STRUCTURED GRADING

Brian Peacock
Doug Stewart
School of Industrial Engineering
University of Oklahoma
Norman, Oklahoma

ABSTRACT

This paper describes a compromise between the idealism of criterion based grading and the pragmatism of a norm based approach. The discussion is supported by a series of computer programs that are deliberately devoid of packaging clutter so that the users can clearly comprehend the processes and adapt the code to suit their particular purposes. The input data can be in numerical, alphabetical or categorical form and the primary output is a matrix of standardized marks. Additional analyses provide frequency distributions, rank order and alphabetic grade matrices, component correlations and categorical item analyses. The programs are designed to provide timely and appropriate information for final grade allocation.

1. INTRODUCTION

Grading is an essential part of the educational process. On occasion it becomes an all important part and overshadows the importance of learning and teaching. It is constantly the subject of criticism regarding its lack of validity, reliability and accuracy (Cheshire, 1975, Wissler, 1975, Work, 1976, de Nevers, 1984, and Boyle & Wright, 1977). Furthermore, the data processing associated with grading is often an unwanted chore and a source of error.

The purpose of this paper is to present a structured approach to grading that is supported by educational theory and easy-to-use software. The structured approach includes the process of assessment component generation, assignment of weights and marks, data processing, report preparation, component evaluation and final grade allocation. The educational theory includes a compromise between the idealism of criterion based grading and the pragmatism of norm based grading. In addition reference is made to well established statistical procedures.

2. ASSESSMENT COMPONENT DEVELOPMENT

2.1 Question Development

A set of examinations, questions or assignments implicitly aims to sample the knowledge of a group of students. This sampling should be as representative and comprehensive as possible regarding its coverage of the course material. Differential weighting may be prescribed according to the perceived or agreed importance of different course components.

An important aspect of assessment material development is the evaluation of that material based on empirical evidence. Most professors have recognized questions that are too easy, too difficult or which contain ambiguities. However, this feedback usually occurs on an ad hoc basis. At the other end of the sophistication continuum, educators have produced voluminous statistical analyses of assessment material including multicollinearity and discriminant analyses (Kilpatrick, 1971 and Flora, 1971).

Immediate feedback can be provided by the development of feedback classrooms (Peacock, 1982) or by using computer laboratories for examination purposes. These hardware and software developments are an inevitable part of the progress of education. However, as with any such technological development, the users' actual information needs should be carefully assessed. With this in mind the assessment component evaluation programs address two basic issues:

a) To what extent are assessment components correlated with each other and with the overall mark?
b) To what extent does a particular component discriminate between good and poor students?

There are many powerful multivariate analysis techniques and the literature in the educational, social science and marketing areas is abundant (Nie, 1981). A conceptual overview is given in Hair (1979) and a mathematical overview is given in Morrison (1976). The input to these analytic programs is commonly a correlation/covariance matrix and the output is commonly statistical support (or rejection) of some hypothesis regarding the complex associations or differences. The outputs of the programs presented here are aimed at cautious subjective interpretation unsupported by statistical significance tests.

2.2 Component Weighting

A relative mark can be given to a particular component/answer without regard to the importance of that component. That is, the

basic measurement and eventual weighting of an element or component are independent concepts.

There are a number of ways of arriving at appropriate weights for a component.

a) Rating. An overall sum of weights (e.g. 1, 10 or 100) is given to the whole course. This sum is divided between the components based on concensus of a group of faculty members and/or students.

b) Ranking. A group of teachers and/or students can, independently, place each component in rank order (least important first) and the sum of the ranks for a component can be used as the weight. This ranking technique may produce different weights from the rating technique and it may be appropriate to check for statistical concordance.

c) Empirical Weighting. The empirical contribution of each component can be achieved by setting the weights equal to the observed standard deviations. This is a common mode of mark amalgamation. However, it must be noted that the resulting component contribution will differ from that prescribed at the beginning of the course.

d) Individual Weightings. It is possible, and may be appropriate, to apply individual weightings to each component. Clearly some caution is necessary with this approach, however, it may be fairer to those students who have different abilities with regard to examinations, projects, etc. (This differential component weighting method is also appropriate for dealing with the amalgamation of faculty evaluation data, where different members of a department have different weightings assigned to teaching, research, administration and service.)

The mathematical implementation of these weighting procedures is simplified when the sets of raw marks from each component are standardized and the sum of the weights is equal to 1.

3. INPUT AND COMPUTATIONS

3.1 Input Forms

3.1.1 Numeric (Interval) Data. This data usually consists of an integer or decimal value on a scale with prescribed upper and lower values. The particular number, however, is usually a result of a subjective judgement regarding the value of one answer relative to another. Consequently, there may be some doubt regarding the claim of interval level measurement - it may only be ordinal. Where a number of such marks are added together then the claim of interval level measurement may be justifiable. In practice it is more convenient to use integer values and move to a larger scale where greater resolution is required. This practice will avoid some of the transduction errors associated with fractions and decimals.

3.1.2 Alphabetic (Ordinal) Data. Ordinal measurement may employ numeric or alphabetic values, with plusses and minusses superimposed where greater resolution is required. The alphabetic form is probably the most common at the initial transduction stage. However, mark amalgamation may involve the allocation of a numeric value to an alphabetic score as follows:

A+	4.3	10	---
A	4.0	9	4
A-	3.7	8	---
B+	3.3	7	---
B	3.0	6	3
B-	2.7	5	---
C+	2.3	4	---
C	2.0	3	2
C-	1.7	2	---
D	1.0	1	1
F	0.0	0	0

These scales are employed in the set of programs described in this paper. However, it is a simple matter to change the values either in a data statement or adapt the program to do this interactively.

3.1.3 Categorical (Nominal) Data. Typically this data arises from multiple choice or recognition formats in which a defined range of possible answers is given to the student to choose from. Common practices employ True/False or the one correct and three distractor forms.

The categorical response technique can be adapted to a great variety of forms:

a) Graded response values: In this form an interval or ordinal scale value can be ascribed to each response. This avoids some of the problems that may be caused by ambiguities or by poorly worded distractors. Furthermore, it allows for the incorporation of a notion of partial credit where the different choices that are presented result from more or less serious mistakes made in the calculations necessary to arrive at the correct answer.

b) Long questions: Typically multiple choice items aim to test the understanding of a particular well-defined concept. Usually this takes a very short period of time. However, multiple choice items can involve lengthy calculations. Alternatively a sequence of multiple choice questions can be introduced at various stages of a problem or a larger variety (e.g. 10) of possible answers

can be presented, each one reflecting a different error or combination of errors in the calculations.

These methods coupled with the graded response approach discussed earlier greatly extend the versatility of multiple choice methods to deal with complex problems.

c) Ordinal or Interval conversion of complex problems: Students' answers to traditional problems, involving calculations and interpretation, can be 'categorized' on various scales. That is, the set of possible ordinal or interval (integer) marks for a particular question can be prescribed. Depending on the resolution required, the scales of 0 to 4 or 0 to 10 will probably cover most individual components or questions. In this case the student's response can be assigned to a particular 'category' which has a prescribed 'value'. The advantage of this approach lies in the facility for analyzing the responses of the class as a whole by the method described in 4.1.

3.2 Standardization

The process of standardization in the statistical sense simply involves the following linear transformation to a set of marks:

$$X_i' = (X_i - \bar{X})/S$$

where X_i' is the standardized mark for the ith student
X_i is the raw mark
$\bar{X}$ is the mean of the set of marks
S is the standard deviation of the set of marks.

The resulting set of standardized marks will have a mean value of zero and a standard deviation of one. Each individual mark will have the same relative value when compared with the other marks as it had in the original raw mark list. The same natural breaks will occur and the distribution will have the same shape (Hastings & Peacock, 1975). A single very low mark will be offset by a slight shift in the positive direction by the higher marks.

The purpose of standardization is to bring sets of marks, with different location and scale characteristics, to a common form. This is particularly important in comparing an individual's performance profile on a series of tests and in removing the empirical weighting (standard deviation) effects which may distort the desired, prescribed weightings of components. A second use of standardization is that such mark distribution forms simplify the analysis of association between marks from separate questions using a correlation matrix.

3.3 Component Weighting

The ultimate objective of most grading systems is the production of a final mark list which is used for a decision regarding the final grade for the course. This final mark will be the weighted sum of the (m) individual components:

$$Y_i = \sum_{j=1}^{m} W_{ij} X_{ij}'$$

where Y_i is the total mark for the ith student
X_{ij}' is the standardized mark obtained by the ith student on the jth component
W_{ij} is the weight ascribed to the jth component for the ith student.

Usually $W_{ij} = W_{.j}$, i.e. the same weight will be given to all students for each component.

In practice it will be appropriate to standardize the final mark list as follows:

$$Y_i' = (Y_i - \bar{Y})/S_Y$$

where Y_i' is the standardized final mark
$\bar{Y}$ is the mean of the Y_i
S_Y is the standard deviation of the Y_i

This final standardized mark list may tend towards normality especially for large classes and large numbers of assessment components. However, in practice the final mark list will contain natural breaks and a certain degree of asymmetry. The process of final grade allocation can be made as follows:

Standardized Mark Lower Limit	Grade
$-\infty$	F
− 2	D
− 1	C
0	B
+ 1	A

These coarse guidelines may be modified by moving the cutoff points to coincide with the natural breaks in the standardized mark list.

3.4 Histograms

Where cutoff points are to be based on natural breaks then it is appropriate to inspect histograms of the standardized mark distributions. Inspection of the individual component histograms can give an indication of their general characteristics. Where a large degree of positive skewness occurs it would appear that only a small portion of the class gave very good answers. Where there is a large amount of negative skewness then clearly a small group of students have been left behind on that component.

3.5 Profiles

A final grade may be assigned directly from the standardized, weighted sum of the component grades. However, it is often instructive to inspect a profile of the students' performance over all the components in the course. For example, where one missed or failed a test in an otherwise good set of marks is sufficient to bring the final mark below a cutoff point, then it may be appropriate to put more weight on the profile. Conversely, where an accumulated set of poor component marks is just sufficient to put a student above a cutoff point with perhaps the help of a good grade on a joint assignment, then it may be appropriate to reduce the final grade.

An essential prerequisite of reliable interpretation of profiles is that the individual components must have the same distribution characteristics. It is almost impossible to reliably interpret a set of raw mark profiles where each component mark set has a different mean and standard deviation. The matrix of standardized marks provides the appropriate base for profile decisions. However, two other matrices are presented which make interpretation easier. First the standardized marks are converted to an alphabetic scale with extra resolution offered by "+" and "-" to help in the subjective process of profiling. A second profiling display contains the rank order of students for each of the components. This ranking allows for ties but is based on the standardized mark list which is rounded to one place of decimals. Consequently, the differentiation between adjacent students is not so fine that the ranks exaggerate close and perhaps non-significant differences.

4. ANSWER ANALYSIS

The collective responses from students provide valuable information to the professor regarding the effectiveness of his teaching and the appropriateness of his assessment methods. In order to provide this information in a useful form two analyses are presented. The first deals with the discriminating characteristics of different categorical alternatives and the second with the level of association between the separate components.

4.1 Categorical Response Analysis

This program tabulates the number of students who gave each response in a multiple choice or categorical question. Where the majority of students give a response that is wrong or has a lower value than the correct or best answer then there is the implication that the material has not been well learned (or taught).

A second part of this categorical response analysis involves the breaking up of the class into different groups, based on the overall mark. Any number of groups can be considered, however, in practice the upper and lower halves are sufficient. The displayed matrix then contains the numbers of students who gave a particular response (category) according to which group they belonged based on overall performance. Again a qualitative interpretation of this array can provide more direct insight than more sophisticated statistical techniques, such as contingency table analysis or discriminant analysis. For example, if five out of a class of twenty gave a wrong response and all those five ended up in the lower half, then it could be deduced that that answer/category was a good discriminator. On the other hand if the five who gave the 'wrong' answer all ended up in the top half of the class, then the teacher would perhaps wish to search for some ambiguity in the question that the more perceptive students detected. This qualitative analysis should be conducted with caution and protected by more sophisticated statistical tests if desired.

4.2 Association Between Questions

The Pearson Product Moment Correlation Coefficient is calculated for all pairs of questions and for all questions with the overall mark. It should be noted that these calculations are simplified by the fact that the input data matrix contains standardized values. The resulting matrix of correlation coefficients must be considered in light of the fact that no sample size, normality or significance test information is given. However the correlation coefficients do give an indication of the association between components. For example, if a very high correlation exists between two components then it is possible that one of them is redundant (i.e. it measures the same underlying characteristic of the students). If a pair of components are negatively correlated (e.g. a test and a project) then these two components may truly reflect different and mutually exclusive characteristics of students. Low absolute correlation values will suggest statistical independence of the components. If the final mark is based on the raw rather than the standardized data then a high correlation of one component with the overall mark will indicate that that component has a high empirical weighting.

As with the item discrimination issue discussed in the previous section, a balance must be sought between statistical sophistication and educational significance. Professors do not want to be faced with a thick computer printout with many derived statistics and tests to interpret the grades from a course that they have been teaching for many years. However, some indication of the statistical nature of their grades might prompt the professor to investigate in more detail his assessment methods.

5. CONCLUSIONS

The answer to the student's predominant question: "What grade am I going to get?", given a norm based mark amalgamation system, is not as straightforward as in a criterion based system. However, the system described in this paper does produce a final mark that accurately reflects a student's relative ability and which avoids the conceptual and practical problems of simplistic criterion systems. Furthermore, the computational processes and the reports that are generated provide the professor with an appropriate basis for his final grade judgement.

REFERENCES

Blankenship, J. "The Gradebook System", Prentice-Hall, 1984.

Boyle, T. A. and Wright, G. L. "Computer-Assisted Evaluation of Student Achievement", Engineering Education, December, 1977, vol. 67, no. 12, pp. 241-244.

Cheshire, S. R. "Assigning Grades More Fairly", Engineering Education, vol. 65, no. 4, 1975, pp. 343-348.

de Nevers, N. "An Engineering Solution to Grade Inflation", Engineering Education, vol. 74, no. 7, April, 1984, pp. 661-663.

Farrow, S. B. and Summereld, J. T. "A Computer-Aided System for Generating Multiple Choice Examinations", Engineering Education, vol. 73, no. 4, January, 1983, pp. 308-310.

Flora, R. E. and Carter, W. H. "The Use of Principal Component Analysis to Increase the Ability of Multiple Choice Examination to Distinguish Among Students", MCV Quarterly, 1971, vol. 7, no. 1, pp. 10-13.

Guilford, P. J. "Fundamental Statistics in Psychology and Education", 5th Edition. McGraw-Hill, 1973.

Hair. "Multivariate Statistical Analysis", PPC Books, Tulsa, 1979.

Hastings, N. A. J. and Peacock, J. B. "Statistical Distributions", Halstead, 1975.

Joreskog, K. G. and Sorbom, D. "Lisrel V", Department of Statistics, University of Uppsala, Sweden, 1981.

Kilpatrick, S. J. "Alternative Methods of Grading One or More Multiple Choice Examinations", MCV Quarterly, 1971, vol. 7, no. 1, pp. 4-9.

Morrison, D. F. "Multivariate Statistical Methods", McGraw-Hill, 1976.

Nie, N. H. et al. "SPSS", 1981, McGraw-Hill, New York.

Peacock, J. B. "The Feedback Classroom", Proceedings of the ASEE Annual Conference, 1982.

Peterson, J. A. and Meister, L. L. "Managing a Test Item Bank on a Microcomputer", T.H.E. Journal, November, 1983, pp. 120-122.

Professional Examination Service. "Multiple Choice Questions", Professional Examination Service, Princeton, N.J., 1983.

Tamara, Y. et al. "Graphical Presentation of Mass Data and its Visual Perception", IFAC Control Science and Technology, 8th Triennial World Congress, Kyoto, Japan, 1981, pp. 1773-1778.

Thomas, C. R. "Examination of a Formula Method for Assigning Letter Grades", Engineering Education, vol. 74, no. 7, April, 1984, pp. 670-672.

Wissler, E. H. "An A is an A is an A, or is it?", Engineering Education, December, 1975, vol. 65, no. 12, pp. 232-237.

Work, C. E. "A Nationwide Study of the Variability of Test Scoring by Different Instructors", Engineering Education, December, 1976, vol. 65, no. 12, pp. 241-248.