Marking and evaluating class tests and examinations by computer

By P. D. Groves*

A program is described which will mark multiple choice type questionnaires, comment on students' progress and provide information about the answering pattern and value of particular questions.

A common method for checking the progress of students is by the regular setting of class tests. Although conventional examination-type questions involve the teacher in the considerable tedium of marking, multiple-choice type questionnaires can avoid this since they readily lend themselves to marking by mechanical means, for example by template or by computer. These methods have frequently been used, particularly in American universities.

A computer however is capable of producing more information from a series of tests than just a set of marks and an ALGOL program has been written for an Elliott 803,computer which will

1. Mark multiple choice type questionnaires.
2. Comment on the progress of each student.
3. Summarize the results obtained by the class, question by question.
4. Provide information on the ability of individual questions to discriminate between able and less able students.

The information so obtained can be of considerable value to student and teacher, particularly in large classes where personal contact is difficult. The comments can act as a stimulus to the student, and the summary of results indicates to the teacher the effectiveness of his teaching and draws attention to any topics which the class has had difficulty in grasping. Information on the discriminatory value of questions indicates their efficiency and assists in the writing of new questions.

Procedure

Students are given a sheet of questions to each of which is provided several numbered alternative answers, only one of which is correct. The student indicates his choice by writing the appropriate number (or zero for "don't know") on a slip at the bottom of the question paper: this is torn off and handed to the teacher at the end of the test. These numbers are transferred to punched tape which is then, together with details of the correct answers etc., input to the computer.

In order to discourage guessing (very necessary when only a few alternative answers are given) wrong answers can be made to carry a penalty. As used currently three alternative answers have usually been given to each question; correct answers have all gained one mark and wrong answers have incurred a penalty of $-\frac{1}{2}$. A possible negative total is avoided by rounding off such to zero. If a larger number of alternatives were given then a penalty might appear to be less necessary (because a guessed answer is less likely to be the correct one); it is, however, still desirable because it is important to differentiate between answers which are not correct (1) because the student simply does not know the answer and so would guess, and (2) because he mistakenly believes an incorrect answer to be the correct one.

Marking of the students' answers is a simple process, correct answers being placed in a one-dimensional array and the students' answers, contained in a two-dimensional array, compared with these. After input of the "answers tape" a "record tape" is input; this contains the names of the students in the class together with a record of the marks that they have obtained in previous tests **(Fig. 1)**. The average mark obtained in the previous three tests (or less if fewer tests have been taken) is worked out for each student and his current marks compared with this. An appropriate comment is then chosen for each student which is dependent on

1. the student's current marks;
2. the comparison between this and his previous average,
3. the number of questions which have been answered with "don't know".

A report is then produced by the computer listing the

```
£  CLASS  TEST  IN  INORGANIC  CHEMISTRY
              HONOURS 1?

 1 £ANDREWS R M      ? 10·0  50·0   60·0     75·0
 2 £BOARDMAN A R     ? 80·0  33·3   85·0     75·0
 3 £CLARK D J        ? 15·0  27·8   50·0     20·8
 4 £COOMBS G E       ? 85·0  83·3   70·0     33·3
 5 £HEALEY L F       ? 70·0  66·7   75·0     91·7
 6 £HUGHES O         ? 70·0  66·7   60·0     41·7
 7 £JAMES H S        ? 40·0  83·3  100      100
 8 £JONES MISS L     ? 90·0  77·8   65·0     33·3
 9 £JONES M I        ? 45·0  33·3   35·0  — 1·00
10 £JONES R W L      ? 30·0  61·1   50·0     70·8
11 £MORRIS J A       ? 40·0  55·6   10·0     12·5
12 £PEPALL J         ? 45·0  44·4   35·0     70·8
13 £RICKETTS I       ? 75·0  33·3 — 1·00     41·7
14 £WALLACE A B W    ? 85·0  83·3  100       91·7
```

Fig. 1. A print-out of part of a typical "record tape". Marks are stored as percentages to allow variations in the numbers of questions in different tests. —1 indicates an absence. The characters £ and ? are the 803 equivalents of string quotes

(*Note. To avoid possible embarrassment, names of actual students have been altered in Figs.* 1 *and* 2)

* *Department of Chemistry, University of Aston in Birmingham, Gosta Green, Birmingham* 4.

```
          CLASS TEST IN INORGANIC CHEMISTRY
                      HONOURS 1
                      TEST NO. 4
                 POSSIBLE MARKS = 12

 1 ANDREWS R M      9·0 GOOD. IMPROVING
 2 BOARDMAN A R    10·5 EXCELLENT. KEEP IT UP
 3 CLARK D J        2·5 NO SIGNS OF IMPROVE-
                        MENT
 4 COOMBS G E       4·0 KEEP UP WITH YOUR
                        READING
 5 HEALEY L F      11·0 EXCELLENT. KEEP IT UP
 6 HUGHES O         5·0 NO REAL IMPROVEMENT
                        YET
 7 JAMES H S       12·0 EXCELLENT. KEEP IT UP
 8 JONES MISS L     4·0 ONLY FAIR. YOU CAN DO
                        BETTER
 9 JONES M I       ABSENT
10 JONES R W L      8·5 GOOD. IMPROVING
11 MORRIS J A       1·5 MARKS STILL LOW.
                        ANSWERS ARE CARELESS
12 PEPALL J         8·5 GOOD. IMPROVING
13 RICKETTS I       5·0 NO REAL IMPROVEMENT
                        YET
14 WALLACE A B W   11·0 EXCELLENT RESULTS
```

Fig. 2. Part of a typical report. Up to 30 different comments can be made available and, being input as data, may be varied from time to time, or, if wished, avoided completely

students' names, their marks in the current test and the appropriate comments (**Fig. 2**). At the same time an updated "record tape" is produced for use with the next test. The information on this tape can also be used as data for a short auxiliary program which will calculate, for each student, total marks to date, average marks per test and number of absences.

The comment is determined by the value of three integers i, j, k. i can have the values 1 to 5 according to which of five categories the current mark falls into. A mark of 80% or more say, would have an i value of 1, a mark between 65 and 80 say, would have an i value of 2 and so on.

The criteria for determining the limits are fed in as data and can be varied from test to test if so desired.

j can have values 1 or 2, 1 indicating that the number of questions answered by "don't know" is greater than or equal to 40%, and 2 that it is less than this.

k can have values 1, 2 or 3. 1 indicates that the current marks are an improvement on the recent average (usually taken over the previous three tests) by an amount greater than 20%. 3 indicates a deterioration (less than 20%) and 2 indicates no appreciable change.

The maximum number of comments is therefore $5 \times 2 \times 3 = 30$ although it may not be possible to call all of these (for example a student obviously could not have 85% marks and 40% "don't knows"). In the current program 27 comments are available. These are input as data using the Elliott ALGOL procedure *instring*. (This is a procedure provided on the ALGOL compiler which enables strings of alphanumeric characters to be read into the computer and stored in an array; such strings can be output using the procedure *outstring*.)

Each particular combination of values of i, j and k clearly defines a comment. For example, marks of 60% with 40% "don't knows" and showing an improvement of 25% on the recent average might call forth the comment "Improved, though more reading necessary in some topics". The criteria by which the comments are called, and the comments themselves, may be varied, if required, from test to test.

```
                   SUMMARY OF RESULTS
                  CLASS AVERAGE = 7·3
 Q    %RIGHT   %WRONG   %DONT KNOW   DISCRIMINATION
 1     61·5     30·8        7·7           0·41
 2     71·8     10·3       17·9           0·47
 3     28·2     35·9       35·9           0·41
 4     69·2     28·2        2·6           0·59
 5     76·9     10·3       12·8           0·24
 6     66·7     23·1       10·3           0·47
 7     33·3     23·1       43·6           0·35
 8     87·2      0·0       12·8           0·35
 9     89·7      0·0       10·3           0·29
10     84·6      0·0       15·4           0·29
11     74·4      2·6       23·1           0·47
12     74·4     10·3       15·4           0·47
```

Fig. 3. A typical summary. The first column gives the question number. The second, third, and fourth columns indicate how each question has been answered percentagewise by the whole class. The final column indicates the value of each question in discriminating between the more able and less able students

Some teachers may feel doubtful about the value of the comments but theories of learning would seem to provide adequate justification for them. They can, by indicating approval or otherwise, provide some stimulus to effort and, particularly in a large class, make the student feel that his teacher is able to take a more personal interest in his progress than would otherwise be possible. The students also appreciate that the comments are reasonably objective and cannot be influenced by any personal bias on the part of the teacher. If, however, they are not desired by a particular user, provision is made for that part of the program to be bypassed. In any case comments will not be produced for the first test of a series.

After the report on the class the computer next produces a summary of the results (**Fig. 3**). The marks gained by each student for each question are stored in a two-dimensional array and these are summed question by question to give percentages of the class getting "right", "wrong" or "don't know". Care has been taken in programming to avoid empty elements of the array due to absent students; counting these would obviously give erroneous results.

The summary so produced is found to be particularly useful and provides valuable "feedback" from the class to the teacher. It indicates topics which the class has had difficulty in understanding (a large percentage of "don't knows") and those which it has misunderstood (large percentage of "wrongs"). It can indicate topics for revision and tutorials and can help the teacher in replanning his lectures.

The discriminatory value of each question is obtained by arranging the students in ranking order according to the marks gained and then, for each question, counting the students in the top third and in the bottom third who obtained a correct answer. The "discrimination" is then given by the difference of the proportion of students in each group. A discrimination of +1 means that all the students in the top third obtained a correct answer, but no students in the bottom third. This clearly discriminates well between the more able and less able students. A value of −1 conversely means that all the less able students were correct but none of the better ones! Such a value is of course highly unlikely but questions having negative discriminations must always be carefully examined since they almost certainly will contain errors or confusing information of some sort. Questions with low positive discriminations might be avoided in actual examinations but should not necessarily be avoided in regular tests since this could lead to the weaker students becoming discouraged.

Conclusions

The program has proved to be very useful in checking the progress of students, particularly those in large classes. It rapidly provides a great deal of information would could otherwise be obtained only with a great deal of effort. Its value is, however, completely dependent on the quality of the questionnaires, and no amount of ingenious programming will produce significant information if these are carelessly prepared.

The program was originally written for use with a large class of part-time students taking a Higher National Certificate in Chemistry course. It has subsequently been used successfully with a first year Chemistry Honours Degree course and has also been used for marking a first year examination in Botany and Zoology.

If the questionnaires are carefully written they can help to *teach* as well as to *test*; the author has found it very useful to follow up the tests by the issue of duplicated sheets of notes which discuss the various alternative answers. These, for maximum effect, should be given out immediately after the completion of the test so that students can straight away check their answers. There is considerable evidence that rapid access to correct answers results in significant reinforcement and contributes materially to the process of learning. This, of course, is one of the principles upon which teaching machines (and other forms of programmed learning) are based.

It is clear that the program could be developed and improved in various ways. It might for example prove very useful to assign the questions in the questionnaires to various categories according to what is required in their answering (factual knowledge, understanding a set of concepts, correlation with other branches of knowledge, imagination etc.). Cumulative scores, in these categories, especially if kept over the several years of a student's course, might well provide some extremely interesting and valuable information both on individual students and on groups. Such information could contribute considerably to the development of teaching techniques.

Various mechanical improvements could be made. For example, it would speed up the running of the program if magnetic tape were used for the record instead of paper tape. Preprinted cards on which students' answers could be recorded and then punched would probably be an advantage over paper tape. A card reader which would directly read hand marked cards would be very valuable.

A print-out of the program is available from the author as is also a set of operating instructions for its use on the Elliott 803.

Book Review

The Mathematical Approach to Biology and Medicine, by N. T. J. Bailey, 1967; 296 pages. (New York; *John Wiley & Sons*, 57s.)

This book is in two parts of roughly equal size. In the first part the author discusses mathematics and statistics within his given context, and follows this with chapters on model-building, operational research, and computing, ending with one on "teams, projects, and organizations". The second part treats five special topics where mathematics have been applied to biology. These are numerical taxonomy, population growth and ecology, the theory of epidemics, genetic linkage and chromosome maps, and mathematical methods of medical diagnosis. A final chapter deals with operational research in medicine. To cover all this ground in less than 300 pages must inevitably mean that only an outline can be given of the subjects in view.

I feel that the two parts are likely to appeal to rather different classes of reader. Many biologists will appreciate Dr. Bailey's clear expositions in the first part, but many will find the second part hard going unless their mathematics is fairly strong. Those in the field of statistics, O.R., or computing will get from the second part a good conspectus of what biomathematics is about, but they will inevitably be dissatisfied by the level of exposition in their own field. Thus readers of this *Journal* who have done large-scale editing of data on a computer may feel less enthusiastic than the author about his statement that "both languages (FORTRAN and ALGOL) are highly suitable for scientific computing".

To sum up, Dr. Bailey has taken on the difficult job of writing a book surveying diverse but inter-connecting disciplines where inter-communication is often poor or non-existent. If he can get even some biologists and mathematicians to talk intelligibly to each other, this will be a great step forward.

J. A. Nelder (Wellesbourne)