Deployment and Evaluation of an Educational Platform for Shared Computing Contexts

by Clint Tseng

A senior thesis submitted in partial fulfillment of the requirements for the degree of Bachelor of Science With College Honors Computer Science & Engineering University of Washington March 2010

Presentation of work given on: __________________________
Thesis and presentation approved by: ______________________
Date: ____________________________________________
TABLE OF CONTENTS

Abstract .......................................................................................................................... 3
Introduction .................................................................................................................... 3
Related Work .................................................................................................................. 4
Study One: Pilot Trial ...................................................................................................... 5
  Study Design ................................................................................................................ 5
  Results .......................................................................................................................... 6
Study Two: Design and Interaction ................................................................................ 8
  Introduction ................................................................................................................ 8
  Study Design ................................................................................................................ 8
  Results: Dominance .................................................................................................... 10
  Results: Collaboration ................................................................................................. 14
  Conclusions ................................................................................................................ 15
Study Three: Learning Outcomes .................................................................................. 15
  Introduction ................................................................................................................ 16
  Methodology ............................................................................................................... 16
  Results ........................................................................................................................ 17
  Conclusions ................................................................................................................ 18
  Future Work ............................................................................................................... 18
Acknowledgements ....................................................................................................... 19
References .................................................................................................................... 20
ABSTRACT

As the need for access to technology in developing regions increases rapidly, the supply of personal computers in these areas is failing to meet the demand. In the context of education and the presence of computers in under-funded schools, the computer-to-child ratio limits equal access to educational material and deprives marginalized children of valuable digital literacy skills. In this thesis, we explore results around a software solution we developed called MultiLearn, which utilizes multiple keypad input, split-screen display, and adaptive questioning to facilitate learning by multiple students simultaneously. We use MultiLearn as a platform upon which to explore issues pertaining to interaction in shared computing scenarios, such as dominance and collaborative effects, as well as assess the MultiLearn system as a whole with regards to learning outcomes.

INTRODUCTION

Where computers are present in low resource classrooms in the developing world, there are often insufficient resources to individually support students. As a result, students often interact with a computer in groups of up to nine or ten. This situation can lead to significant and troubling interaction patterns among the students, wherein the large majority of students in the group never interact with the computer or the educational software [9]. To address this issue, several educational systems have been developed to support multiple input devices, providing one device for each student [8, 9].

This type of multiple input research stems from both single display groupware [5, 11] and shared computer use among children [6]. Recent work with multiple mice on a single computer has found that engagement and learning in certain conditions increased when each student was given an individual mouse to interact with digital learning material [9]. Previous multiple mouse interfaces have largely focused on

Figure 1: MultiLearn setup: four keypads and split screen.
competition-based multiple choice designs, though some attempts have been made to incorporate text entry into the design to allow for more open-ended responses [1, 8].

Building on this research, we developed MultiLearn (figure 1), a multiple-input educational game for math, spelling, and other subjects [4]. As an alternative to the mouse, MultiLearn provides 10-key USB numeric keypads. Due to their near-universal compatibility, low cost, convenient size, and abundance of keys, these keypads provide an effective means of exploring input beyond the mouse, particularly for numeric and text input. MultiLearn supports multiple students at once by dividing the screen into individual sections. It also incorporates several educational design concepts: individual student performance is tracked over time, the difficulty of questions asked of each student is adjusted accordingly (made possible by the split screen), and teachers are provided with tools to create content and examine how each of their students is improving.

**RELATED WORK**

The concept of single display groupware (SDG) is a not a new idea, and researchers have previously explored the potential benefits of collaboration through multiple input devices [5]. Our research follows the SDG model and provides an easily extensible foundation for adding more educational content over time.

Prior research on using multiple mice has demonstrated the benefits of providing each student with their own input device [9]. The overall engagement of the students with the material was higher when each child had their own mouse. We observed similar results in our various studies, described below, with keypads. The chaotic talking and physical contact between students decreased significantly when each child had their own device, and the perceived learning was higher in this scenario as well.

Research with on-screen keyboards has demonstrated and explored the possibility of text entry using a mouse [1, 8]. However, by using a physical keypad, as we do in MultiLearn, we save valuable screen real estate and hypothesize that a physical keypad will allow for faster, more accurate, and more varied types of input.

There has also been research done using a split screen model [7], where the screen was split into two halves and the children worked in teams of two. The results of this study showed that by giving each child their own input device, the common scenario of one dominant child controlling the
mouse or keyboard was reduced and collaboration among the teams increased, and our own work has explored this trend as well [12, 13]. Research has also shown that children pay almost no attention outside of their section of screen [7], which we have found to also apply even when the screen is split into four sections rather than two.

**STUDY ONE: PILOT TRIAL**

To assess the feasibility of our design, and of the USB keypad as an input device, our first study was a fairly straightforward pilot trial. Conducted in March of 2009, we tested the software with 140 fourth and fifth grade students in 35 mixed-gender groups of four. This took place in four government schools in Bangalore, India.

**Study Design**

As this was an initial study, the procedures were kept fairly simple. As students entered the room, they were encouraged to seat themselves at the computer. Keeping in mind prior work showing dominance patterns as they related to the order in which students naturally seat themselves around a computer, we allowed the students to choose their own seating arrangements. We would then run the group through our three conditions.

As our goal was to establish the MultiLearn approach as a feasible and useful means by which to improve the education of students, we set up the conditions as a comparison between shared-input and multiple-input modes. In all conditions, the students completed basic arithmetic questions. As they did so, their progress marker would advance across the screen until a student correctly answered 12 questions, at which point the game would end (Figure 2). Each of the three conditions involved one game round, but varied in the level of individual interaction each student had with the software: in the first condition, each student was provided with a personal keypad, and solved their own problems. In the second, students were put in teams of two, and forced to share a keypad while working on a single problem with their teammate. In the final condition, all the students were asked to share one keypad, while all collaborating on the same problem.

![Figure 2: Original MultiLearn screen layout](image)
The software had a basic level of adaptive capability at this point, adjusting the difficulty level of the problem to the skill level of the student. However, as we only saw each student once as an individual (in the first condition), we did not maintain this state across sessions. We did maintain logging data for the duration of the study. We tracked the questions we gave each student, the difficulty the software judged the problem to be, the times at which the question was given and at which it was answered, and the response the student gave. We also took qualitative observation notes for each group.

As the students involved in our study did not speak English, we hired two local translators to communicate the purpose of the software and the mechanics of the keypads to the students. Afterwards, we would also have them interview the students for us. Questions we would ask included:

- Which round did you like more?
- Did you like racing against each other?
- Were the problems difficult?
- Did you like doing math problems on the computer rather than with paper and pencil?
- What other kinds of questions would you like?
- How often do you use a computer?
- How many people are using the computer with you when you use it?
- Would you use it more or less than other games you play at school?
- What would you change about our game?
- Would you like to know the right answer after you answer a question incorrectly?
- Do you want to know more about computers?
- What do you want to be able to do with a computer?

We would also ask additional questions of specific groups if we saw anything particularly odd or interesting about the group's dynamics.

**Results**

Though there were minor problems with such things as the concept of the enter key, and the fact that the backspace key was labeled merely 'BS', there were in the long term no issues at all with the students grasping the input devices. By the time each group of students got to the end of the first condition, they were fully familiar with the input devices and how they were to be used to answer the math questions. As this was the biggest new concept we were trying to prove as usable, this was the most important result of the study.

There were many other things that we learned from this study, both in terms of lessons to apply in the future as well as in terms of results. While we attempted to introduce both collaborative and
competitive elements in our study conditions, we could not establish a statistically significant difference between the two modes. In fact, we observed much conflicting behavior between groups of subjects — some groups would collaborate even in the purely competitive mode, remaining seemingly uninterested in the competitive aspect of play, while the students in other groups were highly excited and motivated by the prospect of winning the game. It seemed that the difference here lied mostly in the school and the atmosphere it fostered in general.

We also found that, in line with past research [3], student attention rarely wandered away from their particular quadrant of the screen. This meant that the progress indicators we placed along the top of the screen were outside of the scope of the students' view, and students were often unaware of their standing in the game. We corrected this in future versions of the software by placing a vertical progress bar along the side of every student's quadrant, rather than in a communal space.

Finally, we noted a high level of teacher interest — both in their students’ performance in the game, as well as in the game itself. The teachers would often stand by as their students played. While we attempted to stop them if they were clearly helping their students with math, we generally let them hover, as this is how the software would likely be run in real deployments. Afterwards, we interviewed the teachers, drawing some interesting reactions and results. One teacher expressed surprise at learning how poorly one particular student of hers was performing at subtraction, and generally teachers were interested in being able to assess their students through the software.

There was universally more excitement, however, about the content in the application. Many teachers expressed an interest in an expanded curriculum involving English learning and science subjects, but there was also much interest in the possibility for them to generate their own content for the platform. As this had been one of our development goals from the beginning, it was encouraging to hear this.

This being our first field trial of this nature, we also learned many valuable lessons from this first trip. We grew increasingly suspicious of our interview results over the course of the study, and eventually determined that in the future we would need to fool students while interviewing them.
by asking around the question, or else they would simply guess at what we wanted to hear. We were also baffled at the skill level of the students we were sent, as it seemed like the second graders were no less skilled at math than the fifth graders. This was a phenomenon brought about by the fact the teachers were selecting which students to send out to us, and so unsurprisingly they selected only the cream of the crop so that the results reflected favorably upon them.

We also found that it was difficult to draw conclusive results about interaction patterns from the notes we took, as they were wholly qualitative in nature. We would need a way to quantitatively take live observation notes if we wanted to continue to push our research along.

**STUDY TWO: DESIGN AND INTERACTION**

We next wanted to determine how best to design the games that would go into the platform themselves, and so we turned our attention towards how best to achieve better learning results. As a primary advantage of our platform, particularly relative to solutions in the developed world, was the ability to encourage students to work together, we focused on bringing about collaborative learning behaviors through the use of interaction design.

**Introduction**

The significance of collaboration in learning is something that has been studied extensively [3, 8]. In the past, some existing single-display groupware systems would encourage rather than enforce collaboration, by providing tools that would combine to become more powerful when used by multiple students [2], while other systems simply presented students with a single, shared problem to solve [10]. The common thread between these systems is that they rely upon a shared workspace and problem to function — the students would be painting upon the same canvas, or filling in parts of the same answer. Because of our split-screen approach, and the keypad’s lack of physical presence upon the screen, it was much more difficult to apply their results to our model of interaction with any confidence of their effects. Thus, we decided to begin much more simply.

**Study Design**

**Interface Conditions**

Based on our goal of exploring collaboration within MultiLearn, we designed five versions of a math-based educational game for the platform. As with our previous study, in all versions of the game, the system presented students with a series of single arithmetic questions (addition,
subtraction, or multiplication), which students answered by entering a numeric value on a keypad. For the competitive game, each round ended once any student correctly answered 12 questions. Game rounds in the collaborative conditions ended once either team correctly answered 24 questions in total. Unless otherwise stated, each student had his/her own keypad to use. Question difficulty was again adapted to each student based on his/her individual performance. This skill rating of students was not preserved between conditions.

The experimental conditions supported four students at once, and differed as follows:

No collaboration (M4) — As a baseline as well as to introduce students to the concept of the game, we ran this purely competitive and individual mode.

Implicit collaboration (MI) — Students worked in two teams of two. Team members’ individual scores summed to form the score for the team. The team nature of the game was stressed to the students, but they were not explicitly told to collaborate. Since a student could help if their teammate was stuck on a problem, but were not forced by design to do so, we considered this condition to offer implicit collaboration.

Explicit collaboration (ME) — Students were in teams of two, and each team earned a point only when both members correctly answered a question. To reinforce this behavior, the first student to answer correctly was presented with an arrow pointing to his/her teammate’s problem rather than the next question.

Physical collaboration (M2 and M2V) — Each team of students worked on a single question. In M2, the two students shared one keypad, which allowed us to study whether having students physically share a single point of input and problem would encourage constructive behavior. In M2V, both teammates were given active keypads, both of which could be used to input the answer. We introduced this condition to see if it would affect the behavior of the dominant student.
Methodology
The study took place in five public schools in low-income neighborhoods of Bangalore, in August and September of 2009. Applying what he had learned from our first study, this time students were randomly selected from classrooms in groups of four by our study facilitators to mitigate both performance effects from teacher self-selection, and to randomize the social effects that arise when close friends use the software together. Upon entering the study room, they were asked to sit at the computer, whereupon the software would be explained to them in the local language.

Students were always presented with the competitive mode first to familiarize them with the game — the other modes were counterbalanced in order. As each game was explained and played, students were observed for engagement, collaboration, and dominance behaviors, such as using another student's keypad or asking questions of other students. In addition, the software once again automatically logged details about every game session so that we are able to reconstruct and analyze each game as it originally occurred. Once the mechanics of the game itself and each condition was explained to the students, we refrained from interacting with the students unless they were struggling to understand the input.

In addition, we asked teachers separately to rate the students in general class performance. We used this rating to help interpret collaborative and dominance patterns.

Results: Dominance
The data discussed here is from observations of 192 4th, 5th, and 6th grade students, with a total of 8976 responses recorded within the system. For exploratory and logistical reasons, MI was replaced midway through the study with M2V. As a result, data from those conditions is incomplete and we only include M4, ME, and M2 in the following analysis. All statistical analysis was done via repeated measures ANOVAs, with a Greenhouse-Geisser adjustment for non-spherical data. For posthoc pairwise comparisons, a Bonferroni adjustment was applied.

Design Effects
The impact of the study conditions on behaviors related to dominance is summarized in Table 1. Due to logistical constraints, observations were coded in situ by two trained observers. Rules for rating were agreed upon beforehand for measuring the metrics, but otherwise no interrater control was instituted. Since the number of questions per round was not constant, we normalized the total number of instances of each observed behavior based on the number of in-game questions. These figures were then multiplied by 100 for readability.
Use of Another Student’s Keypad

The first metric, physical dominance, refers to the use of another student’s keypad. In the case of M2, we defined this metric as the seizure of the input device and unwillingness to share despite requests from the teammate. This is the most clearly defined of our metrics in terms of dominance — in every instance where we observed this occurring, the student whose keypad was taken away disengaged immediately, not endeavoring to continue work on the problem or question the dominant student. In many cases, it led to the student gradually becoming less engaged over the course of the study, as they accepted the dominance of the other student. Statistical analysis shows that there is an effect of interface condition on instances of physical dominance (F(1.42,55.5) = 34.3, p < .001), and pairwise comparisons showed that significantly more instances occurred in ME mode than M4 or M2 (p < .001 for both comparisons).

The most likely cause for this difference is that, while the stronger student is encouraged to help the weaker one in the explicit collaboration mode (ME), the weaker student still has ultimate control over the input device, and thus the situation — this allows them to stall until they are satisfied with their answer. Dominant students would often lose patience, and attempt to physically rectify the situation. In M2, the weaker student was not empowered by their own input and question, so there was much less need and opportunity for the stronger student to steal input from a teammate.

It should also be noted that the introduction of the shared keypad in M2 brought about a new set of behaviors, wherein students would take turns with the keypad. It is possible that introducing the shared device suddenly triggered existing social expectations that the students be kind to one another and share the input device where possible.

Disengagement

The next metric measures the disengagement level of students. Disengagement was defined as students ceasing participation in the game, generally looking away from the screen and no longer

<table>
<thead>
<tr>
<th>Game Type</th>
<th>Physical dominance</th>
<th>Disengagement</th>
<th>Verbal abuse</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>M4</td>
<td>0.990</td>
<td>2.59</td>
<td>1.74</td>
</tr>
<tr>
<td>ME</td>
<td>5.25</td>
<td>4.73</td>
<td>2.95</td>
</tr>
<tr>
<td>M2</td>
<td>--</td>
<td>--</td>
<td>2.95</td>
</tr>
<tr>
<td>MI</td>
<td>0.878</td>
<td>2.00</td>
<td>2.52</td>
</tr>
<tr>
<td>M2V</td>
<td>1.95</td>
<td>2.57</td>
<td>3.87</td>
</tr>
</tbody>
</table>
answering questions, often until another student intervened. While disengagement may be the result of general disinterest in the game, all students were excited enough about the opportunity to participate in the study that this is an unlikely cause. Instead, disengagement in the paired conditions was likely the result of either the stronger student waiting on the other student to finish before being allowed to continue, or the weaker student waiting for the stronger student to help and possibly finish both of their problems. Both cases are again examples of poor dynamics and a potential imbalance, though the former, which occurred exclusively in ME mode, is less overtly negative than the latter.

Although the analysis on observed instances of disengagement behaviour is not statistically significant, the trend indicates that the team modes (ME and M2) produced more disengagement than the competitive mode (M4). This interpretation is supported by our qualitative observations: the most severe case of disengagement came in the modes where both teammates were solving a single problem (M2 and M2V), since the stronger student sometimes simply answered all questions before the weaker student could hope to make progress, resulting in the weaker student gradually losing interest in the game.

Verbal Abuse

The verbal abuse metric measures the frequency with which one student scolded another. While not as direct as physically using another student's keypad, verbal abuse is still an indicator that one member of the team is working poorly with the other. There was a significant main effect of interface condition on observed instances of verbal abuse \((F(1.53,59.6) = 10.9, p < .001)\). Pairwise comparisons showed that the conditions that used team scoring (M2 and ME) resulted in more instances of verbal abuse than the individual M4 condition (both comparisons \(p < .001)\). Informally, it appears that the frequency was higher in M2 than ME. The most common case of verbal abuse in M2 was when the weaker student was in possession of the keypad, either because it was convenient or because the students were taking turns, and answered a question incorrectly. It is possible that the occurrence of verbal abuse was lower in ME because of the previously mentioned effect of ‘ownership’ of a problem — in M2, a poor answer reflected upon both students,
whereas in ME it was the other student’s responsibility to supply the correct answer.

Environmental Effects
To understand how individual characteristics of the students themselves might relate to dominance patterns, we also conducted an exploratory analysis on dominance patterns and the gender and the teacher-rated proficiency of each student. This analysis is preliminary and will need to be further validated.

For each student, aggression and deference scores were calculated across all conditions based on raw observation notes (see Tables 2 and 3 for a breakdown of scores). For aggression, a student was rated 0 if no aggressive behaviour was noted, 1 if the student had verbally ordered a teammate around or dictated answers, and 2 if the student was physically using their teammate’s input device on a consistent basis. Similarly, for deference, students were rated 0 if no deferent behaviour was exhibited, 1 for passive deference, most specifically waiting for a teammate’s confirmation of each answer before submitting, or 2 if the student actively sought the other teammate to answer the question for them so that the team could move on. These two metrics are not directly linked, since simply being bullied by a teammate does not raise your deference rating, and vice-versa.

We found that male students had higher rankings in both aggression and deference. This is perhaps explainable by the observation that male students tended to be more competitive — they were more emotional about winning or losing the game, for instance. This difference largely explains both metrics — aggression, and thus dominance patterns, tended to appear more often with male students than female students due to their greater interest in competition and thus impatience with teammates. They were also more willing to defer to a stronger teammate and lose their involvement in the game as long as it meant that they would win in the end.

The breakdown by class performance also yields interesting results: more advanced students appear to be far more aggressive than the weaker students. There were some instances of students

Table 2: Dominance metrics by gender. N=167.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
</tr>
<tr>
<td>Aggression</td>
<td>0.686</td>
<td>0.877</td>
</tr>
<tr>
<td>Deference</td>
<td>0.214</td>
<td>0.508</td>
</tr>
</tbody>
</table>

Table 3: Dominance metrics by class performance. N=167.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Strong</th>
<th>Average</th>
<th>Weak</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggression</td>
<td>Mean 0.884</td>
<td>0.571</td>
<td>0.500</td>
</tr>
<tr>
<td></td>
<td>SD 0.868</td>
<td>0.739</td>
<td>0.622</td>
</tr>
<tr>
<td>Deference</td>
<td>Mean 0.192</td>
<td>0.143</td>
<td>0.250</td>
</tr>
<tr>
<td></td>
<td>SD 0.430</td>
<td>0.430</td>
<td>0.623</td>
</tr>
</tbody>
</table>
who were verbally and physically dominant who were actually weaker at math than their teammate, but those students were generally rare occurrences caused by pairing with abnormally timid yet strong students. Weaker students were more likely to defer to others, looking for help while playing. Extremely weak students would receive help even from the other team, as the students were generally aware of each others’ skills.

**Results: Collaboration**

While interaction design can influence the type and severity of dominance, it can also influence the amount of collaboration between students. To look at these effects, we examined a number of observed indicators between the individual (M4) and shared keypad (M2) modes (Table 4).

<table>
<thead>
<tr>
<th>Game Type</th>
<th>Accuracy</th>
<th>Completion Time (seconds)</th>
<th>Collaboration Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>M4</td>
<td>0.60</td>
<td>0.3</td>
<td>26.3</td>
</tr>
<tr>
<td>M2</td>
<td>0.66</td>
<td>0.3</td>
<td>18.2</td>
</tr>
</tbody>
</table>

Due to the fact that M4 was always presented before M2, part of the effect here must be attributed to familiarity with the system. However, the trends are encouraging.

The higher accuracy rate indicates that students are in fact working together, as it is likely the result of students checking each others’ answers. One pattern that we frequently observed emerging was that students would confirm their answer with their teammate before moving on. This may also contribute to the faster completion time when the students worked together. Not all of this can be attributed to improved dynamics, however: cases definitely existed wherein a stronger student would singlehandedly account for the improvement, leaving his/her teammate without interaction with the software.

The collaboration instances metric measured the number of times students physically collaborated on a given problem. This included behavior such as talking about the problem, or sharing hands to count on fingers. While this type of behavior would sometimes occur in the competitive mode, usually as a result of the process of teaching others how to play the game, it was much more prevalent when the design of the game enforced such behavior.

Our observations showed other interesting effects that come with team play. We would observe, for example, a student who was asking for help and being ignored in the competitive mode suddenly
become a valuable asset, if for nothing other than his fingers, in the collaborative mode. With this new dynamic, his teammate was much more receptive to his comments and questions.

Conclusions

Over the course of our interaction study, we found that while all the collaborative modes had some amount of positive effect, it was difficult to balance that with the negative impact of the dominance that arose. We found that the more severely we attempted to enforce collaborative behavior, the more dominant students would feel restrained by the system or held back by their teammate, and in the course of correcting the situation, reduce the learning capability of said teammate.

One curious but potentially far-reaching result to note is that the condition where we physically forced the two students to work together at a single point of entry on a single problem (M2) seemed to produce very positive collaborative results without causing as much negative dominance behavior as the modes where students retained their own keypads. This challenges somewhat the fundamental assumption behind much of multiple input education research, which lies in the belief that no matter what, providing every student with one-to-one interaction with the computer is an absolute necessity for best results. While one does lose the ability for the software to track specific students, it is also possible that there is a balance point between having too many and too few students on a single input device. With such a model, the system gains the social expectations previously mentioned, which compels the students to share with one another. Were we to take such a path, group modeling could be one way to mitigate the loss of individual results.

In all, our interaction design study revealed a lot of strategies that work somewhat well, but have significant drawbacks as well. This is likely due to the simplistic nature of our approach — something which was intentional given the lack of prior work in the domain of collaboration across split-screen interfaces. However, despite the shortcomings of our collaborative strategies, we felt that we had learned enough to comfortably proceed to a larger topic of study.

STUDY THREE: LEARNING OUTCOMES

With our final study to date, we focused exclusively on the big picture, and on the ultimate goal: learning outcomes. While the previous results we had obtained were useful from the perspective of leveraging MultiLearn as a platform from which to study related topics, we wanted to show that our holistic approach to an education system was a potent one — that is, that the combination between
multiple input and associated collaboration, as well as the adaptive nature of our software and content actually make for an effective learning platform.

Introduction

We first turned to Emma Brunskill, a Postdoctoral Fellow with Berkeley, to help improve our adaptive questioning system. Using partially observable Markov decision processes (POMDPs) — modeling the student’s skill level as the hidden state, prompted questions as actions, and student responses as observations — we were able to make much more precise judgments about the specific skill level of each student at the component tasks of solving a problem (Figure 6).

Methodology

The study shared many of the same component parts with our previous work: it involved 121 fourth and fifth grade students from two schools in Bangalore, India. Over its course, we performed both quantitative and qualitative observation of interaction patterns and behavior, as well as logged all information about the students so that we could reconstruct each round afterwards. However, given our focus on learning outcomes, a few critical changes were introduced.

Each student in the study used the software for four days, 30 minutes a day. Due to the high rate of absences in public schools in India, we anticipated that most students would at least participate for three of those days. Any student who participated in fewer days would be excluded from the results. We ran written pre- and post-tests to compare student performance after the use of our system.
Two versions of the written test were given; students were randomly assigned one of the two tests for the pretest, and given the opposite version for the posttest. This was done both to prevent cheating, and to provide a comparison to ensure that the pre- and post- tests were equal in difficulty.

Unlike previous studies, we did a between- rather than within-subjects study. We had three conditions in total: a control group which played an unrelated MultiLearn spelling activity throughout (control condition), a group with played math games within the context of MultiLearn but with randomly rather than adaptively selected arithmetic questions (non-adaptive condition), and a final group which represented the whole of the MultiLearn system, with arithmetic questions coupled with adaptive questioning (adaptive condition).

Table 5: Mean post-pretest score improvement for students across conditions and grades

<table>
<thead>
<tr>
<th>Condition</th>
<th>Control</th>
<th>Non-adaptive</th>
<th>Adaptive</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>SD</td>
<td>Mean</td>
</tr>
<tr>
<td>4th grade</td>
<td>0.070</td>
<td>0.1</td>
<td>0.050</td>
</tr>
<tr>
<td>5th grade</td>
<td>0.047</td>
<td>0.2</td>
<td>0.065</td>
</tr>
</tbody>
</table>

Results

We performed repeated-measures unbalanced ANOVA tests to determine if any conditions improved significantly over the others. We used school, grade, and condition as the between-subjects factors, and the pre- and post-test comparison as the repeated factor. There was a significant main effect on the test, indicating that students in general performed better after using our software (F(1,97)=19.023, p < 0.001).

There was no significant interaction between condition and test, indicating there was no systematic difference between the change in performance among students in the three conditions. There was a significant interaction between condition, grade, and test (F(2,97)=3.623, p < 0.05); however, using the Tukey’s honestly significant difference criterion to examine the pairwise mean comparisons did not yield any significant results. This could be due to a number of reasons, most significantly the short duration of our intervention.

One aspect of our software which we could analyze regardless of student learning, however, was the success of our adaptive system in predicting student performance. Success here means that though we have not yet advanced to the stage of successfully teaching students, we are at least accurately judging the skill of the student. We generated a predicted posttest score for each student
based on our adaptive model and our knowledge of that student, and ran a Pearson correlation between those scores and the actual posttest scores. The correlation was \( r=0.56 \), which suggests a reasonable level of correlation. We then computed the L1 distance between the predicted and actual scores; this showed that almost half (47.6%) of all predicted posttests were within 3 differences from the true posttest score (out of 16 questions). It is reasonable, then, to conclude that our adaptive system is very successfully judging the skill of each student.

CONCLUSIONS

Over the course of our studies, we have shown that the keypad is a viable method of input, that it is possible to encourage collaborative behavior between students with game design, and that POMDPs can be used as a reasonable model by which to build an intelligent tutoring system. However, we have run into many limitations along the way, both in our studies as well as in our system itself. These are issues that we hope to rectify as we continue to look at improving MultiLearn.

Future Work

While we were able to encourage collaboration among students, we also generated a fair bit of dominance behavior as a result. This is likely due to the deliberately primitive approach we took to collaboration. Literature in the area typically relies on approaches such as jigsaw designs, where each student contributes to a larger, shared problem by completing a smaller portion of it. This is difficult to do with our interaction model, which offers no shared space, but it is a better-informed approach than what we currently have. We were able to show that our adaptive system is reasonably good at predicting student performance on any given problem, but we were ultimately unable to show that as a whole the system improves learning. There are quite a few reasons that this could be the case. The one that likely had the greatest effect was the short duration of our intervention. 2 hours with any sort of computer software simply won’t benefit the

Figure 8: Students using MultiLearn
majority of students sufficiently to obtain significant results. We hope to do a much longer deployment and subsequent intervention in the near future.

Before we do so, however, there are more issues we must address. The system is currently purely drill-based. While drill will improve student skills in the case that they already have an understanding of the concepts, it does nothing to actually teach the students new material, or to identify and correct specific problems in students’ processes. While it is possible to add such aspects to the policy calculation that underlies our new adaptive questioning system, there are interaction issues that must be solved before we can attempt such interventions. Under the current model of play, for instance, students tend to answer questions as quickly as possible under the logic that this will earn them the maximum number of points. Such a model is problematic for tutorial-type interventions, as students would simply click through anything that doesn’t directly earn them points; we will have to take a much more subtle approach to presenting such interventions.

Another space we are looking at is in expanding our learning activities to be true games, with a set of interactions that masks the learning aspect from the student. As it is, even with the addition of real tutoring, our system still looks and interacts like a simple drill exercise.

Both of these final tasks are things that will require the assistance of people more familiar with education and learning sciences, and this strongly reflects the direction we feel the project needs to take in the immediate future. We have made many strides in the realm of the technology, the algorithms, and the interaction design, but ultimately our aim is to teach.

ACKNOWLEDGEMENTS

This work would not have been possible without the invaluable assistance, support, and contributions of Richard Anderson, Joyojeet Pal, Sunil Garg, Heather Underwood, Charlotte Robinson, Leah Findlater, Emma Brunskill, Meera Lakshmanan, the UW Department of Computer Science and Engineering, the UW Information School, the UW Undergraduate Research Program, Microsoft Research India, the Azim Premji Foundation, and the National Science Foundation.
REFERENCES


