
Christina Tham-Ng
Ngee Ann Polytechnic

Dr. Peter Lam
Nanyang Technological University

ABSTRACT

This study is an attempt to estimate the change in proficiency levels of students who went through an enrichment program in Additional Mathematics. The test items used were classified by hierarchies using SOLO taxonomy as a guide. The estimation was done with a four-group pretest-treatment-posttest counter-balance design. The tests were administered in two modes, the paper-and-pencil mode and the computer mode. The enrichment program was shown to be definitely effective where students progressed not only content-wise but also in the ability to solve higher level problems. Test modes seemed to have an effect on students' performances favoring the paper-and-pencil mode. However, experience with test on computer possibly took care of the computer mode disadvantage.

Paper presented at the Joint Conference of the Educational Research

Introduction
Like most learning institutions, students of very different abilities join Ngee Ann Polytechnic every year. One can easily find a student who had scored distinctions in both Elementary and Additional Mathematics in their 'O' level examinations sitting next to another who had only done Elementary Mathematics and scored a P6; both attending the same Mathematics lectures. This poses both teaching and learning problems. An enrichment program is being offered to all incoming students with no Additional Mathematics background to bridge this gap. The enrichment program is a sixty-hour lecture-cum-tutorial of which eighty percent attendance is required. An Additional Mathematics textbook written by Talbot J.F. and Heng H.H. is used. Students are taught and strongly encouraged to use Computer Aided Instructions (CAI) Courseware that was designed for them. They are required to take a short-answer pre-instructional test and pass a post-instructional test to be awarded merit certificates of achievements.

Rationale
Much time and money are put into the enrichment course. The end of the enrichment course is not the end of teaching and learning. If weak students and their learning patterns can be identified, it allows for more effective remedial intervention. The need of a fair and effective assessment can never be over-emphasized. Aiming to tap at computer technology is no longer taking a futuristic view. Computer technology is here and now. On top of that, the growth and impact of Computer Adaptive Testing is a direct result of advances in psychometrics, math, cognitive learning theory, educational measurement, human engineering and science technology. This study is an attempt to:
determine the effectiveness of the enrichment program;
determine if there is mode effect on test performance and its direction;
establish computer testing as a viable testing tool in the polytechnic.

Design
To achieve these objectives, students' proficiency changes were measured using two tests. One (pre-test) administered during the registration period of the enrichment program and the other (post-test)
administered towards the end of the program. Students were randomly divided into four groups. Group PP took both tests in paper-and-pencil mode. Group PC took a paper-and-pencil pre-test but a computer-administered post-test. Group CC took both tests on computer mode; and Group CP took a computer-administered pre-test but a paper-and-pencil post-test. The test items were classified using SOLO Taxonomy's hierarchies of difficulty. The responses of the examinees were analyzed using both classical theory and item response theory (IRT). MICROCAT (version 3.0) Item Analysis softwares were used.

Classical/Item Response Theories
The classical number correct score does not take into account the difficulty levels of the items when estimating proficiency levels of students. A student scoring 10 for 10 easy items does not imply equal proficiency to another scoring 10 for 10 difficult items. Logically, getting a more difficult item correct should be given a 'higher' weighting when estimating proficiency. The person ability and item difficulty locating on the same continuum of measure is the essence of IRT. In IRT, the RASCH model is a one parameter model, estimating abilities of examinees based on item difficulty. Other IRT models like the two and three parameter models take into consideration item discrimination and the guessing or chance factor as well.

SOLO Taxonomy

SOLO stands for Structure Of Learned Outcomes. Unlike Bloom's Taxonomy which is what psychological functions are required to answer the item, Biggs and Collis believe in "natural" stages in the growth of learning similar but not identical to Piaget's cognitive developmental stages. The stages and their meanings are:

i. Prestructural: Refusal or inability to become engaged in the problem

ii. Unistructural: Takes note of only one direct relationship or fact

iii. Multistructural: Ability to handle multiple operations with separate pieces of information contained in the stem

iv. Relational: Relates elements within the immediately available concrete system and forms an answer on this basis

v. Extended Abstract: Acceptance of lack of closure, a use of the reciprocal operation, and the ability to work with multiple interacting and abstract systems

Superitems are used to reflect SOLO taxonomy. A stem provides informations and questions that require the student to reason at a different level in order to produce a correct responses using information(s) from that common stem. These questions are called testlets.
An example of a superitem in SOLO taxonomy is shown below:

Unistructural Item : If 14 is put out, what number was put in?
Multistructural Item : If we put in 5, what number will the machine put out?
Relational Item : If we got out a 41, what number was put in?

Related Studies
Olson (1986) compared scores from 3 modes of testing (N=575) namely paper-administration, computer-administration and computer adaptive test (CAT). His study showed no significant difference in measurement precision. However, it showed that CAT required only one fourth the time required by a paper test and is the most precise in ability estimates.

Mazzeo (1991) compared two groups (N=94). One group was given a linear computer version test before a paper version. The other group was given a paper version before computer version test. The students were asked to complete the computer familiarity questionnaire. Mazzeo's results showed no relationship between the computer-version scores and computer familiarity. There was evidence of mode-of-administration effect. However, the mode effect did not affect the construct that was being measured. He found high within group correlation. Mazzeo cited advantages of computer tests as increased test security and reduced costs of production, administration and scoring of tests. Another advantage is the immediate feedback of results.

Perkins (1993) had two groups who did a paper pre-test (N=83). One group took the computer anxiety measure, did the post-test on computer, and took computer anxiety measure again. The other group did a paper post-test. Perkins found no significant difference in scores between computer administered and paper-and-pencil tests. However, computer owners did better and had less anxiety in both pre- & post-tests. He equated computer ownership with computer experience but said that exposure to CAI is not enough computer experience.

Item Calibration
The enrichment program covers nine Additional Math topics. In test construction, writing only two superitems for three levels per superitem would have totaled a number of 54 questions. Two superitems per topic is not enough content coverage. However, to increase by one superitem per topic meant an additional 27 questions. SOLO taxonomy's superitems therefore cannot be adopted here. A modification of SOLO taxonomic classification was used instead. For each topic, 2 'U', 2 'M' and 2 'R' items were written but each item stood on its own. There was no superitem.
Below is an example of a set of 'U-M-R' items on Trigonometry that was used in this study:

'U' item: If \( \cos(\theta - 30) = 0.5 \) for \( 0 \leq \theta \leq 360 \), then \( \theta = ? \)
(1) 240 (2) 60 (3) 270 (4) 330

'M' item: If \( \sin \theta = -0.3814 \) and \( \tan \theta = 0 \) for \( 0 \leq \theta \leq 360 \), then \( \theta = ? \)
(1) 247.6 (2) 67.58 (3) 112.4 (4) 292.4

'R' item: Which of the following is true?
(1) \( \sec x - \cos x = -\sin x \tan x \)
(2) \( \sin x - \csc x = \csc x \cot x \)
(3) \( \sec x - \cos x = \sin x \tan x \)
(4) \( \sin x - \csc x = \cos x \cot x \)

Judges
Two judges, one senior lecturer and one expatriate lecturer, were appointed to comment on the classifications of these fifty-four items. They both had lots of experiences in teaching and designing tests, including MCQ's. The senior lecturer had been the enrichment program coordinator for several years. The 54 items were arranged with all 'U' items appearing before the 'M' items. The 'R' items appear last.

Pilot Test
Sixty-one existing first year students from the Quality Assurance Engineering diploma course took the pilot test. They were given up to one and a quarter hour to complete the test. Microcat softwares for item analysis, using both classical theory and IRT, determined for each item, (i) the proportion of correct answers, (ii) the proportion endorsing each alternative distracter, (iii) the discrimination indices (iv) the difficulty values/ranking, (v) the Chi-square values. The proportion correct confirmed the general difficulty of the three 'SOLO' levels. No alternative distracter were either endorsed by no one or by every one. Nine items were weeded because of they either negatively discriminating, misfitting or having incorrect response correlating with total score more than correct response does. One 'R' item was revised to make it more difficult.

The trend evident in the graph above gave assurance of well classified test items even before pilot testing.

Instrument
The final 45 items were also equally distributed in each 'SOLO' level. The order of appearance of the items remained the same except for twelve items that were re-classified. This became the instrument for a one hour pre-test. The same questions were used in the post-test. However, the items were then classified by topic with the 'U-M-R' order still intact. This rearrangement was to prevent response set. This formed the hour long post-test instrument.

Control Group
Two weeks after the end of the enrichment program, 43 examinees who did not have the enrichment treatment took its post-test in the paper-and-pencil mode.

Analysis of Results and Discussions

Pre-test and post-test scores of all four groups by 'SOLO' levels

<table>
<thead>
<tr>
<th>Mode</th>
<th>Pre/Post Sc</th>
<th>'U'</th>
<th>'M'</th>
<th>'R'</th>
<th>Total</th>
<th>% Inc</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper-computer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-</td>
<td>2.51</td>
<td>2.09</td>
<td>2.20</td>
<td>6.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-</td>
<td>5.04</td>
<td>3.64</td>
<td>3.31</td>
<td>11.98</td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference</td>
<td>+2.53</td>
<td>+1.55</td>
<td>+1.11</td>
<td>+5.19</td>
<td>76.2</td>
<td></td>
</tr>
<tr>
<td>paper-paper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-</td>
<td>3.46</td>
<td>1.95</td>
<td>1.64</td>
<td>7.05</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-</td>
<td>6.55</td>
<td>4.37</td>
<td>3.20</td>
<td>14.12</td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference</td>
<td>+3.09</td>
<td>+2.41</td>
<td>+1.57</td>
<td>+7.07</td>
<td>100.3</td>
<td></td>
</tr>
<tr>
<td>computer-computer</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-</td>
<td>2.23</td>
<td>2.16</td>
<td>2.09</td>
<td>6.48</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-</td>
<td>4.85</td>
<td>3.84</td>
<td>3.37</td>
<td>12.07</td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference</td>
<td>+2.63</td>
<td>+1.68</td>
<td>+1.28</td>
<td>+5.59</td>
<td>86.3</td>
<td></td>
</tr>
<tr>
<td>computer-paper</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pre-</td>
<td>3.75</td>
<td>2.14</td>
<td>1.90</td>
<td>7.78</td>
<td></td>
<td></td>
</tr>
<tr>
<td>post-</td>
<td>6.98</td>
<td>4.29</td>
<td>3.65</td>
<td>14.92</td>
<td></td>
<td></td>
</tr>
<tr>
<td>difference</td>
<td>+3.24</td>
<td>+2.16</td>
<td>+1.75</td>
<td>+7.14</td>
<td>91.8</td>
<td></td>
</tr>
</tbody>
</table>

Every group scored the highest with 'U' items and the least with 'R' items except at one place. This is another evidence of good item calibration.

Also, every group showed greatest improvement at the 'U' level and least improvement at the 'R' level. This is yet another strong evidence of good 'SOLO' level classifications. IRT worked well with item calibration.

Change in Scores by 'SOLO' Levels

<table>
<thead>
<tr>
<th>Mode</th>
<th>N</th>
<th>'U'</th>
<th>'M'</th>
<th>'R'</th>
<th>Total</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>paper/computer</td>
<td>55</td>
<td>+2.53</td>
<td>+1.55</td>
<td>+1.11</td>
<td>+5.19</td>
<td>76</td>
</tr>
<tr>
<td>paper/paper</td>
<td>85</td>
<td>+3.09</td>
<td>+2.41</td>
<td>+1.57</td>
<td>+7.07</td>
<td>100</td>
</tr>
<tr>
<td>computer/computer</td>
<td>75</td>
<td>+2.63</td>
<td>+1.68</td>
<td>+1.28</td>
<td>+5.59</td>
<td>86</td>
</tr>
<tr>
<td>computer/paper</td>
<td>51</td>
<td>+3.24</td>
<td>+2.16</td>
<td>+1.75</td>
<td>+7.14</td>
<td>92</td>
</tr>
</tbody>
</table>

Analysis of the 'SOLO' change in the four groups show greatest percent improvement in group PP (100%) and least improvement in group PC (76%). This was an impressive growth and the effectiveness of the enrichment course cannot be denied.

The success of the enrichment course may be attributed to the following factors:

a. CAI, textbook, instructional examples and tests were of similar
design.
b. Students volunteered themselves into the program and were self-motivated.
c. Few distractions existed since the semester had not started.
d. New enthusiasm in students, first times attending classes in the polytechnic.
e. Students were then unsure of the standard and demand of the polytechnic.
f. Flexibility of switching between lectures and tutorials according to students' pace.
g. CAI was readily available with the computer laboratory opening long hours.
h. Students' sense of handicap make them see the program as a chance to bridge the gap.
i. CAI was user friendly, easy to teach and easy to learn.

Students improvement in all 3 levels showed not only that the students can handle more topics but also solve more difficult questions of the same topics. If only 'U' level showed improvement, there was only low level learning. On each individual, 'SOLO' level changes can be studied to get a better picture of the individual's learning pattern. Such analysis is more telling of the individual's learning pattern than a pass/fail decision using cut-off point. Teachers may group students according to their learning styles. For example, a student who scored six 'U' items correct learns quite differently from another who scored but two 'U', two 'M' and two 'R' items correct. This is a fairer and more effective assessment.

(III) Pre-Tests Mean Scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper-Paper (PP)</td>
<td>85</td>
<td>7.04</td>
<td>2.05</td>
</tr>
<tr>
<td>Paper-Computer (PC55)</td>
<td>65</td>
<td>6.80</td>
<td>2.01</td>
</tr>
<tr>
<td>Computer-Paper (CP51)</td>
<td>71</td>
<td>7.78</td>
<td>2.00</td>
</tr>
<tr>
<td>Compr-Computer(CC75)</td>
<td>75</td>
<td>6.48</td>
<td>2.06</td>
</tr>
</tbody>
</table>

In ANOVA, (prob(F)=0.54 , Fcritical=2.64 , indicating insufficient evidence to say that there was any significant difference among the mean scores.

We can therefore say that group PP was comparable with group PC and group CC was comparable with group CP. Whether all four groups were comparable depended on whether there was a mode of administration effect. This cannot be concluded at this point of time. The analysis of the post-test mean scores would therefore be based on this too.
(IV) Post-Tests Mean Scores

<table>
<thead>
<tr>
<th>Group</th>
<th>N</th>
<th>Mean</th>
<th>Standard Deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Paper-Paper (PP)</td>
<td>85</td>
<td>14.12</td>
<td>2.04</td>
</tr>
<tr>
<td>Paper-Computer (PC)</td>
<td>55</td>
<td>11.98</td>
<td>2.44</td>
</tr>
<tr>
<td>Computer-Paper (CP)</td>
<td>51</td>
<td>14.92</td>
<td>2.17</td>
</tr>
<tr>
<td>Comp-Computer (CC)</td>
<td>75</td>
<td>12.07</td>
<td>2.54</td>
</tr>
<tr>
<td>Control (Ctrl)</td>
<td>43</td>
<td>7.76</td>
<td>3.47</td>
</tr>
</tbody>
</table>

In ANOVA (prob(F)=0.01, Fcritical = 2.64), indicating a significant difference in the mean scores.

Groups PP and CP performed significantly better than the groups PC and CC. It seems reasonable to suspect there was mode effect favoring paper tests. Looking at the growth in mean scores, group PP's growth of 100% was significantly higher than group PC's growth of 76%. This was indeed evident of mode effect favoring paper-and-pencil test. Since group PC was comparable in ability it was therefore evidently disadvantaged in the post-test by the computer test mode. There was a difference in growth of 24%.

Group CC's growth of 86% compared to group CP's growth of 92% confirmed that there was mode effect and favoring paper-and-pencil tests. This time the difference was only 6%. This could be due to group CC's computer experience in the pre-test. Group CC was therefore not so much disadvantaged by the computer mode at its post-test.

It is difficult to compare CP (92%) with PP (100%) because we do not know if the groups were really comparable. The same can be said of CC (86%) and PC (76%) but because there was evidence of interaction we may be able to draw an inference. If mode effect did not exist, the groups were comparable groups, interaction need not be explained. If mode effect existed, the group CC could have done better at pre-test if given a paper test instead. That explains the interaction. Also, although the post-test mode was the same for both groups, group CC had the computer experience in the pre-test that group PC did not have. That explains the interaction too. There was therefore evidence of mode effect favoring tests administered in paper-and-pencils well as evidence of mode effect being removed by computer experience.

Conclusions
From the above analysis, there are evidences to show:

a. That the enrichment program can effectively increase students' proficiency levels in solving Additional Mathematics problems;
b. That test mode does affect performances favoring paper-and-pencil
mode to computer mode.

c. That with every student in the polytechnic becoming computer literate, the computer becomes a viable tool for testings in the polytechnics.

Description of the CAI
The CAI package was jointly developed by Longman Publishing and Ngee Ann Polytechnic. Visual Basic was used in the item-writing. The items were grouped by topic. Within each question, there are thinking steps (guided steps) which are accompanied by self-checks. Hints can be obtained upon requests. Sometimes, more than one attempt is allowed. When the student gives a wrong answer, they'll be given further hints if a second attempt is allowed; otherwise, the right answers complete with explanations are provided. If students answer correctly, they'll be applauded with statements of praise and encouragement, and solutions will also be provided to confirm their understanding. All items were taken from the Longman textbook, and hence it provides additional drill supplementing classroom teaching.

Limitations
Many constraints were encountered:
(1) Test time. The pretest and posttest of the enrichment course are restricted to only one hour each. Therefore the number of items that will eventually be included into these tests must be reasonable. However, content topics were many and there were three levels of classifications to be included.

(2) Sample. Size of samples were considerably reduce at the post-test. The pilot test sample and the control group sample were not big either. Also, students did not have equal computer experience.

(3) Test Administrators. There were more than one administrator involved in the pencil-and-paper post-test. This created a somewhat different test conditions among the examinees.

(4) Judges. Judges were not SOLO taxonomy experts although they are very experienced in test item constructions.

(5) Test Items Since the number of 'U', 'M', and 'R' items had to be the same, it imposes another restriction to the selection of the final set of 45 items for the pre- and post-tests. No superitem was used.

Quotes for Thot
Glaser (1981) outlined the social-educational demands that shape the future of educational assessments:

* the shift from a selective educational system to one designed to
help individuals succeed in educational programs (zero-reject system)
* the requirement for improved levels of literacy and problem-solving
ability in a variety of knowledge and skill domains (minimum competency
and mastery certification)
* the need to understand individual differences in the process of
measurement so that abilities can be improved to facilitate further
learning

"A description of what is learned is more important than a summary of
how much is learned."
- Clarke (1982)

"We believe that the public, and especially educators, parents, and
children, need fair and effective assessment processes that can be used
for diagnosing and prescribing for the needs of individual children.
We also believe that the use of fair, effective assessment practices is
one way of being held accountable for providing quality education for
all students."
- Statement drafted by 25 national organizations including the U.S.
Office of Education during a November 1975 conference on testing,
sponsored by the National Association of Elementary School Principals
and the North Dakota Study Group on Evaluation.

References
Biggs, J.B., Collis K.F. Evaluating The Quality Of Learning. The SOLO
Taxonomy (Structure of
Assessing Mathematical
Problem-Solving Ability. Journal for Research in Mathematics Education,
Vol.17, No.3,
206-221.
Crocker, L., Algina, J. Introduction to Classical and Modern Test
cognitive psychology and psychometrics. American Psychologist, 36(9),
923-936.
IBM. User's Manual for the XCALIBRE, Marginal Maximum-Likelihood
Estimation Program, version 1, USA: Assessment Systems Corporation
Psychological Assessment, Chapter 7, 123-147.
Lam, L.P., Foong, Y.Y. Application of IRT in Estimating SOLO
Proficiency Levels Using Hierarchical Items in Testlets. Unpublished
Article, 1996, NIE, Singapore.
Mao, S.L. Search for Indicative Items in the Developmental Study of
Test Instruments.
Proceedings of the National Science Council, ROC(D), 1991, v1, 1,
41-54.
Mazzee, J., et.al. Comparability of Computer and Paper-and-Pencil
Scores for Two CLEP

By: Tham-Ng Christina . . Dr. Lam Peter