proper correction for the coarseness of grouping in both distributions would operate to reduce the average deviation found for the Hillegas scale more than for the other one. The number of steps in each distribution of judgments by the Hillegas scale is, on the average, a little less than 4, while the number of steps on the percentage scale (each step being 5 units) is, on the average, a little less than 6. From the table given on page 55 of Thorndike's "Mental and Social Measurements," it seems fair

to subtract .04 of a step from the A. D. found for the Hillegas scale distributions, and .02 of a step from the A. D. found for the percentage scale distributions. This makes the corrected A. D.'s .682 steps on the Hillegas scale, and 4.98 units on the percentage scale. (The correction, which is .02 steps, must be multiplied by 5, the number of units in the step, and this product subtracted from 5.08, leaving 4.98.) Reducing the corrected A. D. for the Hillegas scale to its equivalent in percentage units, we have 7.08. This still seems surprisingly large in comparison with 4.98, the average deviation for the percentage ratings.

The conclusion arrived at in the above study pointed to the necessity for further study of the workings of the Hillegas scale. To meet this necessity, the following data were gathered.

The same twenty-eight compositions whose ratings were given in Table 50 were rated by a class of graduate students in Teachers College under the direction of Professor Strayer. The papers were passed to the students who had each a copy of the Hillegas scale. Each one graded the composition in his hands, placing the mark on the reverse side of the paper. On signal, the papers were passed along and graded again. The caution was again urged that each one should have made up his mind definitely what mark the paper deserved before turning it over, and that the mark should not be changed no matter how far it differed from the marks previously recorded. It seems fair to assume that students so interested in education would be able to follow this suggestion.

The papers were passed until each had been marked by sixteen judges. (There were three people who left before the sixteenth judgment was made, thus leaving three papers with but fifteen judgments. These papers are Nos. 12, 25 and 27.) These sixteen successive series of judgments are recorded in Table 52, page 120, as well as the tables of frequency for each paper. The papers are in the same order as given in the previous table, No. 50, so that comparisons of ratings by the two sets of judges may be made.

We note from this table that the variability is greater with this set of judges than it was with the Baltimore County teachers. It will be observed also that the average of the averages is greater by .18 of a step. It is interesting, furthermore, and rather significant, that the averages of the successive series of judgments on

TABLE 52

DISTRIBUTION OF RATINGS ON THE SAME TWENTY-EIGHT SEVENTH GRADE COMPOSITIONS AS WERE USED FOR THE RECORDS IN TABLE 50, BY GRADUATE STUDENTS IN TEACHERS COLLEGE

Each paper (except numbers 12, 25, and 27) was rated by sixteen judges, but not the same sixteen rated all papers. Twenty-eight judges were employed, the papers being passed along from one to another, until each judge had rated sixteen compositions. The judgments are recorded by series, the first column being the judgments rendered first, the second column being the series of judgments rendered next, etc. On the right the ratings upon each paper are grouped into tables of frequency. The Hillegas scale was used, but the record below indicates only the first figure of the assigned value, thus: 2 means 2.60, 3 means 3.69, etc. The numbers of the papers correspond to the numbers used in Table 50.

	SERIES OF JUDGMENTS															TABLES OF FREQUENCY									
Paper	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*	*11*	*12*	*13*	*14*	*15*	*16*	*1.83*	*2.80*	*3.69*	*4.74*	*5.85*	*6.75*	*7.72*	*8.88*	*Avg.*	*A. D.*
1	6	4	5	7	6	6	4	6	4	4	5	6	4	4	7	4				7	2	5	2		5.96	1.02
2	7	6	5	5	6	7	6	5	6	6	5	7	6	5	4	5				1	7	5	3		6.46	.75
3	4	3	4	4	4	5	6	6	4	4	5	6	4	6	7	6			1	7	2	5	1		5.71	1.00
4	6	5	4	6	5	6	6	6	6	6	4	7	5	6	5	4				3	6	7	1		6.13	.75
5	5	4	4	5	7	6	7	7	7	8	7	6	7	6	7	6				2	2	4	7	1	6.93	.94
6	6	6	5	7	6	6	7	6	7	5	6	6	6	5	5	6					4	9	3		6.69	.47
7	5	5	5	5	4	6	6	7	5	5	5	6	5	6	4	6				2	8	5	1		6.13	.58
8	4	5	6	6	5	4	3	3	3	3	4	3	6	5	3	4			6	4	4	2			4.86	.91
9	6	4	4	4	4	4	4	4	4	4	5	5	5	5	6	4				10	5	1			5.23	.55
10	6	6	3	7	7	7	4	4	5	4	6	6	5	6	6	5			1	3	3	6	3		6.25	1.01
11	4	4	6	3	3	4	4	4	5	5	4	5	6	5	6	5			2	6	5	3			5.36	.81
12	3	4	4	4	4	4	4	6	6	6	5	6	6	6	4				1	6	2	6			5.71	.95
13	6	6	6	6	6	6	6	6	6	5	5	5	4	4	6	6				2	4	10			6.30	.62
14	6	6	5	6	6	6	6	7	7	6	7	7	5	6	7	7					3	7	6		6.93	.61
15	3	4	4	4	4	5	7	7	6	6	6	6	6	5	5	6			1	4	9		2		5.71	.69
16	6	4	4	7	6	6	7	7	5	5	6	5	4	6	6	4				4	5	4	3		6.19	.92
17	5	4	4	5	6	5	4	7	6	4	4	4	4	5	6	4				8	5	2	1		5.57	.75
18	6	5	6	6	7	6	6	6	4	6	6	4	6	6	5	5				2	6	7	1		6.25	.69
19	6	7	6	6	7	6	6	6	6	7	6	7	6	6	6	7						11	5		7.05	.43
20	4	4	4	3	6	4	4	3	4	4	3	4	4	4	5	3			4	10	1	1			4.68	.48
21	5	4	4	3	3	2	3	4	4	4	3	4	4	3	4	4		1	5	9	1				4.35	.59
22	6	4	6	6	7	6	5	4	4	4	6	6	4	6	6	6				5	6	5	1		5.96	.78
23	4	4	6	6	4	3	3	4	4	6	4	5	6	6	5	4			2	7	4	3			5.29	.81
24	7	6	4	7	6	4	6	7	4	4	4	4	5	6	6	5				6	3	4	3		6.08	1.03
25	6	7	6	6	6	4	6	6	3	5	6	6	5	6	5				1	1	4	8	1		6.27	.77
26	4	3	3	4	7	5	3	3	6	5	6	4	4	3	5	7			6	4	3	2	2		5.19	1.19
27	3	4	5	6	4	4	4	3	4	3	4	3	4	4	4				5	9		1			4.53	.56
28	4	4	5	7	5	5	7	6	6	7	7	6	4	6	6	6				3	3	6	4		6.47	.82
Averages	5.88	5.48	5.58	6.17	6.17	5.88	6.96	6.14	5.81	6.81	5.81	6.08	5.73	5.98	6.17	5.94	Average								5.87	.768

the whole set of papers varies from a maximum of 6.17 in the fourth, fifth and fifteenth judgments, to a minimum of 5.48 in the second judgment. If such variation is typical among supposedly competent judges, it cannot be held that the scale is very satisfactory as an objective measure, in the hands of unpracticed judges.

Further investigation of the scale was made by examining the ratings upon a set of twenty-eight fifth grade compositions which were marked by about sixteen teachers in Baltimore County, Maryland, and again by the same number of graduate students of Teachers College. The tables of frequency for both sets of judges are given in Table 53.

TABLE 53

DISTRIBUTION OF TWO SETS OF JUDGMENTS BY THE HILLEGAS SCALE GIVEN TO TWENTY-EIGHT FIFTH GRADE COMPOSITIONS BY ABOUT SIXTEEN BALTIMORE COUNTY TEACHERS, AND LATER BY ABOUT SIXTEEN GRADUATE STUDENTS OF TEACHERS COLLEGE

	BALTIMORE COUNTY TEACHERS									GRADUATE STUDENTS								
PAPERS	*1.83*	*2.60*	*3.69*	*4.74*	*5.85*	*6.75*	*7.72*	*Avg.*	*A. D.*	*1.83*	*2.60*	*3.69*	*4.74*	*5.85*	*6.75*	*7.72*	*Avg.*	*A. D.*
1	1		4	5	4	3		4.94	1.03			3	3	5	4	1	5.64	.98
2				7	5	4		5.64	.71				8	7	1		5.36	.56
3		1	7	6	1			4.19	.63			6	7	3			4.54	.61
4		3	8	4				3.76	.50		1	7	6				4.06	.55
5			5	8	2			4.53	.52		1	2	13				4.48	.41
6				7	7	1		5.41	.56			3	6	6	2		5.20	.79
7		5	6	3	1			3.69	.62		1	7	6	1	1		4.35	.75
8			1	9	3	2	1	5.30	.83				7	7	2		5.51	.60
9			4	9	2	1		4.74	.50		1	3	7	3	2		4.88	.79
10			1	4	5	5	1	5.91	.83		1		3	7	5		6.78	.72
11		1	5	6	3	1		4.61	.78			4	9	3			4.68	.47
12				4	6	5	1	6.02	.73			1	5	3	7		5.85	.88
13		1	8	6	1			4.15	.62	1	1	11	2		1		3.82	.66
14					7	9		6.35	.49			2	4	5	4	2	5.85	.95
15					4	9	3	6.69	.47				1	3	11	1	6.52	.50
16					2	12	2	6.75	.25			2	6	2	6		5.57	1.00
17				12	3	1		5.08	.47			5	6	4	1		4.81	.71
18			5	9	1	1		4.61	.55			1	7	4	4		5.37	.81
19		1	6	8	1			4.27	.62			6	8	1	1		4.54	.55
20		1	2	3	8	2		5.30	.87		1	4	4	4	4		5.13	1.08
21			4	7	4	2		5.00	.77			4	7	3	2		4.95	.76
22			4	6	4	2		5.02	.81			1	5	5	5		5.71	.78
23			1	1		10	4	6.79	.59			1	1	5	8	1	6.25	.76
24				6	6	4		5.71	.66			3	3	8	2		5.37	.79
25			3	10	2			4.67	.37				5	4	7		5.96	.77
26		3	5	5	2			4.11	.77			5	9	2			4.54	.51
27				1	11	4		6.02	.41		1		2	8	3	2	5.96	.79
28		3	5	7				3.97	.68			6	8	2			4.48	.56
Averages								5.115	.634								5.184	.718

The averages of the 28 A. D.'s in the case of the Baltimore County teachers and the graduate students, respectively, are .634 and .718. Here again it will be observed that the graduate students vary more in their judgments than do the teachers in the field, although both groups vary less with this fifth grade

set than with the seventh grade set previously examined. This improvement cannot be accounted for by practice either, because it was a different set of judges from those who marked the other set. The difference is probably mere chance, or due possibly in part to a difference in the nature of the two sets of papers which makes one set a little more readily comparable with the scale than the other set.

The average rating upon the set by the Baltimore County teachers is 5.115 and by the graduate students, 5.184, a difference of less than .07 of a step. The remarkable thing, however, is that the difference between the fifth grade set and the seventh grade set is less than .6 of a step according to either set of judges. It will be observed that the average judgments of the two groups of judges on the seventh grade papers differ from each other nearly one third as much as either one differs from the average of the fifth grade papers. Furthermore, the average variation among even the least variable group of judges is seen to be more than the difference between the average values assigned to the two sets of papers. In other words, the variability in steps of the Hillegas scale is nearly twice as great as the difference between the average rating on the fifth grade set and the average rating on the seventh grade set. Or, half the judges, roughly speaking, varied in their judgment on any paper from the average judgment of the group, by more than the difference between the averages of these two sets of papers. This comparison serves, of course, to point out the slight improvement in composition work between this particular fifth grade and seventh grade quite as well as to indicate the extent of variability among the judgments. Incidentally, it may be remarked, that this very service would be impossible even to this rough inexactness without such a scale for measuring the improvement from grade to grade.

Another phase of variability which it is hoped the scale may help to decrease is the variability among the averages of two or more groups of judges upon the same paper. If this decrease is accomplished by the scale, then a supervisor may secure a reliable measure of progress by having several judges rate each paper, even if it were not possible to trust a single judgment.

To determine the extent of this agreement the following table, No. 54, was constructed showing the difference between the averages of two groups of judges upon the same paper. For the

seventh grade set the average of these differences is seen to be .632 steps of the scale, with five differences greater than one step. Twenty-five per cent of the differences are less than .34 and 25 per cent are greater than .84. It will be noted further that the average deviation from the average of the column of average judgments rendered by the graduate students is .55 and for the Baltimore County teachers the same deviation is seen to be .66. Taken together these two figures average less than .632 which is the average of the differences. This signifies that the difference between the averages of two groups of judges upon each paper in this set of papers was greater than the average variation among the judgments given to the different papers in the set.

TABLE 54

DIFFERENCES IN FRACTIONS OF A STEP BETWEEN THE AVERAGE RATING UPON A COMPOSITION BY ONE SET OF ABOUT SIXTEEN JUDGES, AND THE AVERAGE RATING UPON THE SAME COMPOSITION BY ANOTHER GROUP OF ABOUT SIXTEEN JUDGES

The set of seventh grade compositions on the left (from Tables 50 and 52) and the set of fifth grade compositions on the right (from Table 53).

SEVENTH GRADE SET				FIFTH GRADE SET			
Papers	*Avg. Grad. Students Judgments*	*Avg. Balt. County Judgments*	*Difference*	*Papers*	*Avg. Balt. County Judgments*	*Avg. Grad. Students Judgments*	*Difference*
1	5.96	6.01	.05	1	4.94	5.64	.70
2	6.46	5.02	1.44	2	5.64	5.36	.28
3	5.71	5.63	.08	3	4.19	4.54	.35
4	6.13	6.47	.34	4	3.76	4.06	.30
5	6.93	6.41	.52	5	4.53	4.48	.05
6	6.69	6.75	.06	6	5.41	5.20	.21
7	6.13	5.85	.28	7	3.69	4.35	.66
8	4.86	5.29	.43	8	5.30	5.51	.21
9	5.23	5.11	.12	9	4.74	4.88	.14
10	6.25	5.08	1.17	10	5.91	5.78	.13
11	5.36	6.10	.74	11	4.61	4.68	.07
12	5.71	4.54	1.17	12	6.02	5.85	.17
13	6.30	5.49	.81	13	4.15	3.82	.33
14	6.93	7.40	.47	14	6.35	5.85	.50
15	5.71	6.42	.71	15	6.69	6.52	.17
16	6.19	6.60	.41	16	6.75	5.57	1.18
17	5.57	6.07	.50	17	5.08	4.81	.27
18	6.25	5.57	.68	18	4.61	5.37	.76
19	7.05	6.48	.57	19	4.27	4.54	.27
20	4.68	5.29	.61	20	5.30	5.13	.17
21	4.35	4.18	.17	21	5.00	4.95	.05
22	5.96	4.87	1.09	22	5.02	5.71	.69
23	5.29	4.61	.68	23	6.79	6.25	.54
24	6.08	5.37	.71	24	5.71	5.35	.36
25	6.27	5.40	.87	25	4.67	5.96	1.29
26	5.19	4.35	.84	26	4.11	4.54	.43
27	4.53	6.22	1.69	27	6.02	5.96	.06
28	6.47	6.96	.49	28	3.97	4.48	.51
Averages	5.87	5.69	.632		5.115	5.184	.384
A. D. from Avg.	.55	.66			.76	.53	
Middle 50%			.34 to .84				.17 to .50

In the case of the fifth grade set the differences are not so great, being on the average .384 steps, with 25 per cent of the cases below .17 steps and 25 per cent above .50. If we consider the average of the differences found in both groups of papers we find it just slightly above .50 steps. This means that if a child's composition be rated by one set of sixteen judges, and then by another set of sixteen judges (assuming that graduate students and Baltimore County teachers are typical judges, and that these two sets of papers are typical papers) the chances are one to one that the mark will be raised or lowered by one-half step or more.

As a further measure of this agreement between the two sets of judges on the same papers, the coefficients of correlation by the method of unlike signed pairs using the average as the central tendency, were determined between the series of averages, and the following results were obtained:

Seventh grade set:

Baltimore County teachers and graduate students, both using the Hillegas scale .... .33

Baltimore County teachers using the Hillegas scale and same teachers using the percentage scale..... .78

Fifth grade set:

Baltimore County teachers and graduate students, both using the Hillegas scale.. .... .85

It appears from these figures that there is no consistent uniformity in the averages of different groups of judges on the same paper, the coefficient in the case of one set being quite high, but in the case of the other set being quite low.

One more sort of check seems important. Will the same person rating a paper the second time after the lapse of several days tend to rate more consistently than two separate judges? To secure the answer to this question a different method of treatment was necessary from that used previously in the composition study, although similar to that used in the handwriting study. No judge can render many valuable ratings on the same paper because the element of familiarity with the composition soon prejudices his judgment. Consequently, in order to make the most use of two judgments by each judge, it was decided to calculate the difference between each judgment and each other judgment given the same paper by several judges and then compare with the average of these differences the average difference

between a judge's first judgment and his second. For this purpose the data secured during the summer session of 1913 at Teachers College by Mr. R. O. Runnells were given to me. Twenty-three papers were rated by four judges, and then rerated by the same judges after several days. The first and second judgments on these papers are recorded in Table 55, and the differences in judgments in terms of steps of the Hillegas scale are given in Table 56, page 126.

TABLE 55

JUDGMENTS OF FOUR GRADUATE STUDENTS OF TEACHERS COLLEGE UPON EACH OF TWENTY-THREE ENGLISH COMPOSITIONS BY MEANS OF THE HILLEGAS SCALE

A second series of judgments taken several days later by the same judges are recorded for purposes of comparison

PAPERS	JUDGE I		JUDGE II		JUDGE III		JUDGE IV	
	1st	*2nd*	*1st*	*2nd*	*1st*	*2nd*	*1st*	*2nd*
1.....	7.72	7.72	9.37	7.72	7.72	6.75	6.75	7.72
2.....	7.72	7.72	8.38	9.37	8.38	8.38	9.37	9.37
3.....	7.72	7.72	4.74	8.38	7.72	6.75	5.85	7.72
4.....	7.72	7.72	6.75	8.38	7.72	7.72	7.72	6.75
5.....	7.72	7.72	6.75	8.38	6.75	6.75	7.72	6.75
6.....	7.72	7.72	7.72	9.37	8.38	8.38	7.72	8.38
7.....	7.72	6.75	8.38	8.38	7.72	7.72	6.75	none
8.....	7.72	6.75	9.37	8.38	8.38	8.38	7.72	7.72
9.....	6.75	6.75	7.72	7.72	6.75	7.72	7.72	7.72
10.....	7.72	6.75	7.72	6.75	5.85	5.85	5.85	5.85
11.....	6.75	6.75	8.38	6.75	6.75	6.75	5.85	6.75
12.....	6.75	6.75	5.85	6.75	5.85	5.85	6.75	6.75
13.....	7.72	6.75	8.38	9.37	7.72	7.72	5.85	6.75
14.....	7.72	6.75	7.72	6.75	6.75	6.75	5.85	5.85
15.....	6.75	6.75	6.75	8.38	7.72	6.75	5.85	5.85
16.....	7.72	7.72	9.37	9.37	8.38	8.38	8.38	8.38
17.....	6.75	6.75	5.85	6.75	6.75	5.85	6.75	7.72
18.....	6.75	6.75	7.72	6.75	6.75	7.72	7.72	7.72
19.....	6.75	6.75	8.38	6.75	5.85	5.85	5.85	5.85
20.....	6.75	6.75	4.74	6.75	6.75	5.85	6.75	5.85
21.....	6.75	6.75	6.75	7.72	7.72	7.72	6.75	6.75
22.....	6.75	6.75	6.75	6.75	5.85	5.85	7.72	7.72
23.....	6.75	6.75	8.38	6.75	6.75	6.75	7.72	6.75

No clearer evidence of the varying efficiency among judges in the use of the scale could be asked than these tables show. Note, for example, that judge I uses no other points on the scale in either his first or his second judgment than 6.75 and 7.72. No wonder that his first and second judgments show slight differences. On the other hand, judge II, while ranging all the way from 4.74 to 9.37, uses such slight discrimination that there is scarcely any relation between his first and second judgment.

TABLE 56

DIFFERENCES AMONG THE MARKS OF THE FOUR JUDGES WHOSE RATINGS ARE RECORDED IN TABLE 55, IN TERMS OF STEPS ON THE HILLEGAS SCALE

	First Series						Later Series								Each Judge with Himself				
Papers	*Judges I and II*	*I and III*	*I and IV*	*II and III*	*II and IV*	*III and IV*	*I and II*	*I and III*	*I and IV*	*II and III*	*II and IV*	*III and IV*	*Avg. of 12 Differences*	*Greatest Difference among 8 Judgments*	*I and I*	*II and II*	*III and III*	*IV and IV*	*Avg.*
1	2	0	1	2	3	1	0	1	0	1	0	1	1.00	3	0	2	1	1	1.00
2	1	1	2	0	1	1	2	1	2	1	0	1	1.08	2	0	1	0	0	.25
3	3	0	2	3	1	2	1	1	0	2	1	1	1.42	4	0	4	1	2	1.75
4	2	0	0	1	1	0	1	0	1	1	2	1	.75	2	0	2	0	1	.75
5	1	1	0	0	1	1	1	1	1	2	2	0	.92	2	0	2	0	1	.75
6	0	1	0	1	0	1	2	1	1	1	1	0	.75	2	0	2	0	1	.75
7	1	0	1	1	2	1	2	1		1			1.11	2	1	0	0		.33
8	2	1	0	1	2	1	2	2	1	0	1	1	1.17	3	1	1	0	0	.50
9	1	0	1	1	0	1	1	1	1	0	0	0	.58	1	0	0	1	0	.25
10	0	2	2	2	2	0	0	1	1	1	1	0	1.00	2	1	1	0	0	.50
11	2	0	1	2	3	1	0	0	0	0	0	0	.75	3	0	2	0	1	.75
12	1	1	0	0	1	1	0	1	0	1	0	1	.88	1	0	1	0	0	.25
13	1	0	2	1	3	2	3	1	0	3	2	1	1.58	4	1	1	0	1	.75
14	0	1	2	1	2	1	0	0	1	1	0	1	.83	2	1	1	0	0	.50
15	0	1	1	1	1	2	2	0	1	2	3	1	1.25	3	0	2	1	0	.75
16	2	1	1	1	1	0	2	1	1	1	1	0	1.00	2	0	0	0	0	.00
17	1	0	0	1	1	0	0	1	1	1	1	2	.75	2	0	1	1	1	.75
18	1	0	1	1	0	1	0	1	1	1	1	0	.67	1	0	1	1	0	.50
19	2	1	1	3	3	0	0	1	1	1	1	0	1.17	3	0	2	0	0	.50
20	2	0	0	2	2	0	0	1	1	1	1	0	.83	2	0	2	1	1	1.00
21	0	1	0	1	0	1	1	1	0	0	1	1	.58	1	0	1	0	0	.25
22	0	1	1	1	1	2	0	1	1	1	1	2	1.00	2	0	0	0	0	.00
23	2	0	1	2	1	1	0	0	0	0	0	0	.58	2	0	2	0	1	.75
Avgs.	1.18	.57	.87	1.26	1.39	.91	.87	.83	.73	.91	1.00	.64	.93	2.22	.22	1.35	.30	.50	.59

(The coefficient of correlation by the unlike signed pairs method, using the median as the central tendency, is only .16.) In the case of judge III we see a combination of a reasonably wide range of marks with a fair consistency between his successive judgments.

The type of composition seems to have no bearing upon the amount of difference among the judgments. That is, it does not follow that where the judges differ widely one from the other, the judges will likewise differ widely from their previous judgments. The correlation is negative, in fact, between the column of "averages of the 12 differences," and the column "average" of differences between each judge's mark and his own previous mark.

All of the compositions whose ratings have been examined thus far have been those written by elementary school children. The following data are submitted as evidence that the relative variability of marking by the regular percentage method and by the Hillegas scale is not much different when high school

papers are used, and when the rating is done by high school teachers from all the departments of the high school. A group of 24 papers written by the members of a class in English in the Columbus, Ohio, High School,[1] were rated by ten teachers in the same high school on the regular basis of 100. The papers were then given to the same teacher to be rated by the Hillegas scale, with the instructions that they give to each paper the value on the scale assigned to the composition which they considered most nearly equal to it in merit. Both groups of ratings are given in Table 57, page 128.

It will be observed from Table 57 that the average deviation among this group of teachers on these high school papers is greater than the average deviation found for any of the elementary school papers by both the percentage method and the scale method of rating. With these high school papers the average of the A.D.'s for the percentage method of rating is 6.46 units of the scale, and the average of the A. D.'s for the scale method of rating is .875 steps of the scale. If we equate the steps of the Hillegas scale with units of the percentage scale by simply calling equal the range between the average scores of the three lowest papers and the average scores of the three highest papers by the two methods of rating, we get one step of the Hillegas scale equal to 9.49 units of the percentage scale. If we now reduce the average deviation found for the Hillegas scale ratings to its equivalent value in units of the percentage scale by multiplying .875 by 9.49, we get 8.30 as the value in percentage scale units of the average deviation by the Hillegas scale method. This, it will be observed, is considerably larger than the average deviation by the percentage method, that figure being 6.46.

It seems unnecessary to enter into any extended study of this table, since it corresponds in essential respects with what has already been pointed out in the previous tables. One feature, however, is deserving of note. The average of each teacher's ratings on all the papers is given at the bottom of the table. From these averages we may see that while there is still a large range of variation by the scale method, there is a larger range by the percentage method. This indicates that the scale has

[1] The data for this study were procured by Principal A. W. Castle of Columbus, Ohio, to whom my thanks are hereby expressed.

TABLE 57

GIVING THE RATINGS UPON TWENTY-FOUR HIGH SCHOOL COMPOSITIONS BY THE REGULAR PLAN OF PERCENTAGE MARKING BY EACH OF TEN HIGH SCHOOL TEACHERS; ALSO THE RATINGS UPON THE SAME PAPERS BY THE SAME TEACHERS USING THE HILLEGAS SCALE. (IN THE TABLE, 2 STANDS FOR 2.60, 3 FOR 3.65, 4 FOR 4.74, ETC.)

	PERCENTAGE RATINGS												HILLEGAS SCALE RATINGS											
Papers	*Teachers*										*Avg.*	*A.D.*	*Teachers*										*Avg.*	*A.D.*
	1	2	3	4	5	6	7	8	9	10			1	2	3	4	5	6	7	8	9	10		
a....	100	96	90	90	95	80	98	90	95	97	92.9	4.72	9	7	7	6	9	6	7	7	7	8	7.92	.78
b....	60	94	80	90	95	70	85	98	92	94	85.5	9.60	4	7	7	7	8	5	9	8	6	8	7.43	1.16
c....	80	99	80	88	85	80	90	80	90	95	86.5	5.80	7	8	8	6	7	6	6	7	7	6	7.52	.64
d....	75	94	95	85	90	75	80	97	96	99	88.5	7.60	6	8	8	6	6	4	8	8	7	7	7.52	1.04
e....	50	85	75	60	90	60	78	85	85	99	76.7	12.56	5	5	7	6	6	2	8	7	6	6	6.48	1.16
f....	95	92	70	75	75	75	75	90	90	87	82.5	8.20	7	6	6	6	8	4	8	7	5	6	7.04	.96
g....	94	99	85	97	95	85	95	98	98	92	93.9	3.92	8	7	7	7	9	9	6	9	7	7	8.12	.92
h....	91	95	98	70	85	95	70	82	98	90	87.3	8.64	7	8	7	6	8	8	4	8	7	7	7.72	.80
i....	80	85	80	75	85	66	68	80	85	88	79.1	5.76	6	7	7	6	7	8	4	4	5	6	6.75	1.00
j....	80	95	80	80	80	70	82	78	88	80	80.9	4.36	5	7	6	6	6	5	8	6	6	7	6.95	.68
k....	98	97	100	85	80	75	80	87	90	94	88.5	7.00	8	8	8	7	7	6	6	9	7	6	7.85	.84
l....	70	88	70	75	90	70	75	80	80	80	76.9	6.20	6	5	6	5	8	3	4	7	6	5	6.30	1.10
m....	75	92	85	90	95	75	70	98	90	87	85.7	7.36	5	6	7	8	8	4	9	9	7	6	7.72	1.40
n....	70	90	94	85	90	90	75	85	93	85	85.7	5.40	6	7	7	6	8	7	4	8	6	7	7.33	.88
o....	88	99	80	88	95	90	88	85	85	87	88.3	3.92	7	7	6	6	7	6	6	4	6	6	6.85	.54
p....	80	85	70	90	90	75	85	95	80	88	83.7	6.16	6	6	7	6	7	4	6	4	5	6	6.48	.82
q....	98	96	80	80	95	90	80	80	85	90	87.1	6.40	8	8	8	7	8	7	9	9	6	7	8.18	.76
r....	86	98	85	80	95	95	70	90	90	88	87.5	6.00	6	6	7	7	8	5	4	8	6	6	7.04	.96
s....	90	99	80	90	95	97	88	85	94	90	89.9	3.68	7	7	6	7	8	8	6	9	7	7	7.85	.66
t....	92	89	78	93	97	85	85	90	99	93	90.3	4.80	8	7	6	6	7	7	9	8	8	7	7.92	.76
u....	70	85	74	50	85	75	68	70	80	85	74.1	8.20	5	6	6	5	5	6	5	5	5	6	6.12	.42
v....	81	95	80	75	90	75	70	80	80	86	81.1	5.52	7	7	6	6	6	5	3	8	5	6	6.66	.94
w....	85	90	90	70	85	85	85	70	95	90	84.5	6.00	6	6	8	5	6	6	6	7	7	6	7.04	.62
x....	50	80	80	75	80	75	80	88	90	92	78.7	7.12	8	5	6	6	7	4	5	7	5	6	6.48	.76
Avg..	80.7	92.4	82.5	80.7	89.0	79.5	79.6	85.9	89.5	90.7		6.46	7.14	7.37	7.49	6.91	7.83	6.32	7.29	7.68	6.95	7.07		.875

NOTE: The teachers as listed at the top of the table are instructors in the following subjects: 1 and 2, English; 3, science; 4, Latin; 5, geography and history; 6, mathematics; 7, literature; 8, mathematics; 9, German; 10, English.

tended to equalize the standards among the teachers, even though they have varied in their ratings by it more than they varied without it. This may best be seen from a comparison of the deviations from the average of these two lists of average ratings. By the percentage method the average of the ten averages at the bottom of the table is 85.0 and the average deviation from the average is 4.46 units of the scale. In the case of the Hillegas scale ratings, the average of the averages at the bottom of the table is 7.21, and the average deviation is .327. (This is not in terms of steps of the scale exactly, but is nearly enough for practical purposes, since the step of the scale in which this falls extends from 6.75 to 7.72, nearly 1 P. E.) If now we multiply this latter A. D. by 9.49 we get 3.1 as the A. D. among teachers' averages by the Hillegas scale as compared with an A. D. of 4.46 by the percentage method.

In the light of the above findings it is pertinent to examine the derivation of the Hillegas scale to find out whether from its very nature we must expect as wide variability as we have found. In the two criticisms of the scale made by F. W. Johnson in the *School Review* of January, 1913, we have a hint that it must always be found impossible to compare one composition as a whole with each one of a variety of others. He found among the judgments of high school teachers of composition as well as members of a graduate class in educational tests, a wider variability in rating compositions of high school students than is revealed in this study in the rating of fifth or seventh grade papers. He found furthermore a very considerable difference in the average ratings of these two groups of judges upon the same composition. Is there a fundamental reason for this?

In deriving the scale Dr. Hillegas used as the unit of difference in merit that amount which was recognized by 75 per cent of competent judges. The range from 0 merit to the highest, 9.37, represents 9.37 of those units of difference. For the sake of ease in discussion I shall speak of the scale as if it contained 10 steps of equal length extending from 0 merit to 10, and the samples of composition standing at the successive steps of the scale I shall designate by their values as 0, 1, 2, etc. Suppose that a composition of value 6 is being rated by 100 judges. We must expect twenty-five of the judges to rate the paper at 5 or less, and twenty-five to rate the paper at 7 or more. If the judgments

distribute themselves according to the normal surface of frequency, then 2.5 judges will call the paper 9 or better, 6.5 judges will call it 8, 16 judges will call it 7, 16 will call it 5, 6.5 will call it 4, and 2.5 will call it 3 or worse. With this distribution which the derivation presupposes, we have an average deviation from the average of .73 steps of the scale. This, it will be recalled, is larger than most of the A. D.'s actually found with the fifth and seventh grade papers, although smaller than those found with high school papers. It will be found, no doubt, that in the case of a paper possessing marked individuality the ratings will vary even more widely. For example, three seventh grade compositions which were written with the view of meeting the criticisms of the children's classmates when read orally, were rated by fifty graduate students in Teachers College. The distributions of judgments on the three papers are shown in Table 58, where the average deviations are seen to be practically 1.00 on each paper.

TABLE 58

TABLES OF FREQUENCY OF JUDGMENTS OF FIFTY GRADUATE STUDENTS OF TEACHERS COLLEGE UPON EACH OF THREE COMPOSITIONS WHICH POSSESSED STRONG INDIVIDUALITY. HILLEGAS SCALE WAS USED

PAPERS	260	369	474	585	675	772	838	935	AVG.	A. D.
I		1	5	7	14	19	3	1	6.90	.97
II			1	8	11	15	11	4	7.51	1.02
III			3	9	10	17	8	3	7.27	1.07

It is thus seen that the distributions obtained in this study show rather less than normal variability for unpracticed judges (normal being determined by the variability of the judges whose ratings entered into the makeup of the scale). The very effort to define general merit in so complex a thing as a composition by a single example seems to make great variation unavoidable. It does not seem strange, in fact, that a general concept of merit which is *standard for seventh grade*, for example, should become with practice more uniform in the minds of teachers than this definition of merit by a single illustrative composition can be expected to be. An adequate definition of anything so complex as merit in composition work must include separate definitions of merit for the several elements which go to make up the composition. In other words, before we may hope to define merit

so distinctly that teachers will vary little in rating compositions by the definition, we must have a series of scales each one devised for the measurement of a certain feature or phase of merit.

The objection will undoubtedly be raised to the above suggestion that when separate scales are prepared for the standardization of the several elements which enter into the merit of a composition, the scheme will have become so complex that no teacher will have the courage to use the standards. In answer to this objection may it not be well to consider the legitimate use of standard scales? The Hillegas scale was derived on the basis of general merit of compositions. Before such a scale can be applied successfully by teachers they must have a clear conception of the relative values among the elements which constitute merit. It is certain that a single scale of merit cannot give them this conception. It is for the purpose of defining this merit in terms of its various elements that several scales, each based upon a different element, are necessary. With these scales before her she can check up her own concept of merit. Rating compositions is of necessity a subjective process, and the value of a series of scales in composition must be for purposes of definition of the marks to be used, in the mind of the one doing the rating. To measure a set of papers by placing them beside the scale should be a rare exercise of any teacher.

Thus it seems that the simplicity of the Hillegas scale, which commends itself to us so highly, tends to make the operation of rating papers by it very easy, but at the same time ineffective. If we assume that the chief value of scales is the standardization of the concept of merit held by the teaching body, we shall not be afraid of a sufficient degree of complexity to make the scales effective.

In all of this discussion of the Hillegas scale we have not taken account of the effect of practice with the scale. The judges whose ratings enter into the derivation of the scale, as well as all the teachers whose ratings are recorded in these tables, have had no practice with the scale. We have no evidence as to how much a group of teachers would decrease their variability by persistent use of the scale. Such evidence is sorely needed. We have a little evidence in Tables 52 and 56 pointing strongly in the direction of great gain by practice. In Table 52 we have the sixteen series of judgments by the twenty-eight judges. If we compare

the deviations of the first eight with the deviations of the last eight, we find (using simply the averages at the bottom of the table) the average variation of the first eight from the average, to be .21 as compared with .12 for the last eight. Similarly, we have in Table 56 the average of the differences between judgments the first time the judges use the scale, and also the difference between judgments of the same judges with the same papers the second time they use the scale. In the case of the first series of judgments the average difference between judgments by different judges is 1.03 steps of the scale, while with the second series of judgments it is .83 steps. It seems probable then that a considerable amount of the variability will be removed when the judges are practised in the use of the scale.

CONCLUSIONS

1. A given grade or mark means many widely different things to different teachers when they are rating pupils for promotion. As measured by the achievement of the several school groups in their later work this difference amounts in some cases to as much as the difference between a G (good) and F– (fair minus) in elementary schools where the basis of marking includes only the steps P, F, G, and E. In high schools there is enough difference between the standards of schools as wholes that, measured by the achievement of the school groups in later school work, a mark of 70 in one school means more than a mark of 81 in another school having the same passing standard by points. Within the high school and within the college the percentage of pupils which the various instructors fail as a common practice extending over several years varies from 0 to 28, or more.

2. In rating examination papers very great differences of standards appear among supposedly equally competent judges. References to the tables given in the text must be made to determine the extent of this variation for the several subjects and among the several groups of teachers. In the Regents' examinations for New York where only 25 per cent of the papers fall at 75 or above on the scale, and where the passing mark is 60, the state examiners change one fourth of the teachers' marks by 10 points or more, another fourth by from 5 to 10 points, and the remaining half by less than 5. On the whole, the state examiners fail nearly as large a percentage of the papers which the teachers pass as the teachers fail of all the papers written.

3. The effort on the part of Courtis to standardize the ability to do single combinations in arithmetic in the upper grades is a bad educational policy. Probably no uniform test in arithmetic should be given to all ages of pupils.

4. Rating of papers by means of statistically derived scales when the judges are unpractised in the use of the scales but ex-

perienced in marking by the common methods, produces different results for different subjects. In drawing, the variability is greatly reduced by the use of the scale. In handwriting, the variability is about equal with and without the scale. In composition, the variability is somewhat greater with the scale than without it.

BIBLIOGRAPHY

I. *The Theory of Marking and the Derivation and Use of Standard Measures*

AYRES, L. P. A Scale for the Measurement of Quality of Handwriting of School Children. New York, 1912.

——— The Measurement of Educational Processes and Products. New York, 1912.

Baltimore Commission's Report. U. S. Bureau of Education, Bull. 4. 1911.

BONSER, F. G. The Reasoning Ability of Children. Teachers College, Columbia University, Contributions to Education, No. 37. 1911.

BROWN, J. C. An Investigation of the Value of Drill Work in the Fundamental Operations. *Journal of Ed. Psy.*, 2: 81–88.

BROWN, W. The Essentials of Mental Measurement. Cambridge, 1911.

BUCKINGHAM, B. R. Spelling Ability, Its Distribution and Measurement. Teachers College, Columbia University, Contributions to Education, No. 59. 1913.

CATTELL, J. McK. Examinations, Grades and Credits. *Pop. Sci. Mon.*, 66: 367–78.

——— Grading of Mental, Moral and Physical Traits. *Am. Jour. of Psy.*, 14: 310–28.

Cleveland (Ohio) Superintendent of Schools. Measuring Efficiency and Progress. In his Annual Report for 1909, pp. 23–51.

CORNMAN, O. P. Spelling in the Elementary School: an Experimental and Statistical Investigation. Boston, 1902.

COURTIS, S. A. The Comparative Test as an Educational Ruler. *American Education*, 1911, pp. 13–18.

——— Measurement of Growth and Efficiency in Arithmetic. Under various titles in *Elementary School Teacher*, 10: 55–74; 10: 177–99; 11: 171–85; 11: 360–70; 11: 528–39; 12: 127–37; 13: 326–45; 13: 486–504.

——— Standard Tests in Arithmetic; also in Reading, Writing and Composition, with Manual for scoring the tests. Detroit, Mich.

EDGEWORTH, F. Y. The Generalized Law of Error. *Journal of the Royal Statistical Society* (England), 1906, pp. 497–530.

ELLIOTT, E. C. Outline of a Tentative Scheme for the Measurement of Teaching Efficiency. Wisconsin State Department of Education. Madison, 1912.

Experimental Study of Children, including Psycho-physical Measurement. U. S. Bureau of Education, Report 1898, vol. I, pp. 985–1204, and vol. II, pp. 1281–1390.

FALKNER, R. P. Some Uses of Statistics in the Supervision of School Children. *Psychological Clinic*, 2: 227–33.

STRAYER, G. D. and THORNDIKE, E. L. Educational Administration; Quantitative Studies. New York, 1913.

SUZZALLO, H. The Teaching of Spelling. *Teachers College Record*, 12: No. 5.

——— and PEARSON, H. C. Comparative Experimental Teaching of Spelling. *Teachers College Record*, 13: No. 1.

THORNDIKE, E. L. Empirical Study of College Entrance Examinations. *Science*, 23: 839–45.

——— Handwriting. *Teachers College Record*, 11: No. 2.

——— An Introduction to the Theory of Mental and Social Measurement. (Revised.) New York, 1913.

——— The Measurement of Achievement in Drawing. *Teachers College Record*, 14: No. 5.

——— Notes on the Significance and Use of the Hillegas Scale for the Measurement of Quality of English Composition. *English Journal*, 2: 551–61.

WHIPPLE, G. M. Manual of Mental and Physical Tests. 2 vols. Baltimore, 1914.

WHITLEY, M. T. Statistical Study of College Marks. Teachers College, Master's Essay. 1906.

WISSLER, C. The Correlation of Mental and Physical Tests. *Psychological Review*, Monograph Supplements, No. 16.

II. *Studies of Current Marking Systems*

CARTER, R. E. Correlation of Elementary Schools and High Schools. *Elementary School Teacher*, 12: 109–118.

CLEMENT, J. A. Standardization of the Schools of Kansas. Chicago, 1912.

DEARBORN, W. F. Relative Standing of Pupils in the High School and in the University. University of Wis. Bull. No. 312.

——— School and University Grades. University of Wis. Bull. No. 368.

FERRY, DEAN. Grading College Students. Williams College Bull., Series 8, No. 5, 1911.

FOSTER, W. T. Scientific *versus* Personal Distribution of College Credits. *Pop. Sci. Mon.* 78: 378–408.

FOX, W. A. and THORNDIKE, E. L. The Relations between the Different Abilities Involved in the Study of Arithmetic. Columbia Contributions to Philosophy, 11: 138–43.

GRAY, C. T. Variation in the Grades of High School Pupils. Baltimore, 1913.

High School Teachers Association of New York City. Articulation of High School and College. Reports of above association, 1910, p. 49.

HALL, W. S. A. A Guide to the Equitable Grading of Students. School Science and Mathematics, 6: 501–10.

HILLEGAS, M. B. Scale for the Measurement of Quality in English Composition by Young People. *Teachers College Record*, 13: No. 4.

JOHNSON, F. W. A Comparative Study in the Grades of Pupils from Different Elementary Schools in Subjects of the First-Year High School. *Elementary School Teacher*, 11: 63–78.

JOHNSON, F. W. A Study of High School Grades. *School Review*, 19:13–24.

JUDD, C. H. A Comparison of Grading Systems in High Schools and Colleges. *School Review*, 18: 460–70.

——— Reasons for Modifying Entrance Requirements. *Education*, 32: 266–77.

JONES, W. F. A Study of the Problems of Grading and Promotion. Teachers College, Master's Essay. 1908.

KEYES, W. A. The Marking System and Its Influence. Doctor's Diploma Essay, New York University. 1906.

MACMILLAN, D. P. The Physical and Mental Examination of Public School Children in Chicago. *Charities and the Commons*, 17: No. 12.

MEYER, MAX. The Grading of Students. *Science*, 28: 243–52.

MILES, W. R. Comparison of Elementary and High School Grades. University of Iowa, Studies in Education, 1: No. 1.

MILLER, H. I. A Comparative Study of the Grades from Different Elementary Schools in Subjects of the First Year High School. *Elementary School Teacher*, 11: 161–75.

NORSWORTHY, Naomi. The Psychology of Mentally Deficient Children. Columbia Contributions to Philosophy and Psychology, 15: No. 2.

PETTIT, W. A. W. Comparative Study of New York High School and Columbia College Grades. Teachers College, Master's Essay. 1912.

RAFFERTY, W. B. Graduation and Promotion in the Elementary Schools of New York. Teachers College, Master's Essay. 1902.

RICE, J. M. Scientific Management in Education. New York, 1913.

RIETZ, H. L. and SHAVE, J. Correlation of Efficiency in Mathematics and Efficiency in other Subjects. University of Illinois, Bull. 6: 301–04.

RUGGLES, A. M. Grades and Grading. Teachers College, Master's Essay. 1911.

SIES, RAYMOND W. Scientific Grading of College Students. *University of Pittsburgh Bulletin*, Vol. 8, No. 31. 1912.

SIMPSON, B. R. Correlations of Mental Abilities. Teachers College, Columbia University, Contributions to Education, No. 53. 1912.

SLOSSON, E. E. A Study of Amherst Grades. *Independent*, 70: 836–39.

SMILEY, W. S. A Comparative Study of the Results Obtained in Instruction in "Single Teacher" Schools and in Graded Town Schools. *Elementary School Teacher*, 11: 249–65.

SMITH, A. C. Contributions to the Statistical Study of the Abilities of School Children in the Several School Subjects. Teachers College, Master's Essay. 1902.

SMITH, A. G. A Rational College Marking System. *Journal of Ed. Psy.*, 2: 383–93.

SMITH, F. O. A Rational Basis for Determining Fitness for College Entrance. University of Iowa, Studies in Education, 1: No. 3.

Standardization of Education. U. S. Bureau of Education, Report 1910, Vol. I, 99: 84–96.

STARCH, D. The Measurement of Handwriting. *Journal of Ed. Psy.*, 4: 445.

——— Transfer of Training in Arithmetical Operations. *Journal of Ed. Psy.*, 2: 306–10.

STONE, C. W. Arithmetical Abilities and Some Factors Determining Them. Teachers College, Columbia University, Contributions to Education, No. 19. 1909.

STRAYER, G. D. Measuring Results in Education. *Journal of Ed. Psy.*, 2: 3–10.

——— Standards and Tests for Measuring Efficiency of Schools and School Systems. U. S. Bureau of Education Bull., No. 13, 1913.

University of Chicago, President's Report, 1910–11, pp. 91–94.

III. *The Current Examination System*

BAKER, T. O. An Analysis of the Regents Examinations in Relation to Secondary Schools. New York University, Doctor's Diploma Essay. 1896.

BENTON, G. W. Some Problems in Secondary Education. Report of the North Central Association of Colleges and Secondary Schools, 1911, pp. 1–27.

COOLEY, E. G. Value of Examinations as Determining a Teacher's Fitness for her Work. N. E. A. Reports, 1902, pp. 174–82.

DAVIS, H. N. New Harvard Plan for College Admission. N. E. A. Reports, 1911, pp. 567–71.

EDGEWORTH, F. Y. The Element of Chance in Examinations. *Journal of the Royal Statistical Society*, 1890, pp. 460–75, and 644–73.

——— Statistics of Examinations. *Journal of the Royal Statistical Society*, 1888, pp. 599–635.

England, Board of Education. Report of Consultative Committee on Examinations in Secondary Schools. 1911.

FISKE, T. S. Analysis of the Examinations of 1911 of the College Entrance Examinations Board. *Educational Review*, 43: 156–67.

HADLEY, A. T. Use and Control of Examinations. N. E. A. Reports, 1901, 240–50.

International Committee on the Teaching of Mathematics, American Committee, No. 7. Examinations in Mathematics Other than Those Set by the Teacher for his Own Classes. U. S. Bureau of Education, Bull. 1911, No. 8.

KINGSLEY, ~~M. E.~~ Plan for College Admission Proposed by the Secondary Department of the N. E. A. *Education*, 32: 278–83.

LYMAN, C. C. England, Education Department, Special Reports on Examination, 6:107–186.

New York State Department of Education, Annual Reports, Division of Examinations.

NICKOLSON, F. W. Certificate System in New England. *Educational Review*, 42: 486–503.

——— New Methods of Admission to College. *Education*, 32: 261–65.

RUSSELL, J. E. Educational Value of Examination for Admission to College. *School Review*, 11: 42–54.

STARCH, D. and ELLIOTT, E. C. Reliability of Grading High School Work in English. *School Review*, 20: 442–57.

——— ——— Reliability of Grading Work in Mathematics. *School Review*, 21: 254–59.

TUELL, H. E. Public School Teacher and Promotional Examinations. *Education*, 28: 217–23.

WHITE, E. E. Promotion and Examination. U. S. Bureau of Education, Bull. 1891, No. 7.

YOUNG, W. H. The High Schools of New England as Judged by the Certification Board. *School Review*, 5:15.