COMPUTERISED FORMATIVE ASSESSMENT IN VERY LARGE LABORATORY CLASSES

John Cross

Computerised Formative Assessment in Very Large Laboratory Classes

Dr John Cross B.Sc., M.Med.Sc., Ph.D., Dip.Tert.Teach.
Microbiology Dept
PO Box 56
Dunedin
New Zealand
John.cross@stonebow.otago.ac.nz
Tel: (03) 479 7737
Fax: (03) 479 7106

"Feedback . . . is the lifeblood of learning."(Rowntree, 1987)

Key words

Feedback. Formative assessment. CAA. Informing teaching

Abstract

This study was concerned with the introduction of a computer option into a successful laboratory based formative assessment providing immediate bidirectional feedback, and applied to several hundred first year Health Science students. Students view the formative assessment as beneficial to learning, and those using the computer version performed at the same level in summative assessments as those using the paper version. Given free choice between the computer version and a paper-based one, around 37% of students opted for the computer version during the semester, generally when they felt confident about the laboratory material. This allowed experienced teaching staff to spend more time with students in need of more assistance. Student waiting time was substantially reduced. Feedback to staff using the computer version has proved an effective adjunct to improving teaching materials and methods.

Background

This project developed in response to a teaching and learning problem. In establishing a new Health Sciences Biology course, a decision was taken to provide a formative assessment each week to about 400 students in laboratories. The requirement to provide rapid feedback to learners meant the assessments were processed before the students left the laboratory. The number of students actually joining the course in its first year (1994) was 700, and this number has since risen to around 1,000. At the time when many students required assistance to complete the laboratory work, academic staff were occupied with the formative assessment. This lead us to review the value of

the assessment in learning, and to consider the use of computer assessment as a means of making the most effective use of quality teaching staff.

The Assessment

First year Science at Otago is made up of (mainly) basic science papers, and placement in second year courses is highly dependent on first year performance. First year at University has been described as one of the two major 'gateways' in a student's career (Rowntree,1987), and this is clearly the case at Otago with respect to a number of professional Health Science courses, including Medicine and Dentistry, as well as Science degrees. There tends to be quite a strong competitive element among most of the students taking these first year papers papers. The need to maximise performance, coupled with the substantial culture shock encountered by most first year students, creates a stressful environment.
It is a major part of the duty of first year teachers, and that of the University as a whole, to give first year students special consideration, to provide the best possible conditions for the abilities of the student to blossom at this most critical time. The formative assessment appeared to offer a useful tool in achieving that end.

At the time the project began, the Biology for Health Sciences paper consisted of four sections each operated by a different department - Microbiology, Anatomy, Biochemistry and Physiology - and co-ordinated by a small central unit. Each section had a theme which was developed to a reasonable depth, rather than attempting to cover a wide field superficially. As an example, the theme of the Anatomy section was Human Movement, and the bones, joints and muscles involved in walking and running formed the first part of the section, the last part bringing these together in the study of complex movements. Laboratory work and lectures were very closely co-ordinated.

In general, the formative assessment was based on five themes for each of the twelve laboratory sessions, with five questions within each of the five themes, a total of 25 questions in all. In the paper form of the assessment there were five questions per sheet, one from each theme, so that there were five different sheets of questions available. Questions sought responses which were a mixture of fact recall and concept understanding, and concentrated on issues that staff regarded as important information or concepts. Questions are not generally difficult, and were designed to ensure that learners had grasped the essentials of the materials dealt with in that session. Students were free to undertake checkout whenever they wished after a predetermined time, usually after the first hour and a half of the laboratory session. Up to three attempts were allowed, and tutorial help was offered to students having difficulty. Since the assessment was carried out just prior to learners leaving the laboratory, and was intended to ascertain what items of value they may have gained from the class, we referred to the process as “checkout”.

The computer version was designed to mirror the paper version as far as possible. Developed in Hypercard[1], it provided better visuals than the paper version, automatic feedback on time taken to complete the assessment, automatic marking, and flagging of wrong answer choices. If a student did not complete the computer version in two attempts, the computer would not allow them to sit again, and referred the student to a member of staff.

The Problem

The Departments provided roughly one member of staff (including demonstrators) per fifteen students. Students were streamed for laboratories, with about 100 in each stream and six staff members present. Towards the end of a laboratory session the number of students undertaking checkout was quite large, while those still completing the laboratory material included a number who required a little more assistance than others, for a variety of reasons. It is in this period that there were inadequate numbers of staff available to process checkout at reasonable speed and also provide help for other students. This was clear to the senior staff, and also commented on by students. In an end of paper evaluation one of the questions asked students to comment on "the change you would most like to see in this course", produced several comments on this aspect of checkout before the introduction of the computer version. For example, *"You have to wait too long to get marked, I suggest handing in the sheets. Otherwise - everything OK."*
Staff were also well aware from processing the checkout sheets that around 75% of students generally had no difficulty completing checkout, and our interaction with them consisted of a comment such as, "That's fine – well done! Do you have any questions you want to raise about this lab?" It has been noted that, in teaching and learning as well as elsewhere, most people need to have approval from a fellow human being to feel good about something. (Race, 1995). The dilemma for us was that this was at the expense of students needing a substantial amount of help.

The Project

It was proposed that more staff time could be made available by providing a computer version of checkout and encouraging students who felt confident about the laboratory content to use this system. The project was intended to gather information on the value of checkout in learning, and the feasibility and desirability of offering it on a computer platform as well as in paper form.

[1] Ably and amiably programmed by Gordon Yau when CAL and Multimedia Consultant at Otago University, now Multimedia Developer at Melbourne University TeLaRS section.

Objectives

1. To clarify the perception of students on checkout as an aid to learning by gathering more structured information.

2. To ensure that students using the computerised version are not materially disadvantaged, particularly in the major assessment

3. To improve the availability of quality teaching staff to help students in the learning process.

4. To ensure that the gathering of information to improve teaching materials is not compromised.

Methodology

Selection of Student Groups. In deciding on the most appropriate way to go about this work, it appeared that the clearest indicators could be obtained by comparing a group of students who did not undertake checkout, with another group with similar backgrounds and aspirations who did. However, since student feedback to that date suggested that checkout is helpful to learning, this did not appear to be an ethical option, even if a group of students could have been released from the requirement for checkout.

The design involved randomly allocating volunteers to two groups, one undertaking the paper based assessment for the first four weeks of the semester, while the other group used the computer form. In the fifth week the groups reversed roles, and in the last four weeks the members of both groups were given free choice to use either the paper based assessment or the computer based one. Since volunteers were to be used, information was collected to ascertain if the distribution differed from the non-volunteer group in respect of gender, course preference (Health Science or not), ethnicity and frequency of attendance at laboratories. Sixty volunteers were obtained from six different laboratory streams.

The volunteers were asked to fill in a form each week and both they and non-volunteers were surveyed after the fourth and eight weeks, with small group interviews for the volunteers at the end of the course.

After permission was obtained from both the member of staff and the student, video footage with a timescale was taken of the staff-student interaction when processing the assessment.

Following completion of the initial study, students over the last two academic years have been given free choice as to which version of the assessment they would use.

Statistical Methods

Interpolated medians were used to summarise survey results rather than means owing to the discontinuous nature of the data, and to avoid distortions of the central measure that are known to occur with this type of data. (Webb, 1994). Lilliefors' modification of the Kolmogorov-Smirnov (Conover, 1971) was used to test distributions. The Mann-Whitney procedure was also used for Likert scales such as questionnaire responses (Mogey, 1999). Other procedures were standard, analyses being done using "Instat 2.01" by Graphpad Software, except for the Lilliefors' test, which was carried out using 'in-house' software developed by the author.

Results

The Volunteer Group. Testing data from 604 non-volunteers against the 60 volunteers disclosed no detectable differences in any of the parameters tested, that is gender, course preference (Health Science or not), ethnicity and frequency of attendance at laboratories.

Compliance

Of the original 60 volunteers, 43 (72%) completed all of their scheduled computer sessions over the eight week period. The total number of sessions completed out of a possible 240 was 195 (81%). Of the 17 who did not complete their four sessions, six missed through illness, two dropped out stating that they *"don't get on with computers"*, two were having some unspecified general difficulties, attending the lab sessions sporadically. A significant loss also occurred among the group B volunteers, five of whom signed on in week one, but seemed to have decided by week five - when they should have been starting the computers - to stay with the paper version. One was interviewed, and said she felt comfortable with the paper version and decided to stay with it. One other missed one session, stating that he had simply forgotten to go to the computers that week. The last of the 17 withdrew on the grounds that he knew a lot about the subject, and therefore found the questions ambiguous, feeling more able to explain this in the paper version. Therefore eight people had withdrawn, leaving a total of 52 volunteers. The number of completed questionnaires returned for analysis after five weeks was 42 (81%) and after nine weeks was 46 (88%).

1. Checkout as an aid to learning.
Since the paper began in 1994 we have asked students to fill in a general course questionnaire, which employs five-point Likert scales, asking if they find checkout worthwhile, and if they feel it has helped their learning. In the three years before this project began, we obtained the following results for interpolated medians:

	Checkout Worthwhile	Checkout Assisted Learning
1994	1.96	1.96
1995	1.84	1.78
1996	1.68	1.68

Table 1: Interpolated medians for checkout-related questions in a general course questionnaire

This is a very positive response, particularly as the trend was towards an improved perception by students. However, this project presented an opportunity to gather more detailed information, for which we used another five-point Likert scale questionnaire. The questions were largely developed in response to written comments made by students in the end of course evaluation forms in previous years. They were designed to explore the following points: would people prefer the checkout process to be replaced by a practical examination, how stressful do you find checkout, does stress reduce on familiarity, would the process would be as effective if no mark was given, and was everyone treated even-handedly by markers.
These questionnaires were offered to 281 non-volunteers in three of the laboratory streams in week 8, producing 278 (99%) responses.

	Scaling (percent responses)				
Question summary	**1**	**2**	**3**	**4**	**5**
Helps learning	53	39	5	2	1
Preferred to exam	73	16	6	1	3
Not stressful	9	25	42	21	4
Familiarity reduced stress	38	35	18	5	4
Mark unimportant	11	9	2	29	41
Even-handed treatment	58	26	10	5	1

Table 2: Non volunteer students responses (week 8 questionnaire).

These results confirm that students regard this formative assessment as helpful in learning, prefer it to a practical examination and prefer a mark to be associated with it, find it moderately but not unduly stressful, and that the stress reduces with familiarity. Most of the stress reduction probably occurred as they became familiar with their surroundings and gained confidence with the staff and general philosophy of the assessment.

While there was a general perception that students were treated even-handedly, 16% of respondents indicated they were less than happy with this aspect, which caused us to modify marker training.

Objective 2 - Does using the computer version disadvantage students in the major assessment?

This is clearly an important question. Both staff and students regard the formative assessment as helpful to learning, but there are inevitable differences between the two versions. The most obvious way of disclosing a disadvantage, (though one fraught with interpretation issues), would be to see if there is a difference between the marks obtained in the all-important major assessment between regular computer users and others. I have defined "regular" as students using computers more than 80% of the time, and compared their results to students who never used the computers.
Finding the answer to this apparently simple question is not without hazards.
First, analysis of the major assessment results indicates significant differences in performance related to gender, ethnicity and course preference, as detailed below.
Second, data from the project involving volunteers is suspect, since students did not have a free choice about which version to use each week.
Gender and Course Preference Differences. The frequency distribution for the major assessment marks in any year is not Gaussian, as shown by the Lilliefors test, and inspection by eye suggested strongly suggests the existence of a bimodal curve. Separating the marks into those aspiring to the Health Sciences and those not removed the bimodal appearance. The two resulting distributions showed no evidence of being non-Gaussian by the Lilliefor's test. Comparison of Health Science students versus others has consistently showed a substantial performance difference in every year the paper has been taught.

Preference	Males	Females	p
Health Science	129 (N = 185)	123 (N = 317)	0.005
Non Health Science	100 (N = 34)	104 (N = 95)	0.52
p	<0.0001	<0.0001	

Table 3: Gender and Course Preference Differences in the 1999 Major Assessment among European/Pakeha Students. Marks are out of 180.

Ethnicity Differences. We have also found statistically significant differences between ethnic groups, but the pattern has varied. For example, in 1998 Chinese Health Science students substantially out-performed European/Pakeha Health Science students, but in 1999 and 2000 the differences were negligible. Maori and Pacific Island Health Science students have always performed less well than European/Pakeha Health Science students.

Effect of CAA on Major Assessment Performance. In general I have made the following comparisons using students aiming at Health Sciences of the European/Pakeha and Chinese groups, taking account of gender. This avoids known differences between groups, while aiming at as large a group size as practicable to minimise the risk of type II error.

		Computer not used		Computer used>80%		
		Mean	N	Mean	N	p
European/Pakeha	Females	123	224	122	140	0.6693
	Males	125	92	125	83	0.9854
Chinese	Females	121	48	114	18	0.3689
	Males	122	25	132	17	0.19

Table 4: Major assessment mean marks comparing those who used the computer checkout system with those who did not. 1999 and 2000 combined.

At first glance the result for female Chinese students is worrying, even though not statistically significant. However, closer inspection of the data reveals that the lower mean is due to 3 of 9 students performing relatively poorly in 2000. In 1999, the mean for female Chinese computer users was 116.6, and for non-users, 117.7.

In summary, there is good evidence that students are not disadvantaged in the major assessment if they use the computer version of the assessment, either in the volunteer group, or when given free choice to use either version.

Performance in this very important summative assessment is only one aspect of being disadvantaged. Checkout provides a special opportunity for staff-student interaction directed towards learning in spite of the very large classes, and using computers largely negates this. Students were asked what they thought was the worst aspect about the computer version, 17 of 24 replies making reference to the reduced opportunity to talk through problems with demonstrators. Even in this group of 17, however, seven still rated the computer version as being more effective than the paper version in assisting learning, and only two rated it as less effective.

The responses of volunteers to the question “Do you feel that the computer version was as effective as the ‘paper’ version in assisting your learning” were:

	1	2	3	4	5	
Computer much more effective	1	13	16	6	2	Computer much less effective
Percentage	3	35	41	16	5	

Table 5: Student perception of learning effectiveness of computer version

Selection of the correct answer from a list, as used in the computer version, does not call upon the same skills as recalling the terms and constructing an answer, and I was sufficiently concerned about this to supply the volunteers with a booklet of word games to help bridge this gap. Only one student initially considered this to be an issue,

remarking that the worst aspect of the computer version was that "*No recall or spelling of terms is necessary*."

Objective 3. To Improve the Availability of teaching staff to help Students in the Learning Process

Students were asked to note down the lengths of time it took them to carry out the assessment, of waiting for marking, and of the marking itself. The computer programme also kept a record of the time spent on the computer by each student. Overall, including waiting for a computer or waiting for a marker, students using computers took a total of 7.24 minutes, while the paper version took them 8.92 minutes. Video analysis of 178 teacher-learner interactions revealed that an average of 1 minute 58 seconds was spent with each student, with a range of 11 seconds to eight minutes and four seconds. In a class of 96 people all using the paper version, staff would take a total of close to 32 minutes each, or 190 minutes altogether. If around 35% of the students opted for computers, about 66 minutes of staff time would be saved, enabling more interaction with students still completing the laboratory work. This proportion of students using the computer version was based on the fact that, when given free choice in the last four weeks, 50% of the volunteers had chosen to use the computer version. Over the last 2 years, computers have been used for 37% of these assessments.

Objective 4: To ensure that the gathering of information to improve teaching materials is not compromised

The computer programme was designed to gather data on which questions were wrongly answered and which wrong answers were chosen in each case. In the paper version, markers were asked to put a mark to the left of any question not answered correctly initially, and encouraged to report back any problems with question interpretation. The video footage was used to enumerate and analyse errors for type. These devices have been used to improve the structure of questions, and to inform our teaching of the subjects. Video analysis is the benchmark, and revealed 23% errors. Around 10% of these were trivial (e.g. "ECR" instead of "ESR"). The computer recorded 20% errors, but without the detail provided by the video. Returns from markers indicated 4.5% errors – undoubtedly under-reporting. Returns from the computer are therefore now used to detect problem areas, followed by discussions with markers to pinpoint the nature of the problem. The computer version has enhanced our ability to improve the teaching material.

This study was sufficiently encouraging for the Dean to fund 20 computers for use in the laboratory for checkout and CAL use, providing much better access for students over the last two years.

Conclusion

The introduction of CAA in the formative assessment has been beneficial in a number of ways to our teaching and learning aims. It has:

- freed up experienced teaching staff to spend more time with students completing laboratory work
- enhanced the feedback to staff for improving teaching
- markedly reduced the queuing that used to occur, so reducing student frustration
- provided demonstrably even-handed treatment of students.

Possible drawbacks are that it has:

- diminished the opportunity for developing teacher-learner relationships.
- altered the learning requirement from recall and answer construction to answer selection. (We are currently investigating a possible solution following Verloski's 'long menu' idea (Veloski, 1993), which has been adapted for computer use (Schuwirth, 1996).

Most students who use the computer version seem to do so only when they feel confident of the material. As one students put it, *"It's a good solution for someone who knows what they're doing."*

Acknowledgements

I am indebted to the student volunteers, who freely gave up some of the time on which there are such high demands, and to the staff who assisted in the project, particularly Gordon Yau (CAL and Multimedia Consultant, now Multimedia Developer at Melbourne University).
This work was funded in part by a Teaching Technology grant from the University of Otago.

References

Conover, W.J. (1971) *Practical non-parametric statistics.* John Wiley and Sons.

Lilliefors, H.W. (1967). *On the Kolmogorov-Smirnov test for normality with mean and variance unknown.* J. Am Statist. Ass. **62**: 399-402.

Mogey, N. (1999) http://www.icbl.hw.ac.uk/ltdi/cookbook/onfo . . likert_scale/index.html

Race, P. (1995). "The Open Learning Handbook." Kogan Page, London.

Rowntree, D. (1987). "Assessing Students: How shall we know them?" Kogan Page, London.

Schuwirth, L. W. T., van der Vleuten, C. P. M., Stoffers, H. E. J. H., and Peperkamp, A. G. W. (1996). Computerized long-menu questions as an alternative to open-ended questions in computerized assessment.*Medical Education* **30,** 50-55.

Veloski, J. J., Rabinowitz, H. K., and Robeson, M. R. (1993). A solution to the cueing effects of multiple choice questions: The Un-Q format. *Medical Education* **27,** 371-5.

Webb, G. (1994) *Making the Most of Appraisal* London: Kogan-Page.