# A Mark-Scaling Algorithm

**C. G. Broyden**
Department of Computer Science, University of Essex, Wivenhoe Park, Colchester, Essex, UK

**This paper describes an algorithm designed to compensate for the varying difficulty of examination papers when options are permitted. The algorithm is non-iterative, and does not require a common paper to be taken by all students. It may be possible to extend the approach described to the more general statistical 'missing values' problem.**

## 1. INTRODUCTION

Differences between the individual papers constituting an examination, both in their intrinsic difficulty and in the rigour of their marking, form occupational hazards for students and examiners alike. If options are not permitted, no individual student is unduly penalized or rewarded by the occasional maverick paper and a degree of rough justice prevails. If, on the other hand, students have a choice of options open to them, those electing to take easy ones with generous examiners have a distinct advantage over those whose choices are less fortunate or less cynical. In particular, since projects seem to attract higher marks than conventional written examinations, students whose courses include a high project component enjoy advantages over those taking courses of a more theoretical or abstract nature.

Similar problems of assessment arise whenever more than one measure of value is available, where each measurement is made in different circumstances and where some results are 'missing'. Another example would be the analysis of the performance of competing algorithms where the algorithms were tried out on different, but partially intersecting, sets of test problems. If every item to be tested is not subjected to every test there will inevitably be difficulties of evaluation, and nowhere are there difficulties more agonizing than in a final examiners' meeting.

The methods so far proposed to overcome these problems fall into four broad categories. In the first the marks are scaled (with or without the addition of a constant term) in order to produce mark distributions, for particular groups of students, that have comparable means and standard deviations for all papers attempted by that group. Methods of this type have been considered by Peaker[1] and Backhouse.[2] They have been used in many universities for a considerable period of time, and depend for their validity on having a reasonably large class. In the second category the marks scored are used only to determine a set of rankings, the final ranking being determined by a series of pairwise comparisons. These algorithms have been discussed by, among others, Ford,[3] David,[4] Backhouse[5] and Wood and Wilson,[6] with Davidson[7] contributing a tie-break strategy. The third group of algorithms uses regression analysis to estimate the 'missing' data, the final ranking being then determined by assuming that all candidates have taken all papers. Methods using regression techniques have been described by Johnson and Schwartz,[8] Rubin[9] and Backhouse.[2] Finally methods of estimating the missing values using maximum likelihood have been discussed by Orchard and Woodbury,[10] Buck[11] and Beale and Little.[12]

The specific problem of determining which average to use when combining examination marks has been dealt with in an ingenious manner by Griffiths.[13] This paper shows how different averaging processes may be used, sequentially, to give the same limit but does not address specifically the problem of 'missing values'.

It may be inferred from the appearance in the literature of so many types of solution of the compensation problem that no method is entirely satisfactory. The forms of scaling hitherto attempted require at least one paper to be taken by all candidates, and sometimes give the 'wrong' ranking (i.e. a ranking which conflicts with the commentator's prejudices). The regression and maximum likelihood methods require marks to be estimated for examinations that never took place (politically dangerous) whereas the paired comparison methods deliberately and wilfully discard information. The maximum likelihood methods, moreover, are iterative and often require many iterations for their solution.

A spirited critique of many of the existing methods has been given by Wood and Wilson,[6] who justify their own favourite by the statement (p. 211)

> It seems to us that a procedure which simply registers whether individual A scored more (or less) than individual B is truer to the nature of the data we have to deal with in educational measurements than methods which proceed as if marks were unequivocal placings on an equal internal scale.

Whatever one's views on this philosophy, it is adequately refuted by Wood and Wilson themselves (Ref. 6, p. 209) who roundly declare

> What is required . . . is a procedure that uses *all* the available information, however it comes, to estimate what the missing values would have been. (Their italics, not mine!)

Since no agreement on a satisfactory method exists, the proposal of an alternative might prove to be helpful. That described below is a scaling method which works when the data are incomplete in an irregular way. It is non-iterative and computationally simple, although it does give the 'wrong' ranking when applied to certain published data (but see Section 4 below).

CCC-0010-4620/83/0026-0109 $02.00

## 2. THE METHOD

The method is based on the assumption that, on average, a student will tend to perform with a certain degree of consistency in all the papers that he attempts. This is, perhaps, a somewhat hairy assumption but is probably less so than the assumption on which no scaling at all is based, namely that different examiners, teaching in widely different subject areas, set papers of equal intrinsic difficulty. The method works by scaling the marks of each individual *paper* in order to maximize the consistency of each individual *student*. More precisely, let $m$ students attempt a selection of papers from the maximum of $n$ available, with the $i$th student taking $n_i$ papers. Denote the raw mark obtained by the $i$th student on the $j$th paper by $x_{ij}$ and the factor by which the marks on the $j$th paper are to be scaled by $f_j$. Then if $v_i$ denotes the variance of the scaled marks of the $i$th student,

$$v_i = \sum_j (x_{ij} f_j - \bar{x}_i)^2 / n_i \tag{1}$$

where

$$\bar{x}_i = \sum_j x_{ij} f_j / n_i \tag{2}$$

the sum in each case being taken over all papers attempted by the student. Since $n_i v_i$ is a measure of the consistency of the $i$th student, the method seeks to choose the factors $f_j$ that minimize $\sum_i n_i v_i$ over all students subject to the constraint that the total aggregate mark $M$ remains unchanged.

To see how the method works, consider two cases where the marks on a particular paper are low. If this is due to the paper being unusually difficult, the effect of upward scaling will be to bring the marks of each individual student on that paper more into line with his or her other marks, so reducing the variances. Scaling will thus occur, the marks on the other papers being scaled down slightly to preserve the total aggregate. If, on the other hand, the low marks are due to the paper being taken by weak students, any attempt at scaling will have the effect of increasing the variances so in this case no, or only a very slight, adjustment will be made.

To derive the equations for the optimal values of $f_j$, let $\mathbf{x}_i$ denote the vector of the $i$th student's marks and $\mathbf{s}_i$ the vector whose $j$th element is unity if the $i$th student attempted the $j$th paper, and zero otherwise. Equations (2) and (1) may be written, since $\mathbf{s}_i^T\mathbf{s}_i = n_i$ and the matrix $\mathbf{I} - \mathbf{s}_i\mathbf{s}_i^T/\mathbf{s}_i^T\mathbf{s}_i$ is idempotent,

$$\bar{x}_i = \mathbf{x}_i^T\mathbf{F}\mathbf{s}_i/\mathbf{s}_i^T\mathbf{s}_i \tag{3}$$

and

$$n_i v_i = \mathbf{x}_i^T\mathbf{F}\left(\mathbf{I} - \frac{\mathbf{s}_i\mathbf{s}_i^T}{\mathbf{s}_i^T\mathbf{s}_i}\right)\mathbf{F}\mathbf{x}_i \tag{4}$$

where $\mathbf{F} = \mathrm{diag}(f_j)$.

Now, since $\mathbf{F}$ is diagonal, it follows from the definition of $\mathbf{s}_i$ that $\mathbf{x}_i$ will have zeros wherever $\mathbf{F}\mathbf{s}_i$ has zeros. Thus the inner product $\mathbf{s}_i^T\mathbf{F}\mathbf{x}_i$ will remain unchanged if the zeros of $\mathbf{F}\mathbf{s}_i$ are replaced by arbitrary values. In particular, if they are replaced by the corresponding diagonal elements of $\mathbf{F}$, $\mathbf{F}\mathbf{s}_i$ becomes $\mathbf{f}$, where $\mathbf{f} = [f_j]$, and since the inner product is unchanged it follows that

$$\mathbf{s}_i^T\mathbf{F}\mathbf{x}_i = \mathbf{f}^T\mathbf{x}_i \tag{5}$$

Equation (4) may thus be written

$$n_i v_i = \mathbf{x}_i^T\mathbf{F}^2\mathbf{x}_i - \mathbf{f}^T\mathbf{x}_i n_i^{-1}\mathbf{x}_i^T\mathbf{f} \tag{6}$$

$$= \mathrm{tr}(\mathbf{F}\mathbf{x}_i\mathbf{x}_i^T\mathbf{F}) - \mathbf{f}^T\mathbf{x}_i n_i^{-1}\mathbf{x}_i^T\mathbf{f} \tag{7}$$

If we define $\mathbf{X} = [x_{ij}]$ to be the matrix of raw marks (whose $i$th row is, of course, just $\mathbf{x}_i^T$), $\mathbf{N} = \mathrm{diag}(n_i)$ and

$$V = \sum_i n_i v_i \tag{8}$$

it follows from equations (7) and (8) that

$$V = \mathrm{tr}(\mathbf{F}\mathbf{X}^T\mathbf{X}\mathbf{F}) - \mathbf{f}^T\mathbf{X}^T\mathbf{N}^{-1}\mathbf{X}\mathbf{f} \tag{9}$$

Let now $\mathbf{D}$ denote the diagonal matrix obtained by setting all the off-diagonal elements of $\mathbf{X}^T\mathbf{X}$ equal to zero. Then, since both $\mathbf{D}$ and $\mathbf{F}$ are diagonal,

$$\mathrm{tr}(\mathbf{F}\mathbf{X}^T\mathbf{X}\mathbf{F}) = \mathrm{tr}(\mathbf{F}\mathbf{D}\mathbf{F}) = \mathbf{f}^T\mathbf{D}\mathbf{f}$$

so that, from equation (9),

$$V = \mathbf{f}^T\mathbf{K}\mathbf{f} \tag{10}$$

where

$$\mathbf{K} = \mathbf{D} - \mathbf{X}^T\mathbf{N}^{-1}\mathbf{X} \tag{11}$$

The constraint that the sum of the scaled marks must remain unchanged may be expressed as

$$\mathbf{e}^T\mathbf{X}\mathbf{f} = M \tag{12}$$

where $M$ is the original total aggregate and $\mathbf{e}$ is the $m$th order vector whose every element is unity. The value of $\mathbf{f}$ that minimizes $V$ subject to this constraint being satisfied is obtained by solving

$$\mathbf{K}\mathbf{f} + \mathbf{X}^T\mathbf{e}q = \mathbf{0} \tag{13}$$

where $q$ is a Lagrange multiplier whose value is obtained from (12), and which is given by

$$q = -M/(\mathbf{e}^T\mathbf{X}\mathbf{K}^{-1}\mathbf{X}^T\mathbf{e}) \tag{14}$$

Thus, provided that $\mathbf{K}$ is non-singular, the determination of the scaling factors $f_j$ is a relatively straightforward matter.

## 3. THE NON-SINGULARITY OF K

To justify the above method it is necessary to demonstrate that $\mathbf{K}$ is non-singular. Since

$$V = \mathbf{f}^T\mathbf{K}\mathbf{f}$$

and $V$ is defined to be the sum of squares, $\mathbf{K}$ is either positive definite or positive semidefinite, and is singular iff $\exists\, \mathbf{f} \neq \mathbf{0}$ such that

$$\mathbf{f}^T\mathbf{K}\mathbf{f} = 0$$

This occurs if $v_i = 0$ for all $i$, and since $v_i$ is itself a sum of squares, it follows from equation (1) that $\mathbf{K}$ is singular iff a set of scalars $f_j$ can be found such that

$$x_{ij} f_j = \bar{x}_i \tag{15}$$

for all $(i, j)$ pairs for which the $i$th student attempted the $j$th paper.

Let now $p_i = \bar{x}_i$ and $q_j = 1/f_j$. Equation (15) becomes $x_{ij} = p_i q_j$ so that $\mathbf{K}$ is singular only if the elements $x_{ij}$ of $\mathbf{X}$ corresponding to papers attempted (the non-zero elements of $\mathbf{X}$, if it is assumed that no student scored zero marks on any paper) are the $(i, j)$th elements of some

rank-1 matrix. Since, in practice, this is never likely to occur, **K** may be assumed to be positive definite and the constrained minimization problem therefore possesses a unique and easily-determined solution.

Conversely, if

$$x_{ij}f_j = p_i, \sum_j x_{ij}f_j = n_i p_i$$

so that $p_i = \bar{x}_i$. Thus $v_i$ and hence $V$ are all equal to zero, and it follows that **K** is then positive semidefinite. These results may be expressed as the following

**Theorem**

Let $\mathbf{X} = [x_{ij}]$ be an $m \times n$ sparse matrix with $n_i$ non-zero elements in its $i$th row. Let **D** denote the matrix obtained by setting all the off-diagonal elements of $\mathbf{X}^T\mathbf{X}$ equal to zero and let $\mathbf{N} = \text{diag}(n_i)$. Then $\mathbf{D} - \mathbf{X}^T\mathbf{N}^{-1}\mathbf{X}$ is either positive definite or positive semidefinite, and is semidefinite if and only if the non-zero elements $x_{ij}$ of **X** are the $(i, j)$th elements of some rank-1 matrix.

## 4. RESULTS AND CONCLUSIONS

This note presents an easily-implemented method of compensating for differences in difficulty of examination papers when determining the overall mark or when comparing candidates taking different options. The method is non-iterative and works with an irregularly sparse data matrix, and calculates no fictitious scores for papers not attempted.

To compare it with some existing methods, it was applied to an (admittedly trivial) set of data due to Backhouse.[2] The original data are given in Table 1 and the scaled data, together with the resulting aggregates and rankings, are given in Table 2. The final ranking is the same as that given by simple scaling, and has been criticized by Wood[14] for violating 'common sense'. I find the common sense arguments unconvincing, preferring instead to appeal to a clear principle when forced to rank students taking papers whose means differ by as much as a factor of three.

A less trivial illustration of mark compensation is provided by Table 3. This shows a selection of results taken from actual finals papers for three consecutive years and demonstrates forcibly (and quite fortuitously) the distinctions the method was designed to reveal. Column 1 shows the grand average percentage for all papers (the same, of course, for both scaled and unscaled marks), and columns two and three show the unscaled and scaled average percentages of a selected paper with a high project content. In year $n$ the marks are all much of a muchness, indicating that the students taking the selected paper not only were representative of the group as a whole but that the markers of that paper shared the overall view of the worth of the candidates. In year $(n + 2)$ the average mark for the selected paper was 12.5 above the grand average, but the scaled marks fully substantiate the claim that the mark was high because only the best candidates attempted that paper. Year $(n + 1)$, though, is not so reassuring. Again the marks of the selected paper were much higher than the grand average but on this occasion the results show a considerable disagreement between the markers of that paper and their departmental colleagues as to the worth of the students in question.

Since the precise interpretation of the results computed by this and similar methods will always be in some degree contentious, it is unlikely that they will ever be used without modification to produce the final class lists. In particular, *no* algorithm working from the marks alone will be able to distinguish between high marks due to a 'successful course' (one where the blend of personal qualities of those involved leads to higher-than-average performance) and high marks due to over-generous assessment. The value of the algorithm described, and of similar algorithms, lies in their identifying students near a borderline whose positions may have been affected by their choice of options, and in drawing anomalies to the examiners' attention. It could be that the results will then be used on a 'help but not hinder' basis, and if this leads to students selecting options for their academic interest and not for their perceived scoring potential, this will be no bad thing.

**Table 1**

| Candidate | P | Q | R | S | T | U | V | W | X | Y | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper 1 | 20 | 18 | 15 | 12 | 10 | 8 | | | | | 13.83 |
| Paper 2 | 30 | 35 | 28 | 26 | 20 | 5 | 31 | 32 | 27 | 14 | 24.80 |
| Paper 3 | | | | | | | 50 | 37 | 45 | 26 | 39.50 |

**Table 3**

| Year | Grand mean | Selected paper (unscaled) | Selected paper (scaled) |
|---|---|---|---|
| $n$ | 50.8 | 51.8 | 49.8 |
| $n+1$ | 50.3 | 64.9 | 52.2 |
| $n+2$ | 44.7 | 57.2 | 56.9 |

**Acknowledgement**

The author wishes to thank both his colleague, Jim Doran, and the referee for their helpful and constructive suggestions.

**Table 2**

| Candidate | P | Q | R | S | T | U | V | W | X | Y | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Paper 1 | 34.77 | 31.29 | 26.08 | 20.86 | 17.38 | 13.91 | | | | | 24.05 |
| Paper 2 | 29.26 | 34.13 | 27.31 | 25.36 | 19.50 | 4.88 | 30.23 | 31.21 | 26.33 | 13.65 | 24.19 |
| Paper 3 | | | | | | | 32.55 | 24.09 | 29.30 | 16.93 | 25.72 |
| Sum | 64.03 | 65.42 | 53.39 | 46.22 | 36.88 | 18.79 | 62.78 | 55.30 | 55.63 | 30.58 | 48.90 |
| Rank | 2 | 1 | 6 | 7 | 8 | 10 | 3 | 5 | 4 | 9 | |

## REFERENCES


1. G. F. Peaker, Private communication to J. K. Backhouse (1972), cited in Ref. 2.
2. J. K. Backhouse, Determination of grades for two groups sharing a common paper. *Educational Research* **18**, 126–133 (1976).
3. L. R. Ford, Solution of a ranking problem from binary comparisons. *American Mathematics Monthly* **64**, 28–33 (1957).
4. H. A. David, The method of paired comparisons, Griffin, London (1963).
5. J. K. Backhouse, An approach to examining a wide range of ability. A paper presented to the Schools Council seminar on 4 October 1972.
6. R. Wood and D. T. Wilson, Determining a rank order when not all individuals are assessed on the same basis. In *Psychometrics for Education Debates*, edited by van der Kemp, Langerak and de Gruijter, Wiley, New York, pp. 207–230 (1980).
7. R. R. Davidson, On extending the Bradley–Terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association* **65**, 317–328 (1970).
8. B. Johnson and C. Schwartz, The analysis of an unbalanced paired comparison experiment by multiple regression. *Applied statistics* **26**, 136–142 (1977).
9. D. B. Rubin, Selecting the 'best' regression when faced with missing observations. *Research Bulletin 75–10*, Educational Testing Service, Princeton, N.J. (1975).
10. T. Orchard and M. A. Woodbury, A missing information principle: theory and applications. In *Proc. 6th Berkeley Symp. Math. Statist. Prob.* **1**, 697–715 (1972).
11. S. F. Buck, A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Statist. Soc. B* **22**, 302–306 (1960).
12. E. M. L. Beale and R. J. A. Little, Missing values in multivariate analysis. *J. R. Statist. Soc. B* **37**, 129–145 (1975).
13. H. B. Griffiths, Examiners' Meetings and the Arithmetico-Geometric Mean, *Bull. I.M.A.* **18**, 247–251 (1980).
14. R. Wood, Placing candidates who take different papers on the same mark scale. *Educational Research* **20**, 210–215 (1978).


Received January 1982