Using Microcomputers to Assess Achievement and Instruction

Larry R. Nelson
University of Otago,
Dunedin, New Zealand

Most of us go about creating, administering, and scoring tests without using a computer. Our trusty booklets of marks are relied on to record and review student test and assignment results. At the end of the term the same booklet, combined with a small calculator, allows us to render a final grade. Why should we consider the intervention of a microcomputer as an aid? Haven't we gotten along just fine without computers?

Before answering these questions, a list of the steps used in the traditional pattern of student assessment will be helpful: (1) creating tests; (2) giving tests; (3) scoring tests; (5) reporting results to students; (6) recording results in class grade book; (8) deriving final grade at term's end; and (9) reporting final grades to students and to central administration.

I will immediately note two things about this list. The articles by Hambleton and Ward have shown how computers can help with steps 1 and 2. This article will suggest how computers can help with the remaining steps. These comments/suggestions probably will not be novel, but I hope to draw your attention to the two steps missing from the list, namely, (4) determining test quality and (7) reflecting on the quality of instruction.

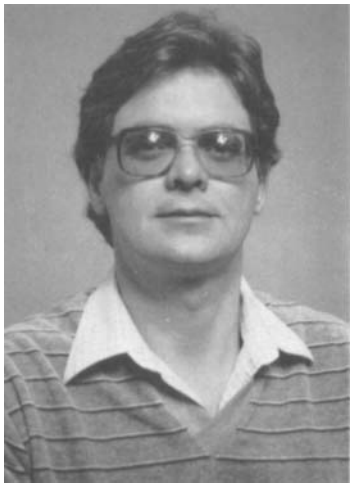
My contention is that steps 3, 5, 6, 8, and 9, that is, our traditional activities, can be considerably assisted by using a computer. Scoring, reporting, and deriving final grades can be done faster and more accurately. Second, I contend that the savings in time, and the computer data base formed, will allow steps 4 and 7 to be carried out as well. The latter two steps are important, but generally difficult to carry out without a computer.

Assumptions

Let us assume that instruction is based on a set of objectives, and that our ultimate goal is summative, not formative, evaluation. The latter assumption will avoid spending time discussing criterion-referenced testing in particular. The reader should *not* conclude that computers cannot assist with formative evaluation or criterion-referenced testing; they most definitely can. The purpose of this paper, however, is not to discuss types of testing, but to point out the general utility of computers in testing.

Raw Material

We have created and given a test. For convenience, let us consider it a forced-response test: true/false, multiple choice, and so forth. Because we have a forced-response test, many readers are likely to assume that we have gathered student responses using special answer sheets or cards. However, we have the original test papers and nothing more. Students have circled the answer they consider to be best for each question.



Larry R. Nelson is a Senior Lecturer in Education at New Zealand's Otago University. Formerly a specialist in research methodology, he is currently interested in computer applications in education.

The test is on basic statistics and covers four topics: measures of central tendency, measures of variability, standard scores, and correlation. We have five items, or questions, for each of these areas, making our test 20 items long. The test spans four pages.

Scoring the Test

We will score the test using a program that runs on a small personal computer, such as an Apple, TRS-80, or IBM PC. The computer generally used with test scoring is on a mobile desk. We put the program and data disks into the computer, and turn it on. Soon, a menu of options appears, and we select one for entering new data. In 20 minutes we have entered, verified, and printed the responses from a class of 35. That is, not only have we typed in the responses, but we have used the computer to double check them and to make a permanent, printed listing of original responses.

We then give the computer instructions for scoring the test. It takes less than 5 minutes to tell it how to create one overall test score — we have to supply the correct answer for each question and indicate how many points apply to each right answer. It takes less than one minute to score all 35 tests. In approximately half an hour we have completely scored the test, double-checking the scoring in the process. Could we have scored 35 20-item multiple-choice tests by hand in the same time? Possibly. With double-checking? Probably not. However, our primary objective is not to concentrate on saving time.

Recall that the test covers four distinct, well-defined subject areas. What we would like to have is a set of five

scores for each person, not one. We would like one score for each area, and one overall score. Creating five scores requires more time but is easy for the computer to do. In less than one hour we have five scores for each of 35 students.

Reporting Results

The test scoring programs we use include several report generators. One of them prints a report that gives a raw test score, a percentage score, and a set of three standard scores (*z, T, CEEB*) for each test area. Thus, we can give students a form showing them how they did in each area, and, assuming they remember how to interpret standard scores, how their performance compares to that of other class members.

If we wanted to, we could use another program to point out exactly which items each student got wrong. Or, we could use the report making program that converts raw scores to letter grades. Obtaining these reports is easy, and results profiles for 35 students can be printed in 10 minutes.

Earlier I said that I would stick to summative evaluation. What we do in some of our papers is actually a bit more than classical summative evaluation; we say that students must get 50% of the items right to pass (with a "C"). Anything less than a "C" means they have to take a parallel form of the test. In this regard, our five-score summary profile proves popular because it indicates the areas students ought to be reviewing.

Could we produce score profiles, broken down by topic areas, by hand? Certainly. Could we, within each topic, also manually create percentage scores and standard scores? Certainly. But it takes too much time to do it by hand.

The Grade Book

Having scored the test and reported results to students, we would next ordinarily turn to our grade book and update it with the students' marks. With the computer, our lists of class marks are stored in a file on a data disk.

After the first test (or assignment), we use a program to open a new grade book for each class. The computer lists student names, the name of the first test, and the mark earned by each student. As the term progresses, additional marks are added. The grade book is now a diskette. Opening it to the right page involves giving the computer the class name or number. Entering new marks is done not by pencil or pen, but through the keyboard; the computer displays a student's name, asks for his or her mark, we type it in, and then the machine moves on to the next student.

Why should we keep grade books on a machine? Scoring tests, especially multiple choice ones, is an activity obviously assisted by a suitably-programmed computer. But why class lists as well? The answer is found in the next section.

Deriving Final Grades

We all know that final grades are formed by adding up the marks earned during the year on course tests and major assignments. What we may also know, but are inclined to overlook, is that combining marks is subject to some peril. The impact each mark has on the final composite depends on the variance of the mark.

Let us suppose, for example, that we have only three test results to be combined. Test 1 had a score range of 15-50, Test 2 a range of 20-42, and Test 3 a range of 25-38. If we want these test results to count equally, do we simply add them up? If we do, Test 1 will sway the final composite its way, because it had more variance than the other tests: because marks were more spread out. A student could get high marks on Tests 2 and 3, a mediocre mark on Test 1, and end up with a total mark that is more mediocre than high. Again, the influence of any particular mark on the final composite depends on the variance of the mark. The more the variance, the greater the impact.

This scaling anomoly can be controlled by boosting the variance of some marks and attenuating the variance of others. This is done by deriving suitable multipliers, or sometimes by standardizing the marks to the same scale before they are added up.

The process of deriving these multipliers is discussed in many text books. See, for example, Hills (1981, p.320); Bloom, Madaus, and Hastings (1981, p. 110); and Stanley and Hopkins (1972, p. 311). The effects of combining marks without taking into account disparate variances can result in final grades that differ from grades that would result if the variances were correctly adjusted.

With this information at hand it should be obvious why we keep our class grade books on the computer. We have obtained a program that permits mark variances to be adjusted. In fact, the program is part of the same package that scores tests and makes printed reports.

Could we accomplish our variance adjustments without a computer? We could, and we might try it had we just one class and sufficient time at year's end to make the many needed calculations. We probably were previously in the same boat as many present readers: We have known about the variance problem for a long time, but we tended to forget its existence when end-of-year pressures demanded final grades in a hurry.

Reporting Final Grades

Final grades can be a bit different from the marks given during the course of the term. The final grade normally requires some sort of summative label, whereas test and assignment marks often require less formal summaries of performance.

How does one take a composite mark, derived by adding up (with variance control) the term's tests and assignments, and convert it to a "B," a "pass," or a label such as "promoted with distinction"? Many different approaches exist, and each major textbook in the area seems to have its own recommendation.

Our approach relies heavily on histograms and correlations. First of all, the components added up to make the final mark should all intercorrelate positively. Our grading package for the computer lets us look at correlations, as well as the reliability of the final composite. If we are satisfied that the composite makes psychometric sense; if it has acceptable consistency, then we turn to a careful examination of the composite's histogram. We look for clusters of scores and award different summary labels to each cluster. In essence, we follow the procedure recommended by Bloom, Madaus, and Hastings (1981).

For many of our courses, we actually derive two final composites. One is a percentage composite that represents the student's final composite as a percentage of the maximum possible composite score, or the score a student would get if he or she achieved the highest possible score on each term test

or assignment. Theoretically, this gives us an absolute scale that we can carry from year to year, and sometimes across classes as well. For example, the median percentage composite for the 1983 educational psychology class was 68. If this slips to 60 in 1984 we will do some head scratching — is the slip explained by less able stuents, by inferior instruction, by harder tests, or by all three?

Our other final composite is standardized, based on a scale with a mean of 65 and standard deviation of 15. This scale immediately lets us pick up the outliers; the lowest students will have composites less than 50, and the highest will be greater than 80.

Again, our computer software enables us to accomplish these tasks with relative ease. It allows us to derive final grades that control for the variance problem, it permits us to look at correlations and reliabilities, and it gives us the chance to create a variety of rescaled composites.

We probably spend as much time accomplishing these tasks as we formerly spent deriving marks with grade books and calculators. The point is that we can accomplish much more; we can produce fairer grades, control for the variance problem, and achieve a vastly superior understanding of student performance.

Assessing Test Quality

Thus far, we have discussed out and out marking. We create a test, students take the test, the test scores are scored, and we record results in a grade book. The year comes to an end; we tally up all the individual tests/assignments and produce a final grade for each student.

We could, and should, do more. Before we turn our assessment scope on students, we ought to make an effort to assess the quality of our measuring instruments. We would also do well to use the information collected on tests and assignments to assess the apparent quality of our instruction.

Assessing test quality generally involves both item and scale analysis. On the item analysis side, we need to identify bad items. Bad items have been inadvertently mis-keyed, have intrinsic ambiguity, and/or have structural flaws that make identification of the correct answer easier than desired.

Classical statistics used to summarize item quality are based on difficulty and discrimination indices. Although procedures for calculating such statistics by hand have existed for years, we suspect that few readers would argue against a computer-based analysis as the preferred procedure for looking at item quality.

An example of the utility of computer-based item analysis is shown in Table I. The data were derived from the administration of a 30-item multiple-choice test to a class of 126 first-year university students. Each item had one best answer and three distractors.

On the left-hand side of the Table, item difficulties are plotted. We see that four items, numbers 2, 4, 17 and 24, were answered correctly by 90+% of the students. Two items were particularly hard: Items 18 and 26 were answered correctly by less than 30% of the students. The information on the right-hand side of the table relates to the discriminating ability of each item, that is, the extent to which students who got an item right tended to have high criterion scores. The discrimination index produced by the microcomputer program we use is a point-biserial correlation coefficient. The criterion can be either internal or external, with the former being the most common. For the items summarized in Table I, the criterion was the total test score; the correlations are automatically corrected for part-whole inflation by the program.

The item discrimination data reveal that three questions had negative discrimination indices (numbers 12, 26 and 27), an undesirable result. Furthermore, seven items had discriminations of less than .10. One third of the test's items need attention.

Such attention is obtained by looking at individual item report cards, shown in Table II. The means columns toward the right of Table II provide the raw criterion mean score for those students selecting each item option, and the same score expressed as a *z*-score. Thus we see that the 30 students selecting option *A*, and the 30 students selecting option *C* had average criterion scores that were above the mean criterion score, something signalled by positive *z*-scores.

The item report card is meant to be readily interpretable. Item 26 was difficult, had a negative discrimination index, and had distractors at less than 100%, that is, it had one or more distractors that either distracted no one or distracted above average students.

Such computer-generated reports

TABLE I

Item Difficulties		Item Discriminations	
Dif.:	Item ID Numbers	Dis.:	Item ID Numbers
.90+ :	2 4 17 24	.80+ :	
.80+ :	7 8 15 19 21 23 25 29	.70+ :	
.70+ :	3 6 10 22	.60+ :	
.60+ :	5 13 16 20 27 28 30	.50+ :	
.50+ :	1 11	.40+ :	
.40+ :	9	.30+ :	10 14 22
.30+ :	12 14	.20+ :	7 13 15 20 23 25 30
.20+ :	18 26	.10+ :	2 3 4 5 6 9 11 16 24 29
.10+ :		. 0 :	1 8 17 18 19 21 28
. 0 :		Neg!:	12 26 27

TABLE II

Statistics for Item 26

Option	Weight	Freq.	Percent	Correlation	Means Raw Score	Means Z Score
A	1.00	30	23.8	-0.009	21.767	0.215
B	0.00	54	42.9	-0.102	20.667	-0.117
C	0.00	30	23.8	0.013	21.133	0.023
D	0.00	12	9.5	-0.022	20.833	-0.067

Difficulty	= 0.238	Discrimination	= -0.009	Distractors at	66.7%
Item Mean	= 0.238	Item Stan. Dev.	= 0.428	Item Variance	0.18

lighten the burden of the item analyzer. Incorrectly keyed items can be found by paging through the item report cards. These items will also generally stand out on the displays of item difficulty and discrimination — they will be too difficult and tend to have negative discrimination.

Items with intrinsic ambiguity include those whose distractors have more plausibility than what the item writer originally designed. Ambiguous items can be expected to have low discrimination and low difficulty.

Test questions with other structural flaws are characterized by items and options having grammatical or contextual clues that make it easy to pick out the correct answer. These items will have distractors which distract no one. They will show up in the .90+ line of the difficulty table, and more often than not in the .00 line of the discrimination table.

The identification of bad items is of obvious importance. It is unfair to students to base test scores on mis-keyed and ambiguous items. Before scores are produced, mis-keyed items should be corrected, and consideration given to double-keying ambiguous items; for example, marking more than one answer as correct. Items with the previously discussed structural flaws should be earmarked for overhaul and should not be used again until they have been repaired. For a particularly lucid example of item revision and repair, see Ebel (1979, Chap. 13).

It might appear that we are limiting our comments to the type of items found in norm-referenced testing. However, we recognize that criterion-referenced items can certainly be misscored, ambiguous, and structurally flawed. Some authors have encouraged the application of classical difficulty and discrimination analyses to criterion-referenced tests, the idea being to provide a ready procedure for uncovering faulty items (see, e.g., Bloom, Madaus, & Hastings, 1981, p. 97).

Supply Items

The application of computer-based routines to the analysis of forced-response items is well known. But it is rare to find tests comprised largely of supply items subjected to any sort of analysis at all, computer or otherwise.

The data in Table III are from a four-item essay test, where each item was worth a maximum of 25 points.

TABLE III

Variable	Mean	S.D.	Low	High
QUES. 1	17.03	1.87	13.0	20.0
QUES. 2	16.11	1.28	14.0	18.5
QUES. 3	11.50	2.06	7.5	15.5
QUES. 4	14.13	5.82	8.0	19.5

	Correlations with Composite		Weights	
Variable	Uncorrected	Corrected	Achieved	Assigned
QUES. 1	.7473	.5404	.2395	1.0000
QUES. 2	.5618	.3815	.1238	1.0000
QUES. 3	.6540	.3687	.2315	1.0000
QUES. 4	.8314	.5262	.4053	1.0000

Thirty students answered each of the four questions. The first part of the table gives the descriptive statistics associated with each of the four questions. The mean column is an index of question difficulty because in this case, all items had the same maximum possible score of 25. Question 3 was the hardest; question 1 the easiest.

The second part of the table provides two important bits of information. The corrected column indicates how each question correlates with a composite formed by adding up scores on the other three questions (uncorrected refers to the correlation between the item and the total composite, a figure boosted by part-whole inflation). The corrected correlation is entirely analogous to classical correlation-based indices of item discrimination.

The achieved column under the weights heading shows the relative effect each question has on the overall test score. The data indicate that question 4 has the greatest influence, and question 2 has the least. This means that a student's final relative standing on the overall test score will be much more dependent on how he or she did on the fourth question than on the second one. This is an undesirable result if the questions were supposed to count equally. Rescaling should be undertaken to bring the achieved weights into line with each other.

We mentioned that supply items are not normally subjected to any sort of item analysis. Our own experience in this area is not extensive; we have had our microcomputer software a relatively short period of time.

A particular difficulty in applying the software comes from the popularity of administering essay tests that allow respondents to select the questions they will answer. As Stanley and Hopkins (1972) pointed out, this makes the comparison of results hazardous if not impossible. For those few essay exams that have not allowed question selection, our analyses suggest that it will be common to find negative correlations among the items, leading to low internal consistency for the test as a whole. Of course, negative correlations among forced-response items are not exactly uncommon. In the forced-response case, however, the number of positive item intercorrelations is usually much greater than the number of negative ones, and coefficient alpha will be strong.

Even when interitem correlations are all positive, and coefficient alpha respectable (above .70), a persistent problem with essay test scoring results from disparate item variances. It is our opinion that the most valuable contribution the computer-based analyses of supply items has to make is to point out the effect of unequal variances, and allow item weights to be readjusted to eliminate variance-related bias. It is a trivial matter to show that the rank order of students will change when bias correction is applied. If effective item weights are not close to those intended by the test scorer, the rank order of student test scores will differ, sometimes dramatically, from what results when the effective weights are adjusted to control for the variance problem.

Scale Analysis

Test scores are linear composites formed by summing a set of items. We are always interested in knowing whether or not our composite makes psychometric sense, if it has acceptable reliability. In addition, if there is some implied hierarchy of skills for the test

item, we generally want to see if item responses follow the pattern implied by the skills.

Again, a computer is an essential commodity. It will provide an overall index of scale cohesion, such as KR-20, Hoyt's index, or coefficient alpha (all three of these indices of cohesion, or internal consistency, can be shown to be algebraically identical). It will also provide the means for validating any implied skill hierarchy, via correlations, cross tabulations, or both.

Feedback on Instruction

Recall the steps in assessment discussed so far — we create a test, administer it, and collect item responses. Before producing test scores, we submit the item responses (or item scores) to a computer program. The statistics produced by the program are used to cull out bad items; mis-keys are fixed, ambiguous questions might be double-keyed, and structurally flawed items are slotted for submission to the repair shop. Test scores are produced, reported to students, and entered into our electronic grade book. At the end of the year the final composite is created, and final grades are applied.

We still need to look at what the item and test statistics can tell us about our instruction. Our undergraduate modules on elementary descriptive and inferential statistics are important for many of our courses. As rough guidelines, we think that students should get at least 80% of the straightforward, knowledge-level items correct. We should look at the printed item summaries we have been saving and review the items. Are we reaching the 80% level? If we can break down our instruction into a series of components with specific objectives linked to specific items, are we getting what we want? Where should revision be undertaken, or where should special tutorials be laid on to correct for apparent inadequacies in our original instruction? What are the implications for next year's planning?

Certainly we want to believe that these are questions we always look at. However as important as these questions are, they are a relatively invisible part of our professional activity; they are, in other words, often relegated to those quiet times outside of normal teaching hours.

There can be little doubt about the merit of employing the assistance of a computer. It will keep all item and test statistics stored away, ready for rapid recall. It will permit us to come back to a particular test, regroup items, rescore items, correlate items, and look at results patterns. And we cannot resist adding what at first would appear to be a relatively facetious comment: Using a computer to query a data base of item and test results has more professional status attached to it than sitting at a desk and shuffling papers. The computer room can provide library-type atmosphere where interruptions are minimized. Furthermore, information retrieval and processing is more stimulating on a computer, and hence more likely to be done.

The Affective Domain

We stray just a bit from the cognitive path to mention assessing student attitudes. When we introduce a new textbook, workbook, film, or design a new lab, are we not usually interested in determining the affective reaction students have to it? Even if we don't make such major alterations to a course, it is probably safe to say that most teachers like to have some candid evaluation of student attitudes at the end of the term.

An example of a typical course evaluation form is given in Stanley and Hopkins (1972, p. 291). The form is comprised largely of Likert questions that ask students to summarize on a 5-or 7-point scale how they feel about aspects of the course.

Again we have a situation where some data analysis could be accomplished manually. However, Likert questions are often best summarized and compared by calculating means and standard deviations. Often one will want to produce some sort of cross-tabulation of results, comparing the answers to a question given by particular subgroups, such as males/females. A computer program will permit these statistics and tables to be readily produced.

Issues and Practice

We have attempted to stress some of the issues involved in practical assessment procedures and to highlight the advantages of using computers. Readers familiar with basic measurement texts will recognize our theme as one that is common to the literature — computer-based processing facilitates processing test results and opens the door to a range of ancillary benefits as well. The issue at stake is really *not* whether or not computers have something to offer us; the answer to that issue is clear. The issue really has to do with our lack of inclination to use them.

Computer software appropriate for achieving many of the activities discussed in this paper has existed since the mid-1960s. Considering the texts that for years have reviewed the advantages of using computers in measurement, why are we again tooting the horn for computers? Until a few years ago, the software and hardware were not always easy to use, were often expensive to use, tended to require bulky storage media such as cards and tapes, and was prone to run under "batch mode." Thus we tended not to use them. But times have changed. We now have computers that fit on a desktop and personal work stations, complete with printers. Machines are user-polite, running software that conducts a dialogue with the user, pointing out what can be done and how to do it. Programs automatically store results and retrieve them on demand. Flexible programs make it possible to regroup items, eliminate bad items, and devise a variety of both test and student report cards. Storage media enable data for hundreds of students to be maintained on a small, lightweight diskette.

The promise of computers has been known for a long time. In practice, however, few of us have been able to realize their potential. But now the promise has become a reality. The potential is there as it has never been before. We can incorporate computers in our teaching, knowing that our students will find machines in their offices and schools when they graduate. We can incorporate computers in our own work, as we have exemplified throughout this paper. The software exists, and we have been using it for more than 2 years. The issue now is to convert the potential into practice. The benefit should be substantial.

References

Bloom, B.S., Madaus, G.F., & Hastings, J.T. (1981). *Evaluation to improve learning.* New York: McGraw-Hill.

Ebel, R.L. (1979). *Essentials of educational measurement.* (3rd ed.) Englewood Cliffs, N.J.: Prentice-Hall.

Hills, J.R. (1981). *Measurement and evaluation in the classroom.* (2nd ed.) Columbus, OH: C.E. Merrill.

Stanley, J.C., & Hopkins, K.D. (1972). *Educational and psychological measurement and evaluation.* Englewood Cliffs, NJ: Prentice-Hall. ■