

RELIABILITY OF GRADES OF TEST PAPERS IN MATHEMATICS.¹

BY D. W. WERREMEYER,

High and Manual Training School, Fort Wayne, Ind.

Teachers of mathematics will no doubt agree that in schools in general much emphasis is placed upon the value that is assigned by the teacher to a test paper. Assuming that the hypothesis is true it is important to know to what degree marks given by teachers are reliable.

Dearborn² in a study of school and university grades points out large inequalities in the standards for grading employed by different teachers. He says that one instructor gave 43% of his students the grade "excellent," and to none the grade "failure." Another one gave no one the grade "excellent" and to 14% of his students the grade "failure." While it is possible that the first instructor had a better class of students than the latter, it is hardly probable that such a difference in grades can be accounted for by saying that it represents actual differences in the classes.

Starch³ in his study on "Grading of High School Work in English," in which he had two test papers graded by one hundred fifty-two teachers, shows great variation in the teachers' estimates of the papers.

He also made a study of the "Reliability of Grading Work in Mathematics."⁴ In this study one hundred forty teachers graded a test paper in geometry. His results showed even a greater variance than in the English papers.

In the present study a definite number of test papers in geometry, algebra, and arithmetic respectively were selected from a regular test given to a regular class. These papers were graded by six teachers in the Ft. Wayne High School for the purpose of obtaining the teachers' estimate of the value of the papers. An effort was made to select the papers so that they would be representative of the best, the medium, and the poorest work of the pupils. The teachers were given their own time to grade the papers so that the work could be done under normal conditions. All papers were graded on a scale of 100.

¹ Read before the Mathematics Section of The Indiana State Teachers Association, December, 1913, Indianapolis, Ind.

² W. F. Dearborn, "School and University Grades," *Bulletin of the University of Wisconsin*, No. 368.

³ Daniel Starch, "Grading of High School Work in English," *School Review*, September, 1912.

⁴ Daniel Starch, "Reliability of Grading Work in Mathematics," *School Review*, April, 1913.

Table I shows the grades given to five test papers in beginning geometry. The papers were numbered consecutively. The teachers are referred to by the letters A, B, C, D, E, and F, respectively. Mr. D and Mr. E were not teaching geometry at the time, but had taught geometry previous to their coming to Ft. Wayne.

A glance at the table makes it apparent that there is a great variation in the teachers' estimate of the value of the papers. The average grade of the six teachers for paper No. I is 80.5; the lowest mark is 63 and the highest 93. This is a range of 30 with a mean variation of 8.5. The average grade for paper No. II is 85.5; the lowest mark is 79 and the highest 90. This is a range of 11 with a mean variation of 3. The average grade for paper No. III is 65.8; the lowest mark is 60 and the highest 74. This is a range of 14 with a mean variation of 4.1. The average grade for paper No. IV is 82.7; the lowest mark is 55 and the highest 96. This is a range of 41 with a mean variation of 9.8. The average for paper No. V is 93.7; the lowest mark is 86 and the highest 97. This is a range of 11, with mean variation of 2.5.

TABLE I.
GRADES GIVEN BY SIX TEACHERS TO FIVE TEST PAPERS IN BEGINNING GEOMETRY.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
A	77	87	61	84	95	-4.2
B	84	88	65	88	95	+11.8
C	63	83	60	92	97	-13.2
D	90	86	70	96	94	+27.8
E	93	90	65	55	95	-10.2
F	76	79	74	81	86	-12.2
Average	80.5	85.5	65.8	82.7	93.7	
Range	30	11	14	41	11	
Mean Variation	8.5	3.0	4.1	9.8	2.5	

Under gross deviation (—) indicates below the average and (+) indicates above the average. Gross deviation takes into consideration all the papers of all the teachers; e. g., the sum of the averages is 408.2.

The sum of Mr. A's grades for all the papers is 404. Hence Mr. A graded 4.2% below the average. This is indicated by -4.2. Two teachers, Mr. B and Mr. D, graded above the average and the remaining four teachers graded below the average. No one teacher estimated all the papers low and no one teacher estimated all the papers high. That is, no teacher had an exceptionally low standard or an exceptionally high standard.

Table II shows the grades for five test papers in 9A algebra. The papers are numbered I, II, III, IV, V, as before and graded by the same teachers who graded the geometry papers mentioned above.

TABLE II.
GRADES IN 9A ALGEBRA.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
A	59	66	87	66	96	+3.3
B	55	65	73	51	90	-36.7
C	64	78	88	68	98	+25.3
D	65	77	80	59	95	+5.3
E	69	78	81	66	96	+19.3
F	54	63	74	69	94	-16.7
Average	61	71.2	80.5	63.2	94.8	
Range	15	15	15	18	8	
Mean Variation	5.0	6.5	4.8	5.4	1.9	

The gross deviation shows that Mr. B and Miss F graded below the average, and the remaining four teachers graded above the average. As in the previous set no one teacher estimated all the papers low and no one teacher estimated all the papers high.

Table III shows the grades for five test papers in 8B arithmetic. This was a class in the Training School connected with the School of Education at Indiana University. The papers were numbered I, II, IV, V respectively, and were graded by seven graduate students at Indiana University. Not all these students were teachers of mathematics, but all were experienced teachers with much practice in mathematics. The teachers are referred to by the letters: T, U, V, W, X, Y, and Z respectively. Mr. Z has a Doctor's degree from Teachers College, Columbia University.

TABLE III.
GRADES IN 8B ARITHMETIC.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
T	86	70	90	98	48	-12.2
U	90	75	95	100	50	+5.8
V	75	75	100	96	60	+1.8
W	95	78	100	100	69	+37.8
X	80	75	90	98	53	-8.2
Y	80	81	94	97	67	+14.8
Z	50	65	85	90	75	-39.2
Average	79.4	74.1	93.4	97	60.3	
Range	45	16	15	10	27	
Mean Variation	9.7	3.8	4.3	2.3	8.6	

The gross deviation shows that Mr. T, Mr. X, and Mr. Z graded below the average, and the remaining four teachers graded above the average. No one teacher graded the highest on all the papers. Of all the teachers, Mr. Z graded the lowest on papers I, II, III, and IV, and the highest on V. This may be accounted for by the fact that Mr. Z had been connected for several years exclusively with university and college work. Hence it was more difficult to adapt himself to elementary arithmetic papers.

Table IV shows the grades for five test papers in an arithmetic class in the School of Education in Indiana University. The papers were numbered I, II, III, IV, and V respectively, and graded by the same seven graduate students as in Table III.

TABLE IV.

GRADES GIVEN BY SEVEN TEACHERS TO FIVE TEST PAPERS IN AN ARITHMETIC CLASS IN THE SCHOOL OF EDUCATION AT INDIANA UNIVERSITY.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
T	68	70	98	38	77	-11.8
U	69	70	35	36	76	-16.8
V	87	79	58	50	86	+57.2
W	65	64	42	45	80	-6.8
X	62	50	30	20	67	-73.8
Y	72	71	43	45	86	+14.2
Z	71	71	67	48	83	+37.2
Average	70.6	67.9	44.7	40.3	79.3	
Range	25	29	37	30	19	
Mean Variation	5.2	6.2	10.1	7.7	5.1	

The gross deviation shows that three teachers graded above the average and four teachers graded below the average. Mr. X graded the lowest on all the papers, but no one teacher graded the highest on all the papers. This would indicate that Mr. X sets a high standard.

Table V shows the results of two gradings of the set of algebra papers referred to in Table II. The second grading was done about seven months later by the same teachers, except Mr. B, who had resigned. The teachers were not told that it was the same set of papers which they had graded before. Under the column "Difference" is indicated how much the grades for the respective papers by the respective teachers differed the second time from the first time; "+" indicates higher than the first time; "-" indicates lower; and "0" indicates that the same grades were given.

TABLE V.

TWO SETS OF GRADES GIVEN BY THE SAME TEACHERS TO A SET OF ALGEBRA PAPERS.

Teacher.	First Grading					Second Grading.					Difference.				
	I.	II.	III.	IV.	V.	I.	II.	III.	IV.	V.	I.	II.	III.	IV.	V.
A .	59	66	87	66	96	62	76	79	64	94	+3	+10	-8	-2	-2
B .	55	65	73	51	90	Resigned					Resigned				
C .	64	78	88	68	98	62	77	86	68	98	-2	-1	-2	0	0
D .	65	77	80	59	95	65	80	80	66	95	0	+3	0	+7	0
E .	69	78	81	66	96	65	89	86	69	97	-4	+11	+5	+3	+1
F .	54	63	74	69	94	67	78	76	66	95	+13	+15	+2	-3	+1

Of the twenty-five grades given to the five papers by the five teachers, five were the same both times, twelve were higher, and eight were lower the second time. The greatest variance was 15%. If the same teacher varies to this extent on the same paper, it is no wonder that different teachers vary more.

Table VI shows the grades for five test papers in 9B algebra. The papers were numbered I, II, III, IV, and V, as before and graded by the teachers of the Department of Mathematics in the Ft. Wayne High School. Three of the teachers are the same as those that graded the papers in Tables I and II, and three are new.

TABLE VI.

GRADES IN 9B ALGEBRA.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
A	53	93	90	70	98	+2.5
B	48	90	85	82	98	+1.5
C	51	90	90	85	86	+0.5
D	58	90	85	85	87	+3.5
E	48	90	73	80	99	-11.5
F	52	92	88	79	94	+3.5
Average	51.7	90.8	85.2	80.2	93.7	
Range	10	3	17	15	13	
Mean Variation	2.7	1.1	4.2	3.8	4.8	

The gross deviation shows that Mr. E graded below the average, and the remaining five teachers graded above the average. No one teacher estimated all the papers low and no one teacher estimated all the papers high.

Table VII shows the grades for five test papers in 9B algebra. These papers were graded by the same teachers as in Table VI, after a discussion of the grades in Table VI, in a teachers meeting.

TABLE VII.
GRADES IN 9B ALGEBRA.

Teacher.	I.	II.	III.	IV.	V.	Gross Deviation.
A	96	58	58	90	92	+1.7
B	95	49	56	85	86	-21.3
C	100	62	53	90	88	+0.7
D	100	60	59	90	93	+9.7
E	100	58	66	90	94	+15.7
F	95	64	54	90	82	-7.3
Average	97.7	58.5	57.7	89.2	89.2	
Range	5	15	13	5	12	
Mean Variation	2.3	3.5	3.3	1.4	3.8	

The gross deviation shows that Mr. B and Miss F graded below the average and the remaining teachers graded above the average. No one teacher estimated all the papers low, and no one teacher estimated all the papers high. A comparison of Tables VI and VII shows that:

1st. While Mr. B graded 1.5 above the average on the first set, he graded 21.3 below the average on the second set.

2nd. While Mr. E graded 11.5 below the average on the first set, he graded 15.7 above the average on the second set.

3rd. While Miss F graded 3.5 above the average on the first set, she graded 7.3 below the average on the second set.

4th. The remaining three teachers did not vary a great deal from the average in the two sets.

5th. The total mean variation was reduced by 2.3.

6th. A discussion of the grades given to any set of papers has a tendency to result in more uniformity.

CONCLUSION.

The above tests show that a group of teachers differ greatly as to the value that is placed upon a geometry, algebra, or arithmetic test paper. It does not show that any one teacher has an exceptionally high standard, nor that any one teacher has an exceptionally low standard. Mr. Z in one set of papers graded four out of five lower than any other teacher, but on the other hand he graded one paper out of the five higher than any other teacher. Even the same teacher finds it difficult to place the same value upon the same test paper twice. (Compare Table V.) It seems that the teacher's own standard is constantly changing, or else he has no definite standard of his own to start with.

The question, "To what degree are marks reliable?" is an interesting problem that we have no data in the present study for

answering. It is also a very important question; e. g., if a county superintendent thinks that a certain paper is worth 50%, then the writer of that paper is not qualified to do high school work, or he is not qualified to teach; but if the superintendent thinks the paper is worth 75% then the writer of that paper is ready for high school, or perhaps ready to take charge of a school. It would seem to be a good plan for the teachers of a department or of a school to get together and discuss the basis for grading so as to arrive at some uniformity.

It is sometimes said that a paper in mathematics is easily graded; that it is either right or wrong. While this is true in a measure, it is not wholly true. Of course a demonstration in geometry cannot be right if it is wrong; it is not wrong if it is right. But there are other things to be taken into consideration; e. g., the figures should be drawn correctly, the special enunciation should be given, facts should be stated with reasons given leading up to the thing to be proved. If the pupil does these things, it is worth something. The question is, "How much?" Some teachers insist upon the shortest proof possible, others will give credit for a longer proof. To what degree should the former take precedence? This very fact accounts for the low grade of 55% by Mr. E for paper No. IV, Table I.

Before any definite and lasting improvements are made in the system of grading, before any degree of uniformity is reached, the systems must be worked over and clarified from the foundation up.

First of all, there must be an established and definite idea of that for which grades are given. At present marks stand for many things. They stand for latent ability, for work actually done, improvement, and so on. Marks are given for various complexes of intellectual, moral, and social qualities. Those given by many teachers are influenced by matters of discipline. These factors, of course, apply to a teacher's own pupils. They cannot enter when pupil and teacher do not know each other as is usually the case in a study of this kind.

Then as to the various qualities entering into the complex for which grades are assigned, each should have a fixed relative value assigned to it, tending toward more uniform and equitable grading.

These are the problems which must be solved before we can expect anything like an equitable system of grading. These must be at least partially settled before there is any use of going into

the question of the best means of determining this "mental ability," this "something" which we are at present grading, although not being sure of just what it is. Then and only then will we be able to tell just what part properly planned and properly conducted examinations play in assisting the teacher and the examiner in determining this ability of which we wish to know.

RED NOT A SATISFACTORY DANGER SIGNAL.

Red has been the sign of danger and a warning signal since the earliest times. Just why it was selected as a danger-warning is a question for the anthropologist and historian to determine. It is unfortunate that this color, which is becoming increasingly important with the growing danger of accidents in civilized life, is the color to which many human eyes are insensitive. Color-blindness is apparently becoming more common. In its most frequent form, it is impossible for the color-blind person to distinguish red from green, yet those two colors, which are the most confusing to the human retina, are the very ones which are in most common use as signals for danger and caution. So common is red and green color-blindness that all licensed pilots, masters of vessels, engineers, firemen, motormen and others employed in directing vessels, trains, trolley-cars and other means of transportation are required to submit to a color-test and to prove that they possess an accurate degree of red-green color perception. The simple expedient of selecting as a sign of danger a color to which practically all human eyes are susceptible has only recently been suggested. *Drugs, Oils and Paints*, in a recent issue, contains an article by Dr. Francis D. Patterson, suggesting a new signal to take the place of the familiar red warning. Patterson calls attention to the fact that the number of industrial accidents is at present enormous and is apparently increasing. As approximately one male in every twenty-five has a deficient color perception and as most of these have an impaired sensibility for red, Dr. Patterson argues that the retention of this color as a danger-signal is simply inviting further increase in accidents. His objection is based on the fact that many persons are color-blind to red and are consequently not only barred from any occupation in which a color perception is necessary, but are also deprived of the protection from accidents and danger supposed to be offered by danger-signals. He also objects to red for practical reasons; it is a fugitive color, difficult to distinguish, fading on exposure to sunlight and requiring frequent repainting. The possibility that red and green color-blindness will increase rather than diminish in the future only serves to emphasize the unfitness of these colors as signs of danger and caution. Experiments with the spectrum and with color-blind persons, as well as with various colors at different distances, leads Patterson to the conclusion that yellow and blue are the best colors for danger signals, as he says that they are the only colors which give rise to a normal color-sensation as soon as they become visible, are the most luminous colors of the spectrum, and are permanent and fast, while color-blind persons react normally to them. It has long since passed into a proverb, says *The Journal of the American Medical Association*, that it is easier to change the laws of the people than to change their customs. The fact that many persons are unable to distinguish red from other colors should alone be sufficient to cause it to be discarded as a danger signal. Whatever color is adopted should be selected after the most careful physiologic and optical investigation.