Title	A computerised adaptive testing approach: What are the benefits to students?
Author(s)	Chew Lee Chin
Source	*ERA Conference, Singapore, 23-25 November 1998*
Organised by	Educational Research Association of Singapore (ERAS)

A Computerised Adaptive Testing Approach: What Are the Benefits to Students?

Chew Lee Chin
National Institute of Education
Nanyang Technological University

Introduction

Singapore schools have embraced computer technology for educational purposes on a large scale. Since the adoption of the IT master plan, numerous initiatives have focused on using information technology to enhance teaching and learning. Accompanying this is a growing interest in computer use for assessing student learning outcomes.

Over the last two decades, the computer has been applied to 'mechanise' repetitive assessment tasks such as the scoring and reporting of tests. Since then further advances in the technology have made it possible to have computerised adaptive testing (CAT). Computerised adaptive testing, where computers are used as devices for delivering a tailored test to each individual student, holds promise as a new and innovative strategy for testing and measuring student achievement.

Research studies conducted in the United States have alluded to several advantages of a computerised testing strategy over its paper counterpart. These studies reported the benefit of reduced testing time (English, Reckase & Patience, 1977; Glowacki, McFadden & Price, 1995; Oslen, 1990; Oslen, Maynes, Slawson & Ho, 1989; Wise & Plake, 1989). Other benefits included ease of data/information collection (Wise & Plake, 1989), immediate test scoring (Bugbee, 1996; Bugbee & Bernt, 1990; Mazzeo & Harvey, 1988), immediate test reporting (Glowacki et al., 1995; Jackson, 1988), improved test security and easy scheduling of test (Bugbee, 1996; Grist, Rudner & Wise, 1989; Wise & Plake, 1989).

Even greater assessment efficiency is possible with CAT because of several unique inherent features. Students are administered test items geared to their individual ability level. The test items are selected on the basis of their psychometric information, and computers are used in the whole testing process. The potential benefits are reduced test length (Smith, Fitzpatrick & Dodd, 1993), reduced testing time, and test adaptability. However, it is purported that the actual benefits derivable from this testing method vary according to the influence of several factors. This research area calls for more supportive empirical evidence from implementation studies on CAT.

This paper will provide an evaluation of the benefits to students that were derived from the use of computerised adaptive tests. The findings and its implications for educational testing and measurement in the Singapore context are then discussed.

Method

Data were collected via implementing a computerised adaptive testing programme in a Singapore secondary school. Computerised adaptive tests on biology were examined as they were actually administered to a sample of 113 secondary students. This was done by drawing upon several lines of evidence regarding the actual benefits to students.

Two mini-CATs on biology were implemented in this study (mini-CAT1 and mini-CAT2). The two adaptive tests were so constructed that each test was terminated on two predefined conditions. One was when a Bayesian posterior variance of 0.2 (or standard error=0.45) was reached, and the other was when a specified maximum number of test items had been administered. For mini-CAT1, a minimum of 12 test items and a maximum of 18 were specified for the testing. For mini-CAT2, a minimum number of 8 test items and a maximum of 12 were specified.

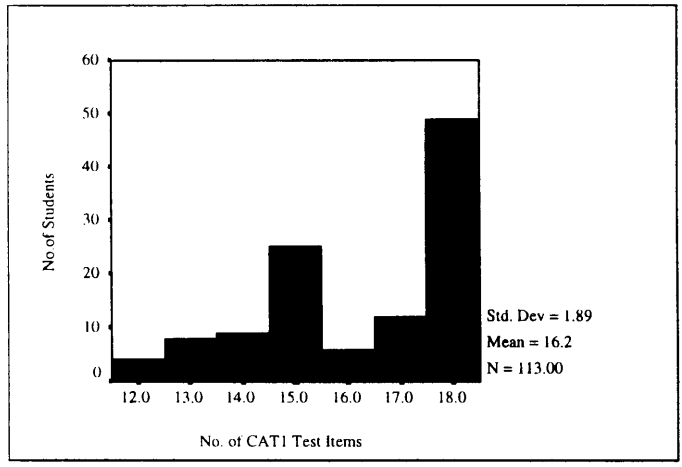
Benefits of reduced test length and time, and of test adaptability, were investigated by examining two aspects of each mini-CAT. These included the total number of test items administered to each student, and the number of test items answered correctly by each student. As the number of test items administered to each student might vary in the adaptive testing, the percentage of correctly answered test items of the total number administered was computed to account for this variation.

Two other benefits that were accrued from a computerised testing strategy, namely, immediate test reporting and the provision of performance diagnosis feedback were also evaluated.

Results

Results show that for mini-CAT1 a mean total number of 16.2 (sd=1.9) test items were administered to students. Based on the maximum number of specified, the average length of the test administered to students was reduced by about 10%. Figure 1 shows the distribution of the number of CAT1 test items administered. Students were administered between 12 to 18 test items. About 43% of them were administered the maximum number of 18 test items.

Figure 1
Histogram of the Number of Mini-CAT1 Test Items Administered to Students



The adaptive test could also be terminated when the posterior variance criterion had been satisfied. A histogram plot of the final standard error values obtained for mini-CAT1 is presented in Figure 2. For 61% of the students, the testing was terminated on this criterion. However, for the rest of the students, the specified standard error of 0.45 was not reached. Their tests were terminated when the fixed maximum number of items had been administered.

The results show that a mean number of 11.8 (sd=2.9) test items were answered correctly by students. Figure 3 shows the distribution. The number range from 6 to 17. Based on the total number of test items administered, the students obtained a mean of 71.7% (sd=11.0) correct responses.

Figure 2
Histogram of the Final Standard Error Values of Mini-CAT1

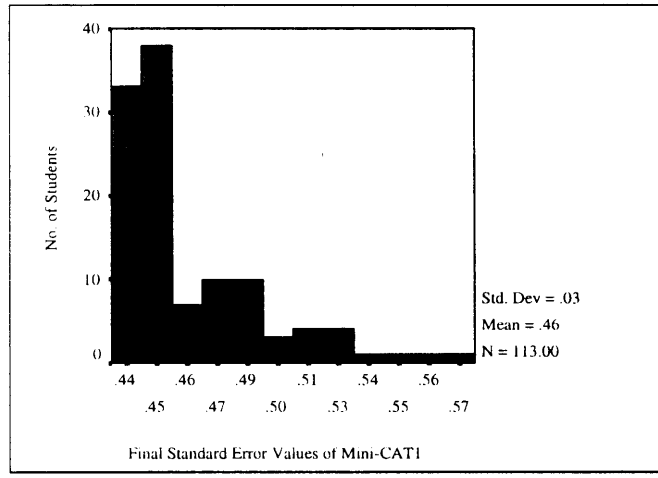
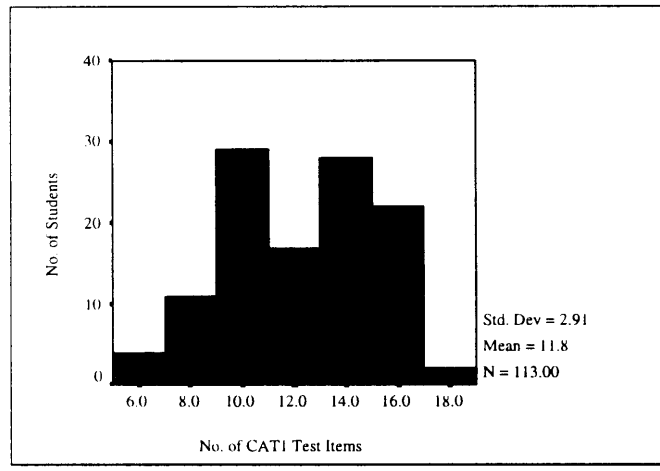


Figure 3
Histogram of the Number of Mini-CAT1 Test Items Answered Correctly by Students



Results show that all students were administered the maximum number of test items for mini-CAT2. That is, no reduction of test length was achieved. A histogram plot of the final standard error values for mini-CAT2 is displayed in Figure 4. For majority of students, the final standard error value of 0.45 and below was not satisfied, and their tests were terminated only when the maximum number of test items was administered. For the 1.8% of the students, their tests satisfied the standard error termination criterion, but this was achieved together with the specified maximum number of test items.

Figure 4
Histogram of the Final Standard Error Values of Mini-CAT2

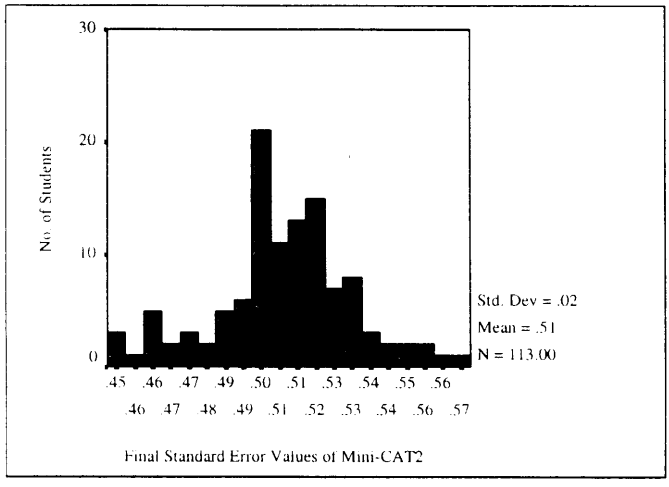
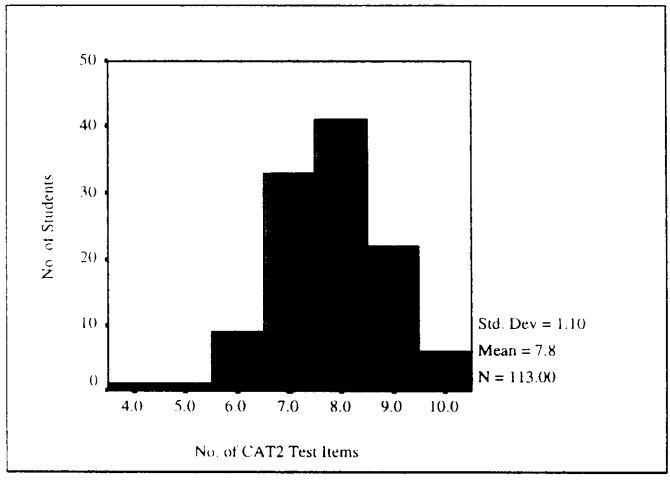


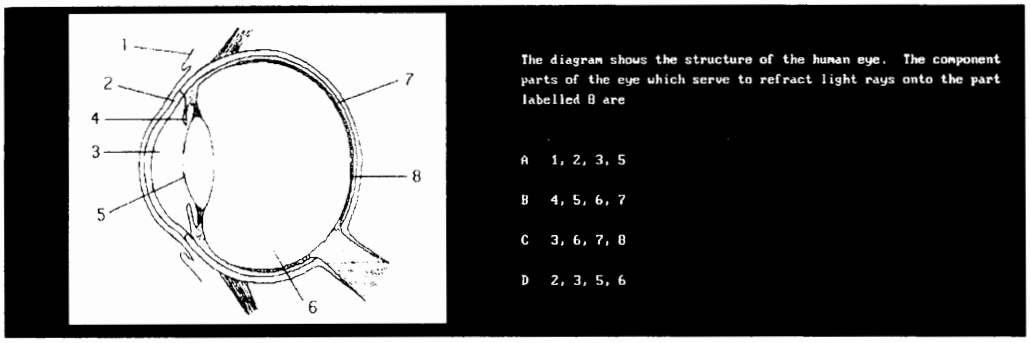
Figure 5
Histogram of the Number of Mini-CAT2 Test Items Answered Correctly by Students



The results show that a mean number of 7.8 (sd=1.1) test items were answered correctly by students in mini-CAT2. Figure 5 shows the distribution. It ranged from 4 to 10 items, but the majority of students correctly answered between 7 to 9 of the administered test items. Out of the total number of test items administered, they answered correctly about 64.9%.

In terms of testing time, results show that the study sample took an average of 17.1 (sd=4.5) minutes to complete the two mini-CATs. In a similar but paper-administered 30-item conventional test, a standard testing time of 35 minutes would have been specified. Thus, with this adaptive testing, a time saving of about 50% was achieved.

Figure 6
A Test Item on Biology and its Specified Performance Diagnosis



A performance diagnosis item was linked to its corresponding test item. Figure 6 shows a test item on biology and its specified performance diagnosis item. This enabled it to be presented to the examinee immediately following his or her response to the test item. The performance diagnosis strategy would give immediate feedback to students about their success or failure on each test item attempted during the adaptive testing. Typically, the computer screen presented the correct of the best answer to an item, together with an explanation of it. An explanation of the wrongness or inappropriateness of each of the other answer options was also given. Half of the study sample was treated with CAT with performance diagnosis, and 97% of them found the immediate diagnostic feedback useful.

Figure 7
A Student's Test Report

Test Report For Student 4C116

Test data for 4C116 on DEFAULT 10/03/95
ID: 4C116

Test Information for Mini-CAT1 of the Biology Test

Item #	Item ID	Resp	Key	Correct	Ability	Std. Error
1	BIO041	2	1		-0.713	0.801
2	BIO083	1	4		-1.199	0.697
3	BIO103	4	3		-1.646	0.631
4	BIO047	2	4		-1.856	0.590
5	BIO091	3	3	X	-1.739	0.558
6	BIO087	1	1	X	-1.622	0.535
7	BIO069	1	1	X	-1.488	0.518
8	BIO101	3	3	X	-1.406	0.501
9	BIO028	2	3		-1.594	0.486
10	BIO053	1	1	X	-1.513	0.473
11	BIO099	4	3		-1.628	0.460
12	BIO079	3	4		-1.770	0.456
13	BIO040	3	3	X	-1.717	0.443

13 items were administered.
Number correct 6
Proportion correct 0.462
The final ModalBayesian posterior mode was: -1.717
The final ModalBayesian posterior variance was: 0.196

Test Information for Mini-CAT2 of the Biology Test

Item #	Item ID	Resp	Key	Correct	Ability	Std. Error
1	BIO132	4	4	X	0.181	0.925
2	BIO212	3	4		-0.239	0.793
3	BIO167	4	4	X	0.012	0.733
4	BIO219	4	4	X	0.197	0.693
5	BIO163	3	3	X	0.389	0.674
6	BIO168	2	2	X	0.561	0.655
7	BIO211	2	3		0.390	0.603
8	BIO209	4	4	X	0.511	0.582
9	BIO218	2	2	X	0.576	0.564
10	BIO190	1	1	X	0.664	0.551
11	BIO146	3	3	X	0.777	0.548
12	BIO125	4	1		0.486	0.532

12 items were administered.
Number correct 9
Proportion correct 0.750
The final ModalBayesian posterior mode was: 0.486
The final ModalBayesian posterior variance was: 0.283

858.051 seconds were taken for the whole test.

Figure 7 shows a student's test report. During the testing, there was tracking of items presented to each student, and also immediate scoring of the test items. Data were collected or recorded by the computer. These included students' responses to each item, the number of items presented, estimates of their ability on the basis of their responses to each item, the number of items presented, and the total testing time. The automated testing system provided each student with an immediate test report.

Discussion and Conclusions

Several interesting findings about the two mini-CATs used in this study were made. Shorter tests were possible with the first mini-CAT. The possibility of an average 10% reduction in test length was observed with this mini-CAT. On the other hand, no reduction in test length was observed to be possible for the second mini-CAT. Two reasons may account for these 'puzzling' results. One is the size of the item pool. An item pool of 162 test items, which was further sub-divided into two smaller pools for the two mini-CATs, might have been inadequate. Another reason may be the unavailability of test items tailored to the 'right' psychometric information at a particular ability level.

In adaptive testing, the test is so tailored to an individual's ability that he/she is expected to obtain 50% of the test items correct. It is noted that at mini-CAT1, students correctly answered between 69% to 79% of test items, and that at mini-CAT2, they answered correctly between 60% to 70% of the items. These results indicate satisfactory adaptability of the tests to individual students' ability. However, better results, and hence better adaptability, was not observed probably because of limitations of the item pool.

The students took an average of 17.1 minutes to complete the computerised adaptive testing. Compared to conventional paper version of the test, which allocated a standard test-time, a time saving of 50% was achieved.

In sum, the adoption of a CAT strategy in school-based testing can benefit students. When students are administered tailored tests that are geared to their performance level, they do not need to take unnecessary too 'easy' or 'difficult' test items. Instead, the computer selects and presents the most suitable test items for individual students, through on-going estimation of their ability during the testing. Each student is thus administered a unique set of test items. Translated into practical terms, it means that the testing is individualised, and this results in the additional benefits of reduced test length and testing time. In the Singapore context, where educational decisions on students' streaming into different courses of study and their promotion therein, are made on the basis of test scores from school-based testing, students implicitly gain from this improved testing efficiency.

Two other benefits, namely, immediate test reporting and the provision of performance diagnosis feedback to students, were accrued from the computerised testing strategy used in this study. A performance diagnostic function in the CAT system can provide immediate feedback and remedial tutorial as and when each question is answered. This allows the student to immediately evaluate his or her own understanding. In other words, the testing takes on, simultaneously, an instructional role. By comparison, a teacher-administered feedback is unsatisfactory as it is typically delivered long after a testing session, if at all. A computer-administered feedback and diagnosis of test performance can thus contribute towards good formative assessment and raise standards of learning amongst students.

It is also advantageous to have a computerised testing system that allows students to obtain their test results and report immediately after the testing. By contrast, under the current testing practice, teachers take some time to score the tests, resulting in a lapse of time before students obtain their test results. From the students' standpoint, the provision of

immediate test results is more effective feedback, and an important motivating factor in learning.

References

Bugbee, A.C. J. (1996). The equivalence of paper-and-pencil and computer-based testing. *Journal of Research on Computing in Education, 28(3)*, 282-299.

Bugbee, A.C. J. & Bernt, F.M. (1990). Testing by computer: Findings in six years of use 1982-1988. *Journal of Research on Computing in Education, 23(1)*, 87-100.

English, R.A., Reckase, M.D. & Patience, W.M. (1977). Application of tailored testing to achievement measurement. *Behaviour Research Methods and Instrumentation, 9(2)*, 158-161.

Glowacki, M.L., McFadden, A.C. & Price, B.J. (1995). Developing computerised tests for classroom teachers: A pilot study. *Paper presented at the annual meeting of the Mid-South Educational Research Association, November, Biloxi, MS.* (ERIC ED391471).

Grist, S., Rudner, L., & Wise, L. (1989). Computerised adaptive tests. ERIC Digest No. 107. *ERIC Clearinghouse on Tests, Measurement, and Evaluation, Washington, DC*: American Institute for Research (ERIC ED 315325).

Jackson, B. (1988). A comparison between computer-based and traditional assessment tests, and their effects on pupil learning and scoring. *School Science Review, 69(249)*, 809-815.

Mazzeo, J. & Harvey, A.L. (1988). *The equivalence of scores from automated and conventional educational and psychological tests: A review of the literature.* College Board Report No. 88-9, ETS-RR No. 88-21. Princeton, NJ: Educational Testing Service.

Oslen, J.B. (1990). Applying computerised adaptive testing in schools. *Measurement and Evaluation in Counselling and Development, 23(1)*, 31-38.

Oslen, J.B., Maynes, D.D., Slawson, D. & Ho, K. (1989). Comparisons of paper-administered, computer-administered and computerised adaptive achievement tests. *Journal of Educational Computing Research, 5(3)*, 311-326.

Smith, N.J., Fitzpatrick, S.J. & Dodd, B.G. (1993). *Results of the administration of the computerised grammar, spelling, and punctuation test to College of Communication students on July 7, 1992.* Texas University, Austin: Measurement and Evaluation Centre. (ERIC ED375150).

Wise, S.L. & Plake, B.S. (1989). Research on the effects of administering tests via computers. *Educational Measurement: Issues and Practice, 8(3)*, 5-10.