3 Engineering Education

Reliability in Cumulative Examination Marks

E.A. JACKSON PhD MIEEE MGhIE
Department of Electrical/Electronic Engineering,
University of Science and Technology,Kumasi,Ghana

ABSTRACT

This paper finds the reliability of the average of examination marks, calculated to two decimal places, as a criterion for the ranking and classification of students. The marks are in percentages and the paper examines the importance of a difference of one point in the unit, the first or the second decimal place.

It is seen from the reliability in marking, even the mathematical courses, that the probability of obtaining a reliable percentage average calculated to one or two decimal places is very low. Thus it is only fair to limit the average marks calculated only up to the units.

Keywords: reliability, true mark, courses, variability

NOMENCLATURE

$F(\bar{x})$ - normal cumulative distribution function

M - true mark, %

S - sum of random variable, e

e - random variable with normal distribution, %

μ - mean of e,

σ - standard deviation of e

INTRODUCTION

Examinations are to measure certain aspects of a student's abilities which are supposed to be relatively permanent. Investigations show, however, that the mark awarded a student is not reliable because if the assessment is repeated in some way, then the second mark is normally different from the first (1,2,3).

Examination questions can be divided into three basic types: essay type, semi-objective type and objective type. Essay type questions require a free response and hence require the marker to pass an opinion. This freedom of response causes difficulties in marking: two markers could give quite different marks for the same essay, even when using an agreed marking scheme.

In semi-objective questions, the response of the candidate can be anticipated to a large extent in the marking scheme. Open-ended one sentence answer, graphs, engineering drawing and mathematical type questions come under this category. The marking of these items can be controlled to a certain degree and the marker can therefore give subjective opinion to an extent.

In objective type questions, the answers follow a well defined form. The marker therefore does not have to pass an opinion. Marking objective questions is much easier and anybody with the marking scheme can mark them reliably. Computers are also used to mark such questions. Random influences such as tiredness and anger do not affect the marker, and objective questions are the most reliable. The mark obtained in an objective examination can be regarded as the true mark.

For examinations with various degrees of subjectivity in marking, the concept of a "true mark" from which any deviation is regarded as an error in marking is not meaningful, unless a final arbiter (chief examiner for instance) whose decision on any script is final and be taken as the true mark. In this paper, the true

mark is taken as the average mark of a number of examiners marking each script.

In the University of Science and Technology, Kumasi, most undergraduate students take about 50 examination courses in a four-year career. The student's final percentage mark is therefore the average of 50 courses.

Other courses of study taken are three-year degree and two-year diploma courses. Here the total number of courses taken is about 40 and 30 respectively. A student graduates with a final cumulative percentage mark calculated to two decimal places. In this paper we shall look at the marking variability of a teacher and its cumulative effect in ranking and obtaining the final year classification of students.

MARKING RELIABILITY

When a student receives a mark of 56%, he might have easily received 53% or 59%, since the teacher could have assigned partial credit differently. A procedure for measuring the marking variability can be based on the following: A mark given as M could well have been M ± e, where e is a random variable whose probability model will be examined.

In an investigation by Hill(4), copies of ten unmarked scripts from each of three B.Sc (Eng) Part 1 examinations were marked by eight different experienced markers, each working to the same marking scheme. The courses taken were Fluid Mechanics, Structures and Thermodynamics. It was found that the mark distribution of a candidate followed the normal distribution curve. The standard deviation of the candidate's marks from the different markers about his mean mark fell within the range 2 to 12 percentage. The average standard deviation was 6 per cent, (Table 1).

A similar investigation conducted by Jackson (5) shows that the standard deviation of marks obtained by a candidate about his mean mark fell within the range 1 to 10 marks. The average standard deviation was 7. The course used for the investigation was Circuit Theory. The student's mean mark from the markers is regarded as the true mark (Table 2). In both Tables 1 and 2 the marks from the principal examiner are in the last column.

Thus in this model, the random variable e is a normal probability function with a mean of 0 and a standard deviation σ.

From the examples of Hill (4) and Jackson (5), it can be seen that even for a mathematical type examination paper, the reliability of marking could vary from $\sigma = 2$ to $\sigma = 12$. Also the random variable e, with the normal distribution can take integer values of 0,1,2,3... n. The smaller the value of σ, the more reliable the marking and the smaller the value of n.

For example, for objective questions, we have a perfect reliability with $\sigma = 0$. The random variable e can only take a value of 0.

Thus n = 0 and Pr[e = 0] =1.
On the other hand if $\sigma = 2$ then n is limited to 8.

$$Pr[e = 0] = 2F(0.5/\sigma) - 1$$
$$= 0.197$$

and $Pr[|e| \leq 2] = 0.789$

where F(Z) is the normal cumulative distribution function. This means that if the true mark of a student is 56% then he will be given this mark with a probability of 0.197. His mark could however range between 54% and 58% with a probability of 0.789.

If $\sigma = 12$, then $Pr[e = 0] = 0.033$

and $Pr[|e| \leq 2] = 0.165$

Hence for the less reliable marking, a person whose true mark is 56%, will receive 54% to 58% with a probability of only 0.169.

With essay type questions, the reliability is quite small even though it could be improved if the following steps are taken.

(a) It is better to ask more specific questions which can be answered briefly rather than fewer broad general questions. The reliability is increased because the syllabus

TABLE 1: PERCENTAGE RAW MARK: FLUID MECHANICS [4]

MARKER / STUDENT	1	2	3	4	5	6	7	8	RANGE	TRUE MARK ≡ AVE.	STD. DEV.
1	44	43	50	69	38	39	67	39	38-69	49	12.6
2	56	55	60	62	50	59	65	60	50-65	58	4.6
3	61	69	64	66	63	62	69	56	56-69	64	4.3
4	69	66	72	74	66	70	73	70	66-74	70	2.8
5	47	43	47	57	40	47	53	43	40-57	47	5.6
6	30	25	26	39	24	22	38	30	22-39	29	6.3
7	43	53	52	56	50	57	57	43	43-57	51	5.7
8	62	45	36	59	38	40	53	33	33-62	46	11.0
9	51	54	53	53	57	53	57	46	46-57	53	3.5
10	68	77	75	81	73	76	75	63	63-81	74	5.6
MARKER'S AVERAGE	53	53	54	62	50	53	61	48			

TABLE 2: PERCENTAGE RAW MARKS+ CIRCUIT THEORY [5]

MARKER / STUDENT	1	2	3	4	5	RANGE	TRUE MARK ≡ AVE	STD DEV
1	71	68	67	67	68	67-71	68	1.6
2	45	44	37	53	56	37-56	47	7.5
3	84	81	85	89	91	81-91	86	4.0
4	39	32	33	33	49	32-49	37	7.2
5	48	56	40	41	52	40-56	47	6.9
6	48	47	47	51	60	47-60	51	5.5
7	57	56	53	53	63	53-63	56	4.1
8	13	35	23	27	33	13-35	26	8.7
9	61	59	43	40	60	40-61	53	10.2
10	64	51	41	57	67	41-67	56	10.4
MARKER'S AVERAGE	53	53	47	51	60			

will be sampled more fairly.

(b) The freedom of choice in the selection of questions to be answered should be limited. If a candidate has a choice of four questions out of twelve, then one candidate can answer four completely different questions than another candidate. The marker then has the task of comparing responses to different questions.

(c) Questions should be framed so that the candidates are completely aware of the topic to be discussed, the direction and the scope of the response required. A question like "Why did you choose to become a teacher?" is open ended and it would be impossible to place the responses in rank order.

MATHEMATICAL MODEL FOR MARKING VARIABILITY

In a four-year career where a student takes about 50 examination courses, his final percentage mark is an average of 50 courses. We are thus interested in the average of 50 values of e.

If the average mark is to be accurate in the second place beyond the decimal point, then the average of e must vary by at most ± 0.005. This implies that an average mark of 56.00 means the average is in the interval 55.995 and 56.005.

To find the probability of achieving this reliability, Kozelka (6) has shown that the Central Limit Theorem can be applied to find this probability with fairly accurate results even for a total of 30 courses.

For reliability in the second decimal place, we require

$$\Pr[-0.005 \leq \text{average } e \leq 0.005] .$$

To use the half unit continuity correction conveniently, it is easier to work with sums than with averages.

A form of the Theorem is:

If $x_1, x_2, \ldots\ldots, x_n$ are independently distributed with identical means μ and variances σ^2 , then the sum

$$S = \sum_{i=1}^{n} x_i \text{ tends to } N(n\mu, n\sigma^2)$$

in the sense that

$$\Pr(a \leq S \leq b) \Rightarrow \int_a^b N(n\mu, n\sigma^2)dx \quad \text{as } n \longrightarrow \infty$$

where $N(u,v) = \left((2\pi v)^{-1/2}\right) \exp[-0.5\,(x-u)^2/v]$

is the normal distribution density function.

Now $\Pr[-0.005 \leq \text{average } e \leq 0.005] = \Pr[-0.25 \leq S \leq 0.25]$ (1)

for 50 courses.

Since S is an integer, $\Pr[-0.25 \leq S \leq 0.25] = \Pr[S = 0]$.

To evaluate the R.H.S we make use of the half unit continuity correction. Also since the random variable e has a mean value

$\mu = 0$ and a standard deviation σ, we have

$$\Pr[S = 0] = \Pr[-0.5 \leq S \leq 0.5] = 2F\left[\frac{0.5}{\sigma\sqrt{50}}\right] - 1 \qquad (2)$$

Generally for 1 decimal place reliability for 50 courses,

$$\Pr[-0.5 \leq \text{average } e \leq 0.5] = \Pr[-2.5 \leq S \leq 2.5] = 2F\left[\frac{2.5}{\sigma\sqrt{50}}\right] - 1 \qquad (3)$$

For reliability in the units,

$$\Pr[-0.5 \leq \text{average } e \leq 0.5] = \Pr[-25 \leq S \leq 25] = 2F\left[\frac{25.5}{\sigma\sqrt{50}}\right] - 1 \qquad (4)$$

Equations 2, 3 and 4 are now evaluated for various values of σ. In order to cater for other disciplines which offer a total of 30 or 40 courses, the probability above has been evaluated

for these number of courses as well.

RESULTS

Table 3 shows reliabilities of the averages of 50, 40 and 30 courses.

Fig.1 is a graphical display of the results in Table 3a. The figure shows the relative magnitudes of the reliabilities in the 1st, 2nd decimal places and the units, in a bar chart, for a total of 50 courses.

It can be seen that for all values of σ, only an average calculated in the units has a reasonable reliability. Calculations of averages up to 1 decimal place or 2 decimal places show very low reliabilities. For example, for 50 courses in Table 3a, when $\sigma = 2$ the probability of a student's average being reliable is .929 in the units. However, in the 1st decimal place or 2nd decimal place, the probabilities are only 0.14 and 0.028 respectively.

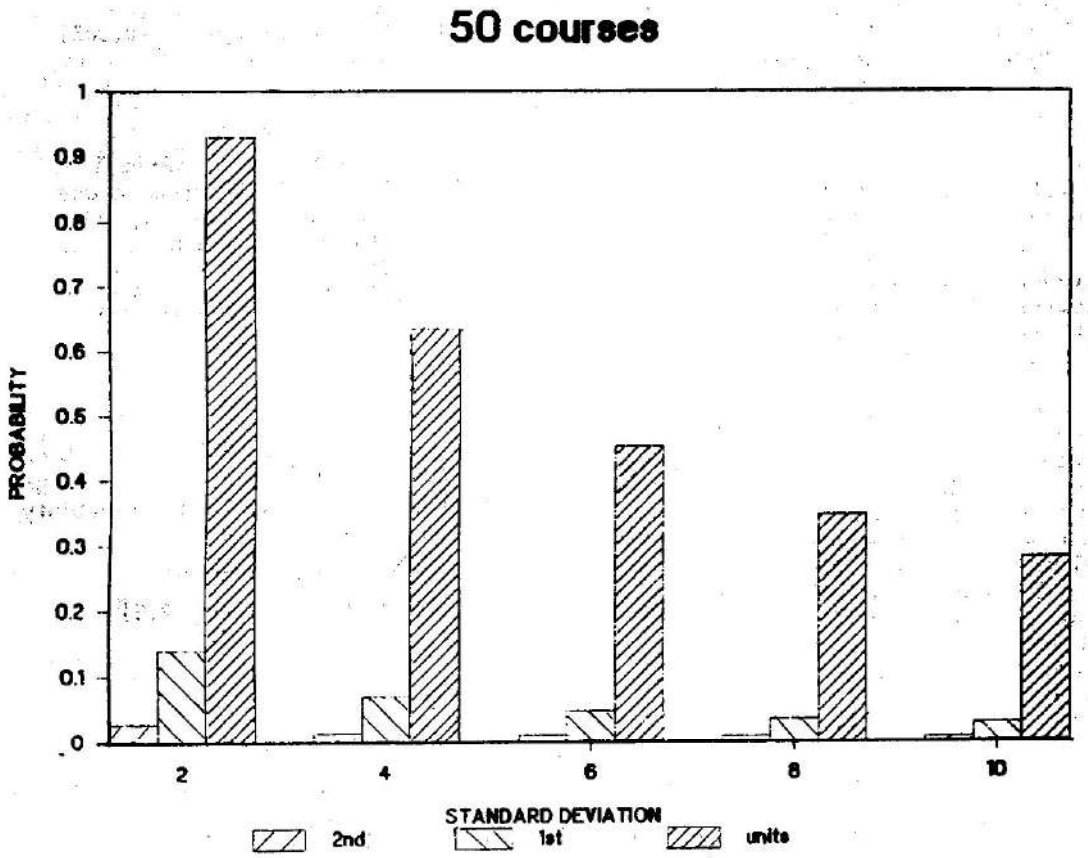

FIGURE 1 RELIABILITY FOR 50 COURSES

TABLE 3 RELIABILITIES IN PERCENTAGE MARKS

(a) 50 COURSES

σ	2ND DECIMAL PLACE	1ST DECIMAL PLACE	UNITS
2	0.028	0.140	0.929
4	0.014	0.070	0.633
6	0.009	0.047	0.452
8	0.007	0.035	0.348
10	0.006	0.028	0.282

(b) 40 COURSES

σ	2ND DECIMAL PLACE	1ST DECIMAL PLACE	UNITS
2	0.032	0.157	0.895
4	0.016	0.079	0.584
6	0.011	0.053	0.412
8	0.008	0.039	0.315
10	0.006	0.032	0.254

(c) 30 COURSES

σ	2ND DECIMAL PLACE	1ST DECIMAL PLACE	UNITS
2	0.036	0.109	0.843
4	0.018	0.055	0.521
6	0.012	0.036	0.363
8	0.009	0.027	0.276
10	0.007	0.022	0.223

Marking with a reliability of $\sigma = 2$ can be achieved in the mathematical type problems, as seen in Tables 1 and 2.

When $\sigma = 10$, the reliability of average mark for 50 courses reduces to 0.282 in the units. This shows that as the questions become more subjective, like an essay type paper, the reliability even in the units can be low.

If we require the reliability of the average in the 2nd decimal place to be realistic, say, equal to 0.5 for 50 courses, then the marking reliability required is given by $\sigma = 0.105$.

With this value of σ, $Pr[e = o] = 0.999999$.

Such marking reliability is beyond the reach of ordinary mortal being when marking non-objective questions.

Tables 3b and 3c show that the order of magnitude of the probabilities for the 2nd decimal place, the 1st decimal place and the units are the same for the 50, 40 and 30 courses.

DISCUSSION AND CONCLUSIONS

It is observed that with percentage marks, only an average calculated in the units has a reasonable reliability. An implication of calculating the average in the units implies that for a pass mark of 45, someone who obtains 44.50 is rounded up to 45, which makes the difference between a withdrawn and a passing student. With the average calculated to 2 decimal places, even an average of 44.99 does not qualify a student to cross the life and death barrier. Thus some students may fail based on a marking reliability which cannot be achieved in practice.

The pass mark of most courses is 40%. Thus with reliability of marking in mind, we should ask ourselves whether a person who obtains 38% or 39% will benefit more by repeating the course for should be upgraded, since he might have easily obtained 42% or 36% with slight adjustment in partial credit.

The difference between the student's true worth and mark obtained can also be due to "student variability". Here a student may make careless mistakes, may be sick during the examination time or may be fortunate to learn only the section of the syllabus from where the questions were set. In fact the "student variability" might even influence his marks more than the instructor's marking reliability.

Subjectivity in marking cannot be eliminated but an effort to reduce it is necessary. The West African Examination Council does this by breaking down the total marks for a question into smaller units. For example, marks are awarded for the method as well as accuracy in getting the correct answer in the mathematical subjects. There are co-ordination meetings to steamline the marking.

The University teacher does not need any co-ordination meeting for his class. However, if more care is taken in the setting of the questions and allocation of marks are broken down into the smallest units, with room for all alternate solutions, the subjectivity will be reduced and reliability will be improved. As an example, for a question with 10 marks, a marking scheme which awards 1 mark each to 10 sections of the question will be more reliable than breaking the question into 2 sections and awarding 5 marks to each of the 2 sections.

REFERENCES

1. Cox, R. "Examinations and Higher Education, a survey of the Literature" University Quarterly, 21, 3(1967)

2. McVey, P.J. "The reliability of examinations in electronic engineering" Report NO.TR 24, Department of Electronic and Electrical Engineering, University of Surrey (1972)

3. Ashworth, A. E. "Testing for Continuous Assessment". Evans Brothers Limited, London (1982)

4. Hill, B.J. "Reliability of Marking B.Sc. Examinations in Engineering" International Journal of Mechanical Engineering Education, Vol.3 No.2 (1975)

5. Jackson, E.A. "Marking Reliability in B.Sc. Engineering Examinations" European Journal of Education, Vol.13 No. 4 (1988)

6. Kozelka, R.M: "Grade Point Averages and the Central Limit Theorem" The American Mathematical Monthly, Vol. 86, No.9 (1979)