Mastery Learning and Its Effect in ASSISTment

by

Zach Broderick

Advisor: Neil Heffernan

Computer Science

Technical Report

Series

WORCESTER POLYTECHNIC INSTITUTE

Abstract: In this paper, we analyze the effect on student learning of introducing Mastery Learning into the Intelligent Tutoring System ASSISTment. 8th grade students at a local middle school were divided into groups and given a pre and post test for several math units, with one group or the other doing Mastery Learning exercises in between for each unit. We found that overall, students in the experimental group improved significantly more between the pre and post tests than did those in the control. Additionally, we found that those students with lower self-discipline and knowledge benefited the most from Mastery Learning.

Introduction

The purpose of this experiment was to analyze the effect on student learning of introducing Mastery Learning into the Intelligent Tutoring System ASSISTment. Also, we hoped to observe whether a student's GRIT or IRT score influenced whether or not and how much they benefited from Mastery Learning. Our hypothesis was that Mastery Learning does in fact positively impact student learning; we were unsure how GRIT or IRT scores may influence this outcome. Before we move on, however, it is necessary to define some of the above terms.

ASSISTment

ASSISTment, a combination of Assist and Assessment, is an online Intelligent Tutoring System developed by Professor Neil Heffernan and his team at Worcester Polytechnic Institute in conjunction with Carnegie Mellon University. Primarily middle/high school students join online classes created by their Math teachers and do problem sets with built-in tutoring. Their performance data is then sent to the teacher (and us, for research purposes) so that she can adapt her lesson plan and teaching methods to better suit her students' needs. Students are also provided with instant feedback on their work. For more information go to www.assistment.org.

Mastery Learning

Mastery Learning is a feature just recently introduced into ASSISTment. Mastery Learning problem sets are different from regular ASSISTments in a few key ways. They are designed to hone in on one particular skill, and thus consist of exercises that are almost identical to each other except for the values used. Students are able to "test out" of these problem sets if they get the first X problems (usually 1 or 2) in the set right, or if they get Y right it a row after that (usually 3 to 5). The idea is that the student is not allowed to progress unless they have "mastered" the particular skill exercised by the Mastery Learning set. To avoid burnout, students who do Z problems (usually 5-10) in one day without testing out are told to try again the following day.

GRIT and IRT

GRIT and IRT measure student self-discipline and proficiency, respectively. GRIT scores are obtained by having students take a short survey in which they evaluate their own discipline, answering questions such as "I often do things that are fun even if they are bad for me" with either "Not at all like me" all the way up to "Very much like me." Each of these answers is assigned a value and then these values are added up across all of the survey questions to arrive at each student's GRIT score. These scores, of course, fall victim to all of the pitfalls one might expect from such a voluntary self-survey, so keep that in mind throughout the rest of this paper.

IRT scores are calculated by ASSISTment based on a student's performance and stored in the database. They are periodically updated as the student does more problems. These scores essentially measure a student's proficiency and can be used to predict the probability he or she will answer a particular problem correctly. In this experiment and several others we use the median of these scores to divide the students into "high knowledge" and "low knowledge" groups.

Method

This experiment was conducted at a local middle school using 8th grade students participating in a program called Math Lab. In the same manner as Art and Gym, each 30 day cycle a group of students goes to the computer lab for an hour each day and does ASSISTment problems. While this is not the intended use of ASSISTment, it has proven a prolific source of data. However, as one might expect, students do not like Math Lab very much, and so many of them put in little effort, game the system, or collaborate with their friends, making the data noisier than a more tightly controlled experiment.

Math Lab spans 4 periods each day, with 20-25 students per period. For the purposes of our experiment, we have divided them into groups of two periods each that alternate between the control and experimental condition. Over the 30 day course of Math Lab, students work on problem sets that fall under 5 different math "units" corresponding to a particular skill set, such as data analysis, arithmetic with negative numbers, linear equations, and so on. For each of these units we administered a pre and post test to both groups. The experimental condition was whether or not the students did Mastery Learning problem sets corresponding to the current unit during the several day period between the unit's pre and post test. For each unit, the experimental group was assigned Mastery Learning in this manner, while the control group did their regular problem sets. The experimental condition was alternated between the two groups over the 5 units. We would like to mention here that the pre-test scores between the control and experimental groups were close enough that we did not expect any extraneous effects related to the particular students in each group.

We would also like to note here a flaw in our experimental design. Our control group was not very well "controlled," in that they were not assigned a particular task while the

experimental group was working on Mastery Learning exercises. In the several days between pre and post tests, these students worked on their practice problems for other units, which is what students in Math Lab do when they are not taking tests or doing Mastery Learning. This means they could have been working on any number of things from a variety of different units, which no doubt introduces a fair amount of variance into our results. In the future, we will likely have the students work on a traditional problem set related to the current unit while the other group does Mastery Learning. For this experiment, we were simply trying to establish, as a baseline, that Mastery Learning was helpful to student learning rather than useless or even harmful.

After the particular Math Lab cycle on which we chose to focus had ended, we computed each student's gain score for each unit by taking the difference of his or her pre and post test scores of that unit. We quickly noticed that some students' scores had actually *decreased* by unusual amounts, such as 50 percentage points. It is highly unlikely that the students "unlearned" that much over the course of a few days—more likely is that they were simply not trying on the day of the test. This perhaps seemingly dubious assumption is highly supported by observational data from Math Lab, and so we decided to eliminate any gain scores that were below -20 percentage points.

We then converted each of these gain scores to a standard Z score by subtracting the mean gain for that unit and dividing by the standard deviation. While we could have simply used the raw gain scores, due to the varying difficulty of the different units we felt that a relative measure would be a more accurate reflection of student learning. We then averaged, for each student, those standardized gain scores for which he or she was in the control group, and those for which he or she was in the experimental. Our results are based on these two groups of standardized means.

Additionally, GRIT scores were computed based on a survey each student was assigned to take. IRT scores were calculated after the cycle was over and extracted from the ASSISTment database. The amount of time spent on Mastery Learning and the number of problems done by each student were also extracted for each unit.

Results

As we expected, our results indicated that the experimental group on average gained reliably more than the control group. Below is a chart of the consolidated data. As can be seen from the table, on average the control group unlearned between the pre and post tests, while the students doing Mastery Learning improved their scores. These results are highly reliable and have a solid effect size.

	Z-Scored Mean Gain
Control	-0.164 +/- 0.072
Experiment	0.186 +/- 0.089
$p < 0.003$	Effect Size = 0.5

There were a few aspects of our results that gave us pause. For one, there was a fairly sizeable gain, most notably in the first two units (not shown here), that seemed to be independent of the Mastery Learning condition—in fact, it seemed to dwarf it by comparison. We caught this early on and hypothesized that it might be the result of tutoring present in the pre-tests; we subsequently removed this tutoring from the remaining units and the overall gain appeared to diminish. While this gain made it more difficult to zoom in on the effect of Mastery Learning, it is a favorable result nonetheless in that it provides additional evidence that our tutoring system reliably increases student learning, which we have established in earlier studies.

After reading the above results, one might legitimately ask whether or not these findings are completely obvious; after all, one would expect that the students who received more practice for a given unit would do better on the post test than those who didn't. This is true. However, as I mentioned, in order to make any other claims of interest a baseline must be established, in that Mastery Learning helps rather than harms or doesn't affect student learning. It should be noted also that doing so in such a noisy environment is not trivial, making the reliability of our results all the more impressive.

GRIT and IRT

The median GRIT and IRT score of all the students were computed and used to subdivide each group into high and low GRIT and IRT groups. The standardized means for each subgroup were then computed in the same manner as previously explained.

GRIT	Control	Experiment
High	-0.21 +/- 0.11	0.00 +/- 0.12
Low	-0.12 +/- 0.10	0.37 +/- 0.13

IRT	Control	Experiment
High	-0.29 +/- 0.08	-0.17 +/- 0.08
Low	-0.04 +/- 0.12	0.55 +/- 0.14

Looking at the above tables, it is overwhelmingly clear that students with low GRIT scores and especially low IRT scores have considerably higher average standardized gains than the other 3 conditions. This suggests that students with low self-discipline and low proficiency benefited much more from Mastery Learning than their peers

Model

Tests of Between-Subjects Effects

Dependent Variable:Gain

Source		Type III Sum of Squares	df	Mean Square	F	Sig.
Intercept	Hypothesis	1.222	1	1.222	1.161	.282
	Error	297.626	282.670	1.053[a]		
Group	Hypothesis	9.064	1	9.064	10.742	.001
	Error	264.106	313	.844[b]		
Last	Hypothesis	128.315	92	1.395	1.653	.001
	Error	264.106	313	.844[b]		
Probs	Hypothesis	.001	1	.001	.001	.979
	Error	264.106	313	.844[b]		
Time	Hypothesis	.882	1	.882	1.046	.307
	Error	264.106	313	.844[b]		

a. .380 MS(Last) + .620 MS(Error)

b. MS(Error)

The above SPSS output is the result of running an ANOVA on a flat spreadsheet of the data, with gain as the dependent variable, experimental condition and name of the student as fixed and random factors, respectively, and the amount of time spent on and problems done in Mastery Learning as covariates. Note that the gain used as the dependent variable was the standardized mean gain of each unit per student before they were combined into one measure.

As the chart indicates, the condition group and last name of the student was highly reliable in determining gain. This reinforces our earlier results that those in the experimental condition learned more than those in the control. The student's last name was used as a factor because it accounts for factors related to the individual student, such as his or her IRT and GRIT scores, that would not be independent if they were used in the model individually. The fact that the student's identity was reliable, we believe, supports that notion that GRIT and IRT play a factor in how beneficial Mastery Learning is to the student. The effect of how much time spent and the amount of problems done on Mastery Learning could not be reliably determined.

Anecdotal Results

Aside from just collecting data from the ASSITment database, we also visited Math Lab in order to gauge student impressions of Mastery Learning, which we feel can be equally important to quantitative data. The strongest sentiment we perceived, on multiple occasions, was that students really enjoyed the concept of "mastering" a problem set. We repeatedly heard yelps of joy when students completed a Mastery Learning set and the "mastered" tag appeared next to it. One student said it felt like "conquering the world." This should not be taken lightly, as morale is notoriously low in Math Lab among the students. Fostering this sense of accomplishment in such an environment is an impressive achievement.

Many students, however, we frustrated with Mastery Learning. Noticing that certain students had been told by ASSISTment to come back and try again on a particular Mastery Learning set 4, 5, or even 6 times, we decided to investigate what the cause might be. Many students did not remember or claimed it was the result of a technical glitch, which is more than plausible. Other, including the instructor of Math Lab, felt very strongly that students were getting discouraged. They would figure out the concept behind a particular Mastery Learning set, but then would make a minor mistake and have to do additional problems or worse, be told to come back the next day. For many students, this pattern repeated over several days, until they simply lose motivation to finish the problem set. Still others, by their own accounts, mostly but not completely grasped a particular concept, and thus were not able to answer 3 problems correctly in a row even though they answered most correctly overall.

When asked directly how they felt about Mastery Learning (and given explicit permission to freely express discontent with it if that was the case), most students were indifferent, viewing it as just another type of problem set. As mentioned, a few liked the feeling of accomplishment given by Mastery Learning, while others disliked being repeatedly punished for minor errors.

Conclusion

Our results seem to suggest that Mastery Learning does in fact increase student learning. Furthermore, it appears that students with lower self-discipline and lower proficiency would benefit the most from Mastery Learning exercises. Further study is certainly encouraged. We would recommend that additional experiments be conducted in a less noisy, controlled environment. Also, a stronger control should be implemented, perhaps having students in that group do a traditional problem set related to the same unit.

Qualitatively, it seems students respond very well to the positive feedback given when they master a problem set. This was so prevalent during our observations of Math Lab that we believe it warrants further study on its own merit. However, it is also obvious from observation that Mastery Learning needs to be improved to prevent student frustration and burnout resulting from continued failure to master a problem set. The ability to reconcile minor errors has been a constantly requested feature by students for the ASSISTment system, and it appears Mastery Learning would benefit from such a feature as well.

Finally, it should be noted that Math Lab is not the type of classroom setting ASSISTment was designed for. In fact, it is not a typical classroom environment at all, but more like a controlled study. This should be taken into account when interpreting the above results. Future studies by us and others might be more beneficial if they were conducted in a normal classroom environment.

Acknowledgements