Reliability and Distribution of Grades

Author(s): Daniel Starch

Source: *Science*, Oct. 31, 1913, New Series, Vol. 38, No. 983 (Oct. 31, 1913), pp. 630-636

Published by: American Association for the Advancement of Science

Stable URL: https://www.jstor.org/stable/1640875

versity of Christiania, has retired. Mr. Askel S. Steen succeeds him in these capacities.

Charles F. Brooks

Harvard University

SPECIAL ARTICLES

RELIABILITY AND DISTRIBUTION OF GRADES

If we consider grades scientifically as a scale of measurements, two important questions arise: (1) How fine a scale of units is distinguishable, and (2) What proportion of persons will ordinarily fall under each unit?

First, let us examine the question as to the size of distinguishable steps. The answer to this question can be determined by the reliability with which marks can be assigned. Recent studies have revealed an exceedingly wide divergence in the grades assigned by different teachers to the same papers. Starch and Elliott[1] found that the grades assigned to two English papers by 142 teachers of English ranged in the case of one paper from 64 to 98 with a probable error of 4.0, and in the case of the other paper from 50 to 98, with a probable error of 4.8. This wide range is not due to the fact that these were language papers, since the grades of a mathematics paper assigned by 118 teachers of mathematics ranged from 28 to 92, with a probable error of 7.5 points.[2]

What bearing do these facts have upon the reliability of marks and how are we to explain such wide ranges of differences? Four major factors enter into the problem which, I believe, fully account for the situation: (1) Differences among the standards of different schools, (2) Differences among the standards of different teachers, (3) Differences in the relative values placed by different teachers upon various elements in a paper, and (4) Differences due to the pure inability to distinguish between closely allied degrees of merit.

How much of the variation is due to each factor? To determine the strength of the first factor we must find out the range of variation in the grades assigned by teachers in the same institution and departments instead of different institutions. To this end I obtained ten papers written in the final examination in freshman English at the University of Wisconsin, and had them graded independently by ten instructors of the various sections of freshman English. An effort is made by cooperation among the instructors concerned to have as much uniformity as possible in the conduct of these sections. The same final examination is given to all.

Table I. gives the marks assigned by each instructor to each paper. The first column contains the grades assigned by the teachers under whom the students took the course. Papers 6 and 10 were obtained from the class of one instructor and all the other papers from the class of another instructor. These ten

TABLE I

Papers	Instructors										Average	Mean Var.	Coefficient of Variability
	1	2	3	4	5	6	7	8	9	10			
1	85	86	88	85	75	80	88	87	85	87	84.6	2.8	.034
2	77	80	87	80	62	82	82	87	85	87	80.0	4.6	.057
3	74	78	78	75	69	84	91	83	79	80	79.1	4.4	.056
4	65	65	62	20	26	60	55	68	55	50	52.6	12.3	.233
5	68	82	78	82	64	88	85	86	78	80	79.1	5.7	.070
6	94	87	93	87	83	77	89	88	88	89	87.5	3.2	.036
7	88	90	95	87	79	85	96	91	87	89	88.7	2.6	.029
8	80	84	73	79	72	83	85	91	77	76	80.0	4.6	.058
9	70	70	68	50	44	65	75	81	79	79	68.1	9.1	.118
10	93	92	85	92	81	83	92	89	84	85	87.6	4.0	.045
Av.	79.4	81.4	79.8	73.7	65.5	78.7	83.8	85.1	79.7	80.2		5.3	.074

General average 78.7.

[1] D. Starch and E. C. Elliott, *School Review*, 20: 442–457.

[2] D. Starch and E. C. Elliott, *School Review*, 21: 254–259.

papers were graded after each instructor had graded the papers from his own sections.

(1) The table reveals an exceedingly wide range of marks, a range just as large as that of the English and mathematics papers referred to above. The average of the mean variations is 5.3 as compared with an average of 5.4 of the English and mathematics papers. (2) The mean variations are fairly uniform for all papers except 4 and 9. These two, no doubt, vary so much more widely than the others because both have an average below the passing grade. Judgments of such papers are more apt to be haphazard since, from the practical point of view, it makes no difference what the grade is, so long as the paper is considered a failure. But the matter is quite serious in case of a paper like number 9 which is considered above passing by six and below passing by four instructors. (3) A third point of interest is the fact that the teachers under whom the students took the course grades in column 1, did not succeed in grading the papers any more accurately than the other instructors who did not know the students at all. The mean variation of the grades in column 1 from the average of each paper is practically as large, 4.7, as the mean variation of all together, 5.3. (4) There is a very noticeable difference in the standard of grading. Two instructors, 4 and 5, graded on the whole very much lower than the average and Nos. 7 and 8 graded higher than the average. These deviations can be found readily by comparing each instructor's average with the general average.

In order to eliminate the variation in the marks due to this difference in standards among the instructors, all the marks in Table I. were weighted by the amount that each instructor's average differed from the general average. The weighted values thus obtained are presented in Table II. The decimals were dropped in the transposition.

The differences in Table II. therefore represent the differences in the relative evaluation of the papers themselves irrespective of whether an instructor marks severely or leniently. It will be noticed that the mean variation is smaller, though not as much smaller as one might anticipate, being 4.3 as compared with 5.3 in Table I.

TABLE II

Papers	Instructors										Ave.	Mean Var.
	1	2	3	4	5	6	7	8	9	10		
1	85	84	87	90	89	80	83	81	86	86	85.1	2.5
2	77	78	77	85	76	82	77	81	86	86	80.5	3.5
3	74	76	77	80	83	84	86	85	80	79	80.4	3.3
4	65	63	61	25	40	60	50	49	56	49	51.8	9.2
5	68	80	77	87	78	88	80	79	79	79	79.5	2.9
6	94	85	92	92	97	77	84	83	89	88	88.1	4.7
7	88	88	94	92	93	85	81	85	88	88	89.2	2.6
8	80	82	72	84	86	83	80	85	78	75	80.5	3.5
9	70	68	67	55	58	65	70	75	80	78	68.5	6.0
10	93	90	84	97	95	83	87	83	85	84	88.1	4.5
Av.												4.3

The next step is to separate the third and fourth factors, *i. e.*, how much of the variation is due to the inability to distinguish between closely allied degrees of merit, and how much is due to differences in relative value placed by different instructors upon various aspects of a given paper, such as form, neatness, clearness, etc.

The accuracy of the ability to distinguish between various shades of merit may be ascertained by having the same person give two or more evaluations of the same papers separated by sufficiently long intervals of time, so that the details and identity of the papers have been forgotten. I have tested this point by determining how closely an instructor is able to agree with his own grades. Table III. gives pairs of grades assigned at different intervals to the same papers by the same instructor. In each case the papers were from the instructors' own classes. The aim was to have ten papers re-graded, but in some instances not that many were available.

Table III. shows that the difference in the marks assigned to the same papers by the same instructor is on the average 4.4 points, or in terms of mean variation 2.2 points. This difference is as large in one sort of papers as in another. It is as large in mathematics as in language or in science. This was to be ex-

pected in view of the fact stated at the beginning that mathematical grades are no more accurate than any other grades. The marks of the second mathematics instructor are so close, not because it was mathematics that he was grading, but because this instructor had a purely mechanical method of grading, of deducting so many points for each kind of error. weighting the second set of marks by the difference between the averages of the two markings. Without giving these weighted values in a separate table it will be sufficient to say that the average difference thus computed is 3.5 as compared with the average difference of 4.4 in Table III., or in terms of mean variation, 1.75 and 2.2, respectively.

TABLE III

Advanced Psychology, Interval 2 Yrs.			Elem. Psychology, Interval 2 Weeks			Math., Interval 9 Mos.			Math., Interval 9 Mos.			English, Interval 6 Mos.			German, Interval 6 Mos.			Elem. Psychology, Interval 4 Yrs.		
1st	2d	Dif.	1st	2d	Dif.	1st	2d	Dif.	1st	2d	Dif.	1st	2d	Dif.	1st	2d	Dif.	1st	2d	Dif.
85	87	2	85	79	6	36	51	15	56	60	4	70	75	5	79	70	9	70	80	10
76	80	4	87	83	4	61	67	6	70	73	3	80	86	6	90	77	13	93	91	2
83	80	3	90	93	3				77	75	2	88	88	0	77.5	73	4.5	82	84	2
89	90	1	90	92	2	61	67	6	88	90	2	74	76	2	85	81	4	75	82	7
84	83	1	83	88	5	73	79	6	62	62	0	77	76	1	78	80	2	75	86	11
93	88	5	78	79	1	81	86	5	89	87	2	85	86	1	70	61	9	78	81	3
84	75	9	93	89	4	71	63	8	82	80	2	65	65	0	72.5	58	14.5	88	90	2
93	88	5	88	88	0	71	79	8	53	56	3	68	75	7	91	86	5	83	78	5
89	85	4	78	76	2	96	87	9	75	75	0				62.5	60	2.5	93	93	0
92	86	6	83	80	3	83	90	7	67	64	3				66	65	1	83	87	4
Av. 86.8	84.2	4	85.5	84.7	3	70.3	74.3	7.8	71.9	72.2	2.1	76.0	78.4	2.8	77.1	71.1	6.5	82.0	85.2	4.6

Average of all the differences 4.4 points.

But this does not mean that his grades were more accurate or just. Another instructor might with perfect justice deduct either more or less for the same kind of error. All that it means is that this instructor was able by means of his mechanical method to match his own marks fairly closely. Furthermore, we must not infer that the other instructors had graded their papers carelessly either the first or the second time, or both times. As a matter of fact, each question had been graded in both markings of all papers except the second and third group of psychology papers and the English papers. And these are not essentially different from the rest. The results, while obtained from only seven instructors (more were not available for the purpose) are quite representative and reliable as any one familiar with statistical methods can determine from the above data. Results from twice or three times as many persons would not be materially different.

We may eliminate one further factor from Table III., namely, the difference due to a change in an instructor's standard after an interval of time. This may be eliminated by

Of the four factors stated at the outset, each contributes the following amount to the total variation: The general mean variation or probable error of grades assigned by teachers in different schools is 5.4 points. The mean variation of grades assigned by teachers in the same department and institution is 5.3. The mean variation of the latter, after eliminating the effect of high or low personal standards, is 4.3. The mean variation of grades assigned at different times by the same teachers to their own papers is 2.2. Hence the largest factors are the second, third and fourth. The fourth contributes 2.2 points, the third 2.1 points, the second 1.0 point and the first practically nothing toward the total of 5.4 points of mean variation.

Now what do all these results mean? How small divisions on our scale are practically usable? As a question of psychological methodology the units of any scale of measurements, if a single measurement with the scale is to have objective validity, should be of such a size that three fourths of all the measurements of the same quantity shall fall within the limits of one division of the scale. For

example, if the marks assigned by 75 out of 100 teachers to a given paper lie between 80 and 90, then the unit of our scale should be ten points. Any smaller division would have little or no objective significance. Of course, almost indefinitely small differences in merit can be measured if an indefinite number of independent estimates is made.

Now what are the actual facts with regard to the size of distinguishable steps in the marking scale? We have seen above that the mean variation of the estimates of a teacher in matching his own marks, after eliminating his own change in standard, is 1.75 points. According to our principle that if a unit is to be large enough in range to include three fourths of all his estimates of the same quantity, then the smallest distinguishable step that can be used with reasonable validity is $2\frac{3}{4}$ times the mean variation (1.75) or probable error, which would be 4.8, or roughly 5 points.[3]

Hence our marking scale, instead of being 100, 99, 98, 97, 96, 95, etc., should be 100, 95, 90, 85, 80, etc. These are the smallest divisions that can be used with reasonable confidence by a teacher in grading his own pupils. This means that on a scale of passing grades of 70 to 100 only seven division points are distinguishable. This substantially confirms the scheme followed in many institutions that the marking scale should be $A+$, $A-$, $B+$, $B-$, $C+$, $C-$, $D+$, $D-$ and failure. No medium A, B, C or D may be used. Letters or symbols are perhaps preferable to such designations as Excellent, Good, Fair and Poor because of the moral implication in the latter.

Even as fine a scale as this might perhaps better be replaced by a coarser one computed on the mean variation of 4.3 points, which is the mean variation of different teachers in the same department and institution after the effect of the personal standard has been eliminated. See Table II. On this basis the range of a division on the scale should be 4.3 times $2\frac{3}{4}$ or approximately 12 points. The reason for this larger step would be that this is as closely as different competent teachers agree on the evaluation of the same papers. One teacher may be as much in the right for grading a paper 80 as another for grading it 90. The only ultimate criterion is the consensus or average of estimates. This coarser scale would allow for only three divisions of passable grades, A, B and C. But the finer scale proposed above can be used with reasonable accuracy by a teacher in grading his own pupils in the light of his own viewpoint.

Of course, any one may use as fine a scale as he pleases provided one recognizes the range of the probable error of the scale used. The fine scale, if conscientiously used, probably tends to stimulate the making of finer distinctions than a coarse scale does. However, the chief objections to a very fine scale are: (1) An illusion of accuracy, (2) injustice to the student of supposed differences where there is no appreciable difference or where the relative merit might be just reversed, (3) embarrassment to the teacher due to this injustice.

If we admit the soundness of our reasoning it may seem to many teachers that even the finer scale of five point steps is rather crude and that the evaluation of a pupil's attainment is very coarse. But not so. As a matter of fact, the steps of the proposed scale are very fine and the measurement of achievement would be fairly accurate.

Apropos of this point we may compare the accuracy of making measurements of a similar type in an entirely different field. A mechanic through constant use has acquired a fairly definite mental image of an inch or a foot. Yet a mechanic's estimate of the length of a rod is not an iota more accurate than a teacher's estimate of an examination paper. I tested this problem by having eleven experienced carpenters estimate in inches as closely as they could the length of five rods varying

[3] To those who may be interested in the basis of this computation I may say that a range twice the size of the probable error includes one half of the series of estimates, and a range $2\frac{3}{4}$ times the mean variation or 3 times the probable error includes approximately three fourths of the series of estimates. In practise the mean variation and the probable error are used interchangeably, but the former is usually a trifle larger than the latter.

in length from ten inches to twenty-three inches. These "measurements" based on visual impressions are given in Table IV.

The validity of these measurements can be readily compared with the validity of the grades in Table I. by means of the coefficient of variability which is computed by dividing the mean variation by the average. The average coefficient of variability of the grades (last column in Table I.) is almost identical with that of the rods, .07 and .06, respectively. Hence measurements made by means of a mental scale are subject to the same amount of inaccuracy in one field as in another. It simply means that the mind can not discriminate any more accurately. If we are attempting to evaluate a paper by a scale of 100, 99, 98, 97, 96, 95, etc., we are attempting the impossible. The mind simply can not discriminate between a paper of grade 85 and another one of grade 86. If the second is appreciably better it more likely ought to have a grade of 90. The situation is analogous to asking a person to estimate the width of a room in inches when you should ask him to estimate it in yards. Estimates in terms of large units, of course, do not have greater absolute accuracy, but they are more apt to be uniform.

TABLE IV

Length of Rods	Carpenters											Av.	Mean Variation	Coefficient of Variability
	1	2	3	4	5	6	7	8	9	10	11			
10	11	10	10	10	8	9	9	9	8.5	9	8.5	9.1	.66	.07
15	14	14	12	13.5	12.5	14	13	14	13	13	14	13.4	.6	.05
17	17	16	15	14	14	16	15	15	17.5	15	17	15.6	1.1	.07
20	20	21	18	22	18	20	19	17.5	20	18	19	19.3	1.2	.06
23	24	24	21	21	20	22	21	22	24.5	24	22	22.3	1.3	.06
Av.														.062

Several criticisms have been suggested to me in discussing the results presented in this paper. For example, some teachers state that they do not attach much importance to the final examination, but grade the student largely by his other work, such as themes, daily recitations, etc., and that the situation is very different in those matters. This objection is beside the point because you are simply shifting the responsibility to something else. You are simply using the same scale for measuring something of similar nature.

Then it has been suggested that the grades in Table I. must necessarily be inaccurate because these instructors did not know the students who wrote the papers. But just on that account they would be all the more able to give an unprejudiced evaluation of the papers as papers. Many teachers have the practise of placing the papers so that when they pick one up for grading they do not know whose paper it is. If then the teacher wishes to raise or lower the mark according to the diligence or negligence of the student, well and good, but that does not mean that the grade of the paper will be any more accurate.

A third suggestion is that with a fine scale of marking the teacher is able to impose a penalty for shiftless work and indifferent attitude. But with a coarser scale on which the steps really mean something it is possible to attach a penalty of real significance.

The second part of this paper relates to the distribution of grades. How frequently should each division of the scale be used when assigning marks to large groups of pupils? By various psychological reasons, which I shall not state here,[4] it can be shown that the distribution of grades among large groups of students who have not been subject to special selection, should follow the probability curve. Thus the distribution of marks of college freshmen, who, strictly speaking, are a more or less select group, should, and in fact does, conform to the probability curve. Fig. 1

[4] See Dearborn, W. F., "School and University Grades," *Bulletin of the University of Wisconsin*, No. 368.

shows how closely the two agree. The curve representing the distribution of marks is based on approximately 5,000 grades assigned to freshmen in the college of letters and science in the University of Wisconsin.[5]

Theoretically, then, on the basis of the probability curve, 3 per cent. of the students should receive $A+$ (97–100), 7 per cent. $A-$ (93–96), 16 per cent. $B+$ (89–92), 23 per cent. $B-$ (85–88), 23 per cent. $C+$ (81–84), 16 per cent. $C-$ (77–80), 5 per cent. $D+$ (73–76), 3 per cent. $D-$ (70–72) and 4 per cent. failure. The percentage of failures is largely arbitrary and should perhaps be higher than here indicated.

The problem of distribution, however, is more complex in the upper classes after considerable elimination has occurred during the freshman and sophomore years. Two extreme positions have been held. Professor Meyer[6] holds that the nature of the distribution in upper classes is the same in spite of the elimination, that although the curve becomes contracted at the base it remains the same in shape. President Foster,[7] on the other hand, holds that the curve should have a very abrupt drop from the middle toward the lower end, on the belief that the university rigorously selects only those in the upper half of the curve. Neither position is entirely justifiable, for the reason that there is elimination during the freshman and sophomore years largely on the basis of intellectual fitness, and that this elimination is not exclusively from the lower half or from the lowest quarter, but is distributed over a large portion of the curve. The only way to determine the form of the curve is by finding the actual facts in the case. That is, in what part of the curve does the elimination occur, and how many are eliminated at each point?

I have computed this on the basis of the curve in Fig. 1 by taking the group of students there represented and finding out which ones dropped out and what their average grades were. Fig. 2 starts with the probability curve and shows what the shape of it is after the elimination in the first two years. The curve shows that elimination is greatest at the lower extreme and gradually becomes less up to the grade of 93, above which there is almost no elimination.

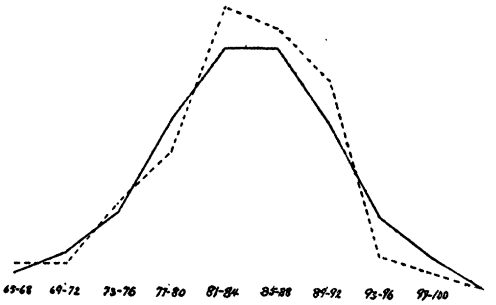

FIG. 1

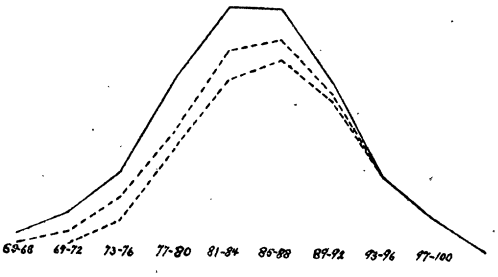

FIG. 2

Theoretically, on the basis of this modified curve, the distribution of grades in the upper two years should be as follows: 4 per cent. of the students should receive $A+$, 10 per cent. $A-$, 20 per cent. $B+$, 24 per cent. $B-$, 22

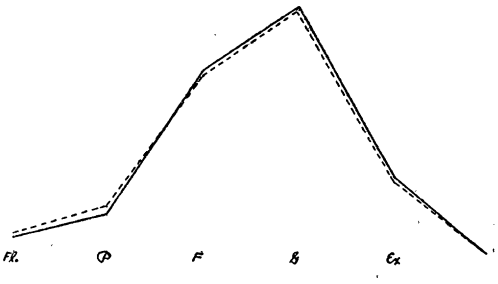

FIG. 3

[5] Dearborn, W. F., "The Relative Standing of Pupils in the High School and in the University," *Bulletin of the University of Wisconsin*, No. 312, plate I.

[6] Meyer, M., SCIENCE, N. S., 28: 246–250.

[7] Foster, W. T., SCIENCE, N. S., 35: 887–889.

per cent. $C+$, 11 per cent. $C-$, 4 per cent. $D+$, 2.5 per cent. $D-$ and 2.5 per cent. failure; or using only the four large steps, 14 per cent. should receive A, 44 per cent. B, 33 per cent. C, 6.5 per cent. D and 2.5 per cent. failure.

Fig. 3 shows how closely the actual distribution of the grades of upper classmen coincides with the theoretical distribution here computed. The continuous line is the theoretical distribution and the broken line is the actual distribution of 5,404 grades assigned to upper classmen in the college of letters and science in the University of Wisconsin. The latter are taken by permission from the unpublished report of Dean Birge.

The adoption of a uniform scale of grades as well as a uniform standard in the frequency with which the different grades are assigned is a pressing need among colleges and secondary schools. These ends could be attained by adopting the scale of eight passing grades, or the coarser one, for reasons given in the earlier part of this paper, and by having each teacher and each institution compare the frequency of the various grades assigned with the theoretical frequency. Then an $A+$ or a $B-$ would have more nearly the same significance under different teachers and in different institutions than they have at the present time.

Daniel Starch
University of Wisconsin

THE AMERICAN CHEMICAL SOCIETY

ROCHESTER MEETING

The forty-eighth annual meeting of the American Chemical Society was held at Rochester, New York, September 8 to 12. This is the first meeting held in September under the newly adopted constitution, and the large number present and the enthusiasm of the meeting amply justify the change in date from the Christmas holidays to the fall of the year.

Below will be found titles of the papers given at the meeting, with such abstracts as could be obtained. A study of the list shows a number of valuable contributions in both theoretical and applied chemistry. Most of these papers will be published in full in the journals of the society.

A complimentary dinner was given by the Rochester Section to the council on the evening of September 8, and following this dinner was held the annual council meeting of the society. Charles L. Parsons was elected secretary of the society, and Dr. A. P. Hallock, treasurer, for a period of three years, under the revised constitution. W. A. Noyes was elected editor of the *Journal of the American Chemical Society*, and the board of associate editors was continued, with the exception of H. P. Talbot and A. A. Noyes, who asked to be relieved of this duty. W. Lash Miller, of the University of Toronto, was elected to the board with special reference to physical chemistry. M. C. Whitaker was elected editor of the *Journal of Industrial and Engineering Chemistry*, and the board of associate editors was continued and the editorial staff strengthened by the addition of two assistant editors. A. M. Patterson was reelected editor of *Chemical Abstracts*, and J. J. Miller and E. J. Crane associate editors.

The first general session was held in the assembly hall of the Eastman Kodak Company, Kodak Park, on Tuesday morning, and was opened by a cordial address of welcome by Mayor Edgerton, and replied to by President Little. Papers were presented as indicated below.

At the conclusion of the morning session the members and their guests were entertained at luncheon by the Eastman Kodak Company. After luncheon the manufacturing department of the Kodak Company was inspected by the members present, who were divided into groups of fourteen for the purpose and placed under the guidance of members of the Eastman Company's technical staff. This opportunity to see one of the most highly developed chemical industries in America was thoroughly appreciated. On Tuesday evening, the members were entertained by the Rochester Section at a smoker, the program for which had been prepared under the able direction of M. H. Eisenhart, assisted by other members of the local section, who provided an extensive program and elaborate feast for the occasion. Each guest was decked out in a commodious white apron, on which was inscribed in bold letters his name and address, and also wore a yellow Chinese mandarin cap with pigtail. The hall was decorated with flags, and contained many small balloons filled with hydrogen, which, as their buoyancy diminished, afforded special opportunities for amusement of the guests. Unusually attractive songbooks had been printed in the works of the Kodak Company, bearing the pin of the society in colors. Three other attractive souvenirs were distributed to each guest.