The AGEM Pilot
at
Maria Niketan School, Bangalore, India

Samar Singh, PhD
Email: samar@agem.in
www.agem.in

November, 2009 to March, 2010

1 Introduction

1.1 Objectives of the Pilot at Maria Niketan School

Our objectives in conducting this program were to:

- establish the robustness of AGEM hardware, software and processes in a school environment
- acquire performance data and determine the efficacy of our systems in a specific discipline i.e. maths
- determine on a prima facie basis, how the system impacts other aspects of the classroom e.g. classroom management.
- gather research data to answer other questions such as the differences in the nature of responses from students with different attributes, e.g. gender, age etc.

1.2 The AGEM system

1.2.1 Characteristics

- Feedback systems in the form of wireless keyboards - one per student - to allow parallel input by individual students to pre-designed questions
- Topic Plans comprising video, audio, instructional content, and embedded questions representing small increments of learning.
- Immediate, short term and long term feedback of performance based on analysis of questions embedded within the topic plan.
- Deficiencies in response of individual students were never made public. All representations displayed were of class response as a group
- Supplementary mechanisms such as activities, projects and reflection sessions provided support to the Topic Plans

1.2.2 Principal design objectives

- Serving large classes typically of 20 to 100 students.
- Providing for a large student base e.g. dozens of classes in hundreds of schools
- Making differentiated instruction practical
- Serving the need for scientific feedback of performance for:
  - design of Topic Plans
  - school management
  - reporting to parents.

1.3 The Maria Niketan School Pilot

The experimental group consisted of 22 students (13 male and 9 female) who had been assigned to an extra-curricular remedial maths program which we took over. The control group constituted the rest of the class which was not in the remedial maths program. This control group comprised an additional 41 students who took the post test. There were an additonal 13 students who did not attend to take the post test.

The test questions for the pre-test were a set of 113 questions selected from the Number Worlds website. Regrettably the control group did not take this test. The post test undertaken by both groups was a 65 question subset of the 113 questions comprising the pre-test.

The instructional component comprised 4 Topic Plans called TP2, TP3, TP4 and TP5 which extended across several classroom sessions. Each Topic Plan contained between 30 and 51 embedded questions which students answered in parallel using the wireless feedback devices. Hence, it was possible to obtain scores for each Topic plan.

The program with the 22 students of the Class 6 remedial maths group serving as the experimental group commenced on 23 Nov 2009. The time spent was as follows:

Time in hours	Topic
3.75	Pre-test with experimental group
18.75	Topic Plans 1 to 5
11.25	Other activities
2.1	Post test for full group
35.85	Total Program

Table 1.1: Nature of Program and Time spent

1.3.1 The Program delivered

An introductory topic plan to show the relevance of maths was followed by 4 topic plans that dealt with:

1. “The concept of number” referred to as TP2 or LP2
2. “Percentages” referred to as TP3 or LP3
3. “Decimals” referred to as TP4 or LP4
4. “Fractions” referred to as TP5 or LP5

1.3.2 Goals of the Topic Plans

In all the Topic Plans, barring TP2, the goals were cited in the form of questions. At the end of the Topic Plan students were invited to indicate on a scale of 1 to 5 whether they believed the goals had been achieved for them personally.

Those responses have not been analyzed as they tended to be overwhelmingly positive. This could be due to the traditional belief that one does not criticize the teacher, which is hard to overturn in the short term.

TP2: The concept of number

1. to understand why we need symbols to represent quantities
2. to understand that we can develop a system to represent quantities using any number of symbols

TP 3: Percentages

1. what are the relationships between integers and percentages when we talk about quantities?
2. what are the characteristics or properties of percentages?
3. how can percentages be added?

TP 4: Decimals

1. What are the relationships between integers and decimal numbers?
2. What are some of the ways we use decimals in real life?
3. How can decimal numbers be added and subtracted?
4. How can decimal numbers be changed into fractions?

TP 5: Fractions

1. Why do we need fractions?
2. How do we represent quantities using fractions?
3. What are the different ways in which fractions are used?
4. How do we add and subtract fractions?

2 Analyzing the data

2.1 The nature of the data collection process

2.1.1 Components

There are three parts to the system.

1. Feedback device: Currently this looks like a standard keyboard but with nothing attached to it. The keyboards are adapted to contain a wireless transceiver that passes data to the second part of the system - the coordinator device.

2. Coordinator: This is responsible for managing the communication with up to 128 feedback devices. It interfaces to a laptop via a USB interface. We have never had the opportunity to test the system with more than 76 devices.

3. Laptop: This has the necessary software and is also used to store the data sent from the feedback devices, as well as to commence capture of input data from the feedback devices.

2.1.2 Operation

Historically, a laptop has been used to display the pre-prepared Topic session with embedded questions. This is used in extended desktop mode on an Ubuntu Linux distribution running on the laptop. This configuration permits any window to be dragged between the projected and laptop displays.

Our Topic Plans have largely been prepared using Latex Beamer which allows us to call any type of application from within the PDF file e.g. a movie file.

Typically, the capture software runs on the laptop screen. At this early stage of development it permits start and stop of capture, monitoring of data as it is typed in, and the ability to generate histograms of data as well as a concordancing of text input by the students as a whole. These displays ensure that no student's response is individually identified in the histogram or the concordance. At any stage this grouped graphical response can be moved to the projected display for students to see. All data that is input by students is automatically logged into a Comma Separated Variable (.csv) file. That file becomes the performance archive source.

2.2 The sources of data

2.2.1 Logged student inputs

Most of the data that has been used to do the analysis in this section has come from the inputs provided by students on the keyboard. The data that is captured is:

- the question number
- the student identity
- the input typed in by the student
- the corrected response where the instructor has corrected, say a spelling mistake in a students input. This is particularly important where concordance is being used

- the date and time
- the delay in seconds between starting capture and the student activating the first key-press after the instructor initiates the capture process

2.2.2 Video and audio record

We tried to maintain a continuous video record of all sessions, as it had been our intention to conduct a Distractor Analysis of the sessions to see which parts of the Topic Plan needed improvement. This has not been possible as there is not an adequately distributed range of Distractors both in terms of intensity and in terms of individuals within the group.

The sessions before 11 January 2010 do not have an audio component. These were recorded using a webcam, and audio quality was too poor to be usable. After the start of the new term on 11 Jan 2010, we were able to deploy a Flipvideo recorder which provides High Definition video and the quality of audio recorded has made background noise quite acceptable. There are some instances where the record is incomplete.

2.3 The Primary observations

2.3.1 Questions

Over the period of the Pilot, students answered more than 400 questions using the feedback devices. These comprised the 4 Topic sessions, activities, projects and reflections sessions but does not include the 113 questions the students answered in the pre-test and the 65 questions answered in the post test.

2.3.2 The Statistical comparison

	Test Score Statistics			
	Mean	**Max**	**Min**	**Std. Dev.**
Experimental Group - Pre Test	24%	50%	5%	9%
Experimental Group - Post Test	75%	89%	46%	12%
Control Group - Test at end of session	61%	82%	35%	9%

Table 2.1: Comparison of group statistics for pre-test and post-test experimental group and the control group post-test

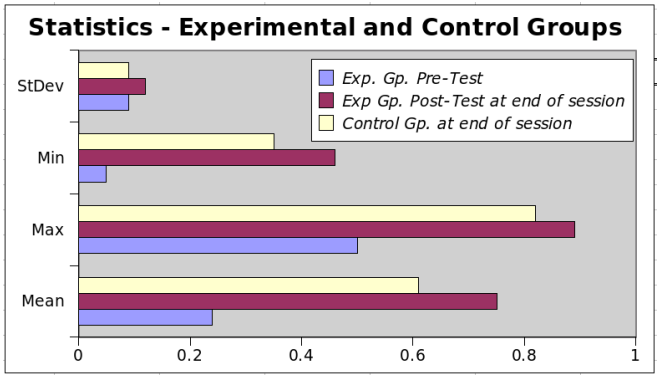

Figure 2.1: Representation of group statistics for pre-test and post-test experimental group and the control group post test

The notable elements are that the experimental group started off at a very low level, through being a remedial group, and at the post test scored higher than the control group in terms of the mean score, the maximum score as well as the minimum score. Unfortunately we have no data for the start in terms of the control group so there remains the possibility that the control group made similar progress during this period.

Potential sources of error

We have used the spreadsheet Average function to determine average score. This can artificially enhance the scores where there are missing values, as these were ignored. Where a student was absent from the class, we have assumed it is legitimate to ignore the missing values. Where a student who was present in class failed to answer a question, it is desirable that the sum of the scores should be divided by the total number of questions to obtain the student's average score.

However, we found that for the experimental group post test, this latter effect amounted to 1.5% of the cases, while for the control group it amounted to 7% of the cases. As any recalculation would tend to further increase the difference between the experimental group and the control group we have chosen not to conduct this re-calculation.

For the experimental group pre-test, we found that there were 26 instances where students were present and failed to enter a value. This amounted to less than 1% of the total opportunities to provide responses.

We have therefore ignored the level of nil responses in all instances.

2.3.3 Normalized gain

On the recommendation of Derek Bruff, Assistant Director, Center for Teaching, Vanderbilt University, normalized gain has been included, as shown in Fig. 2.2.

Normalized gain, usually denoted <g>, equals (post-pre)/(100-pre). So, for instance, if a student got a 40 on a pre-test and a 70 on a post-test, he would have a <g> of (70-40)/(100-40) = 30/60 = 0.5. That means his change in scores from pre-test to post-test was 50% of what it could have possibly been. This measure is a useful way to represent improvement since it factors in the pre-test scores.

As indicated earlier, the pre-test results are only available for the experimental group. The statistical characteristics of the distribution of the experimental group for normalized gain are:

- Mean: 66%
- Standard Deviation: 16%
- Minimum: 27%
- Maximum: 84%

The histogram for the experimental group is depicted in Fig. 2.2 and shows the highest frequency to be in the range of 70% and 80%.

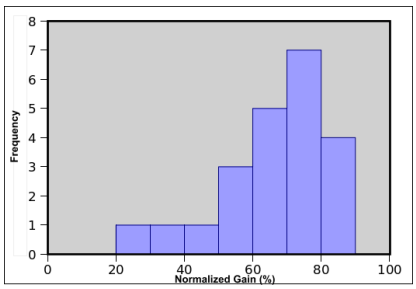

Figure 2.2: Frequencies of Normalized gain for Experimental Group

2.3.4 Measuring progress of the experimental group across Topic sessions

Progression of Topic Plan scores

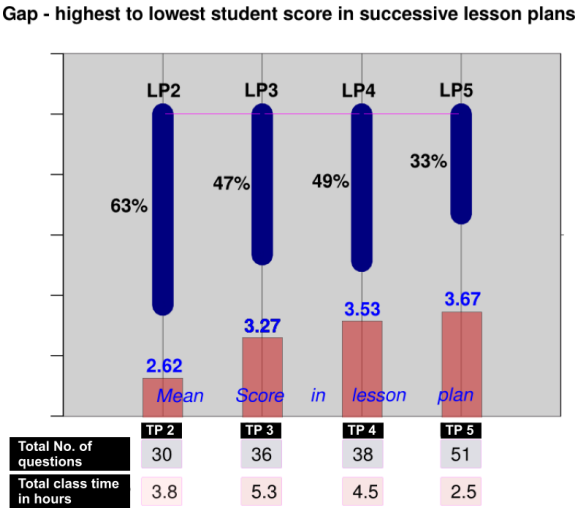

Figure 2.3: Progression of scores and extremes of scores across Topic Plans

In Figure 2.3 two sets of data are visible. The bar graphs in red show the progression of mean scores on successive Topic Plans. These show a gently increasing trend.

The blue bars show the extremes of scores within the experimental group. These were obtained by equating the highest score obtained in a Topic Plan to 100%. In such a normalized scheme the highest score is always 100% in all Topic Plans. Hence the gap between the highest and lowest score can be compared between Topic Plans.

We see a slight increase in the mean score which could also be a function of our increasing familiarity with the student group leading to better design of the Topic Plans, as the analysis of the prior Topic Plans showed us where the steps in learning were too steep. The mean score is on a 1 to 5 scale.

Worth noting is the decreasing gap between the highest and lowest scores taken on a normalized basis. Together these may give etiological pointers to increment in performance over the duration of the pilot.

2.3.5 The potential impact of inferential ability

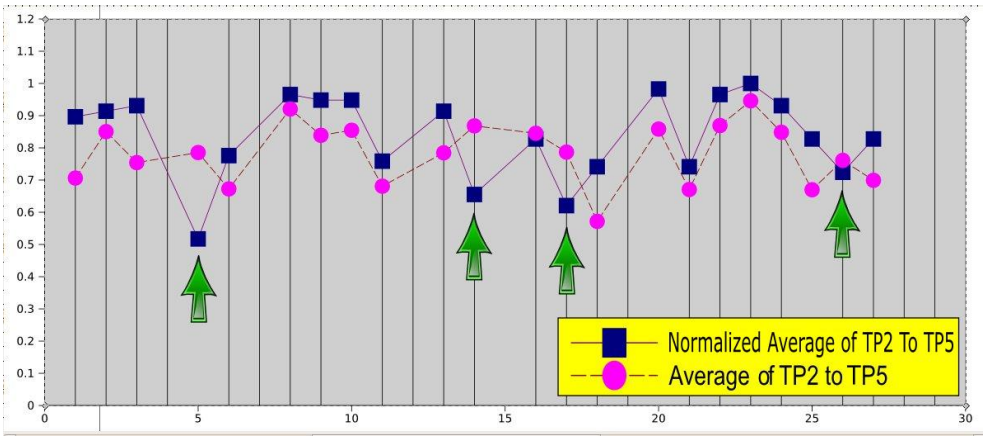

Figure 2.4: Individual student comparison of normalized average of Topic Plan Question Scores against average of Topic Plan 2,3,4 and 5 question scores

The normalized Topic Plan scores, averaged for TP2 to 5, for each student as described in Section 2.3.4 are shown in blue squares for each student. The average scores for Topic Plans 2 to 5 are similarly reduced to a 0 to 1 regime.

If each student had contributed to the increasing compression seen in the upper bars of Figure 2.3, then each instance of the normalized score for each student would have been above the average score obtained over TP 2 to 5.

However, 4 students - indicated by the green arrows - seem to have violated this trend. We chose to investigate if this was related to weaknesses in inferential ability.

Figure 2.5 shows that students 5, 14,17 and 26 represent the lowest results for inference with Student 5 having the lowest which is consistent with the greatest difference in the figure above. Student 18 has a score closely matching that of Student 26 in terms of the Post-test score but there is a relatively high score in the Inference category which is putting the normalized score well above the TP2 to 5 average score.

Against this we need to point out that Student 11 has an inference score which is the same at Student 17 but does not create the same effect in Fig. 2.4. Practically speaking, any correlation with such a small sample does no more than raise the possibility of the need for more research. Student 16, on the other hand, only differs in terms of intensity but has a significantly higher Inference rating.

We would also state our inability to say that these categorizations have adequate rigor as explained in Section 2.4.

Average - Result	QuestionType				
StudNo	Calc	Inference	Obs	Reporting	Total Result
1	0.92	0.64	1.00	0.86	0.80
2	0.96	0.68	1.00	0.71	0.82
3	1.00	0.68	1.00	0.71	0.83
5	0.50	0.43	0.50	0.43	0.46
6	0.67	0.61	1.00	0.86	0.69
8	0.96	0.75	1.00	0.86	0.86
9	0.88	0.79	1.00	0.86	0.85
10	0.83	0.86	1.00	0.71	0.85
11	0.75	0.57	1.00	0.57	0.68
13	0.96	0.68	0.83	0.86	0.82
14	0.75	0.50	0.33	0.57	0.58
16	0.83	0.64	1.00	0.57	0.74
17	0.62	0.57	0.17	0.57	0.55
18	0.67	0.64	0.83	0.57	0.66
20	0.96	0.82	0.83	0.86	0.88
21	0.62	0.64	0.83	0.71	0.66
22	1.00	0.75	0.83	0.86	0.86
23	1.00	0.79	1.00	0.86	0.89
24	0.96	0.68	1.00	0.86	0.83
25	0.75	0.71	1.00	0.57	0.74
26	0.83	0.46	0.83	0.57	0.65
27	0.92	0.61	0.83	0.57	0.74
Total Result	0.83	0.66	0.86	0.71	0.75

Figure 2.5: Post-test breakup per student into C, I, O and R categories and Post-Test Score

2.4 Secondary observations

These observations are based on the Post test measurements. We gave each question one of the four categories below depending on how well it could test the following abilities:

1. Calculation (C): To calculate correctly and reliably
2. Inference (I): To be able to apply previous knowledge in new situations.
3. Observation (O): To observe and measure carefully
4. Reporting (R): To be able to describe the results of calculations.

We are not confident that this was a rigorous classification. We asked two members of our team to provide these classifications and where there was a difference a third member arbitrated to provide a final classification. Hence, we feel that this is a reasonable basis for a tentative assessment.

The results below are indicated in terms of the percentage of instances cited against the total number of instances of that particular category across all participants.

2.4.1 Experimental Group - Pre and Post test comparisons

	Question Category				
Group	C	I	O	R	**Total Result**
Pre-Test (Experimental Group)	24%	24%	26%	24%	24%
Post-Test (Experimental Group)	83%	66%	86%	71%	75%

Table 2.2: Comparison of pre-test and post-test scores for the experimental groups across question categories

In general, we note that students abilities were minimal at the pre-test level with little differentiation across categories. The greatest increase took place in the Observation category, with the least increase in the Inference category. This was consistent with our observations during the class and also with the results of the individual Topic Plan analysis. However, there were marked increases in each category between pre-test and post-test scores of the Experimental Group.

2.4.2 Comparison across Categories between experimental and control groups - performance on post-test

The table below shows the difference between experimental and control group performance

	Question Category				
Group	C	I	O	R	**Total Result**
Experimental (E)	83%	66%	86%	71%	75%
Control (C)	68%	51%	76%	60%	61%
Difference (E-C)	15%	15%	10%	11%	14%

Table 2.3: Comparison of control and experimental groups across question categories

It is notable that the common characteristics between the experimental group and the control group are related to the ordinal progression from being weakest in Inference to increasingly improved performance in terms of Reporting, Calculation and best performance in Observation.

We were surprised by the capabilities of the students when projects were conducted that required taking readings off the screen. For instance, good results were obtained when they were asked to determine the readings of time differences between successive instances of a video of a pendulum reaching maximum amplitude on the same side. However, calculation ability was a relatively severe problem, with children often being unable to divide by ten to obtain the time period from ten observations.

It took some time for students to enter into a "observing-thinking-reporting" process, and this preceded the delivery of Topic Plans 3,4 and 5, where the greatest learning may have taken place.

The experimental group seems to have made the greatest strides in the area of Inference and Calculation compared to the control group. Both results are to be expected as the participating children got considerable experience in answering questions many of which also honed inferential skills and some developed calculation skills. The lower increments in Observation could be based on the initial preponderance of Observational ability. The Reporting results are better than expected as there was no overt attempt at specifically building reporting ability.

2.4.3 Comparisons across Categories between the maths teacher's estimates of ability and scores - experimental group

At the end of the course but before these results were computed, the maths teacher was invited to provide an estimate of the abilities of participating students of the experimental group in maths on a Likert Scale of 1 to 5 where 1 represented the lowest level of ability and 5 represented the highest level of ability in the full class. The table below provides the test scores against the order of ability cited.

		Q Type				
Teacher's Estimate of ability	No of students	**C**	**I**	**O**	**R**	**Total Result**
Likert Level 2	7	0.74	0.63	0.9	0.65	0.7
Likert Level 3	9	0.83	0.67	0.78	0.71	0.75
Likert Level 4	5	0.95	0.69	0.93	0.8	0.82
Likert Level 5	1	0.92	0.61	0.83	0.57	0.74
Total Result	22	0.83	0.66	0.86	0.71	0.75

Table 2.4: Teacher estimate of ability compared to post test scores of experimental group

From this it would seem that while the estimates were generally correct the one student cited as being a Level 5 was probably closer to Level 3.

It is interesting that this student identified as Level 5 scored well in the Calculation category but was weakest in the Inference score. In Indian schools, it is not uncommon for calculation ability to be used as a proxy for competence in maths.

2.5 Term examinations

The school conducted one set of examinations internally at the end of Term 1. These concluded before we commenced our engagement with the school. Term 2 examinations were conducted shortly after we completed our program. We were of the opinion that our program may impact the results of the Term 2 examinations in Maths, English and Science. We asked for and received individual student results for both Term 1 and Term 2

exams. We have no idea of what was taught or examined or the nature of the examination process. We do know that the three subjects being considered had 2 components each - Written and Oral.

It appears there are 75 students who took the Term 1 and 2 examinations out of a class strength of 76. Only 65 students attended our post test examinations which form the basis for our pre and post test assessment described earlier. It is assumed for this exercise that the control group comprised the 53 students who were not part of the experimental group.

Change in Scores between Term 2 (After AGEM Program) and Term 1 (Before AGEM Program)				
		Change (Term 2-Term1) scores		
Characteristic	Group (No. of Students)	English	Math	Science
Average Change	Experimental Group (22)	9.9	7.5	13.8
	Control Group (53)	9.4	2.7	10.3
No change instances	Experimental Group (22)	2	1	1
	Control Group (53)	6	10	4
Negative change	Experimental Group (22)	3	4	3
	Control Group (53)	13	16	12

Figure 2.6: Difference of Term 2 and Term 1 School examination scores - Experimental and Control groups

2.5.1 Principal observations

- The increment shown by the experimental group was more than 2 times that shown by the control group i.e. 7.5% as opposed to 2.7%
- The number of instances where there was no change was limited to 1 for the experimental group while for the control group this was 10 times higher althogh the control group size was less than 3 times larger.
- There were students who got a lower percentage of marks in Term 2 compared to Term 1. There were 4 instances of that in the experimental group and 16 instances in the control group.
- The experimental group recorded a somewhat better performance in the English and Science examinations although this was not as large as it was in Maths.

2.6 Limitations of this study

Sample size

A larger sample would have served the Pilot better.

Control Group Pre-test

The study would have benefited from a pre-test provided to the control group. This would have been possible if our interactions with the students had not led us to believe that they were from several sections of Grade 6 rather than one class of that Grade. This problem was further compounded by the fact that due to software issues we were not able, at that time, to use more than 28 keyboards simultaneously.

Absence of prior studies with multiple keyboards

While numeric clickers have been in use for some considerable time, the capacity to enter alpha numeric text has been more limited. We were unable to find previous studies of this sort of work. We believe further studies may be needed to see if similar results could be obtained by using only multiple choice methods.

However, it appears that the capacity for written expression is a necessary element for most students in India. We are not sure therefore that clickers with limited alphanumeric capabilities would be suitable for that purpose.

3 Conclusions

3.1 Primary issues

The primary outcome of this Pilot was a great sense of personal satisfaction for the AGEM team. The team that delivered the sessions found this to be an enriching experience once TP2 had been conducted and our software problems had been resolved. More specifically:

- While it would have been better to have pre-test results for the control group also, the significant improvement in the experimental group was unexpectedly large. This was supported by the experimental group's performance in the schools internal examinations.
- The feedback from the children indicated they would prefer more questions. We are also of the view that an even slower progression to allow the weakest to learn better would be desirable. We never had a situation where some students appeared to be impatient at too slow a pace of progress.
- We are left with the realization that context is important for young children particularly the urban poor whose exposure to real life events is more limited than it is for the rural poor in many ways.
- Classroom management has been a non-issue although the team that delivered the program was not specifically trained in managing young children. The keyboards kept the children engaged in a relatively permanent state of anticipation for the next question, but the quality of materials on the projected display, though not particularly well done, was an improvement on what prevailed. It is possible both factors conspired to make classroom management an easy task.
- The development of decision making skills is another issue we had targeted at the start. Many of the questions embedded in our Topic Plans did not have a clear cut right or wrong answer. However, all called on the child to make a decision. The very low frequency of cases where children did not respond to a question was possibly due to our emphasis on keeping individual decisions confidential. It was possibly also a factor in developing confidence, and hopefully, decision making abilities.
- There is scope for the Topic Plans to be greatly improved as a result of the analysis.
- We are led to the view that building communication skills in the children should be a parallel activity in any future project.
- We are also of the view that there needs to be a slower progression in terms of building conceptual understanding and in terms of creating inferential skills.