Modes and Mechanisms of Game-like Interventions in Intelligent Tutoring Systems

by

Dovan Rai

A Dissertation

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

in partial fulfillment of the requirements for the

Degree of Doctor of Philosophy

in

Computer Science

April 2016

APPROVED:

_____	_____
Professor Joseph E. Beck Advisor – WPI	Professor Ivon Arroyo Co-Advisor - WPI
_____	_____
Professor Charles Rich Committee Member - WPI	Dr. Kristen DiCerbo External Committee Member - Pearson

DEDICATION

To my parents

ACKNOWLEDGEMENT

I would like to thank the members of my committee for their guidance and support. I am deeply grateful to my advisors Joseph E. Beck and Ivon Arroyo for mentoring me throughout these years and giving me continuous support and encouragement during all phases of my research. I would also like to thank Charles Rich and Kristen DiCerbo for their insightful appraisal and invaluable feedback.

During my dissertation period, I feel fortunate to have worked with and received guidance from different faculty members. I am thankful to Neil Heffernan, Ryan Baker, Janice Gobert, Matthew Kam, Beverly Woolf and Tom Murray. I also received tremendous help from my lab-mates and colleagues who not only helped me with my research work but also enriched my days as a graduate student. Heartfelt thanks to my friends Naomi Wixon, Michael Sao Pedro, Yue Gong, Yutao Wang, Zach Broderick, Adam Goldstein along with various Assistment and Mathspring team members.

Thanks to my family, friends and relatives for their love and support. I would particularly like to express my gratitude to my friends and family members that have helped me during my stay in United States. Thanks to Deepa Rai, Sumugdha Rayamajhi, Dan Cooper, Shila Pradhan, Chama Rai, Sazu Rai and many others.

My thanks also goes to Fulbright Commission for funding me for the three years of my stay in United States and other funding agencies and WPI graduate school for providing me with the funding and logistics in successfully completing my dissertation.

ABSTRACT

While games can be an innovative and a highly promising approach to education, creating effective educational games is a challenge. It requires effectively integrating educational content with game attributes and aligning cognitive and affective outcomes, which can be in conflict with each other. Intelligent Tutoring Systems (ITS), on the other hand, have proven to be effective learning environments that are conducive to strong learning outcomes. Direct comparisons between tutoring systems and educational games have found digital tutors to be more effective at producing learning gains. However, tutoring systems have had difficulties in maintaining students' interest and engagement for long periods of time, which limits their ability to generate learning in the long-term. Given the complementary benefits of games and digital tutors, there has been considerable effort to combine these two fields.

This dissertation undertakes and analyzes three different ways of integrating Intelligent Tutoring Systems and digital games. We created three game-like systems with cognition, metacognition and affect as their primary target and mode of intervention. Monkey's Revenge is a game-like math tutor that offers cognitive tutoring in a game-like environment. The Learning Dashboard is a game-like metacognitive support tool for students using Mathspring, an ITS. Mosaic comprises a series of mini-math games that pop-up within Mathspring to enhance students' affect.

The methodology consisted of multiple randomized controlled studies ran to evaluate each of these three interventions, attempting to understand their effect on students'

performance, affect and perception of the intervention and the system that embeds it. Further, we used causal modeling to further explore mechanisms of action, the inter-relationships between student's incoming characteristics and predispositions, their mechanisms of interaction with the tutor, and the ultimate learning outcomes and perceptions of the learning experience.

CONTENTS

ACKNOWLEDGEMENT iii

LIST OF TABLES ix

LIST OF FIGURES x

1 Introduction 1

2 Background Research 6

2.1 Games and learning: techno-cultural landscape 6

2.2 Games and Learning: threads of academic research 12

2.3 Game: Affordances 18

2.4 Games: Constraints 29

2.5 Empirical evaluation of effectiveness of games in learning 33

2.6 Designing Educational Games 41

2.7 Effective integration of game design and instruction design 45

2.8 Game elements, Game mechanics and Gamification 47

2.9 Intelligent Tutoring Systems and Educational games 52

3 OUR APPROACH 55

3.1 Games as affective, cognitive and metacognitive tool 57

3.2 Web of associations and Causal Mechanisms 59

3.3 Research Questions 62

3.4 Description of three systems 66

3.4.1 Monkey's Revenge: Coordinate geometry learning environment 66

3.4.2 The Learning Dashboard (Student Progress Page) 78
3.4.3 Mosaic: Math mini-games 97

4 Experiments and Analysis 102

4.1 Experiments with Monkey's Revenge 102
4.1.1 Mily's World 103
4.1.2 Pilot study with Monkey's Revenge 107
4.1.3 Monkey's Revenge : Experiment Design 107
4.1.4 Randomized Controlled Study- I 114
4.1.5 Randomized Controlled Study- II 119
4.1.6 Conclusions, Limitations and Future Work 127
4.2 Experiments with Learning Dashboard (Student Progress Page) 130
4.3 Experiment with Mosaic 143

5 Causal Modeling 150

5.1 Causal Modeling of Monkey's Revenge: a case study in Causal Modeling 151
5.1.1 Causal modeling and correlation matrix 153
5.1.2 Causal structure, path orientation and domain knowledge 157
5.1.3 Causal modeling and multiple regression 161
5.1.4 Causal modeling: confirmatory, exploratory and graphical tool 164
5.2 Causal modeling: guide for new experiments 168
5.2.1 Causal Modeling of Wayang OutPost 168
5.2.2 Causal modeling with ASSISTments 179
5.3 Causal Modeling with Mathspring 183
5.3.2 Causal modeling of pre-survey variables 189
5.3.3 Causal modeling with within-tutor variables 193
5.3.4 Causal modeling of Pre-survey and within-tutor variables 202

5.3.5 Pre-Survey, within-tutor, Post-Survey variables 206
5.3.6 What do these causal models say about SPP? 210
5.4 Causal modeling with Mosaic 214
5.4.1 Causal modeling with pre-survey variables 219
5.4.2 Causal modeling with within-tutor variables 220
5.4.3 Causal modeling with pre-survey and within-tutor variables 222
5.4.4 Causal modeling with within-tutor and post-survey variables 225
5.4.5 Causal modeling with Pre-survey , within-tutor and post-survey variables 227
6 Conclusions and Implications 231
6.1 Limitations 235
6.2 Future work 239

LIST OF TABLES

Table 1 Student States 90

Table 2 Two modes of affective intervention 99

Table 3 Linear regression analysis 105

Table 4 Student response on pilot survey 107

Table 5 Four experimental tutor versions with different degree of game-likeness 109

Table 6 survey response on main study-I 117

Table 7 participants who logged in the tutor with pretest score 121

Table 8 Survey Responses across tutors (mean, SD and 95% CI) 122

Table 9 Retention of students in Tutor groups 123

Table 10 Learning gain across tutor groups (mean and 95% CI) 125

Table 11 Student performance across tutors 126

Table 12 Time overload across tutor conditions 126

Table 13 Students in different experimental groups self-report on their experience in Mathspring (mean and SD) 146

Table 14 Students who used Mosaic and who did not use Mosaic self-report on their experience in Mathspring 147

Table 15 Interest and Frustration averaged over participants who used and did not use Mosaic 148

Table 16 Interest and Frustration averaged over participants before and after using Mosaic 149

Table 17 student state variables 187

Table 18 mean and SD of student state variables across all students 194

Table 19 correlation among student state variables 194

Table 20 mean and SD of student state variables across gender 199

Table 21 Student state variables, Mean and SD 221

LIST OF FIGURES

Figure 1 Flow: A state of pleasant frustration—challenging but doable.....23

Figure 2 Zone of Proximal Development.....23

Figure 3 Instructional Effectiveness as degree of overlap between learning objectives and game attributes.....46

Figure 4 Screenshot of Monkey's Revenge.....68

Figure 5 Three possible tradeoff curves for making tutors more like games.....69

Figure 6 Students can help Mike decide name for the monkey.....74

Figure 7 Students can collect badges.....75

Figure 8 Immediate visual feedback for student responses.....75

Figure 9. Screenshot of Mathspring. Learning companions use gestures to offer advice and encouragement. Animations, videos and worked-out examples add to the spoken hints about the steps in a problem. My Progress button allows students to access the student progress page.....79

Figure 10 Math Tree on day 1.....82

Figure 11 The Math Tree: a visual representation of student performance in Mathspring.....82

Figure 12 The Student Progress Page (SPP) encourages students to reflect about their progress on each topic (column 1) and to make informed decision about future choices. The plant (column 2) demonstrates an assessment of student effort and the mastery bar (column 3) assesses students' knowledge. The tutor comments about student behavior (column 4) and offers students the choice to continue, review or challenge themselves (column 5).....84

Figure 13. Clicking a topic in the SPP produces a list of every problem in the system represented as a domino (center) and each problem is shown in detail (below). The topic detail in the SPP provides a view of every problem for each problem set that the student should go through.....86

Figure 14 Student state variables in knowledge-effort quadrants.....90

Figure 15 Sample diagram of how salient aspects of a student's performance are determined 91
Figure 16 Simplified version of the decision tree that delivers feedback based on a variety of student performance states at math problems....92
Figure 17 Effort on each topic is represented by a potted plant....93
Figure 18 rewards and loss....94
Figure 19 Rewards for good help usage. Students who show frequent help usage behaviors have a richer looking pepper plant.....95
Figure 20 Exceptional performance is rewarded with special pepper plants. 96
Figure 21: Solving area and perimeter of rectangular shapes generate the colorful rectangular tiles which fill the Manhattan mosaic. Behavioral momentum of solving the problems and generating the tiles is expected to be enjoyable. 100
Figure 22: When level 1 of mosaic is complete, players get to level 2. In level 2, players solve problems under fixed time. The faster they are, they can solve more problems and get more tiles and hence their mosaic is more intricate. 100
Figure 23: Students solve fraction problems to generate tiles completing London Mosaic. The problems increase on difficulty as students progress while maintaining a rhythmic momentum 101
Figure 24 Overview of experiments with our three interventions....102
Figure 25 Screenshot of Mily's World 104
Figure 26 Screenshot of Monkey's revenge with all game-like elements 110
Figure 27 Screenshot of tutor version without narrative 111
Figure 28 Screenshot of Monkey's Revenge with visual feedback 112
Figure 29 Screenshot of Monkey's Revenge without visual feedback 112
Figure 30 screenshot of Basic tutor 113
Figure 31 Sample of Survey questions....116
Figure 32 Sample of test questions....117

Figure 33 Screenshot of tutorial 120
Figure 34 Students in the experimental condition were offered to see the progress page when they reported low levels of excitement or low levels of interest (boredom). 132
Figure 35 Experimental setup for SPP study 133
Figure 36 Visual representation of the high-level path models for excitement in the no-button, prompt and force conditions from left to right, respectively 137
Figure 37 Miss Pepper is a cartoon character that helps explain different components and messages in SPP 141
Figure 38 Tutor intervention interacts with student learner characteristics to generate outcome variables 151
Figure 39 Causal model from PC algorithm without domain knowledge 154
Figure 40 Graph based on correlation matrix 155
Figure 41 Causal model with Domain knowledge 159
Figure 42 Block diagram of our Causal modeling process 164
Figure 43 Two possible causal models linking LikeMath and %Correct 166
Figure 44 Colors of edges in Doug's version of Tetrad are associated with the strength of the relationship between the two variables 171
Figure 45 Causal model of attitude and affect 173
Figure 46 Knowledge tracing model: Dynamic Bayesian network 180
Figure 47 Causal modeling of self-discipline survey response, performance and Knowledge tracing parameters: Assistment data 182
Figure 48, Causal modeling, pre-survey variables, Mathspring SPP data 189
Figure 49 Knowledge tiers, pre-survey variables, Mathspring data 190
Figure 50 causal modeling with knowledge tiers, pre-survey variables, Mathspring SPP data .. 191
Figure 51 two clusters in causal modeling with knowledge tiers, pre-survey variables, Mathspring SPP data 192

Figure 52 causal modeling, student state variables, Mathspring SPP data 194
Figure 53 causal modeling, Student State variables and affect variables, Mathspring SPP data. 195
Figure 54 Knowledge tiers, Student State variables and affect variables, Mathspring SPP data. 196
Figure 55 Causal modeling, Student State variables and affect variables with knowledge tiers that encourage correlational and causal links from variables towards the top to variables towards the bottom of the figure, for Mathspring SPP data..... 197
Figure 56 causal modeling, gender, Student State variables and affect variables, with knowledge tiers, Mathspring SPP data 198
Figure 57 Knowledge tiers, gender, student state variables, affect variables and test variables, Mathspring SPP data 200
Figure 58 causal modeling, gender, student state variables, affect variables and test variables using knowledge tiers , Mathspring SPP data 201
Figure 59 Knowledge tiers, pre-survey variables and within-tutor variables 203
Figure 60 pre-survey variables and within-tutor variables with knowledge tiers, Mathspring SPP data 204
Figure 61 causal modeling of Assistment data..... 206
Figure 62 knowledge tiers, pre-survey, within-tutor and post-survey variables, Mathspring SPP data 207
Figure 63 causal model with knowledge tiers, pre-survey, within-tutor and post-survey variables, Mathspring SPP data 208
Figure 64 Causal model with Knowledge tiers; Wayang Outpost data..... 209
Figure 65 screenshot of Mathspring SPP 211
Figure 66 Student state variables in knowledge-effort quadrants 212
Figure 67 Causal modeling, pre-Survey variables, Mosaic Data 220
Figure 68 Student State variables, Mosaic data..... 221
Figure 69 Student state and affective state variables within the tutoring session, Mosaic data... 221

Figure 70 Knowledge tiers, pre-survey and within-tutor variables, Mosaic data 223
Figure 71 Causal modeling with knowledge tiers, pre-survey and within-tutor variables, Mosaic data 224
Figure 72 Knowledge tiers, within-tutor and post-survey variables, Mosaic data.....226
Figure 73 Graphical model with knowledge tiers, within-tutor and post-survey variables, Mosaic data 226
Figure 74 Knowledge tiers, pre-survey, within tutor and post-survey variables, Mosaic data 229
Figure 75 Causal modeling with Knowledge tiers, pre-survey, within tutor and post-survey variables, Mosaic data 229

1 Introduction

Educating the younger generation is a core responsibility of our society. Formal institutions of education are now considered a universal right, and making education accessible and inclusive is one of our common goals as members of a civil society. However, as of today, more than 72 million children do not have access to basic education; even among the students who are enrolled, a large fraction of them do not feel included but rather disinterested and alienated, resulting in huge dropout rates. Moreover, we cannot guarantee that the students who are well adjusted and flourishing in our existing educational system today are well prepared to take on the challenges of the future.

The form of education has changed over human civilization based on the priorities and structures of a society and the existing technologies at the moment. The educational system we have today has been criticized as a vestige of the industrial age, a one-size-fits-all approach to create homogenized employees for structured jobs (Robinson & Aronica 2015). The current educational system in our modern societies seems to be inadequate to prepare children for new challenges of the information age.

There is another challenge imposed on the education system due to the changing technology: digital media and entertainment have been prolific, distracting students away from their schools. Digital games, in particular, have been very ubiquitous and effective

at holding young children's attention to the point of being addictive. This phenomenon is disconcerting to parents and teachers but at the same is inspiring to a community of education researchers. It almost cries for opportunism, asking us to answer how can we emulate these games to create engagement in education settings as well.

Games seem to bring their own learning paradigm along within them. Learning is a central aspect of games, as players need to learn to play those games and level up to higher difficulty. Thorough speculations even revealed that good games incorporate good learning principles (Gee, 2007). On these regards, creating educational games seems almost an obvious choice. However, the history and reality of educational games have been rather bumpy, many times giving the word 'educational game' itself a bad reputation among students.

Creating good educational games has been exceptionally hard. There is limited empirical success with educational games. The conundrum of this situation where educational games seem intuitively appealing but deliver low rate of real success brings up the debate on 'learning' vs. 'play'.

Learning and play are two fundamental human activities. They seem complementary overlapping entities and at the same time appear to be on the opposite ends of a dichotomy. Learning evokes fun as well as effort. Children are always learning yet also tend to resist to formal structures of learning. On the surface, play might appear to be an intellectually passive activity, but it is during play that children are most receptive and

willing to put their best efforts. There is even an argument that play has an evolutionary utility towards learning (Pellegrini et al., 2007). If we could do all our learning though play or through a playful activity, that would be a utopia of learning. However, we are required to do complex learning, which requires considerable mental effort, focus and persistence. Play in this context can be helpful or a hindrance. Play incorporates a sense of autonomy and fantasy, which makes it enjoyable. Learning, to be effective, may require for it to be structured, and fantasy can be distracting.

'Limited working memory' is a bottleneck on human learning (Sweller, 1994). Working memory is related to an information-processing approach to the mind, which implies there is limited capacity of memory in our "thinking buffer" in particular. This is a major hurdle when we try to enrich learning by adding engaging material to the learning content, as extraneous information and details may distract and interfere with the main learning content. The use of multimedia in learning faces this challenge and so do discovery learning and other exploratory constructivist and constructionist approaches. Educational games are also trapped in this situation where these ambitious learning approaches fail due to the fact that, even though elements such as novelty, fantasy and discovery are very appealing, they may overwhelm a learner's working memory.

While these learning paradigms struggle with each other, there is another development in digital learning that has been able to deliver impressive learning results. Intelligent Tutoring Systems (ITS) are computer tutors that aim to give customized and adaptive instructions and feedback to students. They have been successful at generating

statistically significant learning gains, and results comparable to one-on-one human tutors (Koedinger & Corbett, 2006). While this makes ITS very promising, ITS on the other hand struggle to keep students engaged over extended periods of time. ITS researchers have been primarily focused on cognitive aspects of learning. But they have realized that affective and motivational aspects are as important as cognitive aspects. They have ventured to incorporate different affective and motivational components to make ITS more robust. Taking surveys and detecting students' emotions and using affective learning companions in tutors are some examples in this direction. Some ITS researchers have been exploring games as well to understand what makes them good at engagement and possibilities to incorporate them within ITS framework.

I belong to the group of researchers who are studying both games and ITSs for their complementary strengths. In this dissertation, I explore the possibility of creating software learning environments that are as engaging as a game and yet can produce quantifiable learning gains. There are some examples of successful intelligent educational games, but these are still very few and far in between and require considerable resources to implement. Taking resource constraint also into consideration, I am taking a very cautious and strategic approach to merging games and tutors, by deconstructing games into game elements. I have chosen three different ways to merge games and tutoring systems, defining three different paradigms to their integration that consider games as cognitive tools, as metacognitive tools and as affective tools. My research involves observing how these interventions change and affect students' interactions and engagement with the tutoring system, as well as cognitive and affective outcomes and

mechanisms of action, trying to understand dependencies among all of these constructs and how they influence each other.

2 Background Research

2.1 Games and learning: techno-cultural landscape

This is an exciting time for digital games, education technology and digital games in education. Digital devices are ubiquitous and so are digital games. Games have emerged with new possibilities taking new intellectual and imaginative, social and physical spaces and appealing to broader populations. While games are primarily about entertainment, vision, design and usage of games have transcended the entertainment sphere and moved towards serious applications in different spheres of people's life, producing the whole new genre of 'serious games'.

Similarly, there has been a revolution in the world of education. As computer technologies are getting more prevalent in classrooms and homes, we see a plethora of innovative possibilities. While some claim that we have moved beyond from the industrial model of education and we need to completely remodel our whole education system, others are trying to use the computers to make education accessible for those who have been denied the existing mode of education, however obsolete. And interestingly, some are trying to do both at the same time. We are stretching the possibilities at both ends, accessibility and creative innovation. Learning is a rewarding activity, but it is not always easy and definitely not for everyone. For every competent motivated learner, there

is a struggling one and there are some who are so alienated that they no longer bother to struggle. It has been a continuous struggle of education community to bring those alienated students back. The education community has worked on creating multiple theories and devising new techniques and incorporating new mediums. Digital games are one of such promising approaches. There are primarily two major factors that make games attractive. The first obvious one is the prospect of games adding fun to learning. The second appeal of games is based on the belief that games are not only vehicles of fun but also constitute superior learning tools.

In fact, the fascination of using games in education is not new. It was a bubble that has come and gone. The 'Edutainment' era of the 1990s had produced some successful titles such as 'Math Blaster' and some acclaimed ones such as 'Oregon Trail', with most titles unused and forgotten. In fact, 'Math Blaster' has been an example of narrow and misleading approach of using games in education, that most learning game designers today tend to shy away. This is also known as 'chocolate on broccoli' approach, where irrelevant 'fun' material is extrinsically added on top of learning content. The learning games community today claims that this approach was based on a limited and superficial understanding of human learning and games. The research community claims to integrate learning theories and game studies to gain deeper insights, innovative designs and more effective implementations, and get things right this time. In fact, it is quite reasonable to be optimistic about these claims. Much has changed now since the 1990s. First, it is not only some educational game companies that create the learning games. Learning games is now a large worldwide community of educators, researchers, designers, companies and

individuals. They are not only exploring wide arrays of games such as Massively multiplayer online role-playing games (MMORPGs), casual games and social games, but are also creating new genres of games. Meanwhile, the learning sciences as a discipline has progressed significantly and constructivism has been a leading pedagogical paradigm. With new tools and technology, creating games has become much simpler, and there are authoring tools that enable learners to create games themselves. Innovation in learning games has accelerated significantly in this interconnected digital age. While there is a lot to be excited about, there are also reasons to be cautioned as well. As researchers, it is our duty to maintain our skepticism. The learning sciences is a growing field, but still a relatively new one. We as a community are still exploring ways to find accurate measurements of something as elusive as learning in real classrooms where noise is the norm. We are also often dealing with young children, who are very vulnerable to our interventions. Therefore, we need to equip ourselves with rational skepticism and restraint along with optimistic excitement.

Interest in academia and acceptance in schools

Using digital games in education is one of the contested issues. There is a mixed attitude among parents and teachers towards digital games. On one hand, digital games are seen as distraction and nuisance. On the other hand, there is a growing interest among parents and teachers towards using digital games as conceptions and priorities of education are changing. On a national survey of 500 teachers who use digital games (Millstone, 2012) , the majority of teachers gave favorable views towards games.

Education researchers are also divided with regards to using games. A lot of academics were dismissive of games and those championing games were a minority on the defensive. However, games as an academic subject has started to gain more popularity and prominence.

Enthusiasm and results

Creating games is a significantly big endeavor. It requires massive resources, money, time and effort. Creating games is not only resource intensive but also highly risky. Creating good games is a very challenging task. Online statistics say that only 4% of games make money and only 20% of games that go to the store shelves make profit (EEDAR, 2008). If creating general games is this difficult, we can easily guess how difficult it would be to create educational games. Most of the learning games were not popular among students. Ted Castranova, in an article in WIRED (Baker, 2008), has been forthright in his failure to create an educational game. His game 'Arden' modeled after a 3D game, looked like a game, but was not able to create 'fun' experience. While it is already a huge struggle to create a 'fun' experiences in 'just' games, it can become even more challenging to create such an experience in educational games. To create learning experiences is yet another challenge. Therefore, the learning games community need to take great caution and study the constraints before they undertake such a risky endeavor.

There are many failed educational games and there are several reasons for this. A game has several aspects that need to go right and together. Creating a game is an interdisciplinary process, and creating educational games requires even more disparate

teams to work together. Content design and game design are challenges in themselves and integrating them is even a bigger challenge. Ideally, subject matter experts and game designers should come together or there is an incredible mix of talent and knowledge of both in one individual or one team. Very often, it is content designers who want to design games as well and it is game designers who want to figure out content as well. When there is lack of appropriate knowledge and skills, we are left with poorly designed and poorly executed systems. We often hear disgruntles among game designers that most of the so called "educational games" have been designed by non-gamers and people do not know the basics of game design. Educational games have therefore received bad reputation and field would benefit a lot if the 'real' game designers are in charge of creating such systems. This discontent has basis, given that most of the systems displayed as educational games are more like interactive quizzes with colorful rewards. The educational games that have been designed by 'serious' game designers are more comprehensive and effective.

Pervasiveness of games

"The growing presence of games in the lives of young people creates perils and possibilities. Games have been a constant source of criticism and alarm among parents, researchers, child advocacy groups, and elected officials. The potential harmful effects of gaming have been linked to society,as understandable concerns about the increasingly sedentary lifestyles of youth and childhood obesity, addiction, gender socialization, poor academic performance, and aggressive behavior. An area of growing concern is the role

of games in the learning experiences and environments of youth." (Everett & Watkins, 2007)

While the gaming community cheers the pervasiveness of digital games and see this as a signal of the importance of digital gaming in this new century (and even claim digital games to be the force for good) some parents, teachers and social theorists see this trend as a social problem. When digital games enter schools and replace books, some people are appalled and see this as a threat to our value system. "Are we trying to create a new generation that is addicted to 'fun'?" "Should we make young people expect to get immediate rewards and feedback all the time?" "Isn't 'delayed gratification' one of the most important values we need to teach the young?"

Games are primarily about action and they are not necessarily reflective mediums. Should people not need to learn to read and reflect before they jump into action and immediate rewards? Learning is an inherent part of games but we need to ask what kinds of learning are facilitated by games. Learning how to make a jump in game is different kind of learning than deep observation and reflection.

Another aspect we need to consider is that games generally employ simplistic dynamics. Since games are about immediate action and consequence, there are clear and crude dichotomies such as good and evil, safety and danger. Real world is full of ambiguities and so many times, the best thing to do is not to act but just observe, listen and understand. There are games that attempt to solve world problems such as hunger in parts of Asia and Africa. Sometimes kids first need to learn to listen and understand. The

action oriented simplistic world view of games can sometimes promote a simplistic perception and misleading simple solutions.

Progress and optimism

Our interactions are becoming more digital, making each action a digital unit. This allows for aggregation, labeling, evaluation and categorization of our actions and interactions. For instance, reading is no longer a solitary activity. We rate and share what we read. Magazines make lists of the 'most read', 'most liked' and 'most shared' articles. We are moving towards more interactivity, more information and more engagement. In a way, our interactions with the world are becoming more game-like. Similarly, games are incorporating more subjects and activities. Thus, we see two trends here: everything is becoming more game-like and games are becoming everything. This expansion and evolution could become a force for good. Players of the game 'Foldit' were able to solve the puzzle of AIDS virus in 10 days that had stumped scientists for 15 years. While we cannot and should not make everything game, there seems to be value in thinking as gamers do.

2.2 Games and Learning: threads of academic research

Education researchers have been studying and employing games in different ways:

- Exploring the educational value of commercial games
- Using commercial games for educational purposes

- Creating educational games
- Using game-based approaches in education
- Students creating their own games

Playing video games has been speculated to help players to develop cognitive skills such as visuo-spatial abilities, and help as a gateway to learning computers and technology. It has been observed that video game players also exhibit non-cognitive behaviors such as persistence and attention to detail that are desired but often missing in schools. Kurt Squire, at the University of Wisconsin, used a strategy game called Civilization in a high school world history class (Squire, 2004). Squire reported that the players mastered many historical facts and developed deep understanding about the intricate relationships involving geographical, historical, economic systems within and across civilizations. Squire has continued his research on exploring the potential of video game based technologies in education.

James Paul Gee, a linguistics and literacy researcher, who claims to have stumbled on games quite late in life, is one of the most persuasive and most cited researchers in learning games. He deconstructed the learning principles in video games and claimed that game designers have been able to hit on the learning principles that are crucial and desirable for any education system. He hails games as ideal learning tools and attribute this as video games being so popular among young children. Both Squire and Gee assume that 'situated cognition' is a cornerstone of good learning.

Researchers like Squire and Gee do not see games as a mere addition to school curriculum to make learning more fun, but rather see games as transformative vehicles to revolutionize learning. In fact, there has been a call for revolutionizing education by theorists and enthusiasts from other areas as well.

2.2.1.1 Games to foster new kinds of learning: 21st century skills and constructivism

One obvious application of games in learning is making boring repetitive aspects of learning exciting and engaging. The majority of the learning games that are available online fall into this category. Students need to practice knowledge and skills to gain mastery. The games encapsulate those learning activities within the fantasy of a game world. This approach can be a double-edged sword. When there is bad integration of learning content and game world, it can result in 'unfun' experiences, may seem exploitative and may even hinder intrinsic motivation. On the other hand, when integration is smooth, students may learn skills in a fun environment. Most of the edutainment-era designers were trying to crack this problem. But, the designers today are not stopping here. They are envisioning games as innovative mediums to foster new kinds of learning. They believe good games have innate affordances to promote and support such novel endeavors.

First, they assert that we have moved beyond the industrial age and education should address new challenges. In this new information age, the skills that we valued in the past

are obsolete and we should instead teach new competencies that are crucial to this age. Broadly named 21st century skills, those skills include problem solving, analytical thinking, systems-thinking, technological fluencies, the ethics of fair play, collaboration in cross-functional teams, and accessing knowledge networks. They claim that young people are practicing those skills in digital games that they play. The games are therefore effective vehicles to train young learners in those new competencies.

Second, games are perceived as active mediums to implement constructivist learning paradigm. Constructivism is a learning philosophy that is based on the principle that learners construct their own knowledge. Constructivism is a reaction to didactic approaches such as behaviorism and programmed instruction. Constructivism encourages discovery, hands-on, experiential, collaborative, project-based, and task-based learning. Constructivism places the learner at the center of learning process rather than the learning content. A learner is seen as self-directed, creative, and innovative. The purpose in education is to become creative and innovative through analysis, conceptualizations, and synthesis of prior experiences to create new knowledge. The educator's role is to mentor the learner during heuristic problem solving of ill-defined problems by enabling quested learning that may modify existing knowledge and allow for creation of new knowledge. Instructors are perceived as facilitators of the learning process and learning is an active social process.

Social constructivism, not only acknowledges the uniqueness and complexity of the learner, but actually encourages, utilizes and rewards it as an integral part of the learning process (Wertsch, 1997). From social constructivist viewpoint, it is thus important to take

into account the background and culture of the learner throughout the learning process, as this background also helps to shape the knowledge and truth that the learner creates, discovers, and attains in the learning process (Wertsch, 1997).

Constructivism

Several researchers previously cited found that learning with well-designed video games adheres to constructivist principles (Dede, Nelson, Ketelhut, Clarke, & Bowman, 2004; Dickey, 2005, 2006; Gee, 2003; Schrier, 2006). In an article describing the multi-user virtual world, SciCtr, (Corbit, 2005) underscored the merits of a constructivist approach for analyzing game-like environments. In SciCtr, students create virtual science worlds, such as rainforests or deserts, that other learners can visit and explore. According to Corbit, these worlds, the paths to navigate through them, and the content embedded in them, are constructed by the developer/learner through meticulous research and thoughtful design.

Constructionism

Designing and developing video games, rather than playing them, applies a contructionist approach to learning with games (Robertson & Good, 2005; Robertson et al., 2004). Scratch, a software platform through which students can program their own interactive stories, games, and animations is one of the successful implementations of constructionism. The constructionist approach to learning involves two activities: the

construction of knowledge through experience and the creation of personally relevant products. Steiner, Kaplan, & Moulthrop (2006) concurred with this constructivist view and contended that children as design partners improve the technologies they consume as well as gain educational benefits from the experience, (p. 137). Burrow and More (2005) applied constructionist techniques in an architecture course by having students render their designs with a game-engine thereby exploring spatial relationships as well as atmosphere, lighting, and other environmental conditions in a 3-D simulation of their architectural designs.

2.3 Game: Affordances

Game designers and academics such as Chris Crawford, Jesse Shell, Katie Salen and Eric Zimmerman have not only advocated for the potential of games to be used in educational content, but have defined games as learning tools in themselves. Chris Crawford even goes far to claim that the fundamental motivation for all game-playing is to learn (Crawford, 1984). Games present novel environments to players to be explored and mastered. Games methodically teach players the skills needed to meet complex challenges. Long, complex tasks are broken down into short, simple components. These components are trained individually before being chained together. Learning is a core aspect of gameplay.

Gee has stated that good video games build into their very designs good learning principles and when young people are interacting with video games–and with other popular cultural practices–they are learning. He identifies 36 learning principles from his observations on video game design that can be extrapolated from the game world to instructional design.

Kirriemuir and McFarlane (2007), in their Futurelab report "Literature Review in Games and Learning" have pointed out that there are two key themes common to the development of games for education, namely:

• the desire to harness the motivational power of games in order to 'make learning fun'

• a belief that ‘learning through doing’ in games such as simulations, offers a powerful learning tool.

In the following paragraphs, I will summarize the game components and design principles that are applicable and supportive for instruction design as well.

Clear goals

Games typically present the player with a series of short, medium and long-term goals. Games give compelling goals to players that are personally meaningful to them. For example: In ‘Lure of the Labyrinth’(www. labyrinth.thinkport.org), players need to find and liberate pets.

The goal of “Oregon trail” (www.oregontrail.com) is to successfully complete the trail by balancing supplies and the health of the family. In the path to achieve this main goal, they have smaller sub goals through the journey. One of the challenges of designing serious games is that the goal of the game and the goal of instruction design should align with each other. If mastering the game does not follow mastering learning content, it promotes performance oriented goals rather than mastery goals.

Immediate feedback

Games provide immediate, appropriate and specific feedback to players. Effective games provide feedback that is (1) clear and unobtrusive, and (2) immediately responsive to the player’s actions (Rigby & Ryan, 2007). Feedback also helps to reinforce motivation (Jones & Issroff, 2005). For example: in the simulation game “Crayon physics”, students

can see immediate feedback to their responses. While feedback on positive responses gives students reinforcement, feedback on incorrect responses provides students with information about what their error was and how it relates to the correct solution (Malone, 1986). Benefits of immediate feedback in learning have been supported by numerous studies. But sometimes, delayed feedback can be better for retention and transfer (Sanders, 1985).

Mastery-based approach

Generally, players are expected to demonstrate excellent performance of a skill before they can advance to using that skill in a more challenging environment. Complex tasks, then, simply require chaining together these previously learned simple skills. In traditional classroom settings, a student who does not master a concept could be left with a gap in their knowledge foundation that challenges later attempts to build to more complex concepts. In contrast, digital games inherently force the player to master a concept in order to advance. Players are able to repeat the same scenario until they master this concept. The same philosophy could extend to the use of digital games in education. A student cannot, in essence, unlock Algebra until a prerequisite knowledge of previous skills has been mastered.

Learning from failure

An attractive element of the gaming experience as a learning tool is that it provides opportunities for continued practice because negative consequences are not typically associated with failure. Rather, failure serves as an integral part of the learning

experience (Gee, 2009; Groff, Howells, & Cranmer, 2010; Ke, 2009; Klopfer, Osterweil, & Salen, 2009). This encourages players to improve through repeated practice either by advancing within a game or replaying parts of a game. Failure with limited consequences, agency and choice are seen as critical elements of a true gaming experience.

Active learning and problem based learning

In games, players learn by actively solving problems, instead of passively reading or memorizing. This active learning paradigm of problem-based learning promotes reflection and deeper learning.

Flow

In his book "Flow: The Psychology of Optimal Experience", Mihaly Csikszentmihaly (1990) introduced the term "Flow" as the state in which people are so involved in an activity that nothing else seems to matter; the experience is so enjoyable that people will do it even at great cost, for the sheer sake of doing it.

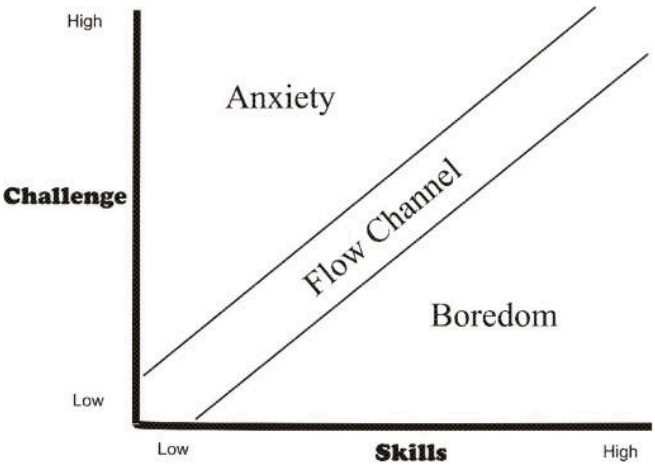
Flow occurs when certain conditions are met, four of which are: clear goals, immediate feedback, focused attention, and tasks that challenge (without frustrating) one's skills

Game designers are the professionals of creating flow-inducing activities (Kiili, 2006). Good games stay within, but at the outer edge, of the player's "regime of competence" (diSessa, 2000). That is, they feel "doable", but challenging, which is a highly

motivating state for learners. This nature of the flow experience supports the ideology of lifelong learning and is a priceless goal in education.

Mattheiss et al. (2009) point that teaching factual knowledge and the need for educational guidance, assessment and other intrusive components impede the creation of a free flowing educational game in contrast to non-educational games.

In general, well designed games as well as well designed education experiences are challenging but achievable. This is similar to Vygotsky's zone of proximal development, which is "the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers". A game is able to provide that opportunity for appropriate guidance or collaboration in order to help players meet the next challenge. The stepwise increase in difficulty reduces frustration and allows players to form knowledge and strategies that will be useful later (Gee, 2003). A state of pleasant frustration—challenging but doable—is an ideal state for learning several content areas such as science (diSessa, 2000)



The Flow. After Mihaly Csikszentmihalyi, *The Flow (1990)*, p. 74

Figure 1 Flow: A state of pleasant frustration—challenging but doable

Murray & Arroyo (2002) have proposed a mechanism of maintaining zone of proximal development in adaptive instructional systems.

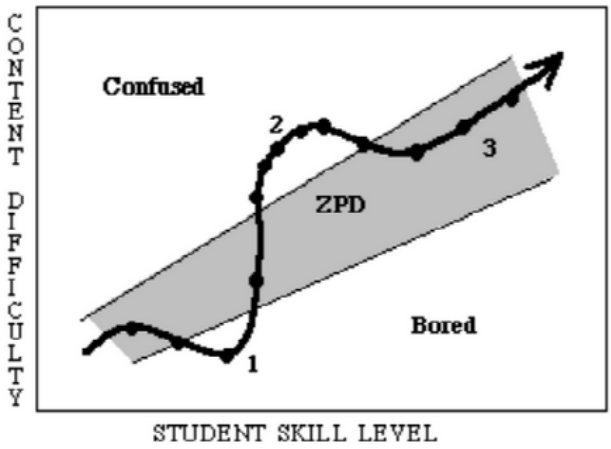


Figure 2 Zone of Proximal Development

Situated meanings

Gee asserts that games always situate the meanings of words in terms of the actions, images, and dialogues they relate to, and show how they vary across different actions,

images and dialogues. People are poor at learning what words mean when all they get is a definition that spells out what it means in terms of other words. Research suggests that people know what words mean and learn new ones only when they can associate them to the sorts of experiences they refer to — that is, to the sorts of actions, images, or dialogues that the words relate to (Barsalou, 1999; Glenberg,1997). This gives words situated meanings, not just verbal ones. And, indeed, words have different situated meanings in different contexts (consider "The coffee spilled, go get a mop" versus "The coffee spilled, go get a broom"). Games always situate the meanings of words in terms of the actions, images, and dialogues that they relate to, and show how they vary across different actions, images, and dialogues. They do not just offer words for words. School should not either. In a symposium on learning theories for the analysis of educational video games, Halverson, Shaffer, Squire, and Steinkuehler (2006) asserted that situated cognition provides a meaningful framework for the study of games, given that games have an ability to situate learning in an authentic context and engage players in a community of practice. Dede, Nelson, Ketelhut, Clarke, and Bowman (2004) outlined both constructivist and situated learning design principles present in effective video games including GST (guided social constructivist design), EMC (expert modeling and coaching) and LPP (legitimate peripheral participation). These authors employed such principles in evaluating game design and applied their findings to future iterations of the design. Lunce (2006) also argued that situated or contextual learning provides the rationale for simulations and simulation games in a classroom environment because of their ability to provide an authentic context in which to situate learning. According to these and other scholars, the authentic, situated context affords greater content mastery

and transfer of knowledge than a traditional classroom learning (Dickey, 2005, 2006; Klopfer & Yoon, 2005; Schrier, 2006). According to Kurt Squire, game-based learning can be understood as a particular kind of designed experience, where players participate in ideological worlds, worlds designed to support a particular kind of reactions, feelings, emotions, and at times, thoughts and identities, which game-based learning designers are leveraging for education and training. Interactive digital media, or video games, are a powerful new economic, cultural, and perhaps educational force. Video games provide situated experiences where players are immersed in complex problem solving tasks. Good games teach players more than just facts, but ways of seeing and understanding problems and opportunities to "become" different kinds of people.

Roles and Identities, perspectives and agency

Players can take different identities and roles. This gives them different perspectives, which they would not get otherwise. This does not only make learning personalized, meaningful and fun but also enables reflective thinking and deeper learning. In virtual games, students act as investigative reporters, environmental scientists, and historians who resolve meaningful dilemmas. Players feel a real sense of agency and control and a real sense of ownership over what they are doing. Such ownership is rare in school. Barab et al. (2010) have put forward the concept of transformational play, "transformational play involves positioning students as empowered actors who must understand and enlist academic content in order to effectively transform problematic scenarios"

In ‘Immune attack’(www. immuneattack.org), players navigate a nanobot through a 3D environment of blood vessels and connective tissue in an attempt to save an ailing patient by retraining her non-functional immune cells. Along the way, they learn about the biological processes that enable macrophages and neutrophils – white blood cells – to detect and fight infections. In ‘Reach for the Sun’ (www.gamesforchange.org/play/reach-for-the-sun), students have to take the role of a plant and balance their resources of starch, water, and nutrients to grow and reproduce.

Games Provide an Environment for Authentic and Relevant Assessment

In Pearson’ review of educational games, McClarty et al. (2014) conclude that games are inherently assessments. Games and traditional assessments share underlying characteristics that provide a means for quantifying knowledge and abilities. The two environments use complementary technologies that can combine to create more accurate models of student knowledge, skills, and behaviors. For example, games provide opportunities for authentic and appropriate knowledge representation of complex ideas, many of which seem under-represented in traditional assessments (Behrens, Frezzo, Mislevy, Kroopnick, & Wise, 2007). In games, the assessment process occurs as the game engine evaluates players’ actions and provides immediate feedback. Players make progress or they don’t; they advance to the next level or try again. Assessment occurs naturally in a game. The challenge is assessing the appropriate knowledge, skills, or abilities (Ash, 2011). Methodologies have surfaced as a means for designing games for assessment and quantifying the knowledge and abilities within game environments. The

opportunity for games to be used as assessments is greatly enhanced because of their capacity to collect deep, rich data about students and then to analyze—through advanced methods (Baker & Yacef, 2009)—their fine-grained interactions. Games can therefore serve as "non-invasive assessments" that provide continuous information which can be analyzed according to several probabilistic techniques (Kickmeier-Rust, Marte, et al., 2008).

Good game design coupled with a robust assessment approach should be the starting point for any research project focused on building a video game for educational purposes. That is, such research should combine game design with assessment methodologies such as Evidence Centered Design (ECD) at the outset of the game design process, rather than considering assessment as an afterthought. These assessments should be grounded in theory, and should start with defining what competencies are important and how a video game can be used to assess and improve these competencies. Finally, more attention should be given to figuring out specifically how video games can help improve important new competencies. Since good video games hold such an engagement value, they are useful (and fun) tools for players to practice skills over extended amounts of time, especially for today's college students who grew up playing such games.

Narrative Context

Based on literature review of Dondlinger (2007), some researchers attribute the compelling nature of some games to their narrative context (Dickey,2005, 2006; Fisch, 2005; Waraich, 2004) while others find motivation is linked to goals and rewards within

the game itself or intrinsic to the act of playing (Amory, Naicker, Vincent, & Adams, 1999; Denis & Jouvelot, 2005; Jennings, 2001).

Dickey (2006) argued that a narrative context that promotes challenge, fantasy, and curiosity, and provides feedback for players is one that promotes intrinsic motivation for play. In another study, Waraich (2004) agreed narrative is essential to motivation but cautioned that, intrinsic rewards are based on a high congruence between the material being taught and the motivational techniques used. Dissonance between the two can decrease learning. In a study of a variety of design elements on game environments for instruction in computer science architecture, Waraich (2004) focused mainly on narrative. This empirical study analyzed the role of both narrative context and game goals as features for motivating and conceptualizing learning in a 2-D interactive learning environment (ILE). The mixed methods design of the study revealed quantifiable knowledge gains in the ILE over traditional instruction. Waraich concluded that, "For any learning task to be meaningful to the learner they must have both a sufficient context for the learning and motivation to perform the tasks that will help them to learn. We believe that game based learning environments that incorporate a strong narrative can meet these requirements if the learning tasks are appropriately designed and tightly coupled with the narrative,".

Fisch (2005) made a similar observation. Although narrative context does motivate learning, for an educational game to be effective the learning content must align with the narrative plotline. According to Fisch's analysis, "research on lectures and textbook

readings has suggested that seductive details do not work; children exposed to such material tend to remember the appealing elements but not the intended educational content," . He found that a far more powerful approach is to place the educational content at the heart of engaging game play, so that children employ the targeted academic skills and knowledge as an integral part of playing the game. Fisch also maintained that selecting appropriate media as well as providing feedback and scaffolding within and outside of the game are essential to effective educational game design.

2.4 Games: Constraints

Pedagogical constraints of using games in education

1. Practical constraint: Time overload

Games and game elements tend to take up time that could have been used for instruction. Game environments can be complex and require students to spend time to learn them first. Besides this, games consume substantial amounts of time via play aspects. Since time on task is an important predictor for learning (Ericsson et al., 1993), students may not learn as much from games as from other material within the same time.

2. Intrinsic constraint: Working memory overload

Cognitive load theory (Sweller, 1994) states that learning happens within constrained and a very limited working (or short-term) memory and unlimited long-term memory. Specifically, a skill can only be learned if all of it can fit within the learner's working memory. Therefore, if there are too many game elements to be learned, then the total cognitive load will exceed the limits of working memory, and there would be less learning. Mayer (2009) has demonstrated that extraneous details in multimedia education can be detrimental to learning. Although details and novelty in a game environment and complexity of the game rules can add excitement and entertainment value in games, they can also overwhelm learners in the case of learning games due to additional memory load of the learning content. Since non-educational games have a sole purpose of entertaining, they can afford to play with novelty, details and complexity to maximize fun. However, learning games have to deliver learning content, and thus have to restrain on the amount of additional details and complexity they might want to add.

3. Goal constraint: Aligning cognitive and affective outcomes

While tutoring systems are primarily concerned with cognitive outcomes (learning gains, retention, transfer, etc.), and computer games are about maximizing fun, educational games have the objective of enhancing both cognitive and affective outcome (fun, attention, engagement, etc.). These two goals are not necessarily in opposition. In fact, they can reinforce each other; students feel better when they learn and they learn more when they feel better. But these two outcomes are not always aligned and sometimes affective and cognitive strategies may be in conflict with each other (Boyer et al., 2008).

As mentioned in the previous section, the elements, which enhance excitement and fun, can overwhelm and overload learners. Similarly, the tutorial practices may seem pedantic and diminish students' sense of choice and control and reduce fun (Easterday, 2011).

4. Design constraint: Integration of learning content and game attributes

Determining what kind of game attributes and characteristics are suitable for the specific learning content, and deciding how we should embed the learning content in a game environment is a very delicate design process. It is more likely that games will be instructionally effective if the specific characteristics of the game (e.g., setting, player roles and activities, rules, etc.) overlap with specific instructional objectives. This overlap must be consciously structured on the basis of a thorough analysis of the reasons for the instruction and the instructional objectives to be met (Hays, 2005). When integration of content and game attributes is unintuitive, it can make learning hard and when the integration is superficial, it may only add extrinsic motivation hindering intrinsic motivation.

One of the reasons for the relative scarcity of successful educational games is that it is indeed harder to design them, as they have to do so many things well. Mattheiss et al. (2009) say that teaching factual knowledge and the need for educational guidance, assessment and other intrusive components impede the creation of a free flowing educational game in contrast to non-educational games.

Social and economic constraints of using games in education

There is still stigma around using video games in education. Teachers may not be familiar with the technology, which can make her aversive or hesitant to use games for teaching purpose. Schools may not have sufficient and proper infrastructure to implement game based learning. Besides implementation and adoption of games in schools, there is a huge economic and resource constraint to produce those games themselves. Games are resource intensive to build, in terms of financial, time and human resources. Given that we still do not have a lot empirical evidence of effectiveness of games in education and do not have sure-fire recipes for effective design, creating educational games can be very risky. Even when teachers are more comfortable and receptive at using digital games, lack of resources can still be a serious constraint. On a national survey of 500 teachers who use digital games (Millstone, 2012) , they report that cost is the number one barrier to using games (50%) followed by access to technological resources (46%). Emphasis on standardized tests also seems to be a substantial barrier (38%).

2.5 Empirical evaluation of effectiveness of games in learning

One of the problems with educational game research is that there are not many quantitative studies, let alone randomized controlled studies. Since the qualitative findings are very promising and quantitative data few and far, there have been some meta-analysis to find whether games are effective and if so, what kind of games and under what circumstances. Empirical results on the effectiveness of educational games is scant in general. However, results from more recent studies are hopeful.

Are games effective?

Connolly et al. (2012), De Freitas (2006), Wouters et al. (2009) point out that despite the optimism about the potential of games for learning, there has been a dearth of high quality empirical evidence to support these claims. In their analysis, Connolly et al. (2012) maintains that the evidence that games leads to more effective learning was not strong. The few papers, which provided high quality evidence to support these claims, presented qualitative rather than quantitative analysis (Mayer et al, 2004; Steinkuehler & Duncan, 2008). Dondlinger (2007) claim that the games might be especially useful in promoting higher order thinking and soft and social skills.

After reviewing a large number of studies on learning from simulations, de Jong and Van Joolingen (1998) concluded, "There is no clear and unequivocal outcome in favor of simulations. An explanation why simulation-based learning does not improve learning results can be found in the intrinsic problems that learners may have with discovery

learning,". These problems are related to processes such as hypothesis generation, design of experiments, interpretation of data and regulation of learning.

Sitzman (2011) made a meta-analysis of instructional effectiveness of computer-based simulation games relative to a comparison group (k= 65, N= 6,476). The post-training self-efficacy was 20% higher, declarative knowledge was 11% higher, procedural knowledge was 14% higher, and retention was 9% higher for trainees taught with simulation games, relative to a comparison group. However, she also mentions that the results provide strong evidence of publication bias in simulation games research.

In a meta-analysis of the cognitive and motivational effects of serious games, Wouters et al. (2013 have found that serious games were found to be more effective in terms of learning (d= 0.29, $p < .01$) and retention (d = 0.36, $p < .01$), but they were not more motivating (d = 0.26, $p > .05$) than conventional instruction methods. Additional moderator analyses on the learning effects revealed that learners in serious games learned more, relative to those taught with conventional instruction methods, when the game was supplemented with other instruction methods, when multiple training sessions were involved, and when players worked in groups.

Clark et. al (2014) made a recent study based on meta-analysis of research published between 2000 and 2012 found evidence for effective of games. Based on 58 studies, digital games were associated with a .33 standard deviation improvement relative to control conditions, even after adjusting for baseline differences in achievement between groups.

Which games are effective?

Clark et. al (2014) have made findings that shed light on what games are more effective. Games with theoretically augmented designs for learning will outperform standard versions of those games. Based on 20 studies, results indicated that augmented game designs were associated with a .37 standard deviation improvement in learning relative to standard versions, even after adjusting for baseline differences in achievement between groups. This finding highlights the importance of design in learning outcomes.

Game conditions involving multiple game-play sessions demonstrated significantly better learning outcomes than non-game control conditions and game conditions involving single game-play sessions did not demonstrate significantly different learning outcomes than non-game control conditions.

Contrary to their prediction that more sophisticated game mechanics, increased variety of player actions, intrinsic integration of the game mechanic and learning mechanic, and more specific/detailed scaffolding will be related to larger effects on learning outcomes relative to non-game conditions, they found that simple gamification as well as more sophisticated game mechanics can prove effective. They recommended that future research and analyses should explore whether or not the "simple gamification" studies more frequently focus on lower-order learning outcomes as compared to studies with more sophisticated game mechanics.

Schematic games were more effective than cartoon-like or realistic serious games and suggested that games with no narrative might be more effective than games with narratives.

Based on their meta-analysis, Connolly et al. (2012) make some recommendations regarding what kind of games and game elements are suitable for learning. According to their findings, simulations were by far the most frequently occurring genre, possibly because their use in education is already established. Puzzles were also used in Game Based Learning (GBL), again because their educational relevance is clear. It may be that the relative lack of other genres in GBL is because educators are unclear about how to utilize the distinctive features of this genre in teaching. If a wider variety of game genres are to be used in learning, better guidance needs to be provided about how the affordances of different kinds of games can support learning in different ways, in terms of detailed accounts of the tasks and activities offered in different kinds of games.

O' Neil (2005) concludes that games themselves are not sufficient for learning, but there are elements in games that can be activated within an instructional context that may enhance the learning process (Garris et al., 2002). In other words, outcomes are affected by the instructional strategies employed (Wolfe, 1997). Leemkuil et al. (2003), too, commented that there is general consensus that learning with interactive environments such as games, simulations and adventures is not effective when no effective instructional measure or support is added.

There appears to be consensus among a large number of researchers with regard to the negative, mixed or null findings of games research, suggesting that the cause might be a lack of sound instructional design embedded in the games (Gredler, 1996; Wolfe, 1997; de Jong & van Joolingen, 1998; Thiagarajan, 1998; Lee, 1999; Garris et al., 2002; Leemkuil et al., 2003; O'Neil & Fisher, 2004). However, as we embed instructional strategies in games, we must consider individual differences. In an examination of the role of training scenarios in video games, Oliver and Pelletier (2005) found that providing training in games can be effective for strategy development, but that players apply those strategies differentially, with some players being more effective than others.

Amory et al. (1999) made a study to identify the game type most suitable to teaching environment and to identify game elements that students found interesting or useful within the different game types. A group of twenty students played four commercial games (SimIsle, Red Alert, Zork Nemesis and Duke Nukem 3D). Results suggest that students prefer 3D-adventure (Zork Nemesis) and strategy (Red Alert) games to the other with Zork Nemesis ranked as the best. Students rated game elements such as logic, memory, visualisation and problem solving as the most important game elements. Such elements are integral to adventure games and are also required during the learning process.

Young et al. (2012) investigated if video games show demonstrable relationships to academic achievement gains when used to support the K-12 curriculum. In a review of

literature, they identified 300+ articles whose descriptions related to video games and academic achievement. They found some evidence for the effects of video games on language learning, history, and physical education (specifically exergames), but little support for the academic value of video games in science and math. They conclude that many educationally interesting games exist, yet evidence for their impact on student achievement is slim. They recommend separating simulations from games and refocusing the question onto the situated nature of game-player-context interactions, including meta-game social collaborative elements.

Moreover, games are not an effective teaching tool for all students; this has partly to do with the pedagogy. Failure is the norm in games; repetition and exploration is how players learn. This contrasts with learning discrete chunks of information which can be found in schools (Squire 2005). Squire found roughly 25% of students in school situations withdrew from his study, which used Civilization to teach geography and history, as they found it too hard, complicated and uninteresting. (to become a competent player takes six to seven hours, and to go through all the stages a hundred hours.) while another 25% of the students (particularly academic underachievers) loved playing the game, they thought it was a perfect,way to learn history.

Ian Bogost (2010) summarizes the limitations and potential of games as educational tools: Games are hard to make. Good games are complex. The real promise of games as educational and political tools is in their ability to demonstrate the complexity and interconnectedness of issues. Simon Egenfeldt-Nielsen (2008) states the most important

consideration from a teacher's perspective is how much the game will make their life easier. Thus the second challenge is to identify whether the selected game will easily enhance teaching, or, if there is not yet a relevant game in that area, what a game would need in order for it to be useful. This consideration involves more than just assuming games will motivate and engage learners, as Squires(2005) points out games in a classroom are not necessarily motivating but includes assessing whether alternative methods of teaching would be better.

Empirical comparison between educational games and intelligent tutors

There is a relative scarcity of evidence directly comparing the educational effectiveness of educational games vs. computer tutors; however, comparisons have found an advantage for tutoring approaches over educational games (Easterday, 2011; Jackson et al., 2011). Intelligent tutors have been able to demonstrate significant learning gain consistently .

Ma et al. (2014) made a meta-analysis of learning outcomes of ITS on 107 effect sizes involving 14,321 participants. They found that the use of ITS was associated with greater achievement in comparison with teacher-led, large-group instruction (g .42), non-ITS computer-based instruction (g .57), and textbooks or workbooks (g .35). There was no significant difference between learning from ITS and learning from individualized human tutoring (g –.11) or small-group instruction (g .05). Significant, positive mean effect sizes were found regardless of whether the ITS was used as the principal means of instruction,

a supplement to teacher-led instruction, an integral component of teacher-led instruction, or an aid to homework. Significant, positive effect sizes were found at all levels of education, in almost all subject domains evaluated, and whether or not the ITS provided feedback or modeled student misconceptions.

The effect size and number of empirical results show that intelligent tutors have an edge over educational games as far as measurable learning gain goes. However, educational games have potential to generate learning gains in the aspects and areas beyond the traditional approach and learning outcome variables. They can be used to enhance students' attitude (O'Rourke et al. 2014), and persistence (Shute et al. 2015). In fact, games can be an innovative platform to measure those very learning outcomes (DiCerbo, 2014 & Shute et al. 2015)

2.6 Designing Educational Games

How to design good educational games that are enjoyable and also effective at teaching has been an ongoing quest for designers and academics. In their paper "Moving learning games forward", Klopfer et al. (2009) point out some basic mistakes that are prevalent in designing educational games. They suggest that instead of grafting academic content into existing game forms, we need to find game in the content. Habgood makes distinction between extrinsic integration and intrinsic games. In extrinsic games, games rely on an extrinsic reward structure, bestowing gold stars for good performance instead of making the incentives internal to the game. Squire(2013) in his paper "Video game-based learning: An emerging paradigm for instruction", talks about shift from content to context. According to him while eLearning focuses on content, saying, "content is king," in a situated view of knowledge would say that it is the context in which learners develop knowledge is king. He thinks games create an emotionally compelling context for the player. Quest To Learn (Q2l), a game-based schooling with it's game-based curricular model, uses internal architecture of games to create game-like learning environment.

Games for Learning Institute (G4LI) focuses on creating good games. They try to investigate key design elements that make certain games compelling, playable, and fun How do game genres differ in their educational effectiveness for specific topics and for specific learners? How do kids learn when they play games? Does the setting (classroom vs. casual) matter? How can games be used to prepare future learning, introduce new

material, or strengthen and expand existing knowledge? How are games designed to best facilitate the transfer of learning to the realities of students' everyday lives? And how can researchers study existing games, identify key design elements and learning patterns, develop prototype "mini games" based on these elements and patterns, test them in classroom and informal learning settings, and evaluate the results. G4LI's initial focus is on digital games as tools for teaching science, technology, engineering, and math—STEM subjects—at the critical middle-school level. Plass et al. (2011) suggest that game mechanics, the essential game play activity, should be distinguished from Learning mechanics and Assessment mechanics. Learning mechanics as patterns of specialized activities, grounded in the learning sciences, that have learning as the primary objective. In contrast, assessment mechanics are patterns of specialized activities, grounded in test theory, that have assessment as the primary objective. Learning and assessment mechanics are design patterns, or meta-mechanics, that can be instantiated onto corresponding game mechanics, following criteria we outlined above to preserve their intended teaching or assessment objective. Variables related to learning that can be measured through game metrics include learning outcomes (cognitive and skills), trait variables, general state variables, and situation-specific state variables. Supplementing log data of game events and user behavior with observational data extends the ECD model and results in more valid assessments of these variables. By using assessment mechanics to measure a series of learner variables, a learner model can be compiled that allows for the design of games that are individualized and adaptive to a learner's specific needs and characteristics. This has implications for the design of effective games for

learning by making games more adaptive and personalized, and, hopefully, more effective.

They come up with these guidelines:

Intelligent tutoring system and educational games

(1) Game mechanic must not introduce excessive amounts of extraneous cognitive load.

(2) Game mechanic must not reduce the amount of the required mental effort by too much.

(3) Game mechanic must not introduce unnecessary confounds (fine motor skills)

Aleven et al. (2013) present a framework for the design of educational games, based on three main components, known to game designers and instructional designers. Each component provides its own analytical angle. Basic assumptions underlying the framework are that an educational game development project is more likely to succeed if the learning objectives of the game have been clearly established early on in the development process, if the designers have carefully thought about how the game's desired aesthetic can be grounded in the game mechanics, through the game's dynamics, and if the game observes well-established instructional design principles. A designer of educational games needs to consider almost constantly how he or she can make the components work in concert. An educational game has to succeed on two fronts: as an educational tool and as a fun game. Fun is judged through success in achieving the aesthetic goals. The mechanics and dynamics are merely tools for getting there. The learning principles work predominantly at the level of the dynamics and mechanics. The

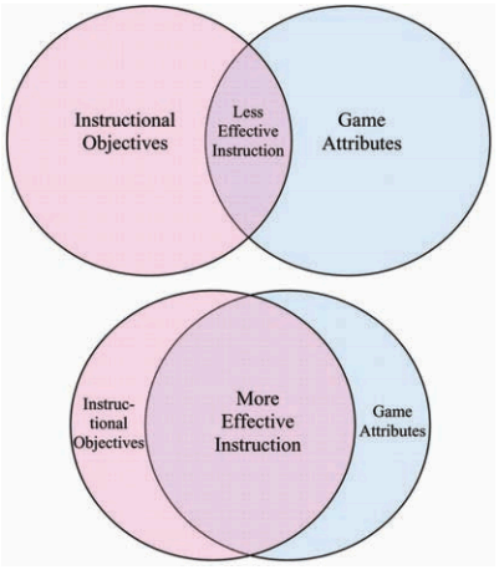
trick is keeping aesthetics in line while tweaking the mechanics and dynamics to work in accordance with instructional principles to meet the game's learning goals

2.7 Effective integration of game design and instruction design

Aligning cognitive and affective outcomes

While tutoring systems are primarily concerned with cognitive outcome (e.g., learning gains, retention, and transfer to novel situations), and computer games are about maximizing fun, educational games have the objective of enhancing both cognitive and affective outcomes. These two goals are not necessarily contradictory. In fact, they can reinforce each other; for example, students feel better when they learn and they learn more when they feel better. But these two outcomes are not always aligned and sometimes affective and cognitive strategies may be in conflict with each other (Boyer et al., 2008). As mentioned in the previous section, the elements, which enhance excitement and fun, can overwhelm and overload learners. Similarly, the tutorial practices may seem pedantic and diminish students' sense of choice and control and reduce fun (Easterday, 2011).

Instructional Effectiveness as Degree of Overlap Between Learning Objectives and Game Attributes



Source: Adapted from Hays, 2005.

Figure 3 Instructional Effectiveness as degree of overlap between learning objectives and game attributes

Integration of learning content and game attributes

Determining what types of game attributes and characteristics are suitable for the specific learning content, and deciding how to embed the learning content in a game environment is a very delicate design process. It is more likely that games will be instructionally effective if the specific characteristics of the game (e.g., setting, player roles and activities, rules, etc.) overlap with specific instructional objectives. This overlap must be consciously structured on the basis of a thorough analysis of the reasons for the instruction and the instructional objectives to be met (Hays, 2005). When integration of content and game attributes is unintuitive, it can make learning difficult and when the integration is superficial, it may only add extrinsic motivation hindering intrinsic

motivation.

2.8 Game elements, Game mechanics and Gamification

Games and game-likeness

One of the main problems with game research is that there is no definitive definition of game or what constitutes a game (Mayer, 2011). Caillois (1961) describes a game as an activity that is voluntary and enjoyable, separate from the real world, uncertain, unproductive (the activity does not produce any goods of external value), and governed by rules. Hays (2005) defines game as an artificially constructed, competitive activity with a specific goal, a set of rules and constraints that is located in a specific context. According to Salen and Zimmerman (2003), a game is a system in which players engage in an artificial conflict, defined by rules, that results in a quantifiable outcome. There are many academic definitions of games, but none of them have been accepted as definitive and all encompassing. In fact, there is an active community of game theorists among whom the debate of exactly how to define a game goes on continuously (Szulborski, 2005). Ludwig Wittgenstein (1953) demonstrated that the elements of games, such as play, rules, and competition, all fail to adequately define what games are. He concluded that people apply the term game to a range of disparate human activities that bear to one another only what one might call family resemblances. While some regard conflict and competition as central to a game, activities without conflict such as *The Sims* (www.thesims.com) and *Farmville* (www.facebook.com/FarmVille), are more popular

than many games that fit the conventional definition. With new media, new demography and new usage, definitions and perceptions of games have constantly evolved. In particular, many educational materials created with an aim to entertain as well as educate have been under debate if they qualify as games or not. There are a lot of poorly designed materials that try to qualify as games without understanding the fundamentals of games. But there are also many carefully designed materials that are game-like but do not fit into the formal definitions of game.

In particular, educational materials created with an aim to entertain as well as educate have always been quite debated regarding whether they qualify as games or not. Game enthusiasts have complained that the educational materials that sell themselves as games are just interactive systems but not games at all. There are a lot of poorly designed materials that try to become games without understanding game's fundamentals and exploiting the benefits of games. But there are also lots of carefully designed materials that are game-like but do not fit into the formal definitions of game. So, why are there so many game-like learning materials that do not qualify as games from conventional definitions of games?

While we acknowledge the necessity and value of clear definitions, ill-defined activities such as *The Sims* have been not only been successful but also reached new demographics that were not addressed by traditional video games. Researchers such as Rieber(1996) have suggested effectiveness of hybrid learning environment combining simulation and games in microworlds. In our research, we are not going to constrain ourselves into the formal definitions and categorization of games, but are exploring a looser and more flexible space of *game-like elements*.

Game-like elements

There have been many attempts to distill game elements, characterize them and study them. Malone and Lepper (1987) mentioned challenge, curiosity, control, and fantasy as integral features of games. According to de Felix and Johnson (1993), games are composed of dynamic visuals, interactivity, rules, and a goal. Thiagarajan (1999) asserts that conflict, control, closure, and contrivance are the four necessary components. Juul (2003) suggested that games consist of six elements: rules, variable quantifiable outcome, player effort, valorization of the outcome, attachment of the player to the outcome, and negotiable consequences. In 2001, Garris and Ahlers (2002) discuss about fantasy, rules/goals, sensory stimuli, challenge, mystery, and control. Marc LeBlanc's taxonomy of game pleasures (Schell, 2008) for participants identifies eight fundamental aspects to fulfilling their emotional needs: sensation, fantasy, narrative, challenge, fellowship, discovery, expression and masochism.

As mentioned earlier, we are not trying to generate formal definitions of games or game elements, but rather we are looking into understanding the properties of game-like elements, which we define as the engaging and interactive aspects of games. Specifically, we are looking into game-like elements such as narrative, immediate visual feedback, visual representation, collecting and sensory stimuli. Even though the game-like elements are defined based on their engaging nature, these elements can have significant pedagogical impact in both positive and negative ways. We want to assess these elements in terms of their pedagogical affordance and constraints and want to select and integrate those ones that can be beneficial pedagogically or at least not hurt the learning.

Game mechanics and Gamification

Marc LeBlanc and his colleagues (2004) wrote a paper proposing a game design framework around the concepts of Mechanics, Dynamics and Aesthetics (MDA). They define those terms as follows:

- Mechanics are the agents, objects, elements and their relationships in the game. They define the game as a rule-based system, specifying what there is, how everything behaves, and how the player can interact with the game world.
- Dynamics are the emergent behavior that arises from gameplay, when the Mechanics are put into use.
- Aesthetics are the emotional response from the players to the gameplay.

According to Salen and Zimmerman (Rules of Play, 2003), core mechanics represent the essential moment-to-moment activity of players. During a game, core mechanics create patterns of repeated behaviour, the experiential building blocks of play.

Hunicke, LeBlanc and Zubek (in MDA Frame Work, 2004) state that mechanics are the various actions, behaviors and control mechanisms afforded to the player within a game context. The mechanics support overall gameplay dynamics.

Game mechanics are principles, rules, and/or mechanisms (much like mechanics in physics) that govern a behavior through a system of incentives, feedback, and rewards with reasonably predictable outcome…Game mechanics are just the basic building blocks. They can be strung together and combined in interesting ways to drive a very complex sequence of actions suitable for different contexts or desired results.

Gamification is the use of game mechanics to drive game-like engagement and actions. Gamification is the process of introducing game mechanics into these regular activities to make them more game-like so that people would want to proactively take part in these tasks.

2.9 Intelligent Tutoring Systems and Educational games

Intelligent tutoring systems (ITS) and educational games are two research areas in educational technologies. Intelligent tutors, which are primarily concerned with cognitive aspects of learning, use adaptive, individualized tutoring to students and have shown evidence to improve learning significantly (Koedinger & Corbett, 2006). On the other hand, education researchers have also been interested in computer games due to their immense popularity and affordance of new kinds of interactions. Games can not only enhance the affective aspects of learning, but can also hold the potential to improve cognitive outcomes of learning as well. There is a relative scarcity of evidence directly comparing the educational effectiveness of educational games vs. computer tutors; however, some comparisons have found an advantage for tutoring approaches over educational games (Easterday, 2011; Jackson et al., 2011). Tutors, though able to effectively produce learning gains, have had difficulties in maintaining students' interest for long periods of time, which limit their use to generate long-term learning (Jackson et al., 2011).

Given these complementary benefits, there has been considerable effort to combine these two fields. ITS researchers want to incorporate elements from games to make them more engaging (Jackson et al., 2011), and games are also using tutorial features such as detailed content feedback (Easterday et al., 2011) to make them more educationally effective. Creating highly engaging educational games, which are as effective as tutoring systems in terms of learning, is a very desirable goal. However, there are several

difficulties in fulfilling this vision. First, research has demonstrated that due to our limited working memory, too many extraneous details can be detrimental to learning (Sweller, 1994; Mayer, 2009; Clark, 2011). Second, there are practical constraints such as time, as games tend to take more time to convey equivalent amount of educational content due to time consumed on gameplay aspects. Thus, the act of combining the best aspects of ITS and educational games is a challenging design goal, as it entails maximizing both engagement and learning. These goals to get both engagement and learning may not necessarily be always incongruent. In fact, they can reinforce each other as engaged students learn more and students get more engaged as they learn. But aligning these two goals is a very delicate design process and the abundance of failed educational games, which can generate neither fun nor learning, and thus resulting in a negative and uncool reputation for educational games, only confirms this difficulty (Clark, 2011).

There have been various efforts in integrating intelligent tutoring systems and educational games. I will briefly summarize the major efforts:

Crystal Island: Intellimedia at Northern California State University (http://projects.intellimedia.ncsu.edu/crystalisland/) have been creating 3D game-based inquiry learning science learning environments. They claim that the additional cognitive load presented by the narrative-centered learning aspect of the game supported the science content learning rather than distracting from the learning. Results demonstrated that students learned problem-solving steps through the game interactions. The research

did not, however, illustrate how students can more readily understand problem-solving steps in relation to non-game environments.

Conati and her students (Conati et al., 2002) have carried out a number of research studies trying to make games more intelligent by adding intelligent pedagogical agents to educational games, modeling learning in educational game, and modeling emotions.

iSTART-ME: iSTART (Jackson et al., 2009) is an intelligent tutoring system designed to improve students' reading comprehension by teaching helpful and effective comprehension strategies. However, these benefits improve with extended practice (taking place over a span of weeks and months rather than the typical experimental intervention of a few minutes or hours). Due to the long-term nature of this interaction, the extended practice module of iSTART is currently being situated within a game-based environment called iSTART-ME (motivationally enhanced). This game-based environment provides students a chance to interact with texts, earn points, advance through levels, purchase in-game rewards, personalize a character, or even play an educational mini-game (designed to use the same strategies as in practice).

3 OUR APPROACH

While games can be an innovative and game-changing approach to education, creating effective educational games is a challenge. Educational games are not only resource intensive, but it is also a big design challenge to effectively integrate educational content with game attributes and to align cognitive and affective outcomes, which can be sometimes conflicting. Unsuccessful implementation may not only result in lack of learning, but may even hamper learning by reducing intrinsic motivation. Given that games may add stimulation and fun but they also pose constraints and overloads in educational usage, whereas computer tutors have proven learning outcomes, integrating game elements in computer tutors can be a good alternative to directly creating an educational game. For this, we need a comprehensive framework to identify different ways of integrating the game elements and interventions into a tutor. We need to empirically assess each element and intervention in terms of its benefits and constraints so that we can create educational tools that have affective as well as cognitive outcomes. Given the complementary benefits of games and tutors, there has been considerable effort to combine these two fields. However, fulfilling this vision is a challenge as it is difficult to effectively integrate educational content with game attributes, and to align sometimes conflicting cognitive and affective outcomes. For example, extraneous details in games can distract and overwhelm students by overloading their working memory. Due to these limitations, there is a search for more efficient and effective alternatives to educational

games. Researchers in computer tutors are trying to make tutors more fun by integrating game elements in tutors (Easterday et al., 2011, Jackson et al., 2011) and there have been efforts to study individual game attributes (Wilson et al., 2008).

3.1 Games as affective, cognitive and metacognitive tool

While the initial effort of intelligent tutors has been focused on cognitive aspects of students, researchers have identified metacognitive and affective aspect of students as important. There has been a lot of effort to improve meta-cognition and affect as well. Du Boulay et al. (2010) make distinction between the different systems aimed at maximizing different gains. They define a cognitively intelligent educational system is a system that is able to deploy its resources and tactics dynamically and effectively to support learning but without modeling the metacognitive, affective or motivational states of the learner. A metacognitively intelligent educational system is a system that is able to deploy its resources and tactics dynamically and effectively to improve the metacognitive capability of the learner. An affectively intelligent educational system is a system that is able to deploy resources and tactics dynamically to provide an educational experience that improves the student's state of well-being during learning.

Games have been primarily viewed as an affective tool, to create experience of fun and delight. It is expected that students using educational game will have better affective state, which makes them, stay more time on task and eventually leading to more learning. For example: Math Blaster (www.mathblaster.com) helps in learning by the fact that the students actually solve the problems in the game. The cognitive support of the game is not particularly superior to a regular pedagogical approach.

Constructivist and constructionist game theorists extend the utility of game as special kind of cognitive tools, which carry specific superior affordances for cognition just because they are games. For example, games set a stage for active learning, where there are clear goals and situated meanings. These affordances of games directly lead to better cognitive gain, even if they are not successful in generating delight and fun and excitement. In this regard, games can act as a direct cognitive tool.

There is a theoretical stance that games are inherently beneficial for metacognition. Gamers need to plan actions, check their initial plan, change strategies if needed and evaluate their actions in game. Games teach critical thinking and problem solving skills. Games nurture an incremental understanding of intelligence. Because players are rewarded for one task at a time — for overcoming one obstacle after another — they learn to understand learning and accomplishment iteratively.

There are some games that are specifically designed to improve metacognitive skills. But even if games are not specifically designed for that, they can help students for self-assessment by their immediate and timely feedback.
Gamification is a practice, where we see game elements implemented in giving feedback to students in the form of badges and leaderboards. This does not actively teach metacognitive skills but give metacognitive support, so that students can make self-assessment and set their goals and strategies.

It can be problematic though to pigeonhole a game as just a cognitive, metacognitive or affective tool, since most of the times, a game serves as all three. Even if it is designed to support only one factor, learner's cognitive, metacognitive and affective channels are interconnected in themselves and intervention in one aspect leading to effect on other as well. For example: a game that helps metacognition makes the student feel more accomplished and generates positive affect, which leads to better time on task and higher cognitive gain.

3.2 Web of associations and Causal Mechanisms

Traditionally, emotion and cognition were considered separate, independent processes. However, it is being realized that, at least in specific types of task, cognition and emotion co-exist in the processing of information and regulation of behavior (cf. Cacioppo & Berntson, 1999). Metacognitive experiences (Efklides, 2006) can make use of both the affective and the cognitive regulatory loops, and this has a series of implications for learning. Metacognitive feelings (Koriat & Levy-Sadot, 2000), have a dual character, that is, a cognitive and an affective one. There are two basic manifestations of the monitoring function, namely, metacognitive knowledge and metacognitive experiences (Flavell, 1979). Specifically, metacognitive knowledge is declarative knowledge about cognition, which we derive from long-term memory (Flavell, 1979). This is a meta-level feedback loop that manifests subjectively as affect and as a hazy sense of expectancy. Positive affect arises when the rate is faster than anticipated and negative when it is slower. If the

rate is as anticipated, no affect is experienced. As Flavell (1979) and other early researchers on metacognition (Brown, 1978; Robinson, 1983) had suggested metacognitive experiences have a direct effect on cognition.

It is a difficult task to tease out these various affective and cognitive aspects of learning, as they are very inter-related. During the learning process, an intervention can impact both cognitive and affective aspects. Also, change in affect leads to cognition and vice-versa. It would be illuminating though to tease these different aspects apart. We would like to see how a particular intervention benefits different aspects of learning. If a student is performing well with an educational game, it is because the student is benefiting from the affective support or despite the lack of ? If we can observe students' affective level improving but no gain in learning, we would like to explore whether it is because the students are being distracted. If we offer students metacognitive support and they have learning gain, is it because students learnt new metacognitive skill or they are benefitting from higher sense of control ?

Besides analyzing how different learning channels work within a student, another important question is how different students interact and are affected by the pedagogical interventions. Students with high knowledge may engage entirely differently than the students with low knowledge. Are games more effective with students with high self-efficacy or the other way around? Do expectancy and pedagogical preference affect students' interaction with the game-like systems?

The same intervention can have a conflicting effect on different students. For example: a low knowledge student would find game-like systems friendlier but may suffer from cognitive overload, whereas a high knowledge student may report the game-like system being silly but still benefit cognitively from the exploratory system.
Similarly, students' prior experience and expectations and pedagogical preference can also influence her interaction and evaluation of the system. For example: one student might be more receptive to multi-media based approaches whereas another student may feel being distracted and overwhelmed.

We want to understand these various associations and causal mechanisms. Getting a better understanding of these mechanisms of learning is important from both diagnostic and prescriptive standpoints. We hope to get a better understanding of our game-like interventions and their impact on students: whether, how and why a particular game-like intervention would work; why an intervention may work for particular students under particular conditions and why it may not work in different population in different settings. This information will give us insights into designing better interventions and developing better systems.

3.3 Research Questions

RQ1: What is the range of activities of game-like interventions, that could impact learning and motivational affect?

Game-like interventions (GLIs) can be used to teach content, act as affective hooks to engage students, or to represent student performance in a fun way. We have analyzed the different ways in which we can use GLIs and we have come up with three broad categories: cognitive, metacognitive and affective. Within each category, the research questions are whether the modes improve cognitive and affective outcomes.

Cognitive mode of game-like intervention: Though the primary connotation of games is "fun," games also have cognitive affordances, which can make them effective teaching tools. We will work to identify game elements (described in section 3.4.1, different than GLIs) that carry these affordances, but avoid adding cognitive overload. We have created a game-like math tutor, *Monkey's Revenge* using game-like elements such as narrative, immediate visual feedback, collecting and building. Our approach is using game-like elements in a very cautious and minimalist way. For example: we want to exploit the learning benefit of narrative by creating a situated learning context but would not like the narrative to be too elaborate as that would distract learners.

Meta-Cognitive mode of game-like intervention: Unlike using GLIs to teach learning content, we are using this mode to communicate metacognitive information with learners.

We created 'Learning Dashboard' for students of *MathSpring*, an intelligent math tutor. Dashboard gives summary of student performance, effort and progress in different math skills along with strategic suggestions for learning. We are using game elements such as progress bars to demonstrate skill mastery. Similarly, students are given a plant for each math skill, as a representation for their math knowledge, which grows, give flowers and fruits and withers depending upon student's effort in the skill. We are not claiming to actively teach metacognitive skill, but we are giving metacognitive support by presenting metacognitive content in more intuitive way that triggers student to take more productive actions.

Affective mode of game-like intervention: In this mode, we are trying to use games solely to enhance fun while leaving the computer tutor responsible for teaching. Our hypothesis is that enhanced student affect will result in more usage of tutor, perseverance and, consequently, more learning. We make use of two strategies: affective repair and affective hook. *Mosaic* is a game where students solve different math problems to create colorful mosaics. We can use this game in two modes, as repair mode and hook mode. In affective repair mode, students first work on math tutor. When they show negative affective behavior such as boredom and frustration in the tutor, they are taken to *Mosaic*. We expect the students to have more positive affective state when they go back to the tutor. In affective hook mode, students first work on the *Mosaic*. If they make certain number of mistakes, they are required to master the skill in the tutor to be able to continue the game. Unlike the cognitive mode of intervention, games here are just a platform to use math skills, not necessarily actively teach the content. These simple

games may not be pedagogically rich but can be helpful to practice the skills that are very basic over and over and help students to get fluent.

We will analyze pros and cons of these different modes, by examining the outcome data such as learning gains, time on task and engagement. We assume that cognitive mode can generate higher learning gain as it directly involves teaching instead of supporting it via metacognitive and affective path. But this mode is also more susceptible to cognitive overload and demands more creative and careful implementation. Metacognitive and affective modes, even if they appear more superficial, are reusable across learning content and may produce learning benefits, particularly over the long-term.

This categorization into cognitive, metacognitive and affective modes is not exhaustive. But we see the value in such categorizations because it lets us compare and analyze advantages and limitations of different ways of using games. For example: we assume that cognitive mode can generate higher learning gain as it directly involves teaching instead of supporting it via metacognitive and affective path. But this mode is also more susceptible to cognitive overload and demands more creative and careful implementation. Metacognitive and affective modes may appear juxtaposed over the learning content but are reusable across learning content. We are also interested in observing how different interventions and game elements lead to different learning outcomes for different students.

RQ2: What are the causal mechanisms of learning outcomes in game-like interventions?

It is one thing to find that GLIs result in increasing learning; we would also like to understand why and how? Why do certain students, but not others benefit from our interventions? If games generate learning gain, it is because they are better cognitive tools or are they effective because students are spending more time on task due to increased engagement?

We are using a causal modeling framework to integrate and analyze student data collected from surveys, logs and tests to understand the interrelationships between different student and tutor variables. We have found causal modeling a superior approach to common statistical techniques such as correlation and multiple regression for generating a plausible set of hypotheses when using observational educational data sets (Rai & Beck, 2011). We can use it not only to confirm our prior hypothesis such as whether the game-like intervention has generated the outcomes expected but also to explore different causal mechanisms of such outcomes. For example: game-like intervention can lead to higher learning outcome only for the students who had higher time on task, or it could be effective irrespective of time on task which suggests that games can enhance learning beyond improving learner engagement. On the other hand, games may enhance engagement but also add cognitive overload. There might not be significantly visible overall learning outcome. But if we are able to measure these mediating variables, we will be able to understand the actual causal mechanisms and effects.

3.4 Description of three systems

In the following sections, I will be describing our three game-like interventions:

- Monkey's Revenge: Coordinate geometry learning environment
- My Progress Page: MathSpring student Dashboard
- Mosaic: Math fluency game

3.4.1 Monkey's Revenge: Coordinate geometry learning environment

Monkey's Revenge is a coordinate geometry math-learning environment with game-like elements. The system is basically a series of 8th grade coordinate geometry problems wrapped in a visual narrative. Students have to help story characters solve the problems in order to move the story forward. Similar to classic computer tutors such as ASSISTments (www.assistments.org), they get hints and bug messages when they stumble upon problem and misconceptions. In the story, a boy, Mike is thrown out of class for playing a game on his cell phone. He is happy to be outside in the sun but the day is going to be a strange one as his world is now mapped into coordinates. As a warm-up problem, students have to find out Mike's height in coordinate units based on the coordinate pair of his head. Mike finds a monkey and, being lonely, Mike wants to befriend him. Students can help Mike give a name to the monkey. Later Mike builds a house for the monkey, but the monkey is not eager to become domesticated (see Figure 4) and destroys the house, steals Mike's phone and runs away. The boy tries to get back his phone by throwing balls at the monkey. To move the story forward, the students have to

solve coordinate problems like calculating distance between the boy and the monkey, the slope of the roof and walls of the house, finding points where the monkey tied to a rope cannot reach bananas and finally figure out slopes, intercepts and equation of the line of the path of the ball. The math content gets more advanced as a student progresses within the story. We are trying to create an emotional dynamics where Mike wants to befriend the monkey but the monkey does not want to be domesticated. Along with this emotional element, we are also trying add humor and mischief and hitting each other with ball is more like a playful banter between siblings or owner/pet which is harmless and non violent.

We have an experimental framework where we make not only theoretical but also empirical evaluation of the game-like elements so as to make careful integration.

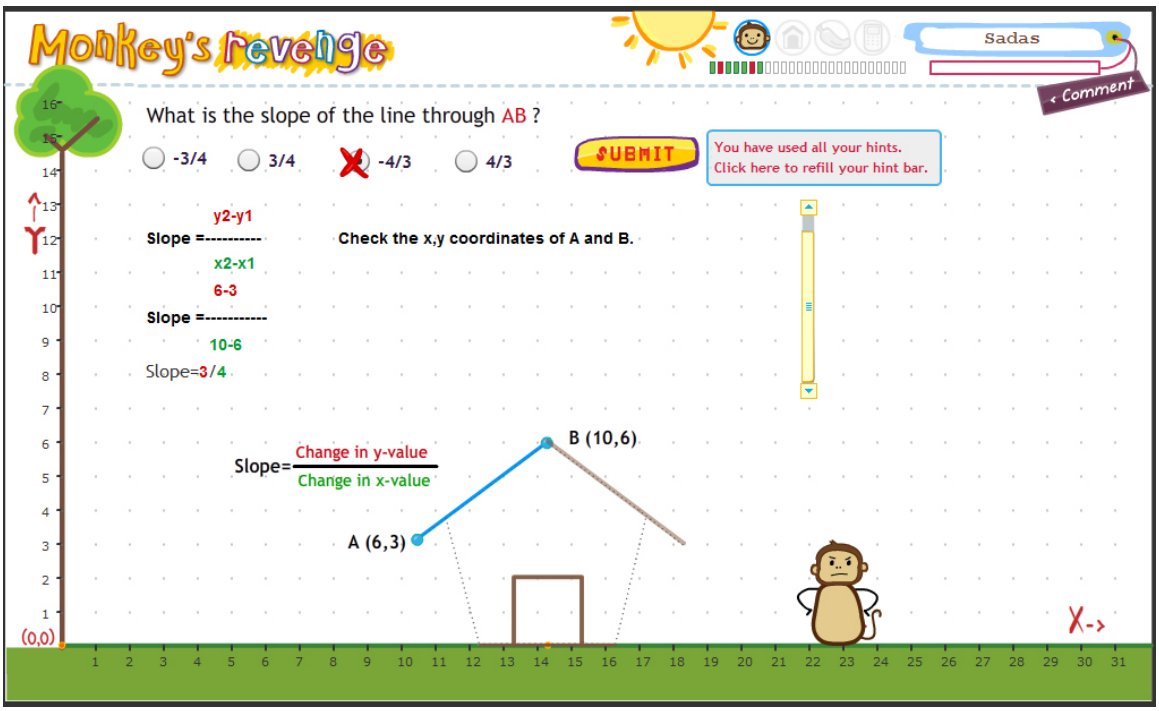


Figure 4 Screenshot of Monkey's Revenge

Integrating game-like elements into tutor

While there have been many analyses to assess the impact of such game elements in learning (Wilson et al., 2008; Aleven et al., 2010), there is still a dearth of controlled experimental studies of individual game-like elements. Therefore, our goal is to analyze and assess each game-like element and their impact on the learning environment.

As we incrementally add game-like elements into a tutor, we may expect to have increased fun (but not necessarily all the time). But given the complicated relation of games with learning as discussed in the previous section, we do not know how learning changes during the process. We have plotted three plausible tradeoff curves of making tutor more game-like in Figure 5.

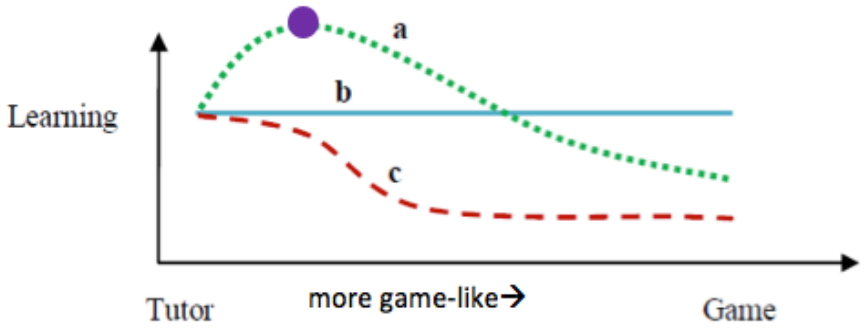


Figure 5 Three possible tradeoff curves for making tutors more like games

a. Some game-like elements can be pedagogically beneficial. For example, narrative can enhance learning by adding meaningful and interesting context to the learning content. But, there can be a tradeoff that reduces the benefit after some point. Once the narrative gets too elaborate and complex, it may make learning process complicated and confusing instead.

b. Some game-like elements may be orthogonal to learning content and may not interfere with, or directly benefit, learning.

c. Some game-like elements can hurt learning. For example: unguided exploration and pedagogically meaningful choices can leave students confused and possibly making suboptimal decisions.

We want to find the sweet spot where the addition of game-like elements maximizes learning. This graph is a simplified representation of the possibilities. It is conceivable

that game-like elements could synergize and enhance the effects, or interfere with each other and reduce their individual effects. Furthermore, the effects might not be constant, and could vary by context and by domain. Finally, the impact of a game-like element will depend on how well it is implemented and integrated into the learning context. We anticipate some potential concerns regarding our framework, which we address as follows:

1. *Is the tutor-game space really a smooth one with one global maxima or a rugged landscape with multiple minima and maxima?*

   The project is an attempt to begin to map the space and not necessarily mapping the entire space or asserting a particular global maxima is generally applicable. It is more about finding a rough map of the terrain.

2. *Is the finding too local around the content and the approach to game-design and implementation? (Was elaborate story really an ineffective game-like element, or was it just a bad implementation of an effective idea?)*

   The study is exploratory rather than prescriptive. For any serious research endeavor, a series of studies are of courses needed to fully explore possible design variants.

3. *Why isn't there a line representing learning increasing continuously as the learning environment becomes more game-like?*

It is implausible that adding more and more game-like elements to a learning environment without limit will result in more learning. After all, even computer games do not incorporate all conceivable game-like elements to avoid overloading the learner.

The “sweet spot” in Figure 5 is not meant to disparage games, and could in fact occur at or near the level of game-like elements found in many games (i.e., we are not yet ready to predict where the maximum will occur).

Game-like elements in Monkey’s Revenge

We carefully picked the game-like elements that we thought to be relevant and cognitively supportive to our content. We made our choice of game-like elements based on the following criterion.

Content and skill: A lot of math games have activities that involve fast reflexes and speeded responses. Such elements would be appropriate for development of skills such as mathematics fluency. But the skills we are trying to teach need more time to think and reflect.

Environment: One of our goals was to appeal to students who have poor self-concept in math and have math anxiety. Therefore, we did not want to have a competitive environment where those students might feel overwhelmed and anxious. Instead, we wanted to create a more relaxed, friendly environment. So, we did not include game-elements such as opponents, points and time pressure. We wanted to create a supportive rather than competitive environment.

In the following paragraphs, we will be discussing the game-like elements we chose and how they hold the potential for enhancing learning.

Embedding domain in a context

Authentic activities: One of the problems math learners face is that math tends to be abstract and they are not able to directly relate what they have learnt in their real life activities. Research on authentic learning has suggested that learning is more efficient and effective when it is embedded in realistic and relevant contexts (Shaffer et al., 1999). Fortunately, our domain of interest, coordinate geometry, has many concrete applications. We tried to incorporate those concrete activities, such as calculating slope of the roof of a house.

Narrative: We see the advantages of narrative in two ways. First, it entertains and engages learners and gives a meaningful context for solving problems. Second, if we use a coherent story, the initial story context can be reused for multiple problems, thus saving effort and cognitive load required reading context for each new word problem, particularly when compared to traditional word problems where the problems tend to have disjoint context.

Visual affordances

Visual problem representation: Graphics not only add appeal but they can help develop mental models, thus reducing the burden on working memory (Hegarty et al., 1995). We used very simple and minimalist visual representation so as not to interfere with the coordinate graph itself. As the problems get harder, they tend to be more abstract and it is harder and unintuitive to have concrete representations.

Immediate visual feedback: We have used immediate visual feedback for student responses to serve both engagement and learning objectives (see Figure 8). Immediate visual feedback makes the interface more interactive, giving users sense of control and reinforcement. When the feedback is appealing and interesting, it adds to sensory stimuli. While visual feedback on positive responses give students reinforcement, with visual feedback on wrong response, students can tell what the error was and how it relates to the correct solution (Malone, 1986).

Other game-like elements

Collection: Students can collect badges after each level as they master a sub-skill (see Figure 7). By tagging those badges with math skills, we wanted to create a tighter bond between the game-environment and content.

Building: Students have to solve different problems to build a house. Using various sub-skills to create a single structure, students can see how different mathematical concepts can be integrated within a single entity.

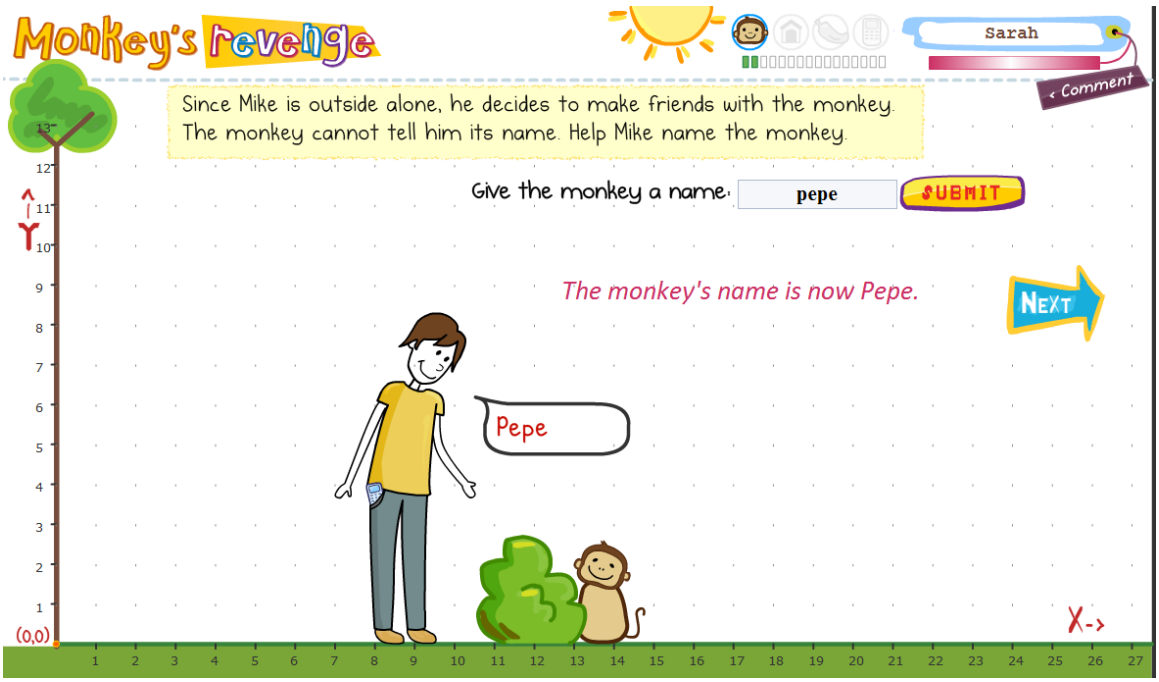


Figure 6 Students can help Mike decide name for the monkey

Personalization: Students can name the monkey (see Figure 6). Though this seems a small addition on the designer's part, students were very excited about this feature.

Sensory stimuli: We have used colorful visuals and animations as well as sound to make it appealing to the users.

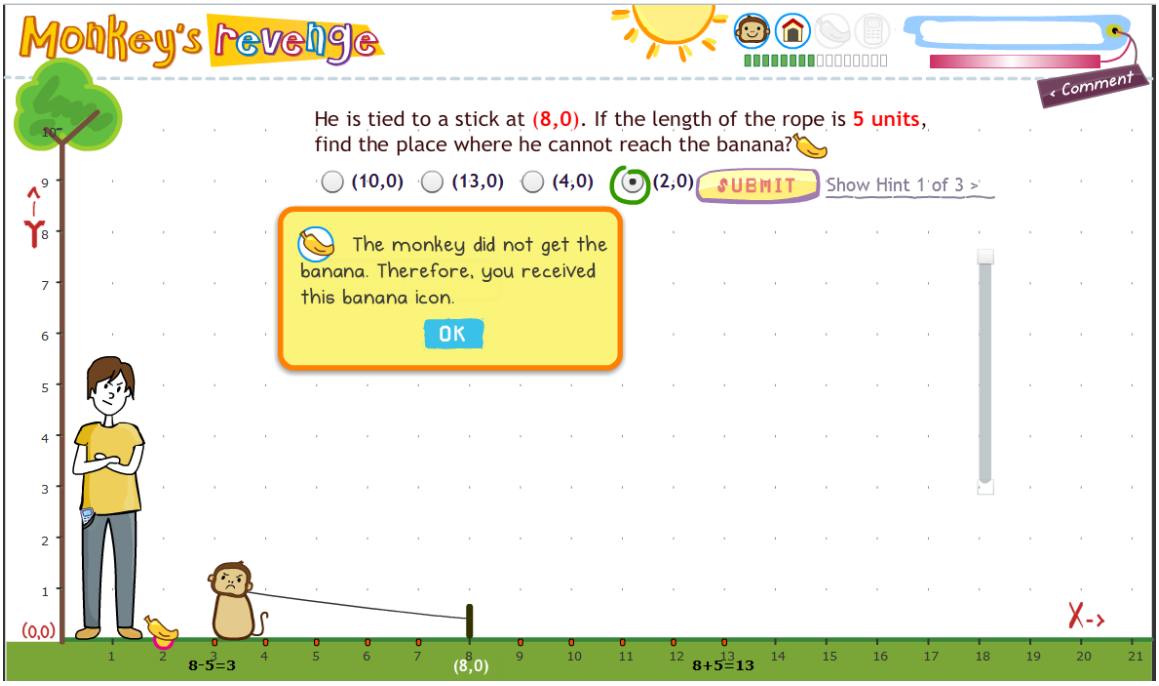


Figure 7 Students can collect badges

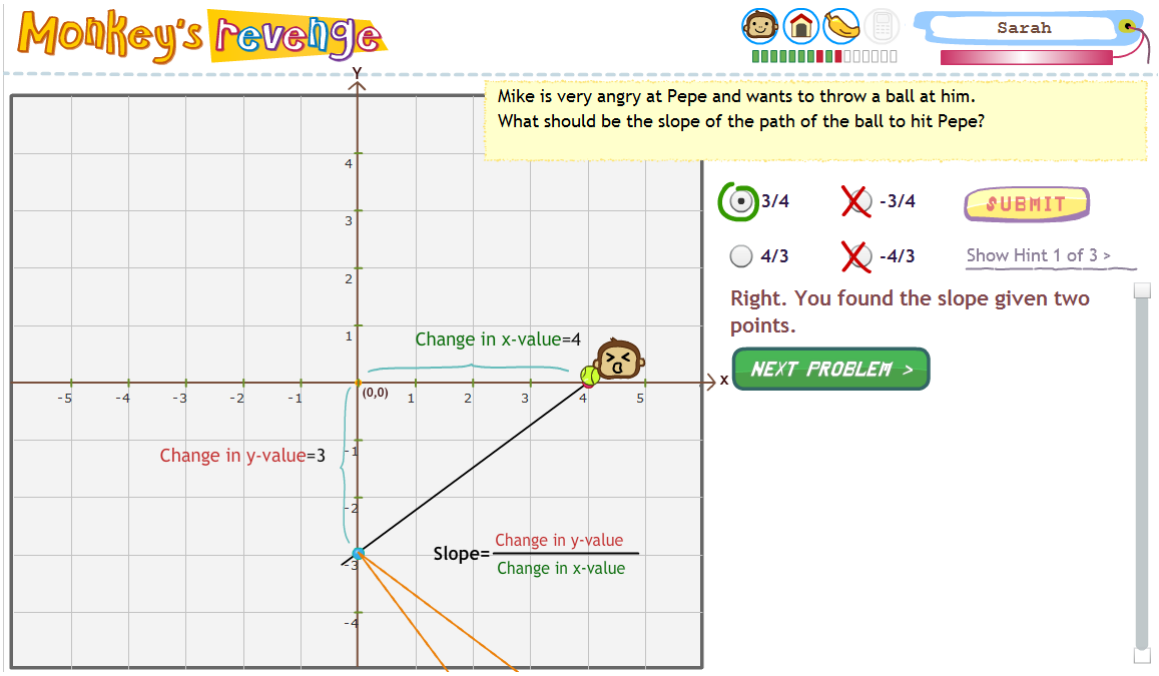


Figure 8 Immediate visual feedback for student responses

Design decisions

Finding fun in the learning

Klopfer et al. (2009) suggest that “finding the fun in the learning" and devising ways to focus on and enhance that fun as a core game dynamic is a good strategy. Solving abstract coordinate geometry problems can be fun for some students but not necessarily for all. But these abstract problems can be situated in concrete context, which can be more fun to the students. So, we have taken a strategy of situating these math problems in concrete scenario posing as interesting, relevant and meaningful challenges to the students weaving them together in a narrative.

Accessibility

Appeal to entire population: Educational materials should appeal to all members in the target population, or at least all the students using the system. We have used a male protagonist but assume that girls will also enjoy the narrative based on emotional dynamics between the protagonist and the monkey. The narrative has different aspects such as emotion, mischief, and humor. We have also tried to make the color theme of the interface gender neutral.

Complexity: Educational games should assume very little or no game literacy among users. Hence, the complexity of interactions should be very simple. We are using a classic tutor interaction in our tutor.

Time Overhead

Details and fidelity: Due to the constraint of limited intervention time, we have put effort to make narrative interesting without adding too many details.

Cognitive Overload

Minimal visual presentation: We have used very minimal visual representation so as not to overwhelm users with too much detail.

Novelty in narrative: Due to concerns of cognitive overload, we have not used very novel scenarios but have rather used very familiar characters and events like a classroom and a mischievous monkey.

Concreteness fading: As the problems get harder, they tend to become more abstract and it is harder and counterintuitive to have concrete representations. Therefore, we have adopted a strategy of making the representations more concrete at first (story characters shown as cartoon image, as in Figure 7) and becoming less so as we proceed (story characters are abstracted to dots, as in Figure 8). Initial concrete grounding facilitates interpretation in later problems (Goldstone et al., 2005).

3.4.2 The Learning Dashboard (Student Progress Page)

MathSpring is an intelligent mathematics tutoring system for grades 5-12 developed at UMass-Amherst (see **Figure 9**). MathSpring targets the mathematics curriculum of grades 6 through 11. It covers a large range of topics including number sense, pre-algebra, algebra, geometry, logical reasoning. The pedagogical approach of the MathSpring Tutor is based on cognitive apprenticeship and mastery learning, and its internal mechanism is based on empirical estimates of problem difficulty and a variety of parameters that can be set by the researchers or the teacher (Arroyo et al, 2010). In this case, the cognitive expert is the computer, who assists the student with tackling challenging tasks.

Metacognitive Support

Metacognition describes student's knowledge about their own strategies for learning, and when and how to apply them. The terms self-regulation and executive control are related to metacognition. It is also referred as "knowing about knowing" (Metcalfe & Shimamura, 1994). In addition to metacognition, there is also the idea of metaemotion, or students' ability and methods to regulate their own emotions (Mitmansgruber, 2009). In the academic domain, students use a variety of coping strategies to regulate their emotions in stressful learning situations, including humor and acceptance, social-emotional coping, abandoning/avoidance, and negation , suggesting some students need support to develop more productive strategies to cope. Prior research showed positive evidence for the impact of basic progress charts showing progress on the last 5 problems

(Arroyo et al., 2007) Students in the experimental condition received charts showing both negative and positive progress and had higher perceptions of the system, higher likelihood of re-engagement after seeing the chart (transitioning from a disengagement to an engagement state), and also higher learning rates and performance in standardized tests.

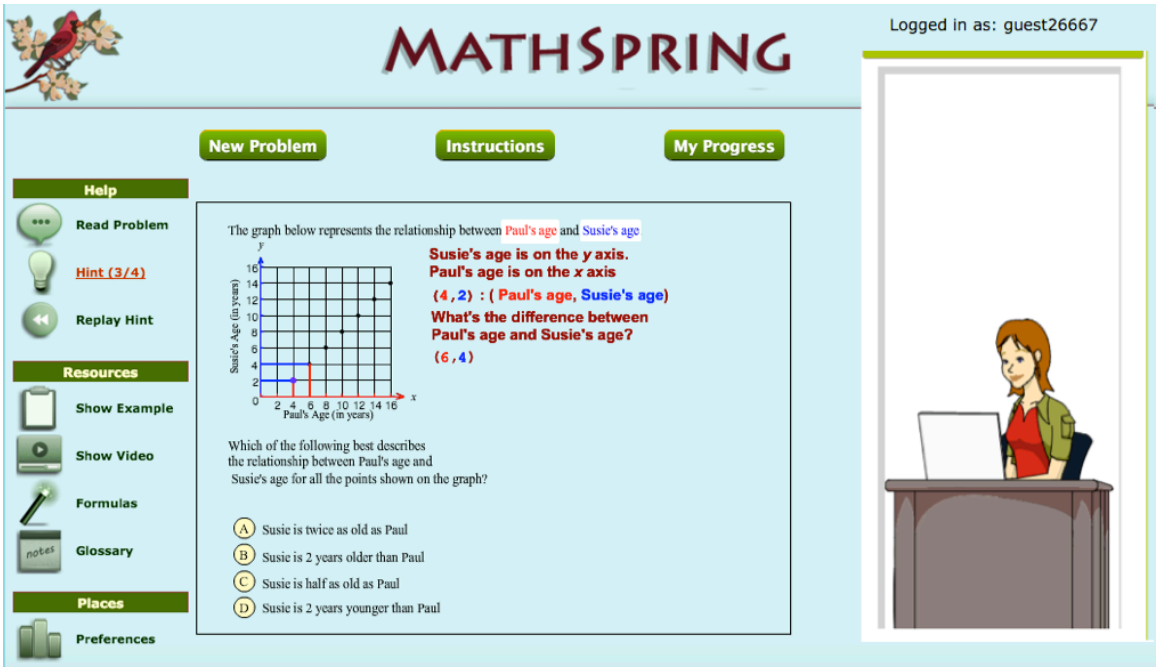


Figure 9. Screenshot of Mathspring. Learning companions use gestures to offer advice and encouragement. Animations, videos and worked-out examples add to the spoken hints about the steps in a problem. My Progress button allows students to access the student progress page.

The Learning Dashboard

A Learning Dashboard is a generic name given to tools that provide visibility into key learning indicators through simple visual graphics such as gauges, charts and tables (Brown et al., 2006). It presents a wide number of different metrics in a single

consolidated view and roll-up details into high-level summaries. Though digital dashboards have been an established practice and teachers' dashboards have been used fairly commonly in computer based education, learning dashboards for students are still novel approach. In Khan Academy (www.khanacademy.org) a students' homepage is a learning dashboard with gamification features. Learning dashboards create an individualized report summary by intersecting content-related and self-related dimensions for each student. Verpoorten (2011) makes the argument that this crisscrossing between content-related and self-related dimensions, arranged within permanent, visual and dynamic displays, is a new phenomenon in the practice of formal eLearning education. Its emergence stands at the cross-section of reflective practice, self-regulation and personalization issues. Teaching learners to engage with learning dashboards may cultivate awareness and coordination of the various personal and contextual dimensions of learning.

I created a Learning Dashboard for Mathspring with the aim of giving metacognitive support to the students. There are three distinct pages comprising our Learning Dashboard. Those three pages display information at different granularity level.

1. A Personalized Math Tree (*Domain* level feedback on the student's overall performance in Mathspring)
2. A Student Progress Page (*Topic* level feedback on each math topic, e.g. "fractions")
3. Topic Details (*Problem* level feedback on each problem within a topic)

The Math Tree: A student's overall performance in Mathspring is represented by a math tree. As each student logs in Mathspring for the first time, they are given a baby tree (Figure 10). As the student works on math problems in Mathspring, the baby tree grows. The tree generates new branches as the student work on new math topics. The tree gives blossoms for the topics that the student masters (Figure 11). Students can observe how the tree grew over different days that they worked in the tutor, by clicking on buttons for each corresponding day that they used the system.

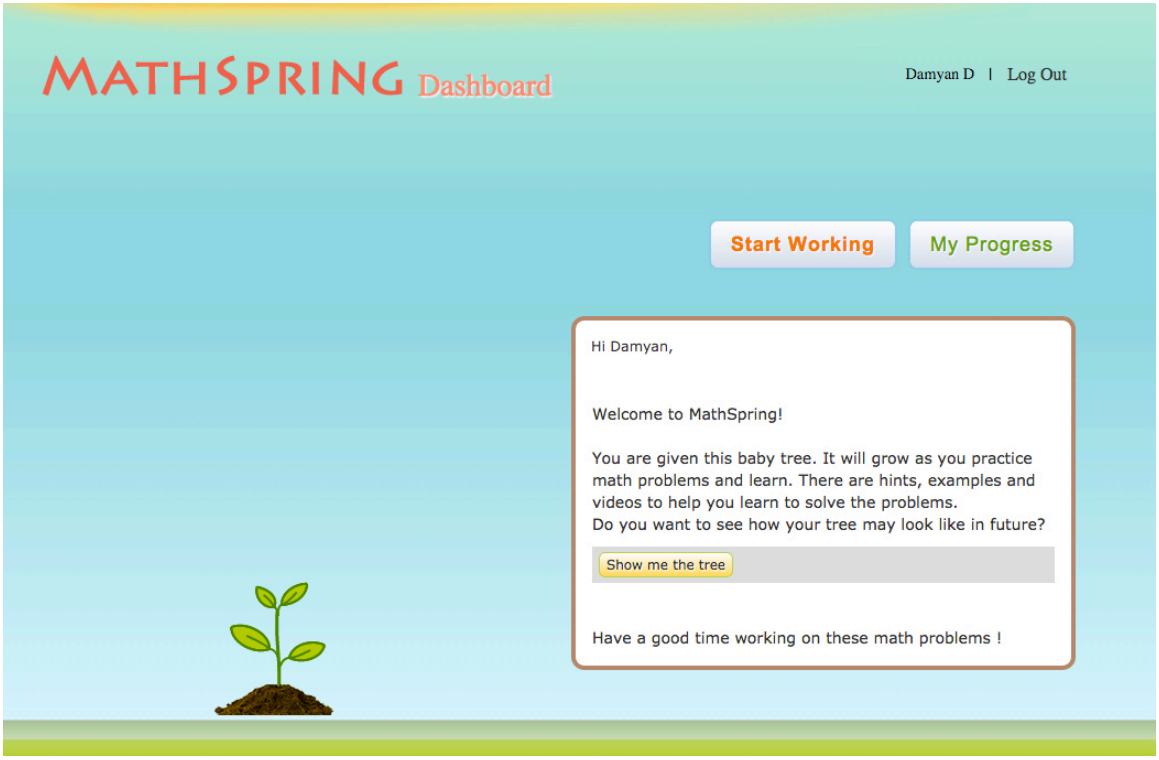


Figure 10 Math Tree on day 1

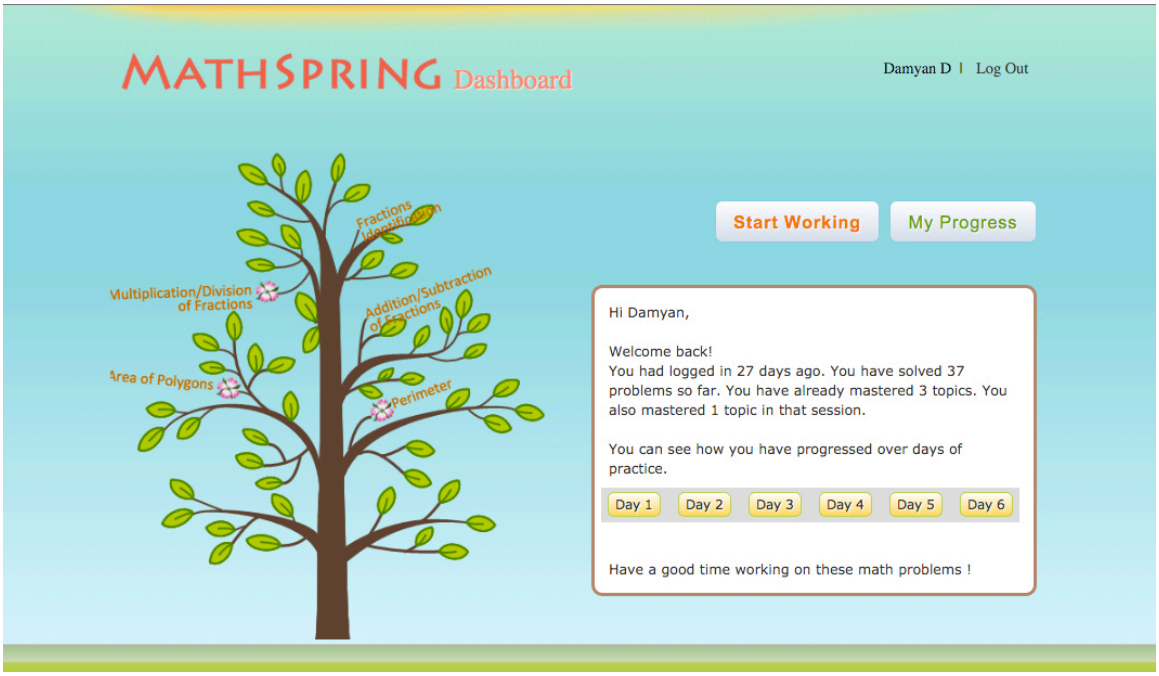


Figure 11 The Math Tree: a visual representation of student performance in Mathspring

The Student Progress Page. The Student Progress Page (SPP) within MathSpring supports students to observe their performance and the tutor's assessment and feedback (Figure 12). The page lists math topics (rows) and provides sophisticated meta-cognitive scaffolding to support students to reflect on the tutor's assessment of their knowledge (column 3) and effort (column 2). For example, it provides an intuitive and appealing representation of effort using a potted plant. The plant grows as students put more effort into solving problems and bears fruit when a topic is mastered. The plant withers when there is lack of student effort. We hypothesized that this intervention would help students stop to think, and then re-engage students in the learning activity, becoming somewhat more active in their learning progress, and hopefully act as a mechanism to address the occurrence of deactivating negative emotions. The page provides a row for each topic consisting of details:

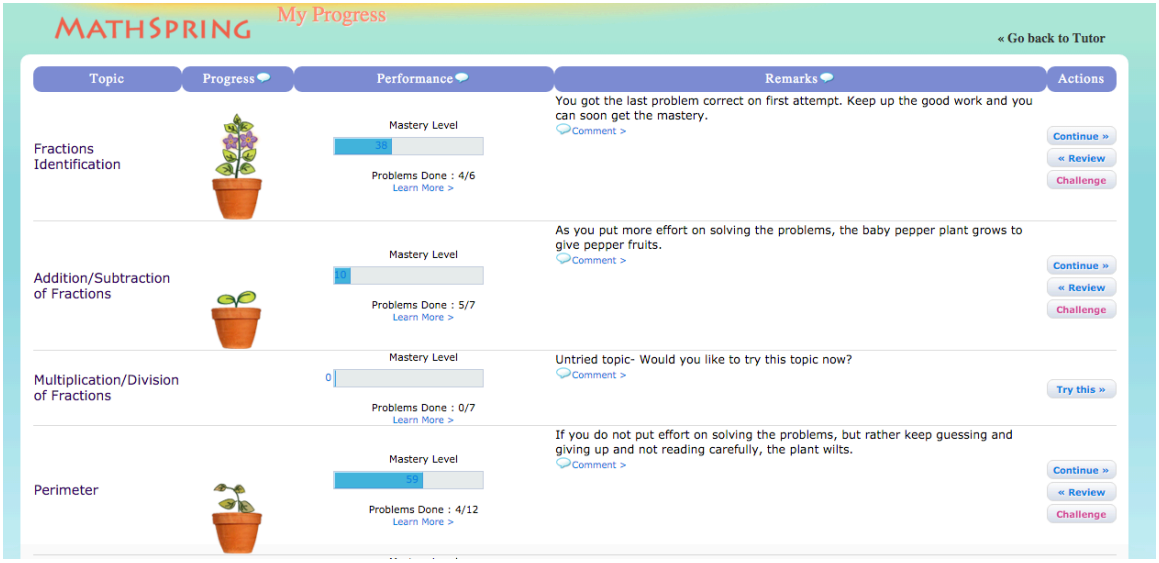


Figure 12 The Student Progress Page (SPP) encourages students to reflect about their progress on each topic (column 1) and to make informed decision about future choices. The plant (column 2) demonstrates an assessment of student effort and the mastery bar (column 3) assesses students' knowledge. The tutor comments about student behavior (column 4) and offers students the choice to continue, review or challenge themselves (column 5).

- Mastery level: This is a probabilistic assessment of students' knowledge in each topic (column 3)
- Progress: This is a measure of students' effort and subsequent progress in a topic (column 2). The tutor makes this inference based on student performance behavior (e.g., solved problem with help aids, not reading problems thoroughly).
- Feedback: The tutor reviews students' overall performance level and behavior in a topic and provides customized feedback (column 4). For example: "That last problem was a hard one. Good work!"; "Are you frustrated? Raise your hand and someone will help you. "
- Navigation: Students can choose different modes of navigation to work on further problems: review older problems or work on higher difficulty 'challenge' problems

(column 5). The tutor also provides recommendations about navigation: "You have already mastered this topic. Maybe you should try 'challenge' problems or a new topic."

Students can also give their own feedback to the tutor on whether they agree with the tutor's assessment and recommendations. They can click on each topic to get problem specific details for each topic and navigate to specific problems.

Topic Details: When student selects a topic in SPP, they go to a "Topic Details" page (***Figure 13***) which shows the details of student performance within the topic. All problems within the topic are listed in increasing difficulty, from left to right. Each individual problem is represented by a domino that is marked according to their performance (for example: a 'star' to represent a problem correctly solved; an 'H' for problems solved with hints and an exclamation mark ('!') to represent disengaged behaviors). Once the student clicks on each problem, details of the problem are shown and the student can choose to work on that specific problem.

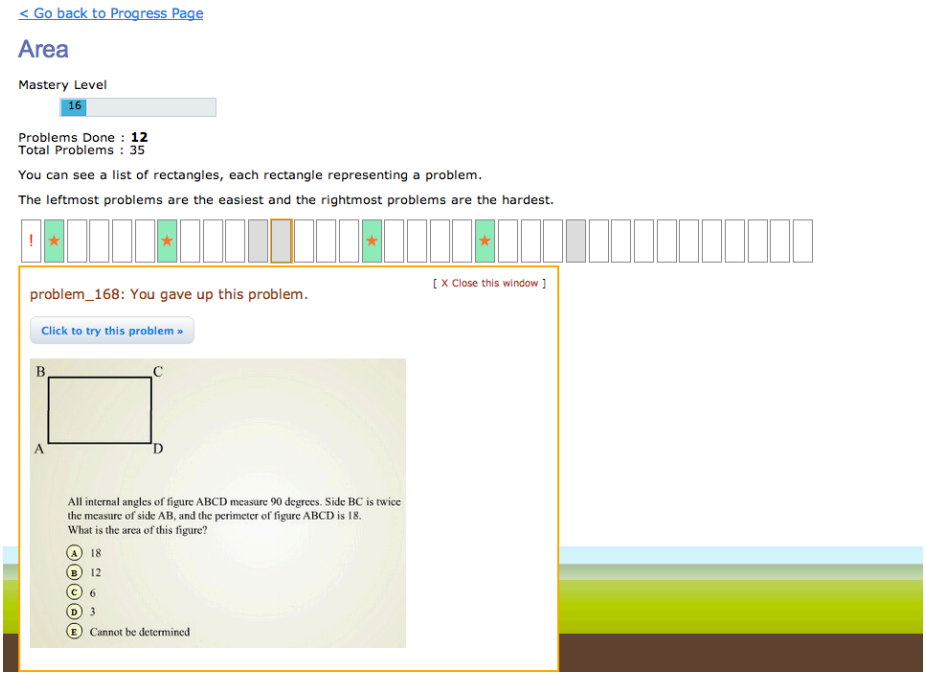


Figure 13. Clicking a topic in the SPP produces a list of every problem in the system represented as a domino (center) and each problem is shown in detail (below). The topic detail in the SPP provides a view of every problem for each problem set that the student should go through.

Learning Dashboards as Open Learner Models: The Learning Dashboard is also a step towards open student models, which are learner models that can be viewed or accessed by learners. Thus, in addition to the standard purpose of a learner model to maintain data that enables adaptation of a tutor to the individual according to their current learning needs, the learner model contents can also be of direct use to the user (Bull & Kay, 2007). This approach has been increasingly adopted by various learning systems (Bull 2012, Matthews et al., 2012). Benefits of open learner models (Bull & Kay, 2007) that apply to our Learning Dashboard include: promoting metacognitive activities (reflection, planning and self-monitoring; supporting learners to take greater control and responsibility over their learning, encouraging learner independence and agency; increasing learner trust in an adaptive educational environment; and increasing the accuracy of the learner model by supporting the user to contribute additional or corrective information.

Besides providing metacognitive support and guidance to students, we expected that the SPP also would directly enhance a student's affective state and promote effective engagement and learning behavior, thereby leading to higher learning.

The Learning Dashboard as a Game-like Intervention: Our primary goal while creating a Learning Dashboard was to provide metacognitive support. But we also wanted the experience to be positive and engaging, which is why we added game elements to the Learning Dashboard.

Gamification is increasingly popular while giving feedback to students on their performance. Game elements such as rewards, badges and leaderboards are used. In the design of Learning Dashboard, I have used Gamification features such as rewards, bonuses and mastery bars. There has been criticism against Gamification that it is shallow and manipulative and it may hinder intrinsic motivation. Keeping this concern in mind, I tried to not just offer rewards in superficial way but make those rewards meaningful. The pepper plants are an intuitive representation of metacognitive information. Plant size, flowers and pepper fruits convey students' performance and knowledge in compact and intuitive way.

Creating the Learning Dashboard

Creating the Learning Dashboard consisted of two major steps:

1. Creating a finer-grained model of student perfomance
2. Presenting the information of student preformance in game-like form

Creating a finer-grained model of Student Performance

In collaboration with other members of the research team at UMass Amherst, we created a finer grained model of student performance so that we would be able to parse and interpret student behavior that we considered crucial to student performance.

First, we calculated a variety of performance metrics (correctness, time taken and help

usage) of student actions within a problem. Then we labeled performance on each problem with one of the following six student states as indicated in shown in table 1.

We interpret a students' state of mind as a reflection of a student's knowledge and effort. For example: SOF (Solved on first) is a good measure of high knowledge. SHINT (answered with some use of hints, but not all hints) is a reflection of engaged behavior. GUESS (answered after several attempts) could be a gaming behavior and NOTR (Not giving enough time to read) is another form of disengaged behavior. The calculation and interpretation of these student states are local to our system, that might not be transferable to other learning systems. For instance, learning systems with open ended questions (instead of Mathspring's multiple choice format) might interpret similar actions to these presented above differently. For example, asking for hints is good evidence of positive learning behavior in Mathspring but may not be as strong evidence of positive learning behavior in another tutor with open ended response. A disengaged student in Mathspring can simply guess (a maximum of four attempts is enough) to move ahead whereas in a tutor with open response, guessing would not be as efficient and a disengaged student would go for bottom-out hinting (meaning that they might go ahead to get all the hints until the answer is revealed).

Table 1 Student States

Student State	Description
SOF	Solved on first attempt without help
ATT	Answered after 1-2 incorrect attempts and self- corrected, without help.
SHINT	Answered with the use of some help, in at most 2 attempts.
GUESS	Answered after several attempts, more than 2 attempts
NOTR	Not giving enough time to read
GIVEUP	Enough time to read, but moved on before answering.

We can trace the student states in knowledge-effort quadrant as in Figure 14.

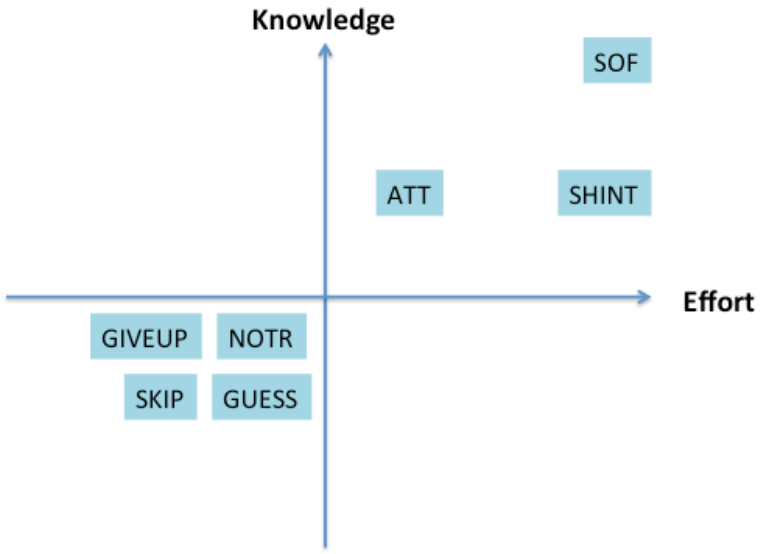


Figure 14 Student state variables in knowledge-effort quadrants

SOF is an indicator of high knowledge and high effort. SHINT is an indicator of high effort. GIVEUP, NOTR, SKIP, GUESS are indicators of low effort. These states do not

give evidence of high knowledge and therefore, we kept them in the lower quadrant of knowledge.

Second, a heuristic was created on how to identify crucial moments in student's performance. Because we don't want to overwhelm students with too much detail, we need to identify salient actions, moments that are crucial to the learning process, important moments to record so as to either highlight or intervene, milestones that allow us to draw meaningful conclusions based on students' performance.

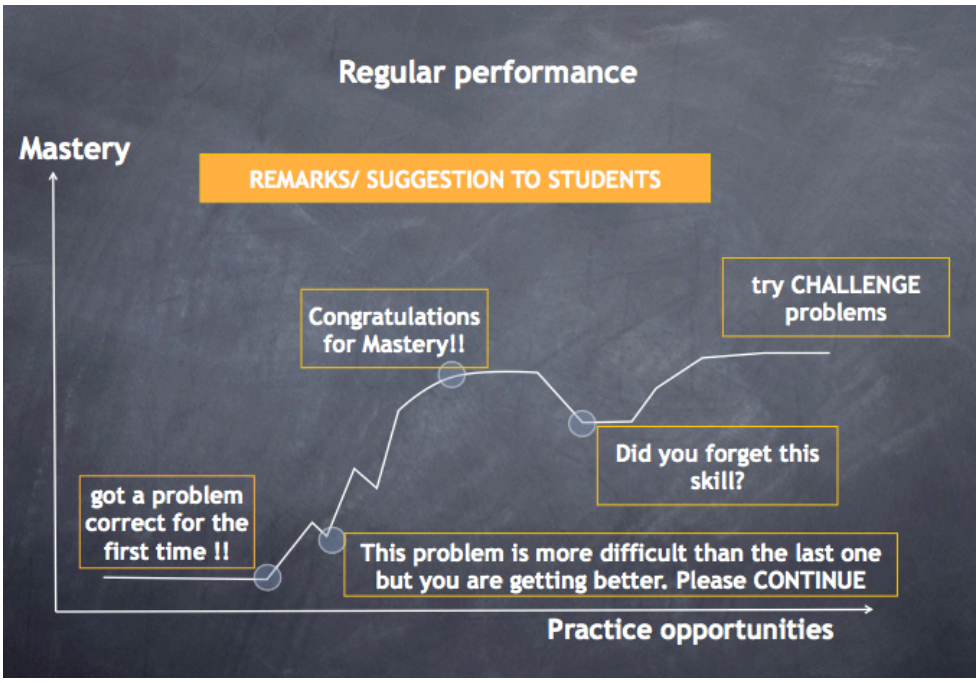


Figure 15 Sample diagram of how salient aspects of a student's performance are determined

Figure 15 is a simplified diagram of how I outlined important actions and major events in the course of learning and performance. For example, we would like to congratulate

students for getting a problem of a kind correct for the first time. Instead, if they have already mastered a skill, we could offer them to try a 'challenge' problem.

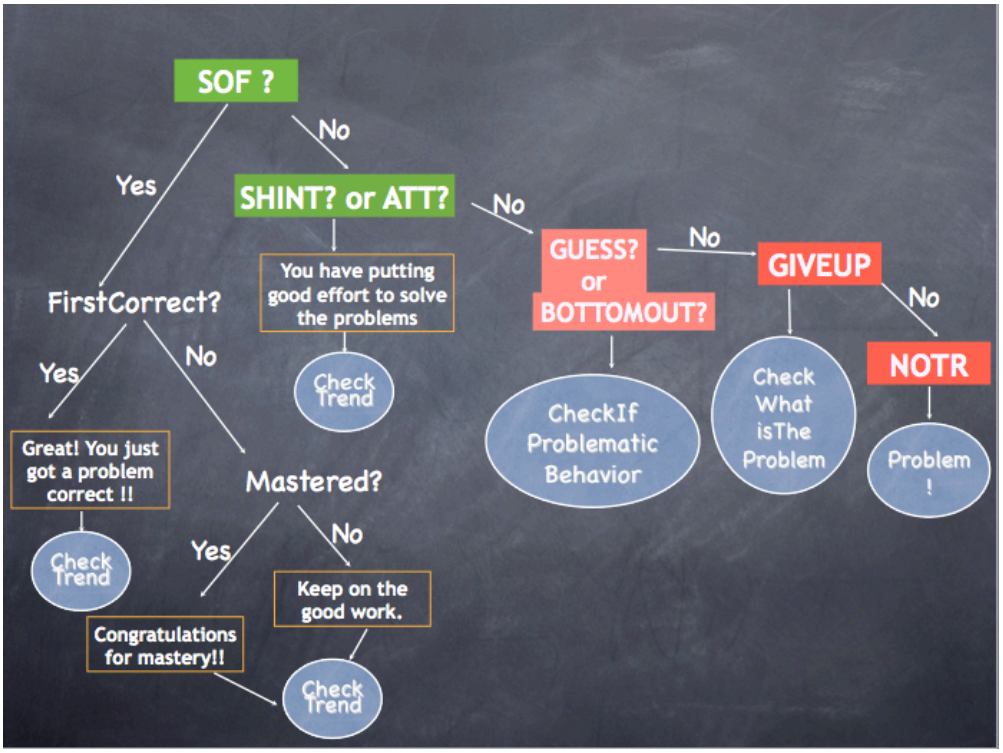


Figure 16 Simplified version of the decision tree that delivers feedback based on a variety of student performance states at math problems

Third, a decision tree was created to connect students' performance with the intervention and feedback we would like to offer. Our decision tree goes over the sequence of actions carried out within a topic/problem set and categorizes student performance into one of a certain kind: regular performance, exceptional performance (good help usage, mastery), disengaged behavior (guessing, skipping, not reading, giving up), etc. and assigns relevant feedback. Figure 16 is a simplified version of our decision tree. The actual decision tree consists of more than thirty nodes.

We give positive feedback for good behavior and alarm students and give them corrective feedback when they demonstrate disengaged behaviors. When we identify disengaged learning behaviors, we try to find out the cause first, and give them appropriate suggestions and recommendations later.

Game-like elements in Learning Dashboard

I wanted to present the information on student's performance in an intuitive and engaging way. Thus, I chose a potted plant as the representation of a students' effort on a topic. The plant grows when the student puts effort to solve the math problem (SOF or SHINT or ATT). If the student does not put effort but shows a disengaged behavior (NOTR or GIVEUP or GUESS or BOTT), the plant will not grow and will eventually wilt.

Figure 17 Effort on each topic is represented by a potted plant

I defined and implemented a reward and loss narrative to encourage good engaged behavior and discourage disengaged behavior that would not be conducive to learning. I wanted to encourage help seeking within the tutor so that the student can actually obtain support and learn in moments when they actually need the help (they do not know how to

proceed in solving the problem). On the other hand, disengaged behaviors cause the plant to wilt.

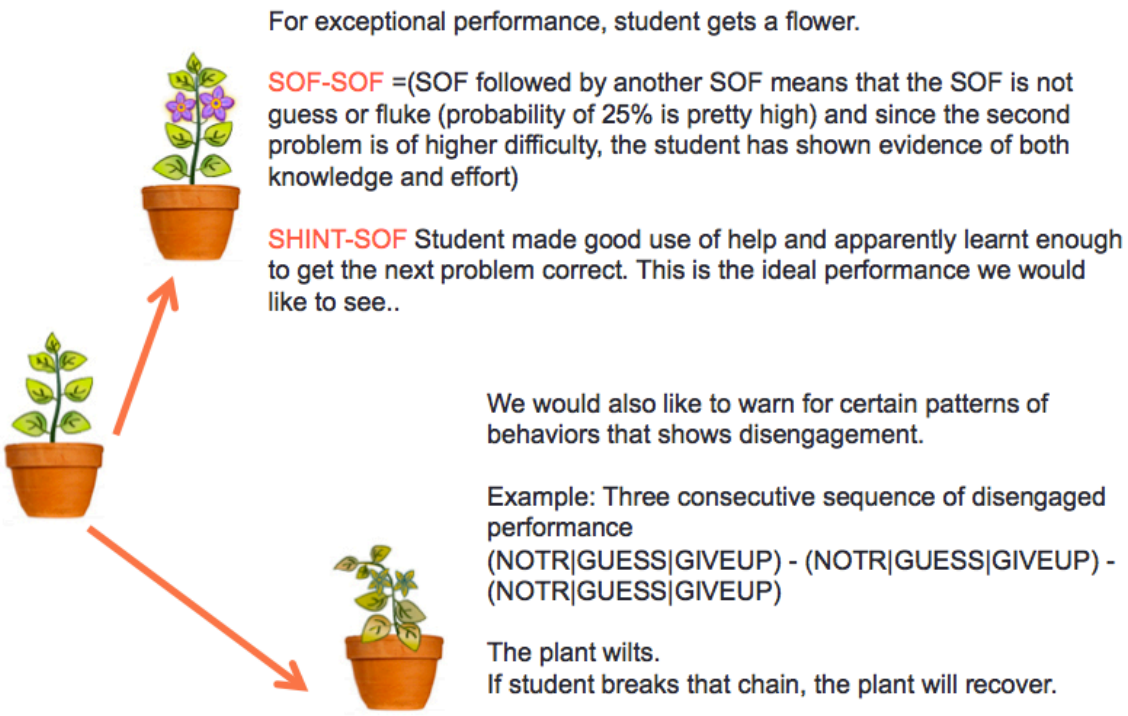


Figure 18 rewards and loss

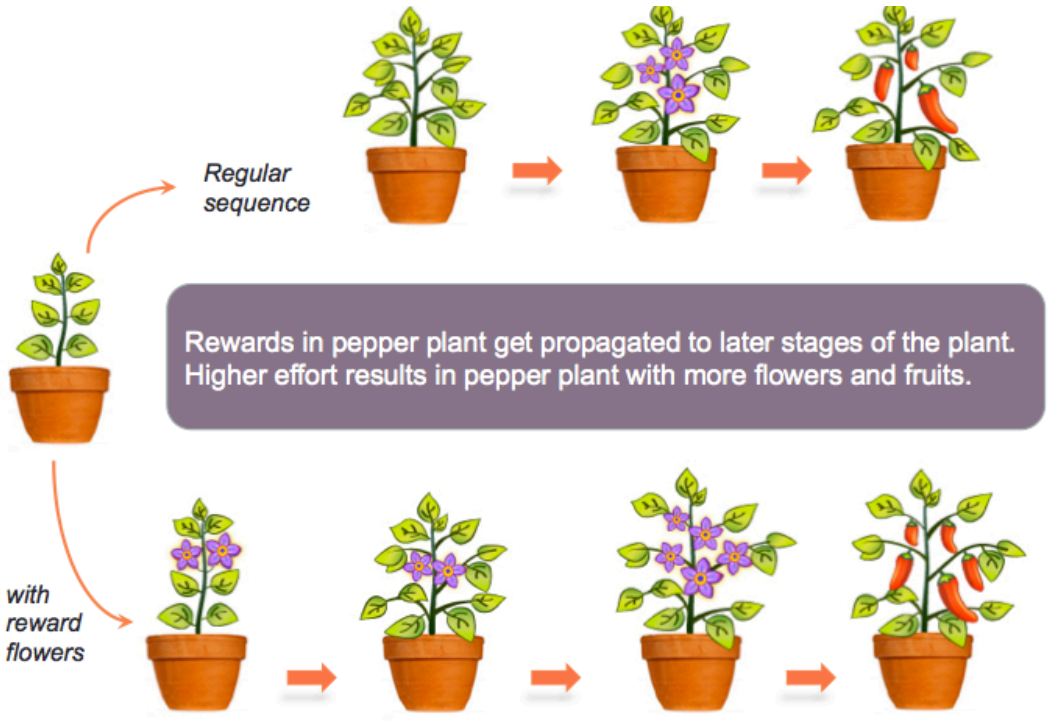


Figure 19 Rewards for good help usage. Students who show frequent help usage behaviors have a richer looking pepper plant.

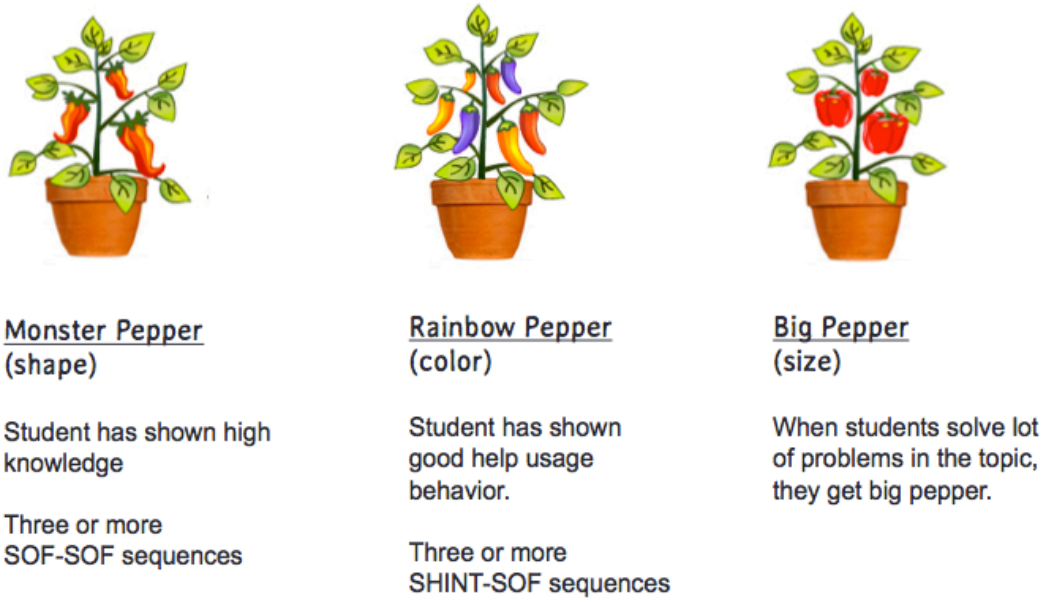


Figure 20 Exceptional performance is rewarded with special pepper plants.

Apart from pepper plants, the math tree (which students can see after log in, or when clicking "home") also has both a metacognitive and affective purpose. The goal is that students get a sense of ownership, a sense of personalized tutor behaviors and personalized rewards, being able to observe their personal progress as the math tree grows with their achievements.

3.4.3 Mosaic: Math mini-games

"Mosaic" is a math game where players solve math problems to generate tiles inside a mosaic. We have created Mosaics of different cities. Each city is associated with a math skill. For example: New York is associated with 'calculating area and perimeter' (see Figure 21), London with 'Fractions'. At first, mosaics are empty with general outline of cities. Once players start solving problems, tiles will be generated and the mosaic gets colored and complete gradually. There are two levels of the game:

Level 1: Players solve problems until the mosaic is filled with tiles.
Level 2: Players solve problems under fixed time interval. The faster they solve problems, they can get more tiles and hence the mosaic is more intricate.

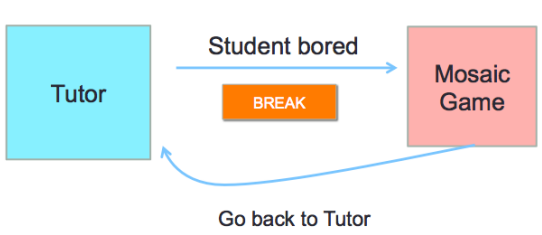
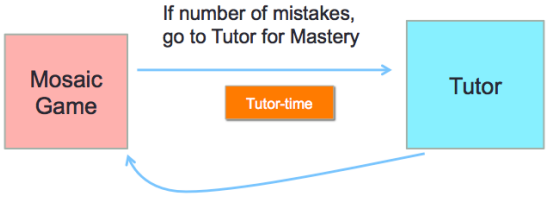
The reason we created those two levels is because we don't want to alienate the students who get anxious under time pressure and we also want to appeal to the students who get thrilled by race against time.

There are different game mechanics used within Mosaic: Tile-laying, completing pattern, structure building, behavioral momentum, race with time and quests.

Mosaic as affective intervention

We are using Mosaic as an affective intervention. Mosaic does not actively teach students at solving math problems but is used as an affective boost/affective hook for a math tutor that would teach the math skills. For example: students are working on perimeter problems in Mathspring and if they are found to be disengaged, they are directed to Mosaic. If the students make certain number of mistakes within Mosaic, they are directed back to the tutor. The students have to master the skill in the tutor to be able to continue playing in Mosaic. Basically, tutor is taking care of teaching the students while Mosaic is an intervention to give students an affective boost by giving them something fun to work on (and taking them out of boredom) and also as an affective hook so that students would work on the tutor to master skill so that they are able to play Mosaic. Ineffective integration of 'game elements' and 'pedagogical elements' may lead to cognitive overload and also may reduce intrinsic motivation. In Mosaic paradigm, students 'practice' in the game while they 'learn' in the tutor. We are therefore keeping 'game elements' and 'pedagogical elements' separate. Our hope is that students will get affective boost from playing the games so that they would be positive and engaged about learning from the tutor.

Table 2 Two modes of affective intervention

Mosaic as affective repair	Mosaic as affective hook
	

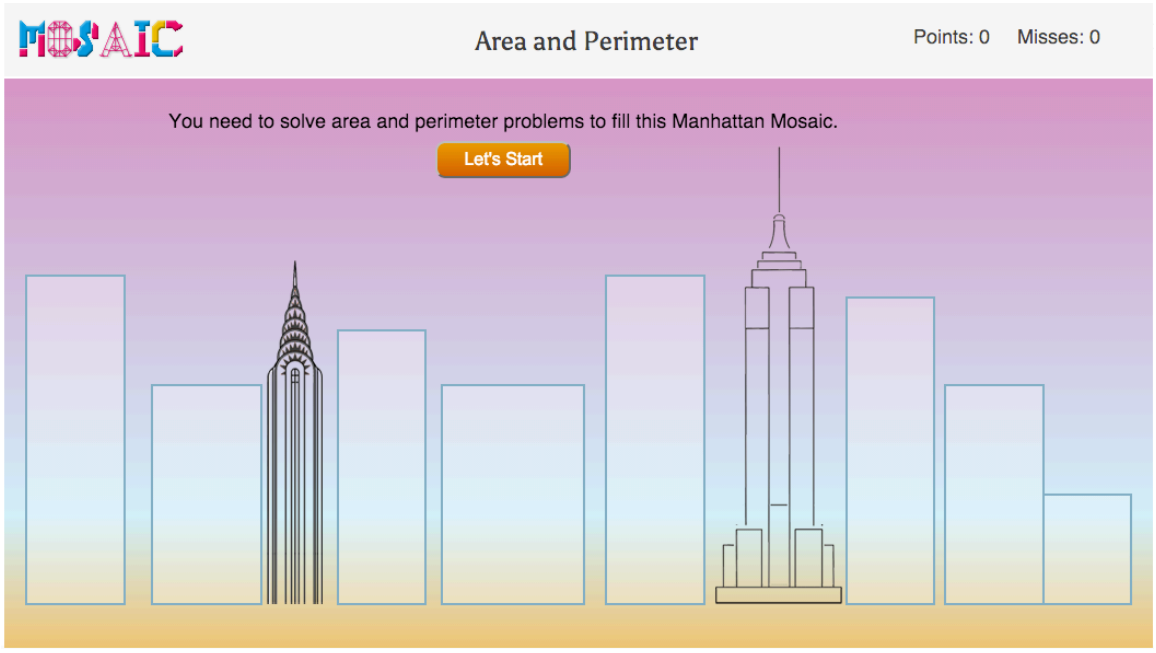


Figure 21: Solving area and perimeter of rectangular shapes generate the colorful rectangular tiles which fill the Manhattan mosaic. Behavioral momentum of solving the problems and generating the tiles is expected to be enjoyable.

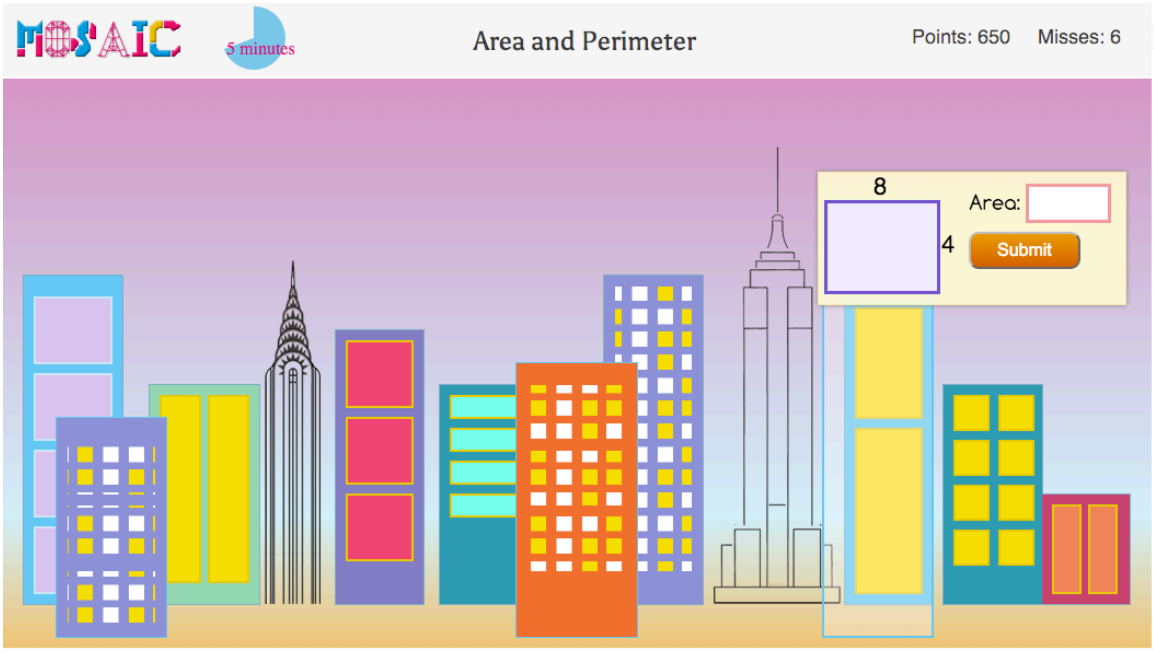


Figure 22: When level 1 of mosaic is complete, players get to level 2. In level 2, players solve problems under fixed time. The faster they are, they can solve more problems and get more tiles and hence their mosaic is more intricate.

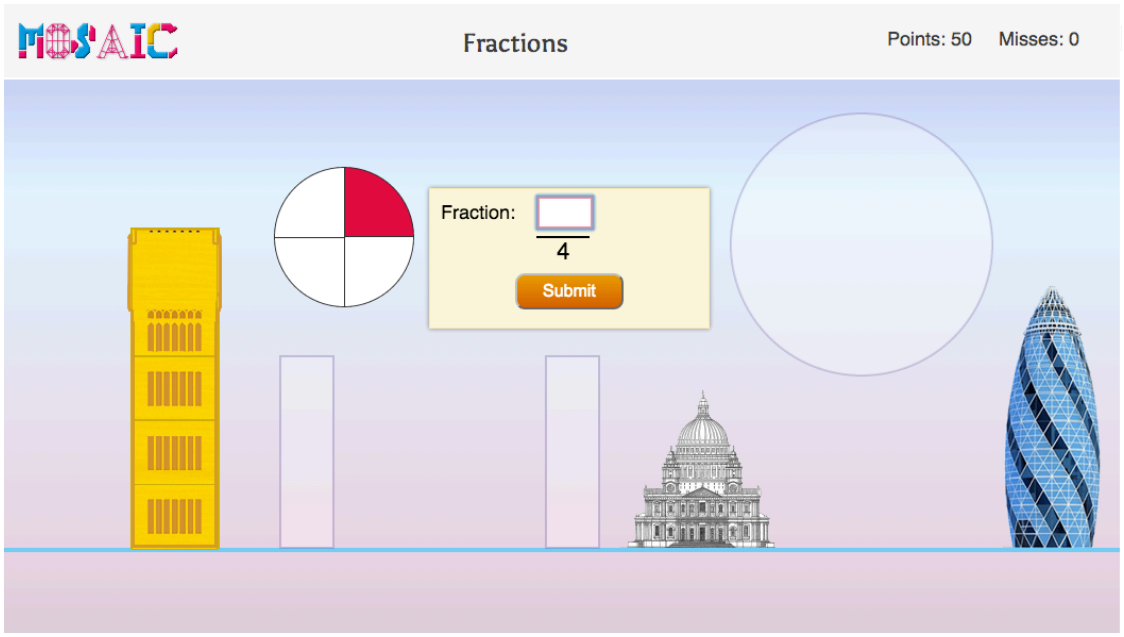


Figure 23: Students solve fraction problems to generate tiles completing London Mosaic. The problems increase on difficulty as students progress while maintaining a rhythmic momentum

4 Experiments and Analysis

We developed three interventions and conducted multiple studies. Figure 24 is a snapshot of the experiments we conducted.

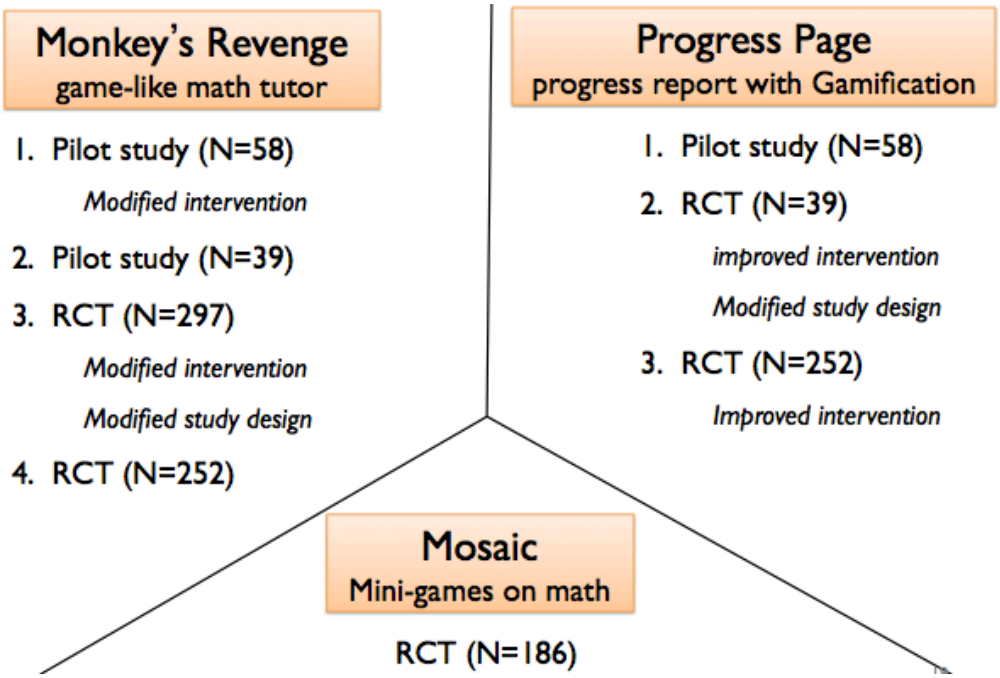


Figure 24 Overview of experiments with our three interventions

4.1 Experiments with Monkey's Revenge

With Monkey's Revenge, we ran a total of four studies: two pilot studies and two main studies.

4.1.1 Mily's World

Mily's World was our first effort in creating a game-like math tutor. In Mily's World, students meet Mily, a 9-year old girl who is the protagonist of the narrative. She has a puppy and some friends with whom she plays soccer. Students are engaged in many different math-related tasks. For example, they calculate Mily's height and the distance between her and her puppy based on the coordinates of their heads. As they proceed, students help Mily decide the name of the puppy and then help create a doghouse (see Figure 25). When students give the correct answer for slopes, the doghouse wall and roofs are built gradually and then a new doghouse pops up. The puppy develops a bad habit of chewing socks; so Mily ties him to a post. Students have to help her find the coordinates of a position to place the socks where the puppy cannot reach them. Afterwards, Mily goes out with her friends to play soccer wearing the socks that the students have kept the puppy from chewing. Here, students have to calculate slopes and equations of the path of the ball as Mily and her friends play.

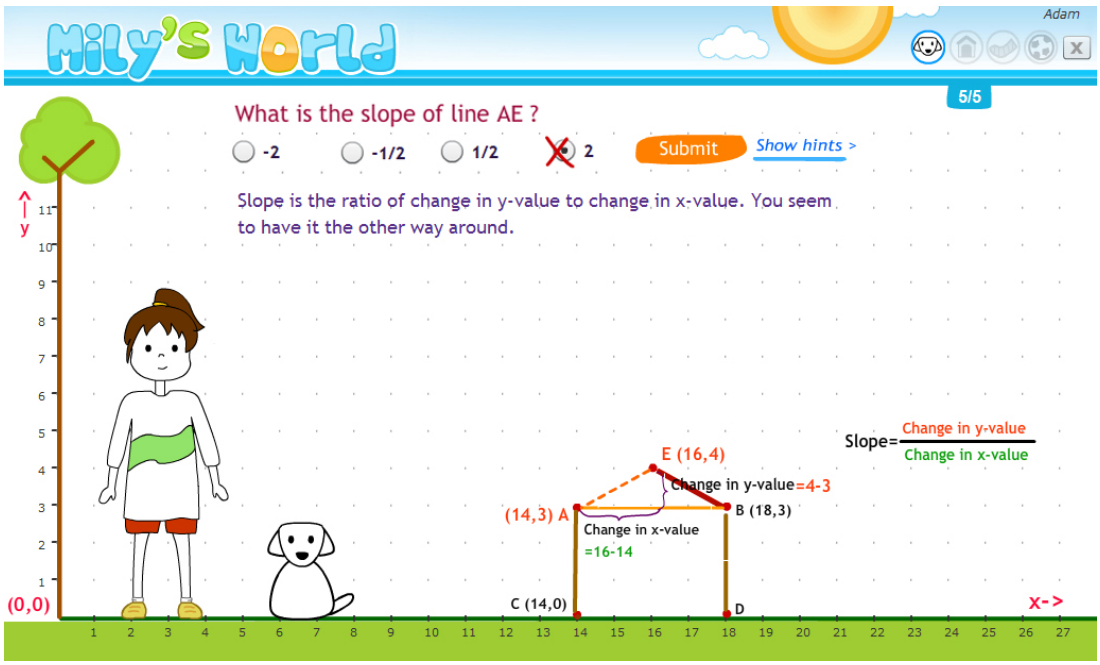


Figure 25 Screenshot of Mily's World

Mily's World was assigned as homework to 8th grade students (12-14 year olds) in a school in the suburb of a small city in the Northeastern USA. Sixty six students started the exercise and 58 students completed it. Those students also used ASSISTment in regular basis. There were 16 math questions and 12 survey questions and one open ended feedback question. Since we considered addition of game-like properties as both a cognitive intervention and an emotional one, we wanted to see if this is preferred by students who have preference for real-world problems and using pictures for learning math. We asked them these questions before using the tutor:

Do you find real-world examples helpful for solving math problem?

a) Yes, examples are helpful b) No, they make it more confusing

Do pictures help you learn math?

a) Yes, pictures help me b) I am not sure c) No, pictures don't help me

We later asked the students about their experience with *Mily's World*. On the question of whether they like *Mily's World*, 20% said they liked it, another 20% said they did not like it and 60% said they find it ok. When we made a regression analysis between liking Mily and students' other survey responses (Table 1), we found that liking is dependent on whether they liked the story and graphics of Mily (emotional interest) and also on whether they find real world examples helpful or confusing (cognitive aspect). The open responses from students also revealed that some students found the mapping of math content to real-world scenario helpful while other found it confusing.

Table 3 Linear regression analysis

Dependent variable: like_Mily'sWorld (R Square= 0.35)

Variable	**Beta** (Standard coefficients)	**Sig.**
Real-world examples helpful/confusing	.31	.007
Pictures helpful/not helpful	.18	.13
Like story and graphics of *Mily's World*	.36	.003

We also asked students about their preference between *Mily's World* and Assistment. 52% preferred *Mily's World*, 13% preferred Assistment and 35% had no preference. This question was asked in the middle of the exercise instead of the end as we wanted to include the students who do not finish the exercise (who are more likely to dislike it, and therefore important to include in our study). So, their preference of *Mily's World* can be a

factor of relative difficulty (questions ordered in increasing complexity in *Mily's World*) along with the novelty effect.

Based on students' open responses, we found that the students generally liked the interactive approach of using pictures and feedback, but felt that the story was not age-appropriate for them. *"The story was a bit childish, but it was clever how everything was incorporated. I found everything easy"*. This was our first iteration of finding the optimal point in the tutor-game space.

4.1.1.1 Lessons learned

This was our first iteration in our quest to find an optimal point in tutor-game space. We had started from very a conservative point with minimal game-like elements. Our first question was if we made this interesting enough as to engage students. Based on students' feedback, we found that we failed to make it engaging enough for all students. The major concern and complaint of students was that the narrative was not age appropriate and appeared rather simple. We had used a younger protagonist (around 10 years old) so that students would be willing to help her solve her problems. However, students did not like this approach and found the character and content too young for them. According to theory on aspirational desire, children like to feel they are more grown up than they really are and prefer to have their character a bit older than they are. When a product seems too babyish, a child will be insulted and will not want to have anything to do with it (Miller, 2014). Based on the students' reviews, we created a new version of tutor called: Monkey's revenge. We created a new character the same age as the target students and added mischief and humor to make the narrative more interesting. We also made the user interface more responsive to user input.

4.1.2 Pilot study with Monkey's Revenge

After we changed storyline to Monkey's Revenge, we made a pilot study. Thirty nine 8th grade students in a small city in Northeastern USA used the tutor in classroom. After using Monkey's Revenge over a class session, we asked them survey questions in 5 point Likert scale.

Table 4 Student response on pilot survey

Survey question	Average response (max 5)
I liked this tutor, Monkey's revenge	3.9
This helped me learn	3.5
I liked the pictures in this tutor	4.1
I liked the story in this tutor	3.9
The problems were boring	2.3

After we got positive qualitative feedback from new design, we proceeded in carrying out randomized controlled studies. We ran two randomized controlled studies.

4.1.3 Monkey's Revenge : Experiment Design

Monkey's Revenge is a tutor with multiple game-like elements. Our approach is to assess each individual game-like element's effects on learning and engagement through controlled experiments. But due to the limitation of the number of students we were able to get for the study, we could not test all combinations of game-like elements. Therefore,

we focused on the two elements we thought would have the most impact: narrative and immediate visual feedback. We created four different versions of Monkey's Revenge with different combinations of game-like elements. All versions had same 16 math problems in the same sequence. Students also get the same hints and bug messages, and the pedagogical help was identical across conditions.

Typically, studies involving the comparison between educational games and regular teaching practices, it is not common to have equivalent pedagogical approach. Therefore, the results tend to be inconclusive and it is unclear if the effect is due to the different pedagogical approach or due to the game-like elements themselves (Sitzmann, 2011; Jackson et al., 2011). By making all the tutors pedagogically equivalent and changing one individual game-like element at a time, we are just looking at the affective and pedagogical impact of the particular individual game-like element. In current study, we are investigating learning gain between pretest and post-test as our cognitive outcome and students' liking of the tutor and satisfaction as affective outcome.

Table 5 Four experimental tutor versions with different degree of game-likeness

Tutor Version	Game like elements		
	Immediate visual feedback	Narrative	Other game-like elements
a: Monkey's revenge	Yes	Yes	Yes
b: Monkey's Revenge without visual feedback	No	Yes	Yes
c: Monkey's Revenge without narrative	Yes	No	Yes
d: Basic tutor	No	No	No

Condition a: Monkey's revenge

This is the full version of Monkey's Revenge with all the game-like elements. Figure 26 demonstrates a problem in the tutor where Mike finds a monkey hiding behind a bush that looks like he escaped from a zoo. Students have to calculate the distance between Mike and the monkey based on the coordinates of their heads.

Condition b: Monkey's Revenge without visual feedback

This tutor version (Figure 29) has no visual feedback. In the full version (Figure 26), there would be visual feedback on students' correct and incorrect responses. For example, if student gives an incorrect value of slope, lines with the wrong slope would be drawn on the graph and if the student gives the correct response, a ball would hit the monkey and he would make a 'hurt' face. In the version with no visual feedback, students receive only text-based feedback.

Condition c: Monkey's Revenge without narrative

This tutor version had all the activities and pictures but the activities were not tied together in a story. For example, students have to calculate the distance between Mike and monkey based on the coordinates on their head (Figure 27). But there is no narrative element (Mike finds an escaped monkey hiding behind the bush) present in this tutor version as illustrated in Figure 26(full version) and Figure 27 (this version).

Condition d: Basic tutor

This is a basic tutor (Figure 30) without any game-like elements. The problems are abstract math problems without any context, pictures and narrative. Students receive the same hints and feedback as in the other tutor versions.

Figure 26 Screenshot of Monkey's revenge with all game-like elements

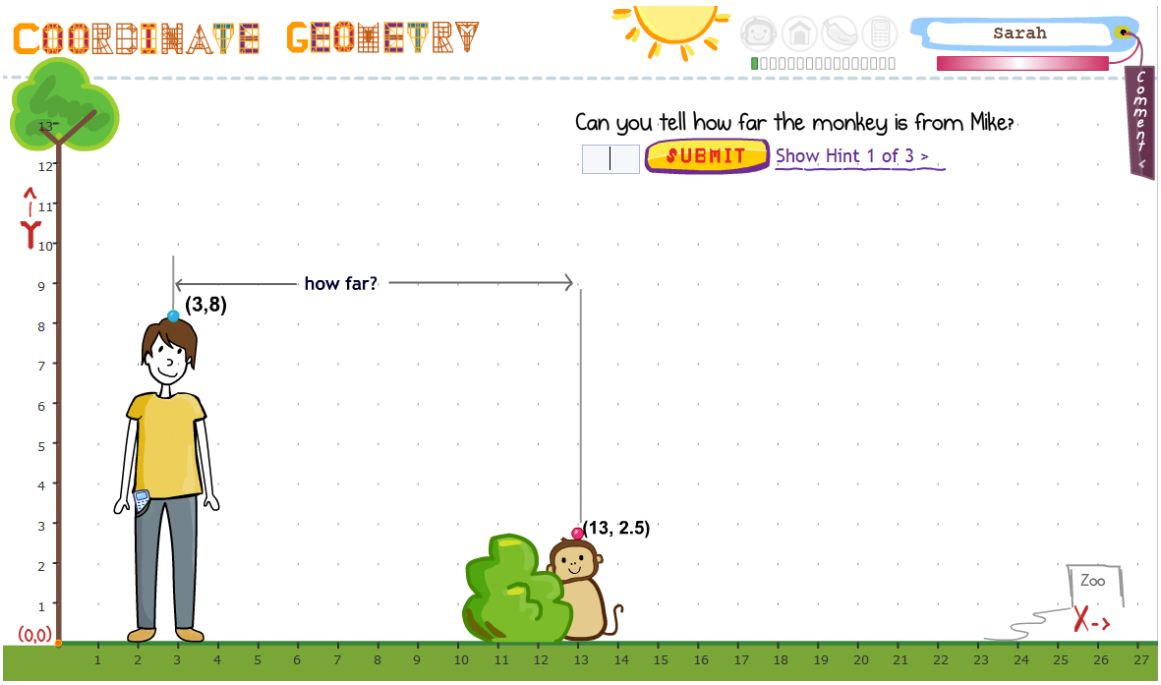


Figure 27 Screenshot of tutor version without narrative

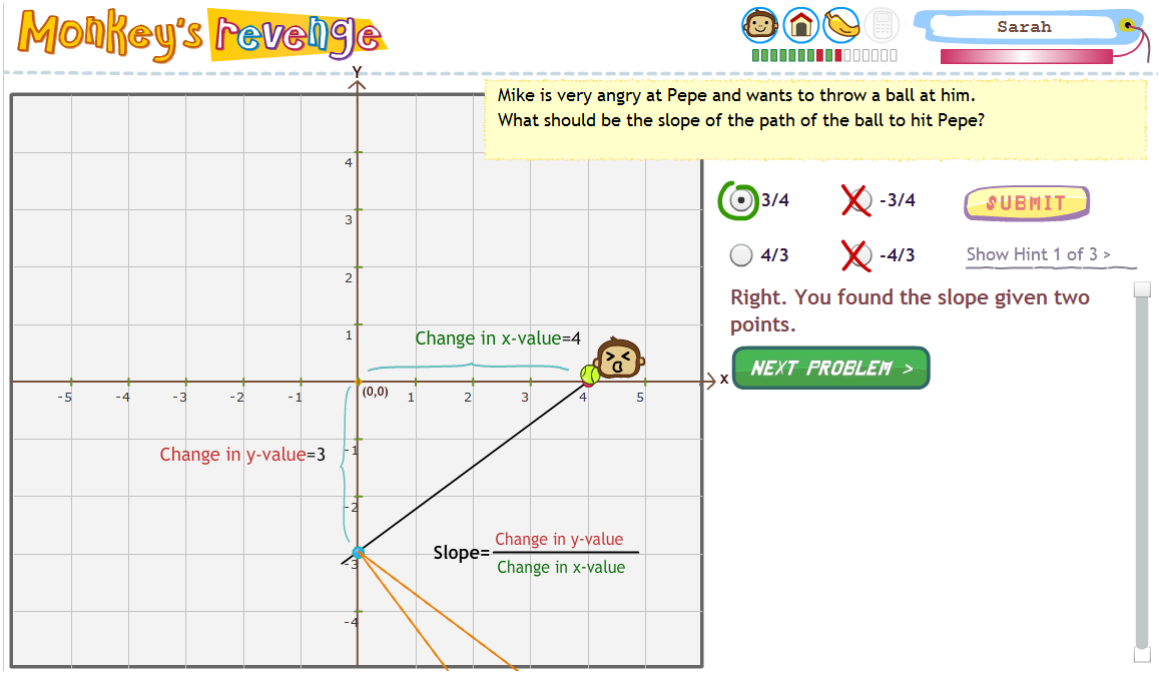


Figure 28 Screenshot of Monkey's Revenge with visual feedback

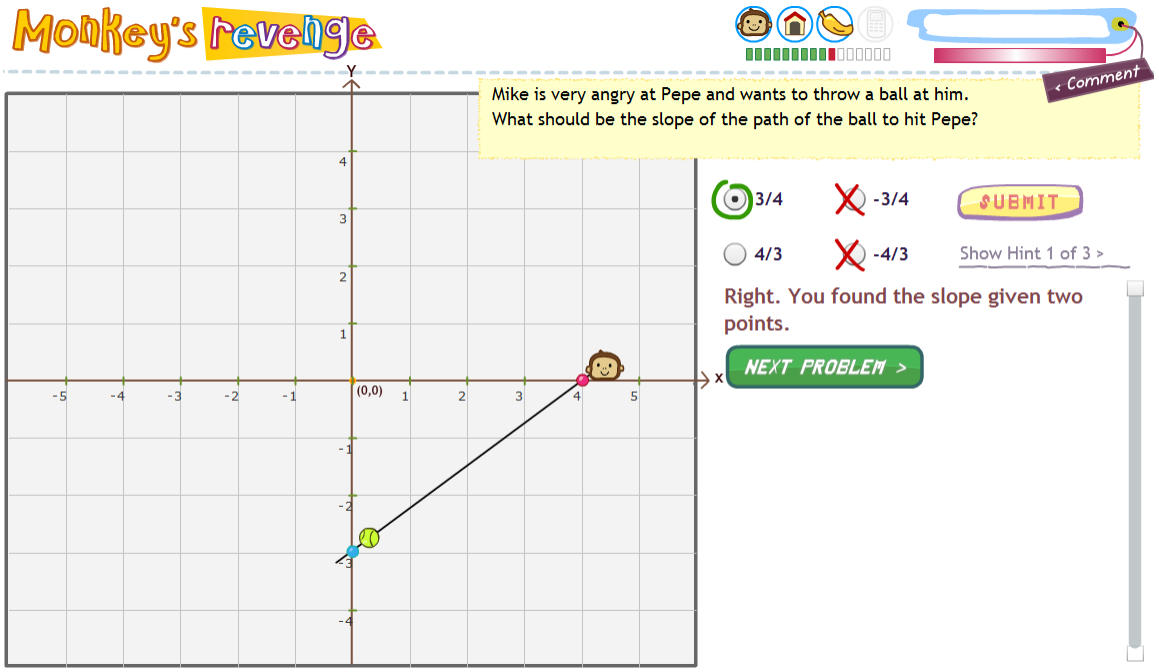


Figure 29 Screenshot of Monkey's Revenge without visual feedback

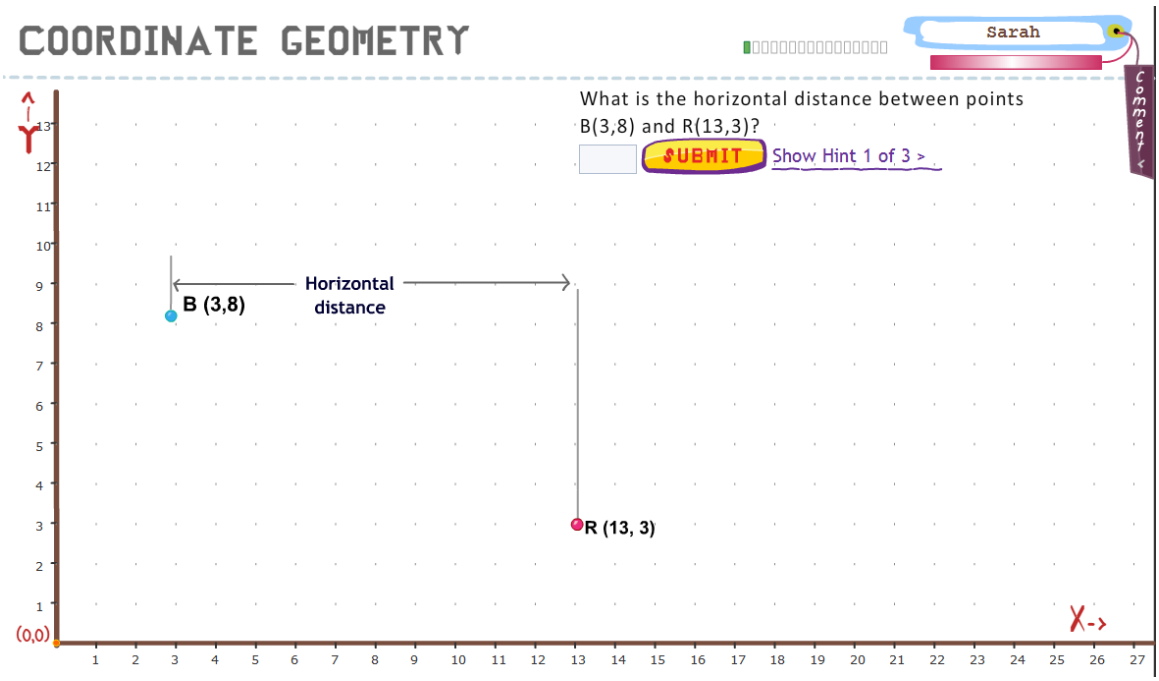


Figure 30 screenshot of Basic tutor

Hypotheses

We had four main hypotheses for the experiment.

I. Versions of the tutor with game-like elements lead to higher student engagement and satisfaction compared to the basic tutor.

II. Versions of the tutor with game-like elements lead to higher learning gains compared to the basic tutor.

III. Individual game-element such as narrative and visual feedback lead to higher student engagement and satisfaction and learning gain.

IV. Versions of the tutor with game-like elements generate higher learning gain and engagement without compromising on cognitive overload and time overload.

To put it another way,

Is our game-like intervention:

- Engaging (higher liking and satisfaction)?
- Effective (higher learning gain)?
- Efficient (no cognitive overload and time overload)

4.1.4 Randomized Controlled Study- I

Participants

A total of 297 middle school (12-14 year olds) students from four Northeastern schools of the United States participated in this study. Among the students, 157 were female and 140 were male. Students were randomly assigned to the four groups, where the randomization was within each class (thus the experiment is not confounded by differences in teacher effectiveness).

a. Monkey's revenge (N=62, 56% female)

b. Monkey's Revenge without visual feedback (N=69, 52% female)

c. Monkey's Revenge without narrative (N=63, 60% female)

d. Basic tutor (N=67, 40% female)

We excluded data from the students (9, 7, 9, 11 from conditions a, b, c, d respectively) who did not complete the exercise.

Data collection

We collected data in the following categories.

Survey questions: We asked the students 16 survey questions with a 5 point likert scale from "strongly disagree"(1) to "strongly agree"(5). The survey involved questions on students' attitude towards math, pedagogical preference, experience within tutor and their liking and satisfaction with the tutor. We computed the corrected split-halves reliability (Crocker & Algina, 1986) by splitting questions into even and odd, correlating students' scores, and applying the Spearman-Brown prophecy formula ($2\varrho /(1+ \varrho)$). For the 16 questions related to students' attitude and experience with the tutor, the corrected split-halves reliability is 0.87. In addition to the likert scale items, the students were also allowed to leave open feedback on the tutor.

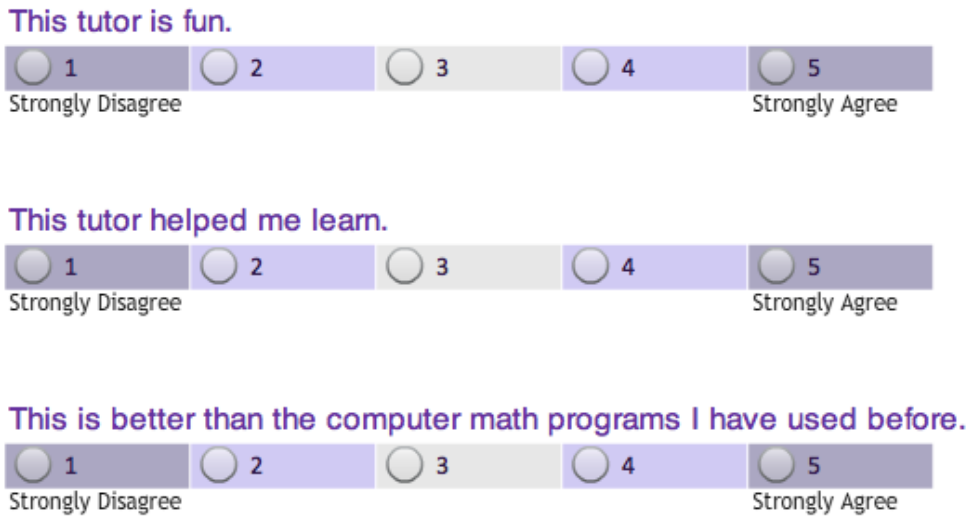


Figure 31 Sample of Survey questions

Performance data: We logged students' activity and performance within the tutor such as the number of hints asked, attempts made and attempt time.

Pre-test and post-test: The students were asked 8 item open-response questionnaire as a pre-test and the same set as a post-test. We collected pre/post-tests from only 216 students and 51 students did not complete the post-test. Thus, we had data from 165 students, which was graded based on automated grading rubric, blind to the student's tutor condition. The mean pre-test score was 5.8 and mean post-test score was 6.28 out of total 8 points. The correlation between pre-test and post-test is 0.6 ($p<0.01$) and correlation between pre-test and pre-post gain is -0.48 ($p<0.01$), suggesting either regression to the mean or a ceiling effect.

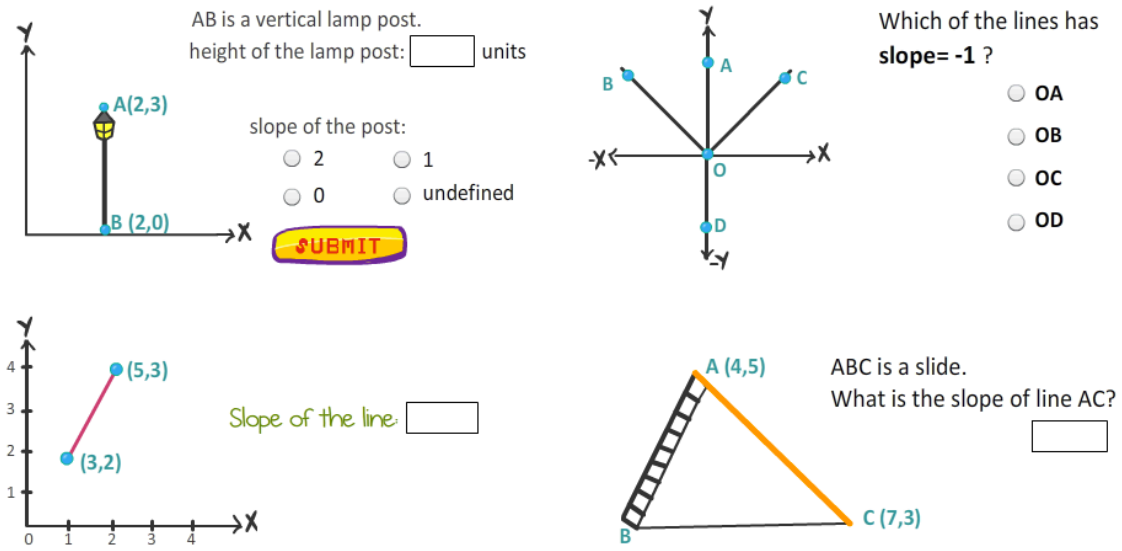


Figure 32 Sample of test questions

RESULTS

Table 6 survey response on main study-I

Tutor	**Like tutor (max 5)** mean (SD) (95% CI)	**Learning gain (max 10)** mean (SD) (95% CI) (N)
Monkey's revenge (N=62)	3.9 (1.2) (± 0.3)	0.41 (1.8) (±0.6) (N=34)
without visual feedback (N=69)	3.8 (1.3) (± 0.3)	0.88 (2.1) (±0.6) (N=46)
without narrative (N=63)	3.6 (1.2) (± 0.3)	0.31 (2) (±0.6) (N=41)
Basic tutor (N=67)	2.8 (1.3) (± 0.3)	0.45 (2) (±0.6) (N=44)

Liking and Learning gain

We found a gradient across increasing levels of game-likeness where liking the tutor increases as the tutor becomes more game-like. However, statistically, the three groups with game-like elements are similar to each other and different from "Basic tutor". Based on students' rating of the tutor and game-like elements, we can conclude that adding game-like elements increased students' liking and satisfaction with the tutor relative to the basic tutor ($p<0.01$). Though this finding may seem obvious, we had made a very conservative progression from tutor towards game and were concerned that we would not be able to engage students.

Learning gain

We were not able to find any conclusive results or patterns in students' learning gains. We think there are a couple of main reasons for this inconclusive result. First, the intervention was brief and it involved a variety of skills. Hence, students did not have enough time to practice any one skill in any depth. Second, we used the same set of questions as pre-test and post-test which might be a reason that students were negligent while doing the post-test as they had recently done the pretest. Finally, the large standard error suggests students were not taking the test seriously, that the test was not long enough to estimate student learning, or some combination of both.

Cognitive overload

We were concerned that adding narrative and pictures may pose cognitive overload among students. On the survey question, "I found the problems difficult because of the

story and pictures", students' mean response was 1.9 (N=187). The mean correct responses among the experimental groups are almost the same (9, 10, 10, 9). So, we are assuming that pictures and story might not have added difficulty, at least for solving the problems that students had prior knowledge on.

Time overload

One of our goals is to make narrative captivating without making it detailed and long. Students in all three groups spent around 13 minutes on solving the problems. Students in the narrative condition spent 2 more minutes in additional story.

4.1.5 Randomized Controlled Study- II

For our new experiment, we made some changes in our intervention while maintaining the same experiment design. We increased the overall problems within the tutor.

We also added extra-tutoring sessions. When students make multiple errors in selected problems, they are taken in special tutor-mode screen. Students go through tutorial on particular skill before they resume their activity in Monkey's Revenge.

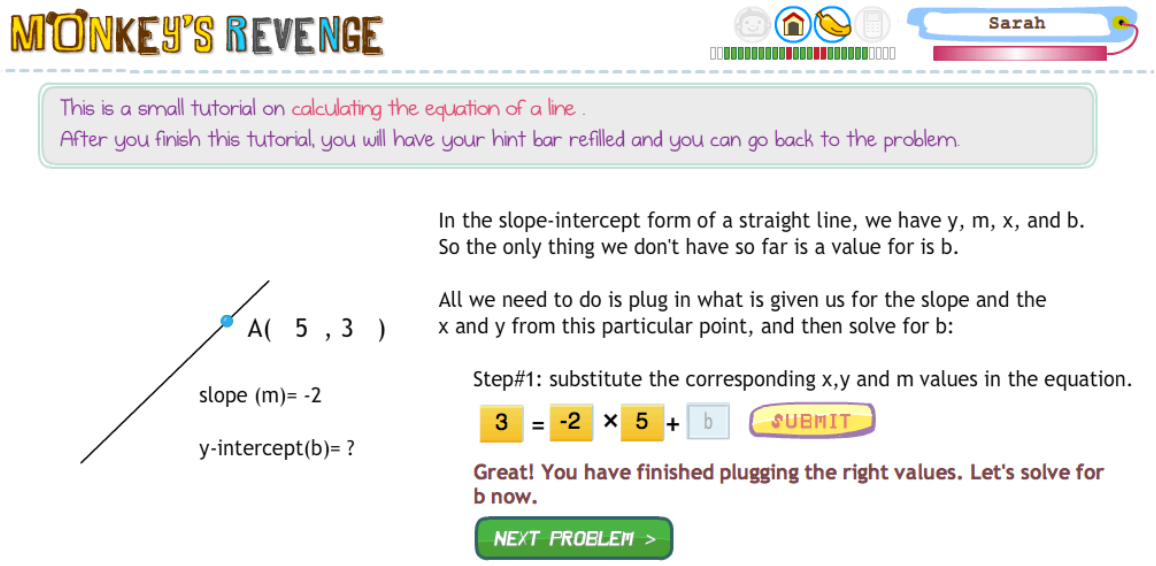


Figure 33 Screenshot of tutorial

We made some changes in our pre-test and post-test design. We created two sets of equivalent questions that were randomized among students. We also increased the number of questions in each set from 8 to 11.

4.1.5.1 Participants

A total of 252 middle school (12-14 year olds) students from four Northeastern schools of the United States participated in this study. Students were randomly assigned to the four groups, where the randomization was within each class.

Table 7 participants who logged in the tutor with pretest score

Tutor version	**N**	**preTest % correct** mean (SD)
Monkey's Revenge	53	53(23)
without visual feedback	46	52 (23)
without narrative	44	52 (24)
Basic tutor	47	52 (26)

RESULTS

We want to analyze the results on the paradigm whether our game-like intervention:

- Engaging (higher liking and satisfaction)?
- Effective (higher learning gain)?
- Efficient (no cognitive overload and time overload)

Similar to our previous study, we found a gradient across increasing levels of game-likeness where liking the tutor increases as the tutor becomes more game-like. The three groups with game-like elements are similar to each other and different from "Basic tutor". We also asked if students liked specific elements such as story and graphics. The mean responses were 4.0 (N=101) for story and 4.0 (N=158) for pictures.

Is game-like intervention Engaging (higher liking and satisfaction)?

Table 8 Survey Responses across tutors (mean, SD and 95% CI)

Tutor version	N	Like tutor	Had fun	Tutor helped	Better than other programs
Monkey's Revenge	34	4.0 (0.9) (±0.3)	4.1 (1.2) (±0.4)	3.9 (.9) (±0.3)	3.9 (.9) (±0.3)
without visual feedback	25	3.9 (1) (±0.4)	3.9 (1) (±0.4)	3.6 (1) (±0.4)	3.7 (1) (±0.4)
without narrative	27	3.6 (1.3) (±0.5)	3.3 (1.3) (±0.5)	3.2 (1.3) (±0.5)	3.8 (1.3)(± 0.5)
Basic tutor	28	3.0 (1.3) (±0.5)	3.0 (1.3) (±0.5)	3.1(1.3) (±0.5)	3.4 (1.3)(± 0.5)

We looked at student's profile on who completed the study. We found that students with higher incoming knowledge were the ones who persisted in all groups. But this trend was more prominent as the tutor is less gamse-like. This suggests that the basic tutor was able to retain only high knowledge kids, losing the low knowledge kids. If we accept this hypothesis, we would expect to see higher overall retention in game-like condition, which is not the case now. A possible explanation would be that students in game-like condition could have run out of time rather than dropping. We cannot make any claim on the basis of data we have, but retention is a worthwhile variable we need to look at.

Table 9 Retention of students in Tutor groups

	Students who logged in		**Students who completed study**	
Tutor version	**N**	**preTest % correct** mean (SD)	**N**	**preTest % correct** mean (SD)
Monkey's Revenge	53	53 (23)	34	66 (19)
without visual feedback	46	52 (23)	25	69 (20)
without narrative	44	52 (24)	27	70 (18)
Basic tutor	47	52 (26)	28	74 (18)

We also collected open feedback from the students to get a qualitative assessment of the tutor. Though we did not quantify the open feedback, we found the feedback to be favoring game-like aspects in general. The following is a sample of students' open comment feedbacks:

"I liked how the monkey was brought into the story and how I got to give him a name. Also I liked how the story went with the coordinates and it wasn't too difficult but helped me learn. Some of the problems were confusing though."

"You made this exercise fun by putting in pictures, words and a story! These problems made me want to do more; I was always excited for what might happen next!"

"I liked the pictures, but some of the questions were pretty confusing. You could word the words a little better."

"I think that the problems are challenging, but they could be harder. The storyline is great, same with the pictures. It would be great if the game was more interactive in a learning manner."

"The monkey was very cute and usually I have a problem focusing but this helped me stay focused because I liked it a lot. Thank you!!! :)"

"This was very fun. I enjoyed playing it. I liked being able to customize my characters name, and it made it more fun to play. Also it made learning a little more interesting. The monkey was mean though."

"I can't do these problems. I didn't like the pictures or scenario. I already have low self esteem."

Is game-like intervention Effective (higher learning gain)?

This time again, we are again not able to make any conclusion on learning gain. Students in Monkey's Revenge condition got significant learning gain. But the effect size is too small and variance too high. There could be a number of reasons for this. The study ran only for one class session of around 60 minutes. The tutor covered a range of skills rather than focusing on few measurable skills. A lot of students dropped post-test which shows their unwillingness and also lack of adequate time.

Table 10 Learning gain across tutor groups (mean and 95% CI)

Tutor	**N**	**Pretest % correct** mean (SD)	**Learning gain (in %)** mean (SD) (95% CI)
Monkey's Revenge	34	66 (19)	10 (27) (±9)
without visual feedback	25	69 (20)	5 (18) (±7)
without narrative	27	70 (18)	7 (29) (±8)
Basic tutor	28	74 (18)	3 (19) (±7)

Is game-like intervention Efficient (no cognitive load and time overload)?

One of our major objective is to minimize cognitive overload on students that could result from extraneous details. We did not have sophisticated rubric to measure cognitive overload. We are however observing students' performance across tutors to make sure that cognitive overload is not causing lower performance. Students in Monkey's Revenge condition are having more problems correct despite having slightly lower incoming knowledge. Though this does not guarantee that there is no cognitive overload (could have been offset by increased attention), cognitive overload is not prominent enough to cause hamper student performance.

Table 11 Student performance across tutors

Tutor	**N**	**Pretest %correct** mean (SD)	**Problems correct in the tutor (max=27)** mean (SD) (95% CI)
Monkey's Revenge	34	66 (19)	20.3 (3) (±1.1)
without visual feedback	25	69 (20)	19.8 (5) (±2)
without narrative	27	70 (18)	18.6 (3) (±1.2)
Basic tutor	28	74 (18)	18.5 (4) (±1.5)

To measure time overload, we calculated time spent on non-tutor activity (narrative sequences and general instruction). While total time spent on the study is similar for all tutor groups, students' in Monkey's Revenge spend 5 more minutes in non-tutor activity. Our effort for minimalist approach seems to have been effective.

Table 12 Time overload across tutor conditions

Tutor	**Total time (in minutes)**	**Non-tutor time (in minutes)**
Monkey's Revenge	50	10
without visual feedback	47	13
without narrative	42	9
Basic tutor	56	5

4.1.6 Conclusions, Limitations and Future Work

It may appear that, while creating educational games that generate learning gain is difficult, creating an educational game that is likeable should be easy. But in the realm of such games, creating a fun experience can be a challenge in itself. Ted Castranova (Baker, 2008) has been forthright about his failure to create a fun experience despite the significant investment in creating an educational game. In our first attempt at such a game, we also struggled with this issue. When we surveyed the students using Mily's World, they reported a less than average rating in regard to liking the tutor. Their main concern was that the narrative was not age-appropriate, so we changed the narrative from 'cute' to 'cute and humorous' and created Monkey's Revenge. While we made an effort to make it interesting and resonate more with the age group, we stuck with our design approach of making the system minimalistic. We wanted students to like and enjoy the system, but not at the expense of creating too many extraneous details. To our relief, students responded well to Monkey's Revenge. In our personal observation in the classroom and in interviews with students, we found they enjoyed the game-like tutor. We created four different versions of the tutor with different degrees of being game-like and ran randomized trials comparing those versions. We observed significant and consistent results demonstrating game-like intervention resulting in more student enjoyment.

However, we were not able to get any conclusive results on learning gain. There was no significant difference between different versions of the tutor. Only the full version of the game-like tutor was able to generate learning gain significantly higher than zero (10% ±

9%). While content design is an obvious area of improvement, there are multiple reasons for this compromised result.

The study ran only for one class session lasting about 60 minutes. We believe this short intervention duration is one of the major reasons for the lack of significant learning gain. In their meta-analysis of educational games, Clark et. al (2014) found that game conditions involving multiple game-play sessions demonstrated significantly better learning outcomes than non-game control conditions, while single game-play sessions did not. We need to extend Monkey's Revenge to multiple sessions in order to observe significant learning gain.

There is an improvement we need to make on the experimental design as well. Many students dropped out before taking the post-test due to lack of time or sheer unwillingness. This is one of the common problems in education studies. We need to design experiments in such a way that students will care about taking the post-tests.

Besides enjoyment and learning gain, our other cognitive processes of interest were 'cognitive overload', 'distraction' and 'confusion'. While we did not have robust or sophisticated tools and metrics, we did look at students' log records to get a reflection of those constructs. We found that students using Monkey's Revenge were performing marginally better than the students using the basic tutor. It is possible that since the students using the former are more engaged, any overhead due to increased cognitive overload is canceled out. It might also suggest the cognitive overload is less pronounced

and thus doesn't hamper performance. Identifying cognitive overload is a crucial concern in education, and there are new efforts to measure it in educational games (citations). In the future we would be interested in using more exact tools to gauge this and other constructs. We should also be measuring and maximizing other learning outcomes such as effort and persistence.

We started this study with the aim of creating a game-like tutor that can generate not just enjoyment but also empirical learning gain. We created a game-like tutor that generates more enjoyment and marginally better performance and learning gain. From that perspective, we have come up short of our own goal. However, we see the value in our overall theoretical framework and design approach. We have taken a cautious and minimalist approach to making the tutor more game-like. We carefully picked game-like elements that we assumed would add to both overall experience and learning. When confronted with design decisions about whether to add 'attractive' elements that might detract from or hamper learning, we made the choice to maximize learning while compromising on enjoyment. This approach can be an alternative to creating resource-intensive immersive systems.

4.2 Experiments with Learning Dashboard (Student Progress Page)

Learning Dashboard is a metacognitive support that we added in Mathspring. Within Learning Dashboard, there are three different pages (Math Tree, Student Progress Page (SPP) and Topic Details) that give visualized information on student performance in different granularity. Math tree gives a summary of overall performance within the tutor. SPP gives information of the students' performance on topic-level. The students can click their way from SPP to Topic Details to see detailed information of their performance on each problem within a topic. Math Tree is the landing page for Mathspring users. Due to development constraints, we were not able to make randomization with Math Tree. So, we ran randomized controlled study using SPP as our intervention rather than the whole Learning Dashboard.

We conducted three studies with middle school students from public schools in Arizona and California. The first two studies, a pilot study in May 2012 and a main study in January 2013, were later found to have contaminated data-logging. We therefore discarded the results from these two studies. We are now describing our third study.

We developed SPP as a metacognitive intervention that will help students' metacognitive process. But we do not have experimental means of observing the students'

metacognitive process as of yet. We expect that SPP will generate overall learning outcomes with direct or indirect result of metacognitive gain. As students use SPP, they will have a better sense of agency and control over their learning process. This will improve their self-efficacy and perception of tutor. Graphical representations of their learning process and gamification in SPP will enhance students' attention and interest. All these cascading effects can lead to improved engagement and performance. In our study, we are explicitly measuring student affect while trying to gauge engagement and performance from log records.

Two hundred and nine grade seven students from public schools in Arizona and California participated in the study. They used MathSpring over three consecutive class sessions. On part of the first and last day, students filled in an pre- and post-affect survey, respectively, which included questions related to various types of affect, including interest and excitement, and so provided baseline data on affect. To obtain information on affect as students were solving math problems, MathSpring prompted students to self-report their affect every five minutes, or every eight problems, whichever came first, but only after a problem was completed to avoid interruption. The prompts were shown on a separate screen and asked students to report on a target emotion (interest or excitement) via a 1-5 point Likert scale (e.g., 310 K. Muldner et al. for interest, "How interested are you feeling right now? Not at all interested (1) … somewhat interested (3) … extremely interested (5); an analogous question appeared for excitement).

Figure 34 Students in the experimental condition were offered to see the progress page when they reported low levels of excitement or low levels of interest (boredom).

The software cycled through the two emotions and students typically self reported several times on each emotion. The study used a between subjects design with four conditions that ranged in terms of degree of access to the SPP tool: (1) no-button (N = 49): the SPP button was not present in the MathSpring interface (the only way to access SPP was through a convoluted set of steps that students were not informed about), (2) button (N = 53): the SPP button was present and prominent but MathSpring did not encourage SPP use, (3) prompt (N = 52): MathSpring invited students to view the SPP immediately after they self-reported low interest or low excitement (< 3), but students could ignore this invitation, (4) force (N = 55): same as in prompt except that MathSpring took students to the SPP page and viewing it was not optional. Students within a given class were randomly assigned to one of the four conditions.

Prior to data analysis, as a manipulation check we verified that SPP access indeed increased across conditions, from no-button to force: M = 1.3, M = 3.1, M = 6.0, M = 8.8. We also confirmed that there were no differences between conditions in terms of baseline interest and excitement as measured by the pre-affect survey (ns).

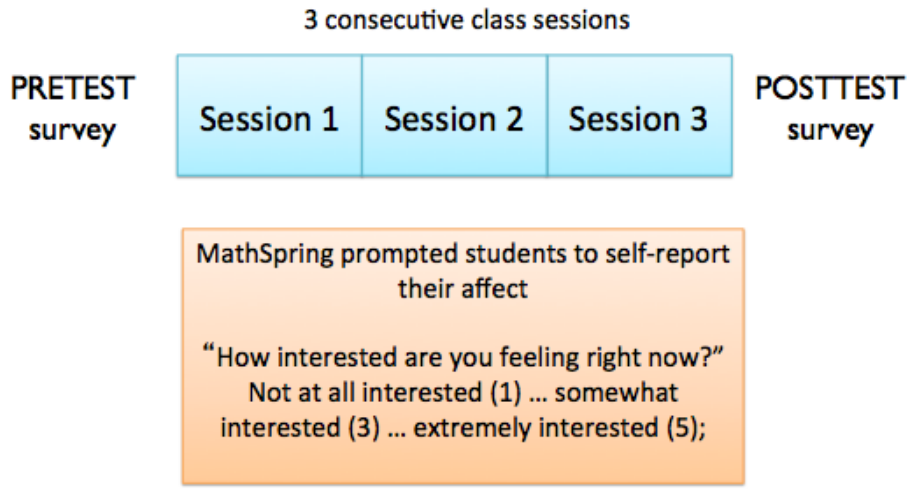


Figure 35 Experimental setup for SPP study

Does the Student Progress Page as experimental condition Impact Student Affect?

To analyze the impact of SPP on affect, we obtained a mean value of self-reported interest and excitement for each student using the student's self-report data. For excitement, there was little difference between the middle two conditions (M = 2.6 for both), while the force and no-button conditions had the highest (M = 2.8) and lowest (M = 2.5) reported excitement, respectively. In contrast, for interest, the force condition had the lowest value (M = 2.5), and there was little difference between the remaining

conditions (M = 2.7 for all three). Neither affective state produced a significant overall effect or follow-up pairwise comparisons as reported by an ANCOVA with the target emotion as the independent variable, the corresponding pre-affect survey emotion as the covariate baseline, and condition as the independent variable (ns).

Thus, overall we did not find an effect of the various experimental conditions.

Is Student Progress Page Usage Associated with Positive Affect?

Another way to analyze the impact of SPP is to check for associations between its

usage and affect, and in particular to evaluate if higher SPP usage is associated with less deactivating emotions (boredom, lack of excitement). However, this analysis is complicated by the fact that MathSpring encouraged SPP usage in two of the conditions (prompt and force) when low interest or low excitement was self reported. Thus, SPP usage could be correlated with negative emotions in these two groups. In contrast, in the other two conditions (no-button and button), students were not encouraged to view the SPP and so it was up to them to access the tool or not. To take these considerations into account, we checked for correlations between SPP usage and self reported affect separately in each of these two groups.

For the SPP not promoted group (no-button, button conditions), interest was positively associated with SPP usage ($r = .24$, $p = .023$) – excitement also was positively associated with SPP but this did not reach significance ($r = .13$, $p = .26$). One explanation for these findings is that in the SPP not promoted conditions, students who had positive affect to begin with (high interest and excitement) used SPP more because they were more motivated, and so SPP usage did not impact affect per se. To check for this possibility we controlled for students' pre-existing affect as derived from the pre-affect survey by running partial correlations. We found that the results held, i.e., interest was still significantly associated with SPP usage ($rp = .25$, $p = .036$) and the result for excitement did not change ($rp = .14$, $p = .3$). Overall, these results suggest that SPP usage may have improved student affect, but given the correlational nature of this analysis these results should be interpreted with caution.

In contrast, for the SPP promoted (prompt, force conditions), as predicted interest was negatively associated with SPP usage ($r = -.32$, $p < .01$); there was also a trend for excitement being negatively associated with SPP but this did not reach significance ($r = -.15$, $p = .16$). These results held after controlling for the pre-affective survey data ($r = -.31$, $p = .012$ for interest; excitement-SPP correlation negative and ns).

How do Conditions Impact Affective State Transitions?

While the above analysis uncovered interesting indications of SPP impact, it did not shed light on how students transitioned between affective states (e.g., if they got "stuck" in the negative deactivating states in some conditions). Addressing this question requires information on student affect more frequently than provided by the self-reports.

Wixon and Muldner (2015), two graduate students working with Mathspring team, generated affect predictions using two user models built from the data, one for each target emotion. They did not use the models during the study to obtain affective information because that would have required having the data from this target population prior to the study, in order to construct the models (or alternatively having a model that was proven to generalize to the present population, which they did not have).

Affect Models. The affect models generate a prediction of a given student's target affect (interest or excitement) after each problem the student solves. While the two

models were created specifically for this analysis, the methodology for their construction comes from their prior work (Wixon et al., 2014). The models were trained using 4-fold student level batch cross validation over the target data set. Each model employed a total of 10 features to predict students' self reports. The excitement model used 2 features based on student's interactions with MathSpring; the interest model used 3. The models' performance (excitement R = 0.43, Kappa = 0.18; interest R = 0.46, Kappa = 0.28) is comparable with existing sensor free affect detector results (Baker et al., 2012). Using the affect model predictions, we followed the procedure in (Arroyo et al., 2007) and generated Markov Chain models for the two target emotions for each condition. These high level "path" models provide the probabilities of transitioning between levels of a given affective state (e.g., from neutral to excited) –we restricted this analysis to three levels for a given affective states (e.g., interest: bored, neutral, interested).

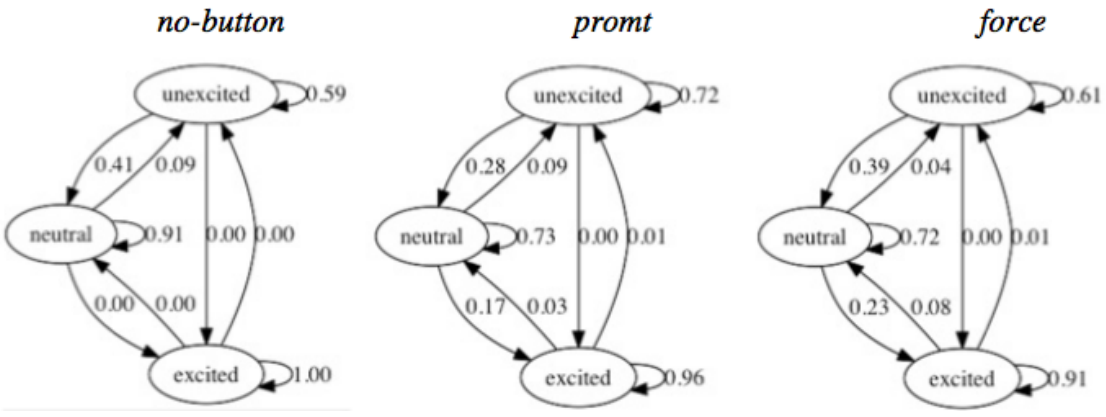


Figure 36 Visual representation of the high-level path models for excitement in the no-button, prompt and force conditions from left to right, respectively

The path models provide a high level view of how a student transitions between levels of an affective state. For instance, we can ascertain that for excitement, overall the probability of transitioning from neutral to excited is the highest in the force condition (Figure 34). However, these models are difficult to interpret and compare between conditions. This can be addressed by computing the joint probability of a student's affect undergoing particular transitions (i.e., following an affective path). For instance, given the condition forcing SPP usage, what is the probability that a student starting in a neutral state ends up excited?

Muldner and Wixon (2015) have described about the affect models in detail in their paper 'Exploring the Impact of a Learning Dashboard on Student Affect.'.

The affect models showed that no-button condition fared worst compared to all other conditions. This suggests that in general, having the SPP present resulted in positive affective paths (ones that led to excitement). For interest, again the no-button condition was the least effective at promoting interest, compared to the other conditions. However, the other conditions were not highly effective in promoting the beneficial affective paths (ones that led to interest), except for the condition that left it up to the student to choose when to see the progress page (i.e., button).

In conclusion, affect models show that SPP is affectively beneficial for students, promoting excitement, and decreasing the likelihood of paths that lead to boredom.

Conclusions, Limitations and Future Work

In general, we found that SPP usage was associated with more positive interest in conditions where MathSpring did not prompt for SPP usage. While the opposite pattern was found for the conditions that MathSpring did prompt SPP usage, this was expected given that the prompts were triggered by negative student affect. When considering all four conditions, however, overall we did not find significant differences in terms of affect. This was somewhat unexpected. On the one hand, students are not good at monitoring their own progress and this can negative affective consequences, so one might expect the conditions that encouraged or even forced SPP usage would improve affect more. On the other hand, however, some theories of motivation argue that having control over ones' activities increases intrinsic motivation, which is related to interest and possibly excitement (Ryan and Desi, 2000) .

Thus far, we have been discussing our analysis related to overall affective differences. However, exploring more fine-grained implications of affective interventions is also paramount. This level of explanation was accomplished by analyzing how students transitioned between levels of affective states, such as from bored to excited, as well as how likely certain affective paths were in the four conditions. This analysis focused on affective paths of length two, and in this context, the SPP promoted positive changes towards excitement in students, but was less effective at promoting interest. One possibility for these results is that excitement is a short-term affective state, which would be captured by the short paths we confined our analysis to, while interest

might take more time to develop, and so was not captured by the particular length of affective paths we focused on.

We see a lot of avenues for future work:

Measuring metacognition: One of our major limitations was that we were not able to create an experimental framework to measure metacognition. We are interested to see how metacognition, affect and learning are related. SPP, as a metacognitive support, is aimed at improving metacognition. Improved self-efficacy should enhance students' affect. SPP, as a game-like intervention, is assumed to enhance student affect directly as well. Enhanced affect and metacognition, both should result in learning gain. We would be interested to see those direct and indirect pathways of our intervention.

Robust measure of constructs: We have used surveys as a primary means to measure the student constructs. Self-reports are limited as a means of gauging into student emotions. With self-reports, our sampling of affect is limited as we cannot present surveys too frequently. We would like to incorporate non-intrusive tools such as eye-tracking. We also need to observe students over longer periods of time so that we can devise data mining tools to make accurate and reliable inference of the learning process.

Improve SPP to make more accurate and accessible : Our assessment of a student's affective and cognitive states are based on their log records. Our inference is just an approximation of the true student state. We should therefore work continuously to make

our inference algorithm more accurate and reliable. Unless the students are assured that their true performance is being reflected, they are not going to be enthusiastic about SPP. They might even be resentful. We also need to present the information in accessible and intuitive form. The information we provide might be too overwhelming to the students. They might not be able to navigate and inspect the elements in Learning Dashboard. As a future work, we want to add SPP characters. SPP characters are animated characters (Figure 37) that reside in SPP. These characters, when prompted, will help explain the SPP elements. For example: if students click on flower pot, the character will show a dialogue box that explains what the particular pepper plant says about their performance. For example: the plant is wilting due to disengaged behavior and the student needs to be more engaged do in order to bring the plant back to health.

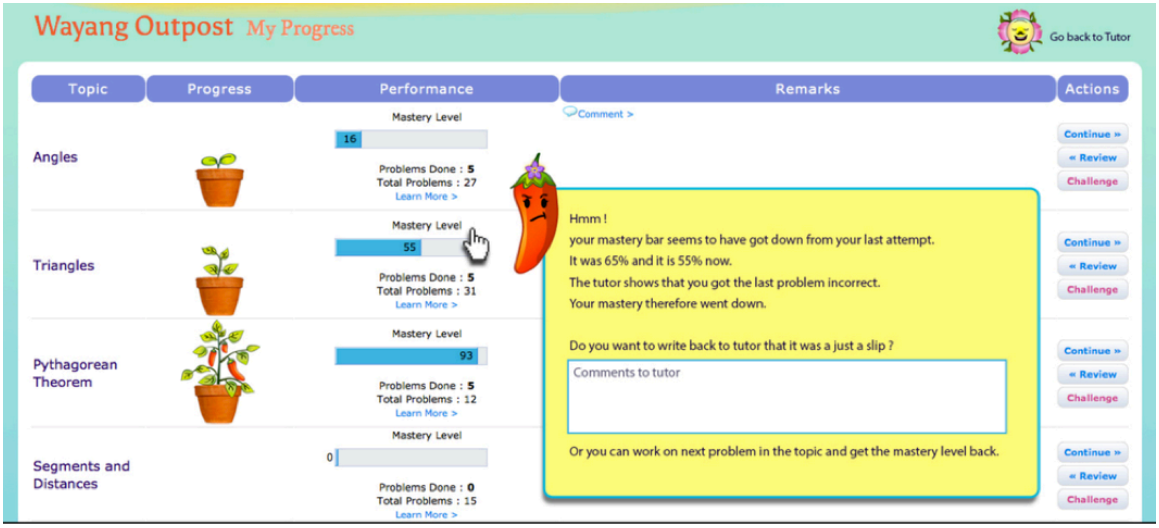


Figure 37 Miss Pepper is a cartoon character that helps explain different components and messages in SPP

Longer studies: We need longer studies to see how students use SPP over long duration. As the students get more familiar with the interface, their usage and reaction will evolve.

We imagine different scenarios. At first, they may be enthusiastic and excited or confused and overwhelmed. In our optimistic projection, as the students get acquainted with SPP, they will learn to use SPP in a productive manner. They will be able to monitor their performance and progress. If they are getting bored in a particular topic, they can select challenging problems, a new topic or work on 'fun' activities. If they are struggling, they can go to 'Topic Details' page and hand pick easier problems to practice the basics. An intelligent tutoring system's goal is to offer problems with optimal difficulty level. Games strive to keep their players in 'flow' zone, balancing challenge and skill level. Learning Dashboard can be a tool that students can use to take control of their learning, observing their performance and making appropriate choices. We have added game-like elements in Learning Dashboard to make this process more intuitive and engaging.

4.3 Experiment with Mosaic

We created two Mosaic mini-games 'London' and 'Manhattan' on math topics 'Fraction' and 'Area and Perimeter' respectively. We embedded those mini-games within Mathspring. As students work on these math topics, corresponding mini-game will pop up randomly. We hoped that as students play these mini-games, they will get affective boost. To study whether Mosaic mini-games can uplift the students' affect within Mathspring, we ran a randomized controlled study. One hundred and eighty six students from urban schools in Northeast USA participated in the study. The participants used Mathspring over one class session. They were randomly assigned to one of the three experimental conditions:

a. No-Mosaic: The students assigned in this condition do not get Mosaic mini-games during the whole Mathspring sessions (N=60).

b. Mosaic Prompt: While using Mathspring, students will be asked randomly whether they want to try Mosaic mini-game. Students are free to accept the offer and play mini-game or reject and continue with Mathspring (N=62)

c. Force Mosaic: Students in this group are taken to Mosaic directly while working on Mathspring (N=64)

The participants took a pre-survey before they started the session and they took a post-survey after completing the experiment. To obtain information on affect as students were solving math problems, MathSpring prompted students to self-report their affect. The prompts were shown on a separate screen and asked students to report on a target emotion (interest or frustration) via a 1-5 point Likert scale (for interest, "How interested are you feeling right now? Not at all interested (1) … somewhat interested (3) … extremely interested (5); an analogous question appeared for frustration).

We expected that using Mosaic mini-games will improve student's affective states. Due to the boost in positive experience within the tutor, the students who used mini-games will also have better perception of the tutor. We assume that in the long run, the boost in affect and perception will increase students' engagement and overall learning. But our study was just for one class session and we did not expected learning gain for the study.

Results

We were interested to study whether playing Mosaic enhanced student affect, perception of the tutor and overall satisfaction and enjoyment. We were also interested to see if improved student affect can lead to improved enjoyment of the tutor.

At the end of the experiment, we asked students to rate their experience in Mathspring.

- *Do you think you have performed well in the math problems in Mathspring ?*

- *Do you think that you have learned a lot on math topics using Mathspring ?*
- *Did you enjoy using Mathspring?*

Table 13 Students in different experimental groups self-report on their experience in Mathspring (mean and SD)

Group	Total participants	Participants with complete survey	Performed well (max 5)	Learned a Lot (max 5)	Enjoy using Mathspring (max 5)
No Mosaic	60	34	3.3 (1)	2.3 (1.1)	2.6 (1.2)
Prompt Mosaic	62	42	3.5 (1.2)	2.5 (1)	2.9 (1.2)
Force Mosaic	64	41	3.4 (1.4)	2.4 (1.2)	2.9 (1.3)

We did not find any statistical difference between these experimental groups.

Next, we divided the participants in two groups: those who used Mosaic (Used Mosaic) and those who did not (Did not Use Mosaic). A fraction of participants in 'Prompt Mosaic' and 'Force Mosaic' did not receive Mosaic mini-games at all as the mini-games are offered randomly. The participants from those experimental groups who did not receive the mini-games are practically similar to the participants in 'No Mosaic' group. Therefore, we regard dividing the participants in those two groups reasonable. We found that the participants who used Mosaic (Used Mosaic) reported enjoying Mathspring significantly higher than the participants who did not use Mosaic (Did not use Mosaic) (Table 14). Their self-reports on whether they enjoyed Mathspring and whether they learnt more were significantly higher than that of the participants who did not use the mini-games. The students who did not get Mosaic were aware that of the fact that the mini-games were offered to some of their fellows. The students were told that the mini-games appear randomly. The difference between the self-reports between these two

groups might have been contributed partly by the disappointed of the students who did not receive the mini-games.

Table 14 Students who used Mosaic and who did not use Mosaic self-report on their experience in Mathspring

Group	Total participants	Participants with complete survey	Performed well (max 5)	Learned a Lot (max 5)	Enjoy using Mathspring (max 5)
Did not use Mosaic	88	54	3.11(1.2)	2.17 (1)	2.57 (1.2)
Used Mosaic	98	63	3.59 (1.2)	2.62 (1.2)	3.08 (1.3)
p-value			0.04*	0.03*	0.02*

We also looked at student's self-report of their affect (Interest and Frustration). We expected that the students who used Mosaic would report higher positive affect and lower negative affect. However, we did not find any statistical difference between the two groups (Table 15). The self-report of the students who used Mosaic is slightly more positive. But the effect size is too small and not statistically significant to make any claims. We also noted that 18% of participants who did not use Mosaic skipped the affect survey. The ratio was 10% for the participants use used Mosaic.

Table 15 Interest and Frustration averaged over participants who used and did not use Mosaic

Group	Interest (max 5) mean (SD) (N)	Frustration (max 5) mean (SD) (N)	Participants who skipped affect survey
Did not use Mosaic (N=88)	2.4 (1.1) (N=55)	2.5 (1) (N=66)	16 (18%)
Used Mosaic (N=98)	2.5 (1.1) (N=65)	2.4 (1.2) (N=60)	10 (10%)
p-value	0.4	0.08	

We were interested to see whether the students' affect changed positively after using the Mosaic mini-games. Our first approach was to aggregate a students' self-report on an emotion before using the mini-games and compare that with the aggregation of self-report on the emotion after using the mini-games. However, there were not enough samples. The study was only over a class session of 50 minutes. The students had less than 40 minutes to work with the tutor. One mini-game could take 10-20 minutes. Mini-games would appear randomly. Some students saw Mosaic too early that we did not have opportunity to observe their affect before and some students saw Mosaic too late that we were not able to observe their affect afterwards. There were not enough instances where a student got the survey on the same emotion before and after using Mosaic. Therefore, we were not able to observe how the student affect changed on individual level. Next, we averaged the self-reports across all participants under the categories of before using Mosiac and after using Mosaic (Table 16). For the students who used Mosaic, we

averaged affect self-reports for all the participants before they used Mosaic. We also averaged the reports across those participants after they used Mosaic.

Table 16 Interest and Frustration averaged over participants before and after using Mosaic .

	Before using Mosaic mean (SD) (N)	After using Mosaic mean (SD) (N)
Interest	2.39 (1.19) (N=27)	2.53 (1.23) (N=46)
Frustration	2.33 (1.19)(N=23)	2.34 (1.23) (N=43)

Conclusions

A good study design was again our major limitation. We ran the experiment for a single class session. The impact of the intervention was limited. We were not able to sample affect sufficiently. There are some signs that hint that Mosaic mini-gams have positive impact on the students' affect and experience. A new study where we can observe students over multiple sessions would be more illuminating.

5 Causal Modeling

Student learning is a complex process. We introduced different educational interventions in hopes of generating positive learning outcomes, but the effects of those interventions aren't always straight forward. An intervention might generate learning gains but students may not choose to use the intervention. On the other hand, students may choose an intervention that produces no measurable learning gain. Furthermore, the same intervention can have different effects on different students. When studying the impact of particular interventions, measuring the resultant learning outcome alone provides an incomplete picture; we must understand finer elements in the learning process.

Engagement, learning, and use of interventions such as tutors vary as a function of student gender, prior knowledge, and pedagogical preference. Knowledge of these interrelationships provides a clearer picture of student learning and guides intervention refinement. Therefore, we can conduct exploratory analyses of student data to understand the interrelationships between student characteristics, tutor intervention, interaction and learning outcomes.

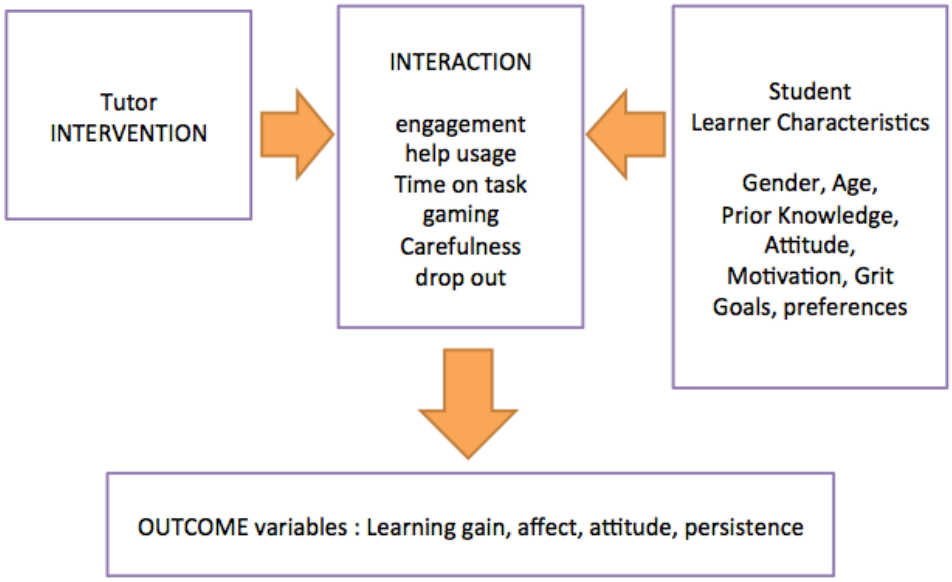


Figure 38 Tutor intervention interacts with student learner characteristics to generate outcome variables

Deriving causal inferences from statistical associations within data has been a contested field, with researchers such as Pearl (2009) and Sprites et al. (2001) advancing the field of causality and detractors claiming that drawing causal inferences from statistical data is impossible (Freedman, 1987; Rogosa, 1987; Denis, 2006). Since causal modeling isn't widely used within the educational technology community, we have decided to first run a case study of causal modeling itself.

5.1 Causal Modeling of Monkey's Revenge: a case study in Causal Modeling

Causal models: Causal models are graphical models that make the additional assumption that the links between nodes represent causal influence. By causal, we mean that a link A→B means that if we intervene and change the value of A, then B will change. Based on the conditional independencies within the data, causal modeling makes causal inferences among the variables. Causal modeling is a generic name used for statistical

methods like path analysis and Structural Equation Modeling (SEM), and represents both the technique used and the assumptions underlying the analytic approach

We used TETRAD, a free causal modeling software package (Glymour et al., 2004), to generate causal graphs. It supports both Bayesian networks and SEM.

Causal model has basically four types of associations:

i. A→C (A has *direct effect* on C)

ii. A→B→C (A has *indirect effect* on C through *mediating* variable B)

iii. A←B→C (A and C have *spurious association* since they are correlated but not causally related, and B is the *confounding* variable)

iv. A→B←C (A and C are *independent* of each other)

We had run a study with Monkey's Revenge where a total of 297 middle school (12-14 year olds) students from four Northeastern schools in the United States participated. We dropped 71 students due to missing data. We had asked 16 survey questions To perform analyses such as this, we first simplified our survey questions. We used factor analysis to reduce the 16 survey questions into six categories:

likeMath: "Mathematics is interesting."; "I enjoy the challenge presented by Math problems."

mathSelfConcept: "I am afraid of Math."; " I am afraid of doing word problems."; "I enjoy the challenge presented by Math problems."

pedagogical preference: "I like to learn from Computers rather than books."; "I find real world examples helpful for learning Math."

tutorHelpful: "This helped me learn."; "I found the hints helpful."; "These problems helped me learn about slopes."

tutorConfusing: "I find the questions very confusing."

likeTutor: "This tutor (Monkey's Revenge) looks interesting."; ""I liked this tutor.""; "I will recommend this tutor to a friend learning coordinate geometry."; "This is better than the computer math programs I have used before."; "The problems were boring."

From students' log data, we calculated variables like ***per_correct*** (ratio of correct problems to total problems); ***avgAttemptTime*** (average time student spent on each attempt) and ***avgHints*** (average number of hints students asked on each question). Along with other variables gender, game-like, ***preTestScore*** (students' score on pretest) and ***prePostGain*** (students' gain score from pre-test to post-test), we had a total of 13 variables.

5.1.1 Causal modeling and correlation matrix

Based on the data we collected, we used TETRAD with the PC search algorithm to generate a causal graph (Figure 39).

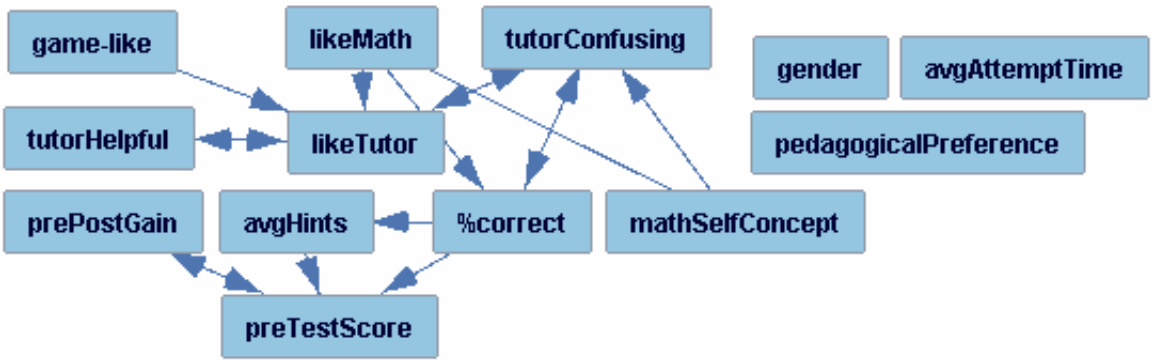


Figure 39 Causal model from PC algorithm without domain knowledge

We also generated a graph based on correlation matrix (Figure 40). We computed the correlation of every variable against each of the other 12, and added a link in the graph whenever the correlation was statistically reliable.

Correlation is relatively lenient about making associations whereas causation is strict, as it only puts a link after controlling all other variables in the model. In other words, the link from *game-like* to *likeTutor* in Fig 1 indicates that there is no variable, that when used to compute the partial correlation, that can remove this relationship. From Figure 39 and Figure 40, we see that, due to ensuring no variable(s) can remove the link, causal modeling has far fewer links than the correlation model. When causal model does not link two nodes, it might have correctly identified absence of link, we would call that a true negative. On the other hand, it might have missed a link that should be there which we would call a false negative.

True negatives (indirect and spurious associations): Correlation is not causation as there might be possible confounders causing the spurious association (see definition iii, above), and causal modeling controls for all third variables regarding them as possible

confounders. From the correlation matrix, we see that *likeTutor* and *%correct* are correlated which would suggest that students who like the tutor performed better. This result would have been an evidence for student engagement, since students who liked the tutor are presumably more engaged while using it. But the causal model (Figure 39) infers that this is a spurious association confounded by *likeMath*. Students who like math tend to like tutor more and to have better performance. Once we control for *likeMath*, there is no relation between *likeTutor* and *%correct*.

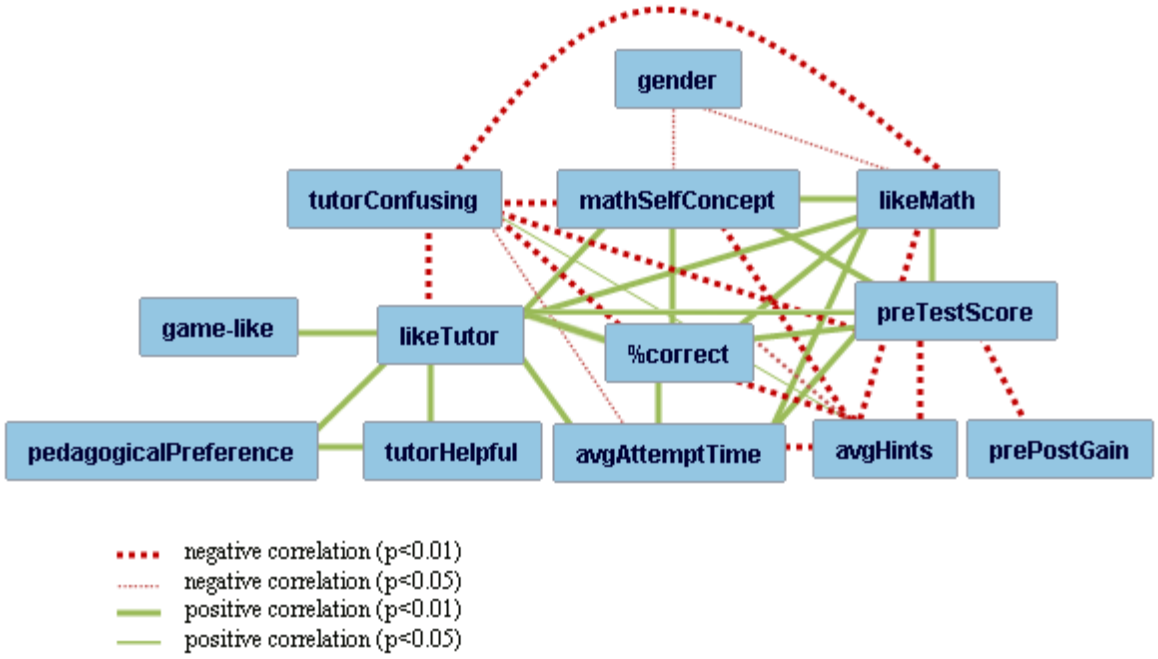


Figure 40 Graph based on correlation matrix

Still, the causal model is limited to assertions about the observed variables as there might be other confounders which we have not observed. After controlling for all possible confounding variables within the system, the causal model has inferred that

likeMath→*likeTutor*. But it is possible that being agreeable on survey questionnaire might be an unobserved confounder affecting both variables.

Causal modeling makes distinction between direct and indirect association. *likeMath* and *avgHints* are negatively correlated (-0.3**) which suggests that the students who like math ask fewer hints. But once we control for *%correct*, that correlation is gone (see Figure 39). So, we can conclude that the students who like math ask for fewer hints only because they already know the correct responses and so do not need as much help. The students who like math and have few correct responses will ask for as many hints as a student who does not like math and has few correct responses.

False negatives (reduced statistical power and multicollinearity): Controlling on third variables reduces statistical power and we might get false negatives if we have few data. We made a small simulation and found that adding more data removes false negatives without adding false positives. But when the independent variables are correlated among themselves, we face the problem of multicollinearity. Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. That is, a multiple regression model with correlated predictors can indicate how well the entire bundle of predictors predicts the outcome variable, but it may not give interpretable results about any individual predictor, or about which predictors are redundant with others.

For example: *avgAttemptTime* is correlated with both *%correct* (0.3**) and *preTestScore*(0.3**). But since, *%correct* and *preTestScore* are highly correlated among themselves (0.6**), *avgAttemptTime* is conditionally independent to both of them. We

can see that *avgAttemptTime* is an isolated node in figure 1; in contrast, the correlation graph (Figure 2) indicates *avgAttemptTime* is related to both *preTestScore* and *%correct*.

5.1.2 Causal structure, path orientation and domain knowledge

Beyond false positive and false negatives, which simply deal with the presence or absence of a link, we can also examine whether the link *orientation* is plausible or not. Some of the links had plausible orientations, such as *likeMath* →*likeTutor*←*game-like*, which suggests that students who like math also liked the tutor more, and students who had more a game-like tutor reported greater liking. Using the information that *likeTutor* is correlated with both *likeMath* and *game-like,* but *likeMath* and *game-like* are independent between themselves, the search algorithm correctly identifies that it is not *likeTutor* influencing *likeMath* and *game-like* but the other way round (Pearl, 2009) for a discussion of "colliders" such as this). However, we see that there are other edges which are incorrectly oriented such as *%correct*→*preTestScore*; student performance on the tutor cannot have influenced a pretest that occurred before students began using the tutor. Correlation underdetermines causality as covariance in statistical data is rarely sufficient to disambiguate causality. Therefore, even after we use search algorithms to find some structure, there are a number of "Markov equivalent" structures. For example, given a data set with just two variables A and B which are correlated with each other, true causal structure can be A→B or A←B, and there is no way to tell which model is correct. However, we can narrow our search by adding domain knowledge. In TETRAD, we can

add domain knowledge in the form of knowledge tiers which represent the casual hierarchy. Causal links are only permitted to later tiers, and cannot go back to previous tiers. We used the following knowledge tier based on our knowledge of assumed causal hierarchy and temporal precedence.

i. Gender

ii. Game-like, mathSelfConcept

iii. likeMath, Pedagogical preference

iv. preTestScore

v. %correct, avgAttemptTime, avgHints, tutorConfusing, tutorHelpful

vi. likeTutor

vii. prePostGain

We are taking the temporal order of when variables occurred, which is not necessarily when they were measured. For example: we asked students‟ experience with tutor *tutorConfusing*, *tutorHelpful* after they finished the tutor activity. Still, we have placed them in the same tier as the tutor activities like *avgAttemptTime*, *avgHints* since students‟ experience would have affected their tutor activities. Since the pairs (*likeMath, mathSelfConcept*) and (*tutorHelpful, likeTutor*) are highly correlated, we placed them in different tiers even though we cannot specify which one precedes which.

We see from Figure 1 and Figure 3 that adding domain knowledge not only fixes the path orientations (*preTestScore→%correct*), but have changed the whole causal structure

adding some new causal links (*gender→mathSelfConcept, pedagogicalPreference→tutorHelpful, correct→avgAttemptTime*).

At first, it may appear that knowledge of causal hierarchy only helps to orient the edges specifying which one is cause and which one is effect. I.e. If A is higher than B and we found that A and B are correlated, then A→B

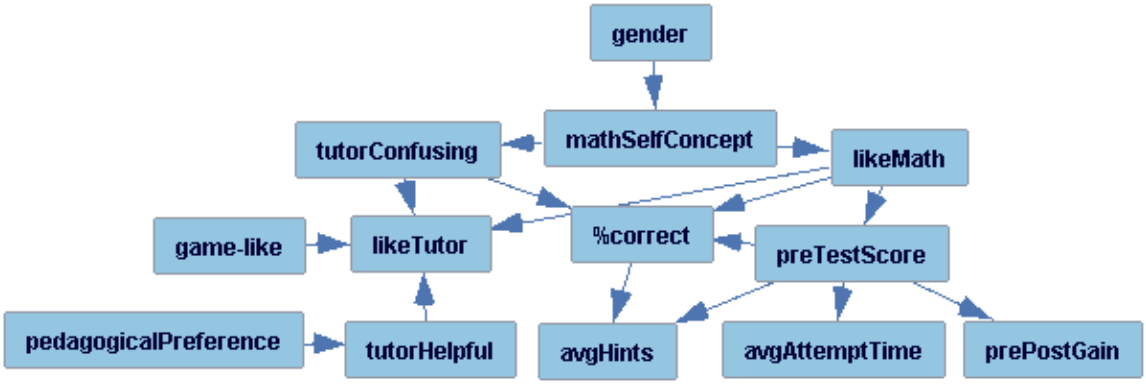


Figure 41 Causal model with Domain knowledge

However, besides distinguishing variables as potential causes and effects, the domain knowledge also restricts the set of variables to be considered as confounders and mediators. Aside from improving efficiency, this approach also results in stronger inference. Let us consider an example where we are interested to know the relation between two variables A and B. We have the following knowledge tiers:

Tier 1: C Tier 2: A Tier 3: M

Tier 4: B Tier 5: E

We should control on variable C to consider it as a potential confounder, and on M as a potential mediator. But variable E cannot be a confounder or a mediator and conditioning on E is not required. In fact, we should not condition on E as we might get a false positive. If the true causal model of A, B, and E is A→E←B, where A and B are independent but have E as a common effect. However, if we compute the partial

correlation of A and B, controlling for variable E, then we have produced a statistical correlation between A and B.

Sometimes, we do not know about the causal hierarchy of the variables we are trying to analyze and may not know which is the cause and which is the effect, but having information of the causal hierarchy of third variables, such as whether they are a potential confounder or a potential mediator, can help infer if there is any causal path between the variables of interest. We can illustrate this with a concrete example in education. Suppose we have observed that engagement and learning are correlated, but want to understand the causal relation between them. Imagine there are two other variables, prior knowledge, a potential confounder (since it is a possible cause of both), and performance, a potential mediator (since it co-occurs with both). Consider two scenarios: if controlling for prior knowledge removes the correlation, then we know there is no causal relationship between engagement and learning, and the causal structure is engagement←prior knowledge→learning. On the other hand, if partialing out performance removes the correlation between engagement and learning, then there is still an *indirect* causal effect between the two, either engagement→performance→learning, or learning→performance→engagement. So even though we were unable to provide information about the causal direction between engagement and learning, by providing information about other variables we are able to better differentiate if there is any causal relation.

Interestingly, adding domain knowledge can also address the problem of multicollinearity. *preTestScore* and *%correct* were correlated with each other (Figure 40).

Therefore, we did not see their effect on *avgAttemptTime* in Figure 1 because when it calculated both partial correlations (*preTestScore, avgAttemptTime | %correct*) and (*%correct, avgAttemptTime | preTestScore*) there was no statistically reliable correlation remaining due to the colinearity of *%correct* and *preTestScore*. However, providing the domain knowledge provided powerful information: since we have set *preTestScore* on higher causal tier than *%correct, %correct* cannot be a possible confounder or mediator and therefore, the partial correlation (*preTestScore, avgAttemptTime | %correct*) is not calculated. As a result, the link from *preTestScore* to *avgAttemptTime* is placed based on correlation (*preTestScore, avgAttemptTime)* while controlling for other variables aside from *%correct*. Thus, by excluding *%correct* as a confound or mediator, we are able to infer additional causal links.

5.1.3 Causal modeling and multiple regression

Causal modeling is a sophisticated extension to multiple regression and basically adds two things to multiple regression.

a) Two-dimensional graphical representation instead of flat one-dimensional

b) Causal assumptions to direct inference algorithm

We are using an example of multiple regression to illustrate this.

likeTutor = 7.8**tutorhelpful* + 5**game-like* - 3.2**tutorConfusing* + 3**likeMath* + 2.2**pedagogicalPreference* -0.5 (Equation 1)

Causal model employs a series of multiple regression and is two-dimensional rather than one. Addition of one more dimension offers the following benefits:

Direct and indirect effect: Multiple regression only looks at direct effect but fails at identifying indirect effects. For example: we can see from causal model (Figure 41) that *mathSelfConcept* affects whether students find the *tutorConfusing,* which in turn affects *likeTutor*. Thus, there is an indirect effect between *mathSelfConcept* and *likeTutor*. We can see this indirect effect in the correlation graph but not in the multiple regression (eqn 1). While multiple regression can be equally robust when it comes to predictive accuracy, causal modeling provides a better representation and framework to *understand* interrelationships of variables. In educational domain, we are interested to know the relationships between variables not just in the predictive accuracy of our models.

Using domain knowledge in the form of causal hierarchy: Since causal modeling allows multiple layers of associations of variables, it adds affordance to insert domain knowledge in the form of a causal hierarchy. As mentioned earlier, this knowledge helps to deal with false negatives and multicollinearity.

Causal assumptions: Statistical methods employ statistical assumption such as normality, independence, homoscedasticity, etc. On top of these statistical assumptions, causal modeling adds causal assumptions (Sprites et al., 2001):

- Causal Markov assumption: A variable X is independent of every other variable (except X's effects) conditional on all of its direct causes.

- Faithfulness: independencies within data is generated not by coincidence but by structure
- Causal sufficiency: the set of measured variables M include all of the common causes of pairs in M

As a consequence of making these assumptions, causal modeling approaches can use more powerful inference algorithms. However, these assumptions are also the ones most criticized and scrutinized by the critics of causal modeling (Freedman, 1987, Rogosa1987, Denis, 2006). There are situations where these causal assumptions do not hold true and may be unreasonable. Stronger assumptions add more analytical power but also higher chances of inaccuracy. Certain assumptions have to be made to gain valid conclusions in any analysis procedure. It is up to researcher to select these assumptions based on their data and domain. We have accepted the causal assumptions made by TETRAD since they seem reasonable for our data and purpose.

Figure 4 provides an overview of our causal modeling process. We can use our domain knowledge and inference algorithms to generate a set of possible models consistent with the data we collected. Both the data and our domain knowledge are based on the Real Model of the phenomenon, but are not assumed to be identical (the error component). Even if we assume that data and domain knowledge are generated by the real model without error, there are possible sources of error due to statistical sampling issues, resulting in type I (false positive) and type II (false negative) errors.

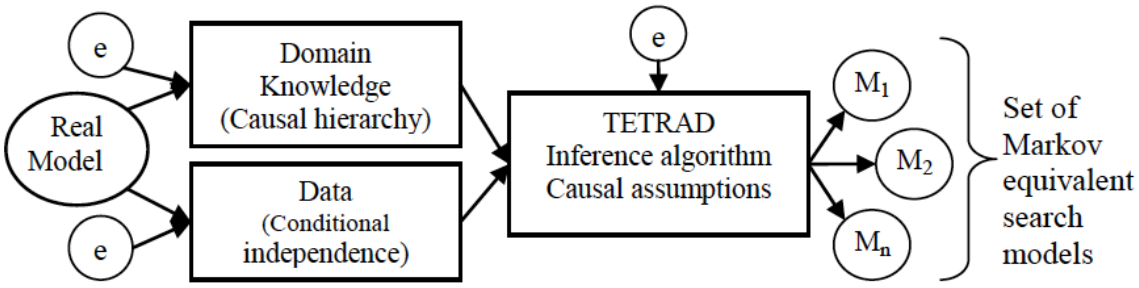


Figure 42 Block diagram of our Causal modeling process

5.1.4 Causal modeling: confirmatory, exploratory and graphical tool

We made a randomized controlled trial on the tutor's degree of being *game-like*. Other than this variable, the inferences we are making from our causal models are solely based on statistical independencies within the data, on the domain knowledge we added, and on the causal assumptions of the inference algorithm. The inferences from the causal model from Figure 41 has not only confirmed some of our prior hypotheses but also unraveled some new interesting patterns that we would like to explore, such as whether *likeMath* really has direct and indirect effects on performance. Although we can make causal claims only with controlled manipulations and all other inferences will be questionable, we are faced with the fact that we cannot always make the controlled interventions due to issues of time, cost, and the impossibility of directly intervening on variables such as *likeMath*. In this scenario, causal modeling offers the best possible tools to make causal inference from statistical observation. We see three uses of causal modeling.

Confirmatory tool

The most common and accepted practice of causal modeling is using as a confirmatory tool, to support or reject the theory based model. In TETRAD, we can create a graphical model and then fit the model with data and measure goodness of fit. As we have only conducted one initial study and are still creating our theoretic framework, we have not tried this approach. However, the causal model generated has supported some of our prior hypotheses. We were interested to see how different student subpopulations would react to our intervention. We basically looked at pedagogical preference and students‟ self concept in math. We found that students who have preference to learn from computers and find real world examples helpful reported that they found the tutor helpful and liked the tutor more (*pedagogicalPreference→tutorHelpful→likeTutor*). Similarly, students who had lower self concept in math found tutor more confusing which made them like the tutor less (*mathSelfConcept→tutorConfusing→likeTutor*).

Exploratory tool

Using causal model as an exploratory tool has been criticized and warned against as we cannot build theory from non-experimental data. As mentioned earlier, possibility of unobserved confounders and under determination of causality from correlation pose serious limitation to generate new valid conclusions. But, conditional independencies in data and domain knowledge can offer some new inferences which can be helpful in guiding us towards further analyses and examination. Like a less than100% accurate test (and to be fair, no randomized controlled trial is 100% accurate either), it cannot establish a claim but at least direct to what further explorations we need to make.

For example, in our causal model, we found that likeMath has both direct (likeMath→%correct) and indirect (likeMath→preTestScore→%correct) effect on %correct. Based on this, we are considering two possible causal models as shown in Figure 43.

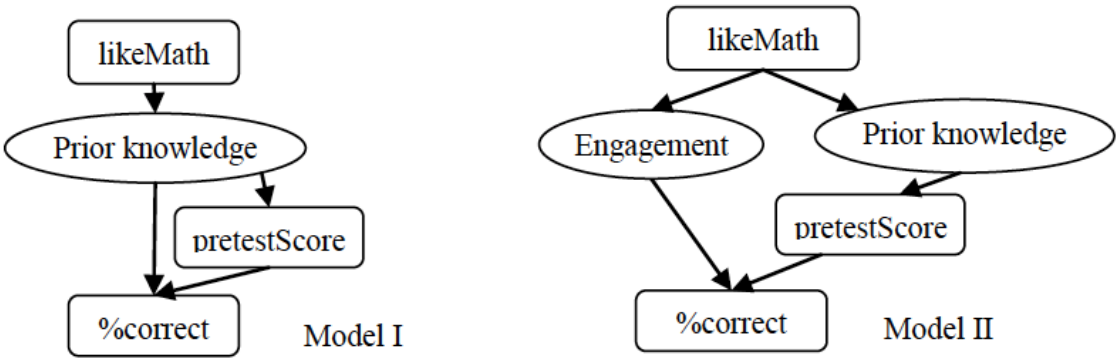


Figure 43 Two possible causal models linking LikeMath and %Correct

Model I suggests that *pretestScore* does not capture all of the variance in prior knowledge of the student, as represented by the latent node "Prior knowledge." So, students who like math and have high prior knowledge may have a low pre test score but they have high performance nonetheless. In other words, likeMath only affects student knowledge but does not affect engagement.

Model II on the other hand suggests that students who like math both have higher prior knowledge and are more engaged, and have therefore higher performance. In other words, likeMath affects both prior knowledge and engagement.

One approach for evaluating these models is to consider other effects we would see if they were true. If Model II were correct, and engaged students perform better, we might expect that students who also like the tutor to also be more engaged. However, in our

causal model, we do not see a directed path from *likeTutor* to *%correct* though they are positively correlated (Figure 41).

Again, we are faced with two possibilities:

Possibility I: Though there is not direct path from *likeTutor* to *%correct*, there are two paths between them *liketutor*←*likeMath*→*%correct* and *likeTutor*←*preTestScore*→%correct. Perhaps the correlation between *likeTutor* and *%correct* is lost once we control for the two possible confounders and this might be a case of reduced statistical power while making a partial correlation.

Possibility II: Students who like the tutor may be more engaged but this engagement may not necessarily lead to better performance. Students might like the tutor and instead of focusing on solving the problems, they might just engage with game-like aspects of tutor like narratives and pictures. This inference is very important for us as we are trying to improve engagement by making tutor more game-like so as to improve their performance and learning in addition to arouse sensory interest among students.

We were not able to make any conclusive findings with causal model but this has at least made interesting inferences and raised questions that are very important for us. It has directed towards the possibilities that we would like to make further examination and possibly run some controlled randomized trials.

Graphical tool to make compact visual representation of associations

Even if researchers are skeptical of the domain knowledge we have brought to bear and are dubious of the causal modeling assumptions, it is still possible to consider Figure 1 without the assumption that the edges represent causality. This graph would be a compact

representation of the partial correlation relationships among the variables. For example, we know there is no relation between likeMath and avgHints once %correct is controlled for. This relationship is purely statistical in nature, but there is no convenient notation in traditional statistics to represent the necessary set of variables to make two other variables independent. Therefore, we think that causal modeling can be useful as graphical tool to make a compact visual representation of association within the observed variables.

5.2 Causal modeling: guide for new experiments

As we discussed earlier, causal modeling is a useful tool for exploratory analysis. While we cannot always make causal claims, causal models can give us some intuitive insight into the data. Even though we cannot draw conclusions, these models tells us where to look next: which variables look more decisive and need to be measured more robustly, and which variables are confounding and need to be teased apart. Before we started studying our new interventions in Mathspring, we created causal models of the existing data from previous studies.

5.2.1 Causal Modeling of Wayang OutPost

Wayang OutPost is an earlier version of Mathspring, an intelligent math tutoring system. We utilized data from 94 students in grades 7 and 8 from a rural-area public middle school in Massachusetts. These students were part of a mathematics class that used the Wayang Outpost Math Tutoring system for a period of one week. As part of the activity,

students took a survey on the first day that assessed baseline achievement level as well as affective and motivational factors related to mathematical problem solving. Student responses were collected on a 5-point Likert scale.

Variables

Based on the survey data, we created the following variables. For example: MathLiking is a variable created from a student's response in a scale of 1-5 for the survey question: '*How much do you like doing math?*', with possible answers 1=not at all, 2=a little, 3=somewhat, 4=quite a bit and 5=extremely. Some variables are averaged across multiple survey responses that represent the same construct, (example: MathValue).

Attitude towards Math

MathSelfconcept: *How good would you be at learning something new in math?*

MathLiking: *How much do you like doing math?*

MathValue: *Some things that you learn in school help you to do things better outside of class; that is, they are useful. For example, learning about plants might help you to grow a garden. In general, how useful is what you learn in math?*

Affect

Students were asked questions on four affective variables while using the tutor.

1. Confident: *How confident do you feel when solving math problems?*
2. Frustrated: *How frustrated do you feel when solving math problems?*
3. Interested: *How Interested do you feel when solving math problems?*
4. Excited: *How Excited do you feel when solving math problems?*

Pedagogical intervention

MathFluencyTraining (training on basic math facts (e.g. multiplication tables) and retrieval speed) Students were randomly assigned for the math fluency training.

Perception of tutoring system

PerceptionWayang (Students' perception of the tutor)

Do you think you learned how to tackle math problems by using the system?

How much did you like the system?

What did you think about the help in the system?

Perception_LC (Students' perception of the learning companion)

How much did you like the learning companion?

Was the learning companion helpful?

Pretest Score and learning gain

Students took the MCAS (state standardized) test before using the tutoring system. We used this test score as the pre-test score. We calculated the difference in test scores between the MCAS tests students took before and after using the tutor and designated that value as learningGain.

Gain and outcomes

Students also took identical surveys after they completed the session. Gain outcomes were calculated from the pre and post data. These outcomes are expressed as the "base" and "gain" over the intervention.

Doug's version of Tetrad

Doug Selent, a graduate student at WPI, made an extension to the current version of Tetrad created at CMU (https://sites.google.com/site/dougstetrad/tetrad). The Tetrad program uses edges to show relationships between two variables; however it does not show the strength of the relationship. Doug's extension adds weights to all edges in the graphs generated by the search function. The color of the edges represents the strength of the relationship between two variables. In addition to colors, a positive "+" or negative "-" symbol is shown next to each edge to indicate a positive or negative relationship. The weights of the edges are determined by the partial correlation between the two variables connected by the edge. Each partial correlation is taken with a set of variables that disconnect all indirect paths from one variable to another.

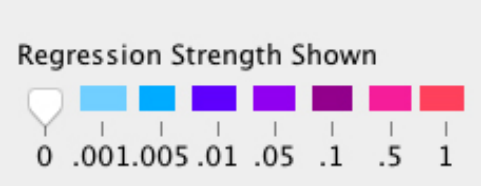


Figure 44 Colors of edges in Doug's version of Tetrad are associated with the strength of the relationship between the two variables

Knowledge Tiers

To narrow our search among the Markov equivalent models and avoid the multicollinearity, we added domain knowledge in the form of knowledge tiers. We have taken temporal precedence as the main basis for categorizing the variables in different

tiers rather than analyzing the inherent causal mechanism. For example, we can argue that math liking may affect students' performance on the pretest rather than the other way around. But since we measure their liking of math after the students took the pretest, we are putting the pretest in a higher causal tier than math liking.

Tier 1: gender, preTest, mathFluencyTraining,

Tier 2: Math appreciation (mathSelfConcept; mathLiking; mathValue)

Tier 3: Affect within Tutor and Perceptions of Tutor (Confident, Frustrated, Interested, Excited; perceptionWayang; perception_LC)

Tier 4: gain variables (e.g. learningGain, mathSelfConceptGain, confidenceGain, etc)

Tier 1 consists of the student's gender, pretest score and pedagogical approaches that were collected before they took the survey. We placed the math attitude variables in tier 2, affective variables and perceptions in tier 3 and the gain variables in tier 4. We could have collapsed the variables in tier 2 and 3 but we did not. Math attitude variables and affective variables are tightly linked and putting them in the same tier would cause multicollinearity, whereas when we put the affective variables into the lower tier they act as mediating variables coming from the math attitude variables.

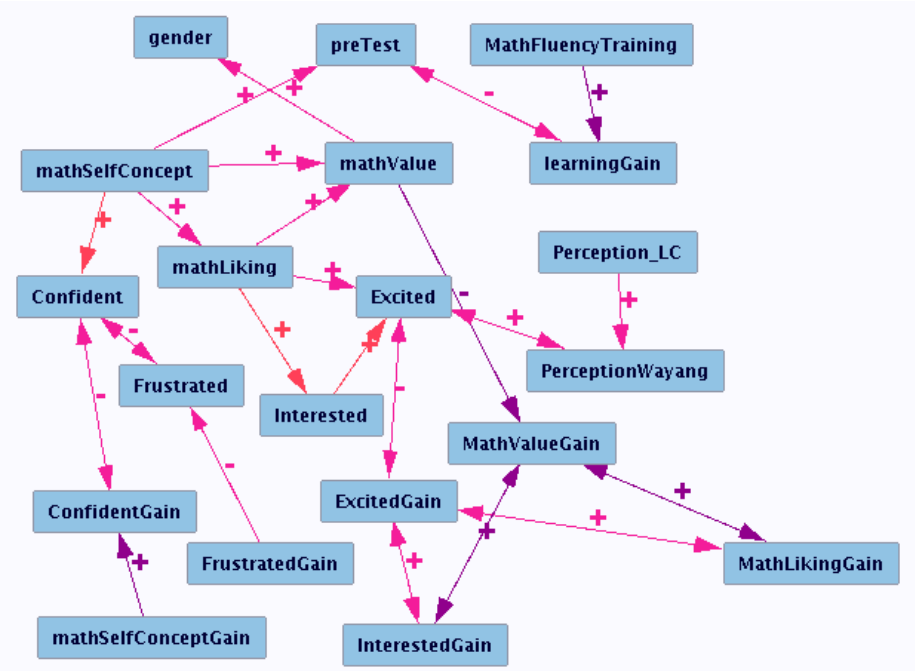


Figure 45 Causal model of attitude and affect

Attitudes and Affect

We observed very strong relationships between student attitude towards math and their affective states within the tutor. Students who had higher self-concept in math reported being more confident. Students who liked math reported being more interested and excited while using the tutor. While Interested, Excited and Confident are more tightly coupled with attitude variables, Frustrated is relatively separate and connected to that web via Confident.

In the correlation matrix, all three math attitude variables were related to Interested, Excited and Confident, with only mathSelfConcept and MathLiking negatively correlated to Frustrated (mathValue had no relation to Frustrated). The three math attitude variables were also highly correlated among themselves. The causal model teased apart this dense correlation web into a sparser directed structure. Since this is a dense web, it has multiple Markov Equivalent Models and the structure we received is only one of those possible models. We are not claiming that the causal structure we produced is the true one. But it is helpful in making a compact representation with reasonably plausible inferences.

Gender

From the correlation matrix, we found gender to be positively correlated with math attitude variables mathValue (.33**) and mathLiking (.28**) but not with mathSelfConcept. Gender is also correlated with emotions while using the tutor such as Interested (.3**) and Excited (0.24*) but not with frustration and confidence. From the causal model, we can see that gender directly affects mathLiking and mathValue. But emotion variables Excited and Interested are indirectly affected by gender mediating through mathLiking. The causal structure gender→mathLiking→Interested states that female students like math more, which makes them more interested while using the tutor. There is no direct link from gender→Interested, which implies that the female students who like math as much as male students do not necessarily have any higher Interest level. This again could be a case of multicollinearity since correlation (mathLiking, Interest)=0.85**. Therefore, we are more inclined to believe the direct correlation between gender and Interested.

One point of concern is whether the association between gender and math attitudes/affect has more to do with 'being agreeable'. Female students have been demonstrated to give more positive responses about liking in general. But we do not see gender related to a more positive perception of the tutor or learning characters. So we could really be uncovering gender and math attitude dynamics. However, we maintain our general suspicion regarding such elusive constructs, especially when they are self-reported.

Pedagogical Intervention.

Math fluency training has a significant impact on improving learning gain of students as shown by the link Math Fluency Training → learningGain. This indicates that a group of students who received math fluency training achieved higher improvement in math tests after using the software. A paper has been published about this (Arroyo et al., 2011); however, it is interesting to see the strength of this causality compared to other factors.

Pretest Score

Pretest score has a direct effect on math attitude variables (MathSelfConcept, MathLiking and MathValue) in our causal model. In the correlation matrix, we found pretest to be correlated to affect variables as well. However, it is only indirectly related through attitude variables as mediators. Basically, this states that higher student knowledge has to be internalized into higher self-concept and math liking in order to eventually manifest into their enjoyment of the tutor.

We also see a negative causal link going from pretest to learning gain. A naïve causal interpretation of this would mean that students who have higher knowledge learn less, but such causal interpretation would be fallacious since it is just the statistical phenomenon of regression to the mean.

Perception of Tutor

From the causal model we see that students' perception of the tutor is directly related only to their perception of the learning companion and whether they are excited while

using the tutor. In the correlation matrix, we see this variable significantly related to attitude variables mathValue (0.4**) and mathLiking (0.34**) as well as learningGain (0.29**). Apparently, these relations are all being channeled via the indirect path Excited ←→PerceptionWayang. Since perceptionWayang is such a crucial variable, we would be interested in observing all significant relationships related to it rather than tracing possible indirect paths. In such instances, we recommend directly observing the correlation matrix.

Outcomes and gain variables

We see causal links between base variables and their gain parameters (example: confident→confidentGain), which are just the statistical phenomena of regression to the mean as mentioned earlier. However, we also observed those gain variables being correlated among themselves (example: mathValueGain→InterestedGain). This could be explained in various ways. This could again be regression to the mean since the students who have lower mathValue have lower interest and thus have more space to grow. Or it could be that the students who are susceptible to being positively influenced by the tutor experience are also prone to reporting higher math value. Or it could also be that the students who gained higher interest levels while using the tutor were positively influenced so much that they ended up with a higher mathValue. This would be a great outcome.

In terms of affective variables, we can see the following two clusters:

Performance oriented (incoming math ability) student descriptors:

preTest, math self-concept→confidence and frustration

Students who have higher prior knowledge and better self concept in math reported higher confidence and lower frustration.

Liking and Appreciation:

Math value, math liking, perception of LC, perception of Wayang → interest and excitement

Also, students who reported a gain in confidence also had higher gains in self concept in math, and those who gained in interest and excitement also ended up with higher liking for math and had greater value for math.

Basically, among the four affective variables, confidence and frustration are more tightly linked with performance and ability whereas interest and excitement are more related to attitude and appreciation for math and the tutor.

5.2.2 Causal modeling with ASSISTments

We would also like to describe a study we ran with data from another math tutoring system, ASSISTments. ASSISTments is used by middle school students in United States, the same demographics of students that we are studying.

We used the data from 171 twelve- through fourteen-year old 8th grade students. These data consisted of 74,394 problems solved along with response data recorded by ASSISTments. This includes performance records of each student across time slices for 106 skills (e.g. area of polygons, Venn diagram, division of fractions, etc).

In order to measure individual differences in self-discipline, we employed a survey called the Brief Self-Control Scale (BSCS; Tangney et al., 2004) in December 2008. BSCS is a 13-item questionnaire that measures self-regulatory behavior in four domains: thoughts, emotions, impulses, and performance.

Each question (e.g. “I am lazy”, “I am good at resisting temptation”) asks the respondent to choose from a 5-point Likert scale-based answer list: a. Very much like me, b. Mostly like me, c. Somewhat like me, d. A little like me, e. Not like me at all. We assigned each response -2, -1, 0, +1, or +2 points respectively.

For each student, we had a 12 dimensional vector representing their responses to each survey question. We calculated the sum of the responses, *sum score*, and performed a factor analysis to reduce data dimensions, giving us 2 principle factors. We found the correlation between the first factor's score and the *sum score* to be 0.99. For simplicity, and for consistency with prior research using this scale, we used the *sum score* as the student's measure of self-discipline.

Knowledge tracing model

We used knowledge tracing in a Dynamic Bayesian Network (DBN; see Figure 46 Knowledge tracing model: Dynamic Bayesian network), which makes inferences about a student's knowledge based on their performance.

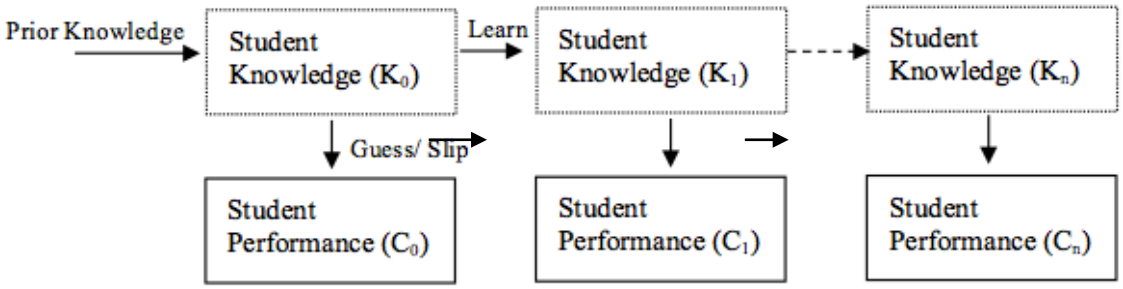


Figure 46 Knowledge tracing model: Dynamic Bayesian network

Student performance is assumed to be a noisy reflection of their knowledge mediated by two performance parameters, *guess* and *slip*. The guess parameter represents the fact that the student may sometimes generate a correct response in spite of not knowing the correct

skill. For example, some tutor items are multiple choice, so even a student with no understanding of the question could generate a correct response. The slip parameter acknowledges that even students who understand a skill can make an occasional careless mistake. *Prior knowledge* refers to the probability the student knows the skill before he begins working with the tutor. The *learning rate* parameter estimates the probability that the student learns new knowledge that they did not know beforehand.

$Guess = Pr(C_n=True \mid K_n=False)$

$Slip = Pr(C_n=False \mid K_n=True)$

$Prior\ Knowledge = Pr(K_0=True)$

$Learning\ rate = Pr(K_n=True \mid K_{n-1}=False)$

We used the Bayes Net Toolkit for Student Modeling (BNT-SM, Beck et al., 2008), which takes as inputs data and a compact XML specification of a Bayes net model that describes causal relationships among student knowledge and observed behavior. BNT-SM gives us the knowledge parameters (prior knowledge and learning) as well as the performance parameters (guess and slip).

We then input the data to TETRAD using the following knowledge tier.

1. Gender
2. selfDisciplineSurvey, inconsistency
3. KT parameters (knowledge, guess, learn, slip)
4. Performance (percent correct), problemsSolved

Since the KT parameters are highly correlated among themselves (correlations as high as 0.8), and since the effects of multicollinearity can make model interpretation difficult , we forbade causal links between the four variables in tier #3.

These causal inferences (Figure 47) are in fact very consistent with our conclusions so far. Self-discipline impacts a student's incoming knowledge and guess rate, but it has no effect on learning rate. Students who were less consistent on the survey were more likely to make slips. Self-discipline also had an impact on the students' performance and behavior (problems solved), but this is an indirect impact through knowledge. This implies that if we can directly observe a student's knowledge, knowing a student's self-discipline does not add much additional information except through the performance parameters.

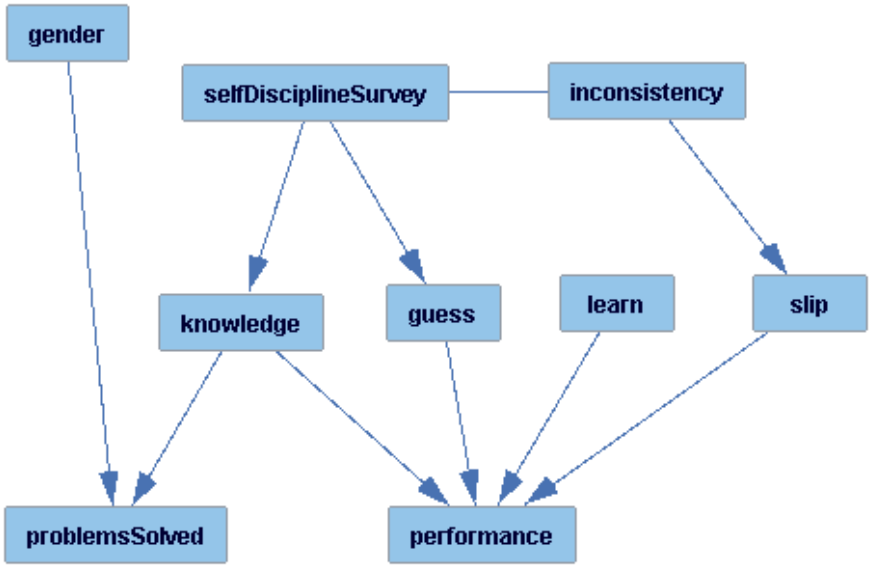


Figure 47 Causal modeling of self-discipline survey response, performance and Knowledge tracing parameters: Assistment data

One odd link is from gender to number of problems solved. Girls solved more problems than boys did, but apparently for reasons other than incoming knowledge. This additional path, and the lack of connection between gender and self discipline and knowledge is perplexing. One possibility is that we need to find a better method to calculate student's self-discipline other than just survey measures. A second possibility is that there could be another causal path that influences performance related to gender but not knowledge or self-discipline. A third possibility is that the constructs of self-discipline and engagement are less related than they first appear.

Self-discipline seems an interesting variable when it comes to performance of middle school students. We therefore decided to add this construct in our Mathspring experiment.

5.3 Causal Modeling with Mathspring

We ran a randomized controlled study with 209 seventh grade students from public schools in Arizona and California. The students used MathSpring over three consecutive class sessions. On the first and the last day, students took pre and post surveys. The students were asked on their attitude towards math, learning orientation , their affect and their perception and enjoyment of the tutor. The students were asked about their level of interest and excitement, that gave baseline data on affect. To obtain information on affect as students were solving math problems, MathSpring prompted the students to self-report their affect.

We are grouping the data variables in three groups:

1. Pre-survey variables: variables created from the student responses in the survey before the study.

2. Within-tutor variables: variables created from the log records in the tutor and the survey while using the tutor.

3. Post-survey variables: variables created from the student responses in the survey after the study.

6.3.1 Pre-Survey variables

We are creating variables from the survey responses. For example: MathLike is a variable created from a student's response in a scale of 1-5 for the survey question: '*Do you like your math class?*', with possible answers 1=not at all, 2=a little, 3=somewhat, 4=quite a bit and 5=extremely. Some variables are averaged across multiple survey responses that represent the same construct.

Survey questions regarding Attitudes towards Math

We used survey questions to understand students' attitudes towards math. These are based on Wigfield and Eccles (2000) Expectancy–Value Theory of Achievement Motivation.

MathLike : *Do you like your math class?*

MathValue : *Compared to most of your other activities, how important is it for you to be good at math?*

MathDifficult: *Do you worry that math class is much too difficult for you?*

Survey questions regarding Learning Orientation

To understand student's learning orientation, we used Carol Dweck's (1999) theory of motivation.

Pre_LOR: *When you are doing math exercises, is your goal to learn as much as you can? Is it your goal to demonstrate that you are better at math than your classmates?*

HelpSeekPre: *Do you often ask for help when you are doing math even when you are not stuck? Do you prefer to learn on your own, without being offered help?*

GiveupPre: *When solving math problems, do you prefer to give up?*

Competitive: *Do you work hard in school so that you can beat everyone?*

PerfAvoidPre: *When you are doing math, do you want to avoid making errors so that you don't look or feel incompetent?*

Survey questions regarding Self-discipline

To measure student individual differences in self-discipline, we used a questionnaire survey, Brief Self-Control Scale (BSCS) developed by Tangney et. al. (2004) and also

used successfully by grit researcher Duckworth (2005). BSCS is a 13-item questionnaire to measure self-regulatory behavior in four domains: thoughts, emotions, impulses, and performance. Since we could not use all of these questions from the questionnaire, we selected two questions which we assumed to be relevant for our study.

Impetuous: *Do you often act without thinking alternatives?*

DiffConcentration*: Do you have difficulty concentrating?*

Survey questions regarding Baseline affect

We wanted to measure student's baseline affect before they started using the tutor. Our measure is on Pekrun's theory of achievement emotions (2007), in particular coming from his AEQ-Math instrument (Pekrun et. al, 2005).

IntePre**:** *In general, do you feel interested when solving math problems?*

ExciPre*: Do you feel that solving math problems is so exciting?*

EnjoyPre*: Do you enjoy your math class so much that you are strongly motivated to participate?*

FrusPre*: Does solving math problems make you feel frustrated?*

AnxiPre*: Do you get anxious while solving math problems?*

BorePre*: Does it make you feel bored to just think of your math homework assignments?*

6.3.2 Within-tutor variables

Student State Variables

We have labeled student's record on each problem with one of the six student state variables (SOF, ATT, SHINT, GUESS, NOTR, GIVEUP). For each student, we are generating those six variables, by looking at the actions and timing excerpted by a student on a specific math problem. For example: SOF for student X, on a specific math problem Y, would be the number of times the student solved problems on the first attempt without help. We also calculate the total number of math problems that a student saw, where that specific behavior was observed. Table 1 shows the different specific behaviors that we summarize for each problem-student interaction. This variable then summarizes how often a student excerpted that specific behavior in math problems they encountered.

Table 17 student state variables

Student State	Description
SOF	Solved on first attempt without help
ATT	Answered after 1-2 incorrect attempts and self- corrected, without help.
SHINT	Answered with the use of some help, but not all, in at most 2 attempts.
GUESS	Answered after several attempts, more than 2 attempts
NOTR	Not enough time to read
GIVEUP	Enough time to read, but moved on before answering.

***Survey questions regarding* Affect variables**

Students were asked about to self-report their affective state, in particular how they felt in terms of Interest and Excitement, while working within the tutor. We averaged their responses on those questions and created the two affect variables.

INTE: average value for "How Interested are you?"

EXC: average value for "How excited are you?"

We also calculated the number of times a student accessed the Student Progress Page (SPP).

SPP: number of times SPP accessed by student

5.3.1.1 Post-Survey variables

After the students complete their experiment with Mathspring SPP, we asked them about their experience and perception of using Mathspring.

performedWell: *Do you think you have performed well in the math problems in Mathspring?*

learntLot: *Do you think that you have learned a lot on math topics using Mathspring?*

enjoyMathSpring: *I enjoyed using the system*.

hintsHelpful: *The hints and other help features were useful to me*.

The participants also took 7-item pre-test before the experiment and the same set of questionnaire as post-test at the end of the experiment.

5.3.2 Causal modeling of pre-survey variables

We decided to create a causal model from pre-survey variables to observe how the student variables are inter-related before our intervention.

We ran the pretest survey data in the Tetrad causal modeling software, in search for dependencies and potential causal links. There were 167 rows of data, where every row corresponded to a student. We excluded the data from the students who did not filled their survey. The result is in Figure 48.

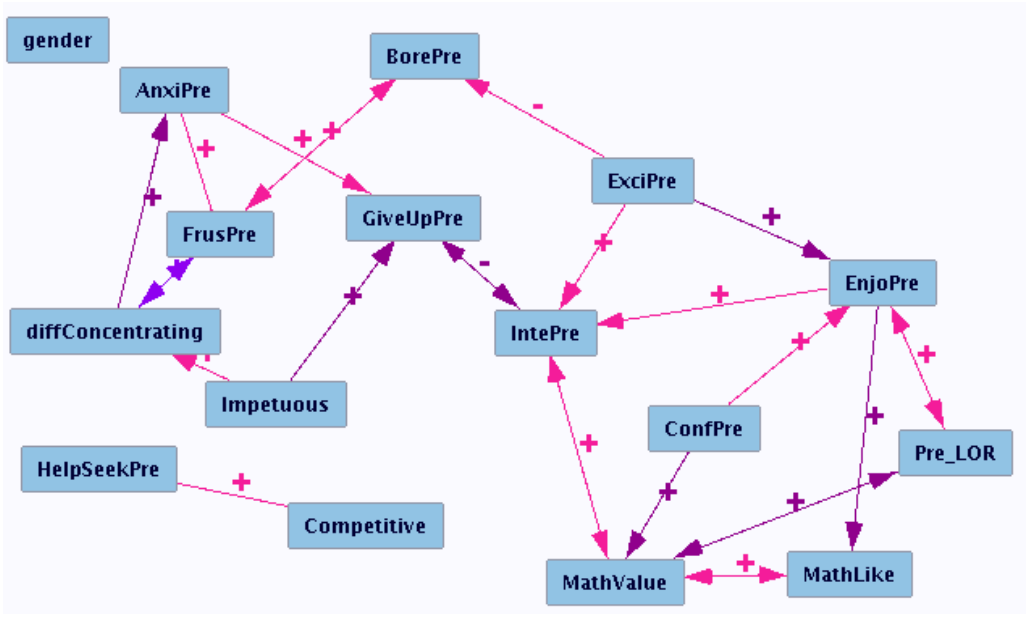


Figure 48, Causal modeling, pre-survey variables, Mathspring SPP data

Knowledge Tiers: We added knowledge tiers, which allows causal links to go from the higher tiers to the lower tiers but forbids the links going from lower tiers to higher tiers. We kept trait-like variables assessed at pretest time (e.g. Impetuous, liking of math) in tier 2 and state-like variables (AnxiPre, ExciPre) in tier 3, as shown in Figure 49. We are using our human knowledge of the domain to narrow our search to fewer Markov equivalent models. We have described the usage of knowledge tiers in detail in section 5.1.2.

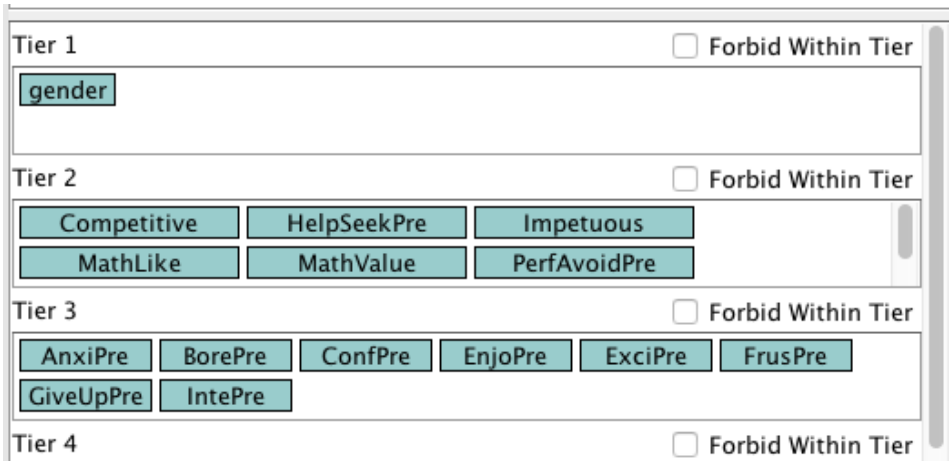


Figure 49 Knowledge tiers, pre-survey variables, Mathspring data

After we added knowledge tiers, we generated the causal model in Figure 50.

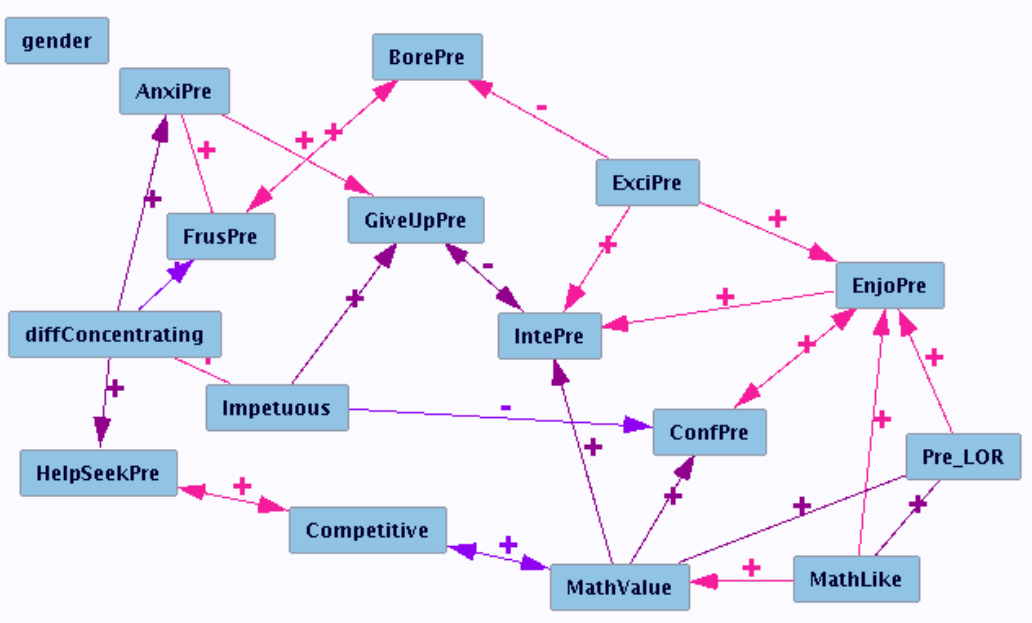


Figure 50 causal modeling with knowledge tiers, pre-survey variables, Mathspring SPP data

Next, we visually arranged the variables in the graph until we managed to notice two apparent clusters of variables (Figure 51). Variables that indicated positive learning behavior such as mathvalue, exciPre, Confpre are in one cluster while variables indicating negative learning behavior such as anxipre, boredPre, giveupPre are in another cluster. Gender seems to be unrelated to any of the variables. The clusters we have identified are logical groupings rather than actual statistical clusters.

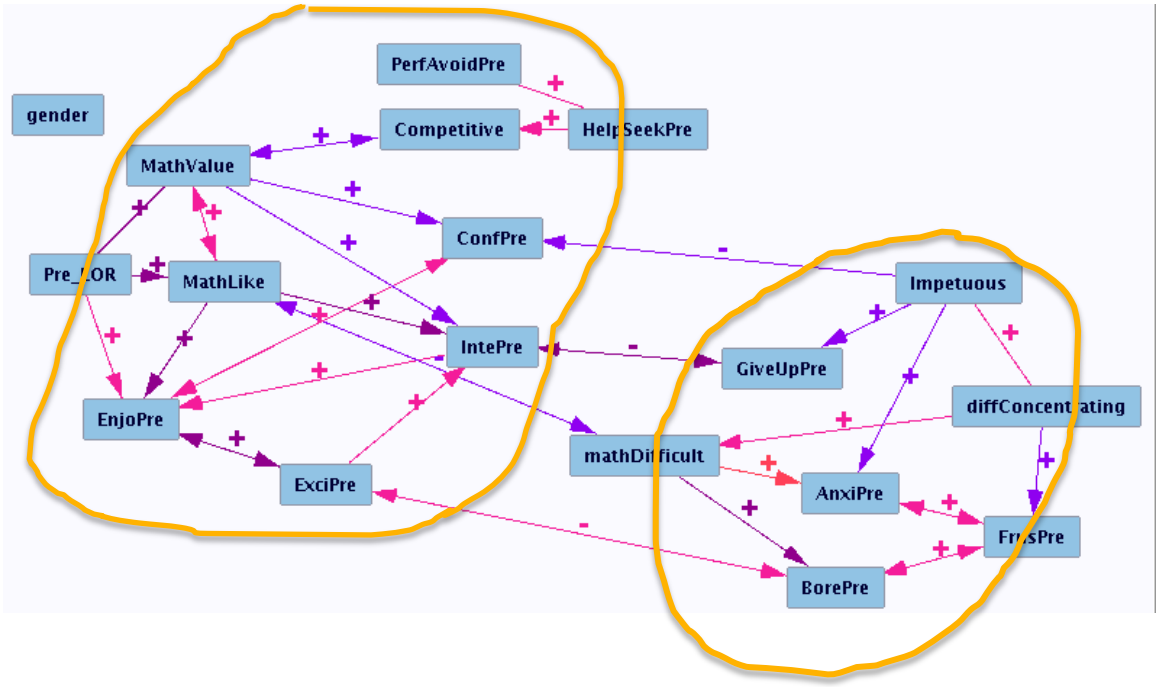


Figure 51 two clusters in causal modeling with knowledge tiers, pre-survey variables, Mathspring SPP data

If we look at the left cluster, we can see that the students who like math (MathLike) and value math (mathValue) tend to have more positive learning orientation (Pre_LOR). They also tend to work hard to win (competitive). Their reported higher level of prior confidence (confPre), interest, excitement and enjoyment.

On the right cluster, we see that the students who found math more difficult (mathDifficult) also have higher difficulty to concentrate (diffConcentrating) and often act without thinking (Impetuous). Those students also reported having higher level of anxiety (AnxiPre), boredom (borePre), and frustration (FrusPre) and they have higher tendency to give up (GiveUpPre).

We see the two clusters connected by negative links. The students who get excited get less bored. The students who like math find math less difficult. The students who are interested more have lower tendency to give up. The students who often act without thinking are less confident.

This dynamic between the student variables is as we have expected. Still, there are some interesting observations. Whether a student find math difficult (mathDifficult) seems to be affected both by her attitude towards math (MathLike) and her personality trait (diffConcentrating). Similarly, whether a student have higher tendency to giveup is dependent both on her interest in math and her impetuousness.
We can see that the students already have prior disposition to enjoy or get bored and frustrated, which are dependent on their prior attitude and experience with math learning and their personality traits developed over time.

5.3.3 Causal modeling with within-tutor variables

Next, I created a causal dependency graph from variables that describe students' state within MathSpring. The data consisted of 211 rows, where every row corresponded to a student, and each variable represents the total number of problems in which that specific behavior was exhibited in a math problem.

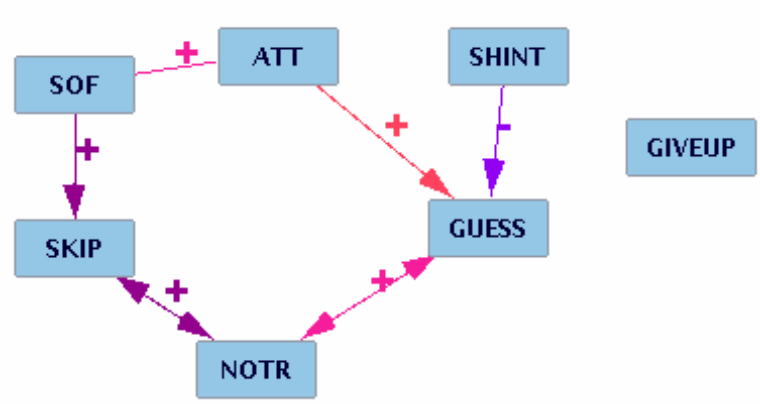


Figure 52 causal modeling, student state variables, Mathspring SPP data

We also analyzed general descriptives of student state variables and correlation values between them.

Table 18 mean and SD of student state variables across all students

SOF	SHINT	ATT	NOTR	GUESS	SKIP	GIVEUP
11.3 (8.9)	7.2 (6.7)	5.1 (4.6)	0.9 (2.6)	9.8 (9.5)	5 (4.0)	1 (1.8)

Table 19 correlation among student state variables

	SOF	SHINT	GIVEUP	NOTR	GUESS	SKIP	ATT
SOF		0.01	0.07	.143*	.407**	.401**	.557**
SHINT			0	-0.04	-.219**	-0.02	-.162*
GIVEUP				0.04	-0.07	.166*	-0.04
NOTR					.368**	.271**	.327**
GUESS						.270**	.811**
SKIP							.306**

From graph in Figure 53, we see that SOF, SKIP, NOTR, ATT, GUESS are all positively related. We do not interpret them as causal links. Solving a problem in the first attempt (SOF) does not make a student more prone to skipping a problem (SKIP). Rather, the students who have more SOF behaviors also happen to make more SKIP as they are solving more problems in general. It is however interesting to note that SHINT and GUESS are negatively correlated. SHINT is engaged behavior, whereas GUESS is a gaming behavior. There are too few instances of GIVEUP, which can be the possible reason why that behavior is not related to any other variables.

Next, we analyzed affective variables, namely students report of interest within MathSpring (INTE) and students' reports of excitement (EXC) and obtained the causal model in Figure 53.

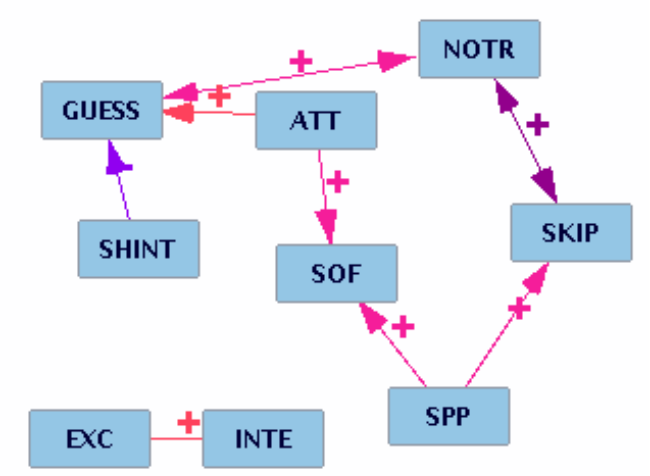


Figure 53 causal modeling, Student State variables and affect variables, Mathspring SPP data

Affect self-report variables EXC and INTE seem to be unrelated to all other student state variables. We need to note that EXC and INTE are highly correlated among themselves and so are the student state variables. This causes multicollinearity and false negatives. We have described multicollinearity and causal modeling in detail in section 5.1.1. Therefore, we added knowledge tiers that separate those variables into different tiers and forbid links among the variables within the tier. By forbidding correlational links among the variables within the tiers, links that connect different tiers are encouraged.

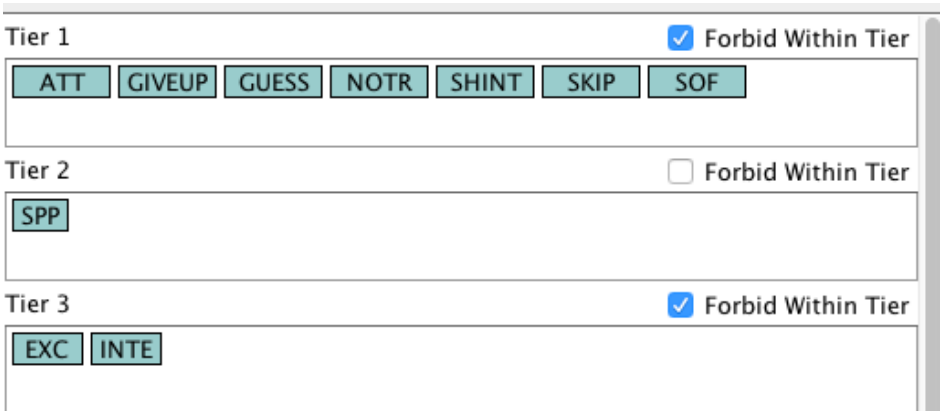


Figure 54 Knowledge tiers, Student State variables and affect variables, Mathspring SPP data

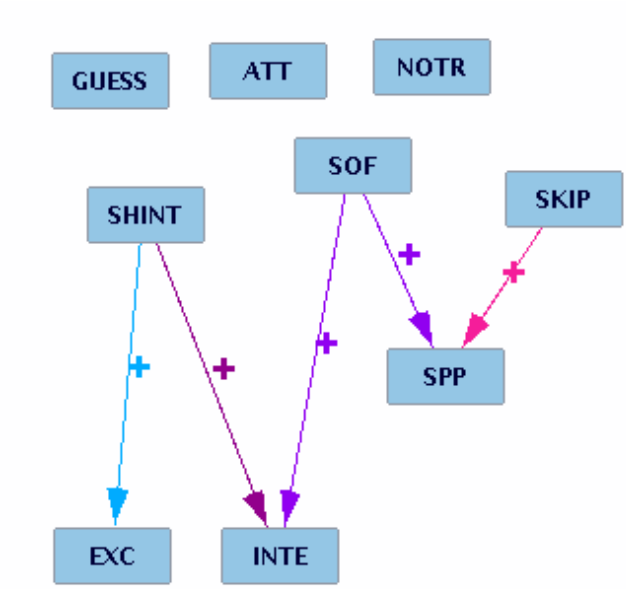


Figure 55 Causal modeling, Student State variables and affect variables with knowledge tiers that encourage correlational and causal links from variables towards the top to variables towards the bottom of the figure, for Mathspring SPP data

We see that the variables SOF and SHINT are related to EXC and INTE. This relationship is reasonable, as students solving problems correctly in the first attempt (SOF) and using help to solve problems (SHINT) are showing signs of engagement, and this is an indication of feeling positive in general, thus being more excited (EXC) and interested (INT). At the same time, students who are more excited and interested should in turn demonstrate more positive engaged behaviors that are conducive to learning, solving more problems and asking for more help. We also see a link SKIP → SPP, which is probably due to the experimental manipulation, in which students who skipped problems might have been further encouraged to use the SPP by offers from the Math Tutor. This link will be removed from further graph presentations, as this link has no real causal meaning.

We are next interested in whether gender has any role among those within-tutor variables. Thus, we ran the causal modeling algorithm adding gender to the within-tutor variables. We have assigned the gender variable a value of 1 for male students and a value of 2 for female students (thus, higher=female).

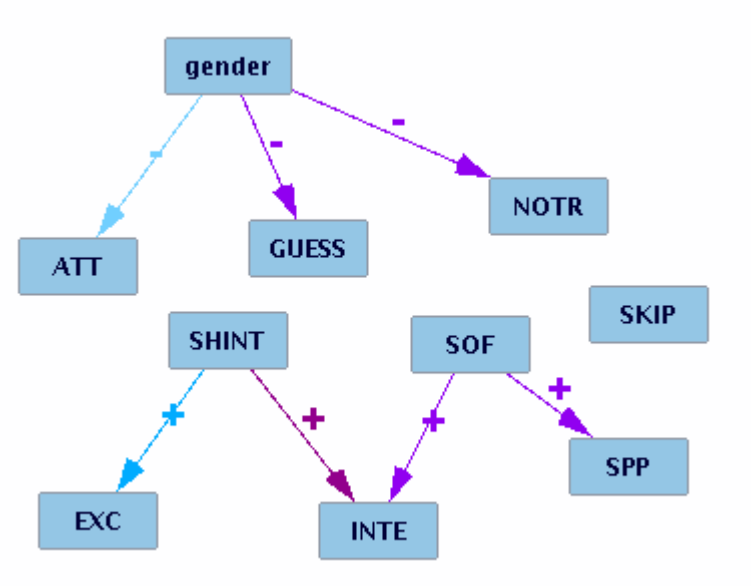


Figure 56 causal modeling, gender, Student State variables and affect variables, with knowledge tiers, Mathspring SPP data

Quite interestingly, gender is negatively related to ATT, GUESS and NOTR. Female students are solving less problems with multiple wrong attempts (ATT), making less quick guesses (GUESS) and having less instances where they are not reading the math problems (NOTR). We also looked at the average values of student state variables across gender (Table 20). It seems that male students are more disengaged (more GUESS and NOTR).

Table 20 mean and SD of student state variables across gender

gender	N	SOF	SHINT	ATT	NOTR	SKIP	GUESS	GIVEUP
1 (male)	104	12.8 (5.2)	6.7 (1.4)	6.2 (1.7)	1.4 (1.9)	5.5 (2.7)	12.3 (3.8)	0.8 (1.9)
2 (female)	87	10.8 (4.6)	7.6 (1.2)	4.3 (2.1)	0.4 (2.6)	4.9 (2.8)	7.9 (4)	1.3 (1.6)
p-value		0.14	0.3	0.06	0.01*	0.32	0.00*	0.01*

Next, we include the pretest, posttest and learning gain data and generated a new causal graph. We used the knowledge tiers as in Figure 57. We again forbid correlational links among variables within tiers to avoid multicollinearity.

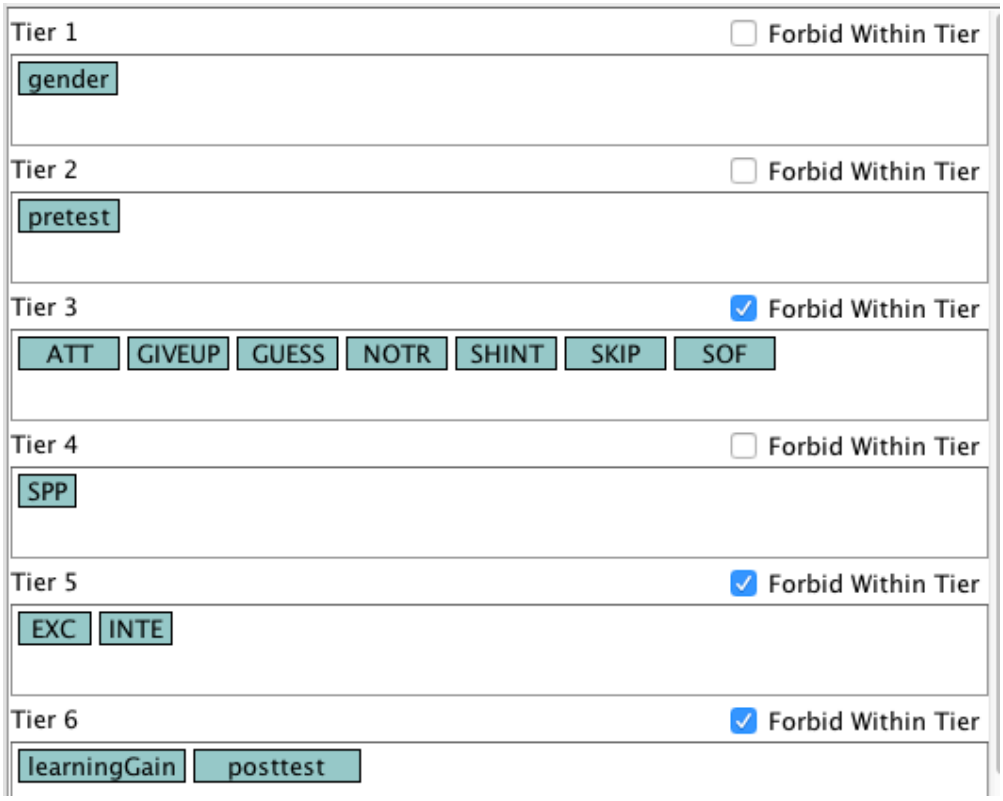


Figure 57 Knowledge tiers, gender, student state variables, affect variables and test variables, Mathspring SPP data

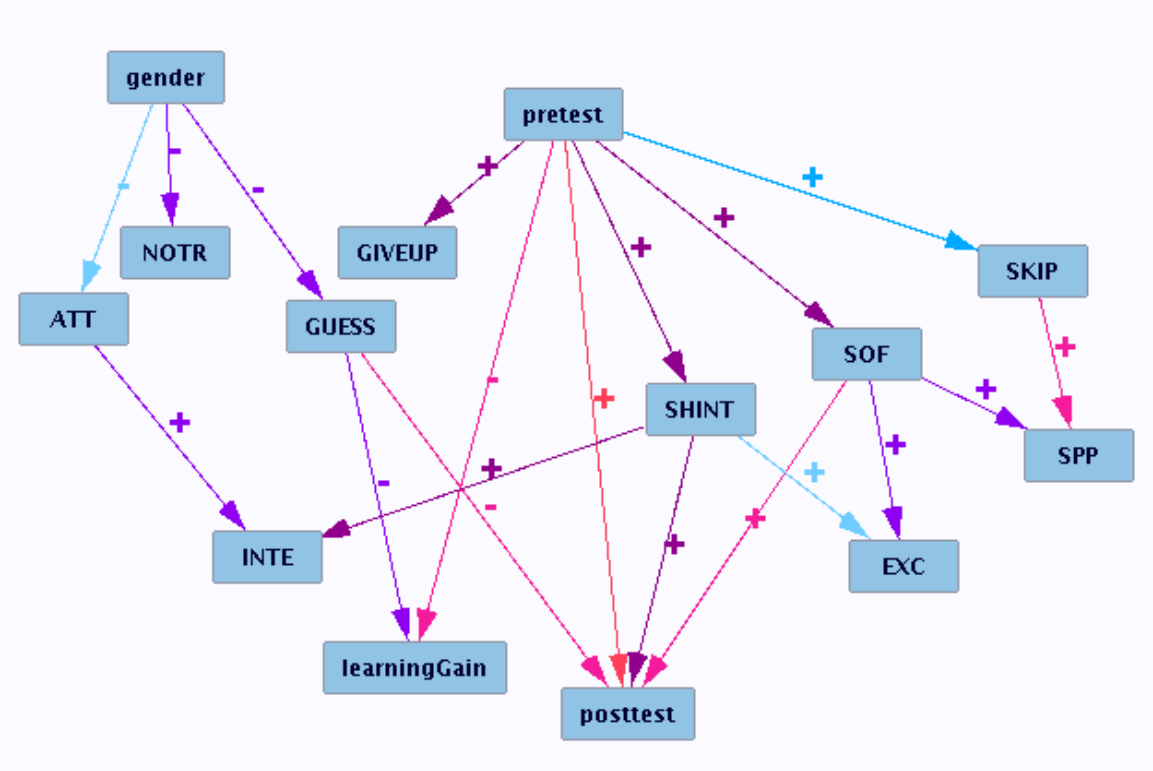


Figure 58 causal modeling, gender, student state variables, affect variables and test variables using knowledge tiers , Mathspring SPP data

Math incoming ability (Pretest) is positively related to math posttest outcomes and negatively related to learningGain. This is generally expected, as students who have lower ability have more room for growth. Students who have a higher pretest score have higher instances of SOF, SHINT, SKIP and GIVEUP. Students who had higher instances of SHINT and SOF also had a higher math posttest score and those students who had higher instances of GUESS also had lower posttest scores, as well as lower learning gains.

5.3.4 Causal modeling of Pre-survey and within-tutor variables

Next, we included all pre-survey and within-tutor variables to the overall model. When we included all 16 variables from the pre-test survey, the resultant graph was too dense to be intelligible. We played with different combinations of variables and decided to select only 9 pre-survey variables. We included baseline affect variables (incoming confidence, interest, boredom, excitement and frustration towards mathematics problem solving) as well as mathLike (representative of attitude towards math) and pre_LOR (representative of learning orientation). We decided to exclude trait-like variables such as competitive and impetuousness as they only had indirect links to within-tutor variables.

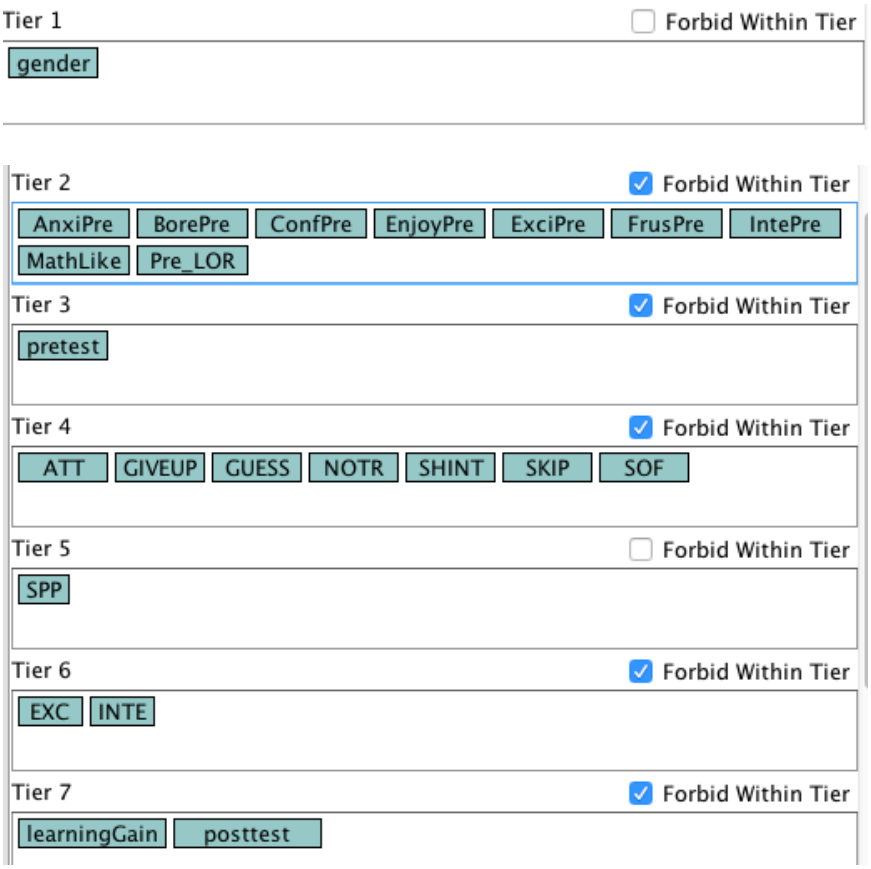


Figure 59 Knowledge tiers, pre-survey variables and within-tutor variables

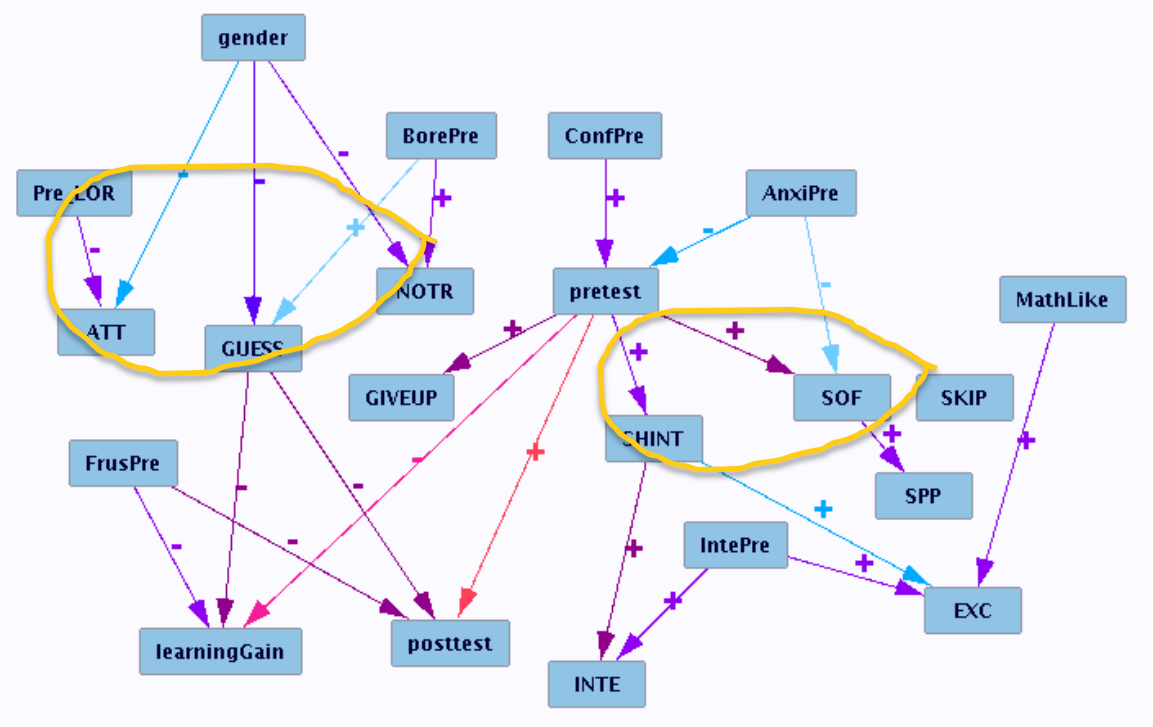


Figure 60 pre-survey variables and within-tutor variables with knowledge tiers, Mathspring SPP data

From this graph in Figure 60, we can observe two clusters of highly interconnected variables. We would like to emphasize again that the assignment of these clusters are logical demarcation based on our domain knowledge rather than actual statistical distinction. We are identifying two clusters in this graph based on our intuition of positive learning behavior and negative learning behavior. The cluster on the right comprises of SOF and SHINT and higher pretest, higher confidence and lower anxiety. They are also associated with higher interest and excitement. The cluster on the left is comprises of ATT, NOTR and GUESS. They are linked to pre-survey variables that indicate negative learning behavior (boredom and negative learning orientation). Gender is linked in this cluster. This suggests that there could be gender differences regarding

student learning behavior. However, we do not see any gender differences in how male and female students are reporting their affect and learning orientation and performance.

Duckwoth and Seligman (2006) , in their famous paper 'Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores' have found that female students have higher self-discipline, which gives them edge in academic achievement. We would like to recall the model from Assistment data (Figure 61) that we have descried in section 5.2.2. We found that female students are solving more problems despite having higher prior knowledge and despite reporting higher self-discipline. While analyzing the causal model, we had faced a confound: are female students more self-disciplined and solving more problems due to their higher self-discipline, but are under-reporting their self-discipline? Or solving more problems is not a reflection of self-discipline in the first place?

We are in a similar confound with the causal model with Mathspring. Are female students showing fewer instances of disengaged behavior (ATT, GUESS, NOTR) because they have better learning orientation (even though they are not reporting better learning orientation)? Or female students showing fewer instances of disengaged behavior is unrelated to having better learning orientation?

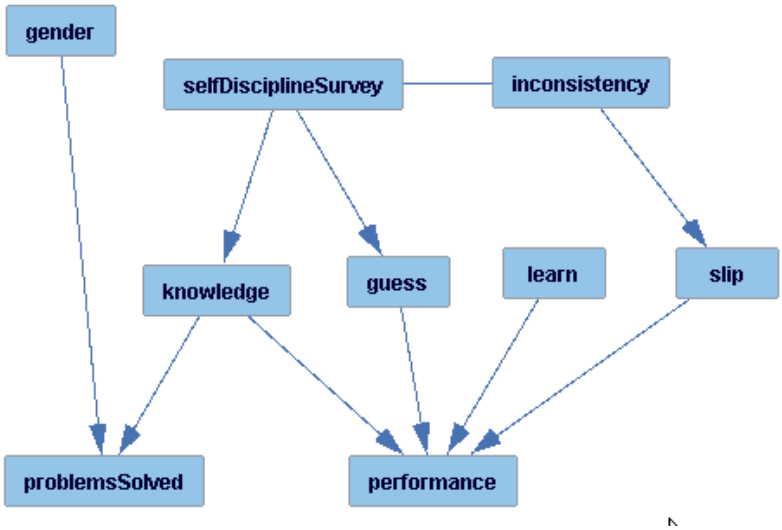


Figure 61 causal modeling of Assistment data

5.3.5 Pre-Survey, within-tutor, Post-Survey variables

We created a causal model combining pre-survey, within-tutor and post-survey variables. We did not include all variables from pre and post survey variables, so as not to make the graph too crowded. There are 130 rows of data. We discarded data from students with incomplete data.

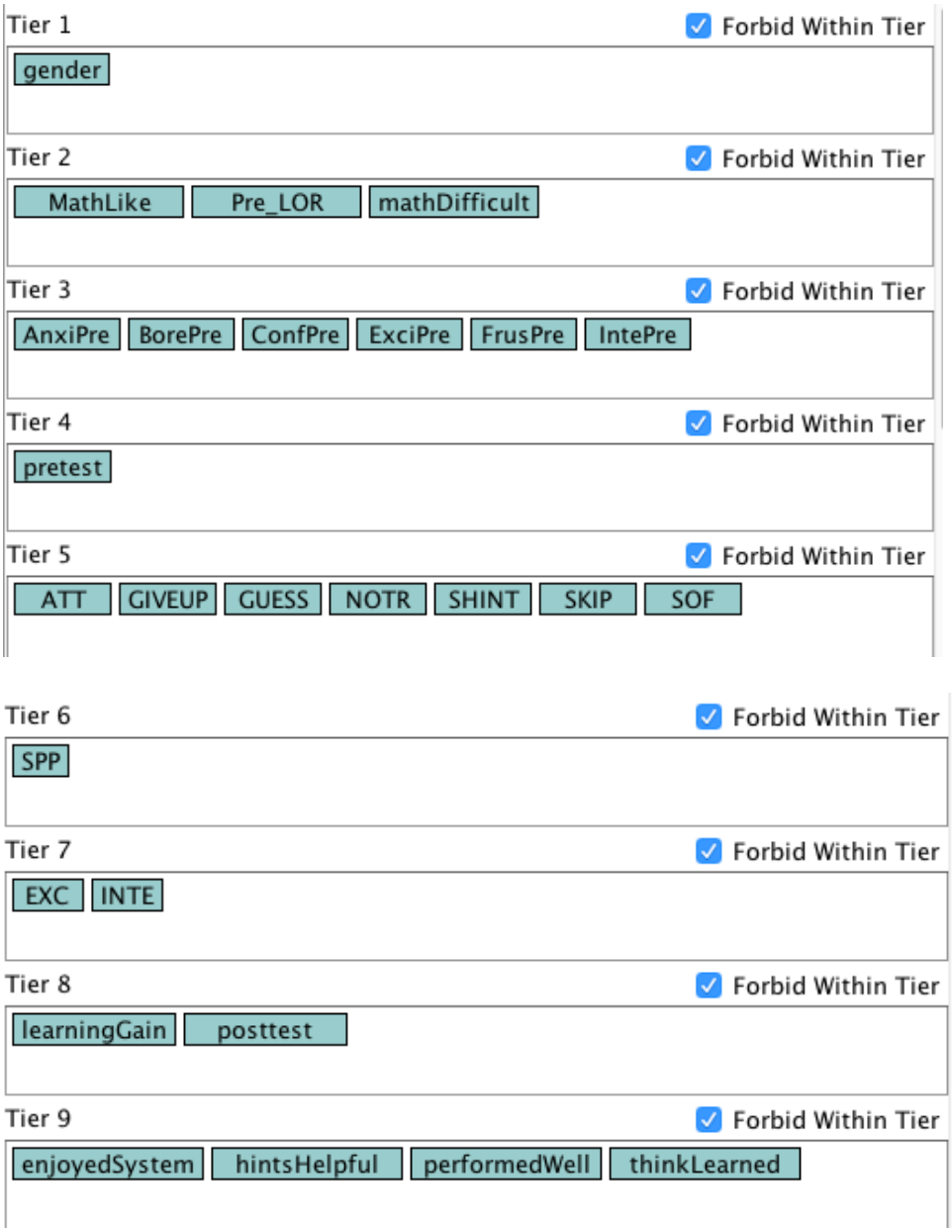


Figure 62 knowledge tiers, pre-survey, within-tutor and post-survey variables, Mathspring SPP data

We used knowledge tiers in Figure 62 and created causal model in Figure 63. We arranged the variables so that we can divide the variables in two clusters: performance oriented and enjoyment oriented.

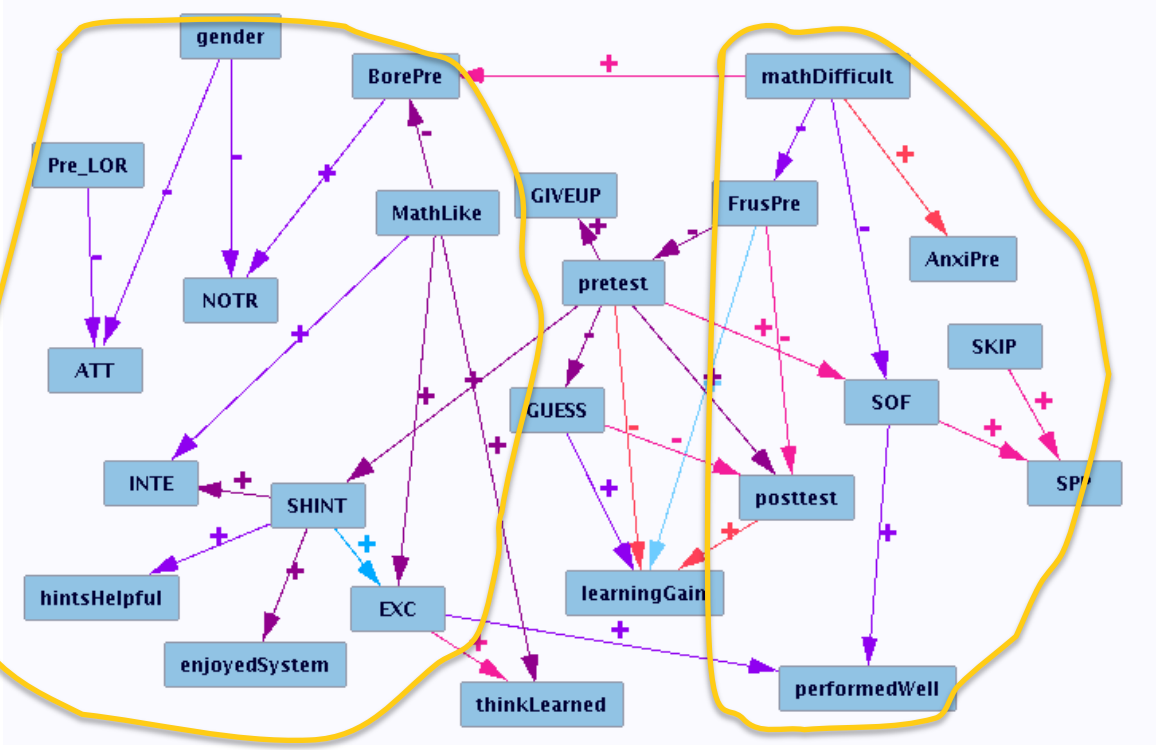


Figure 63 causal model with knowledge tiers, pre-survey, within-tutor and post-survey variables, Mathspring SPP data

Before explaining this causal model, we would like to recall the causal model (Figure 45) that we created for Wayang Outpost, earlier version of Mathspring. We have described the process in section (5.2.1). In that causal model also, we had identified two clusters of variables. Performance oriented variables (Students who have higher prior knowledge and better self concept in math reported higher confidence and lower frustration) and

variables related to liking and appreciation (liking math and perception of the tutor related to interest and excitement).

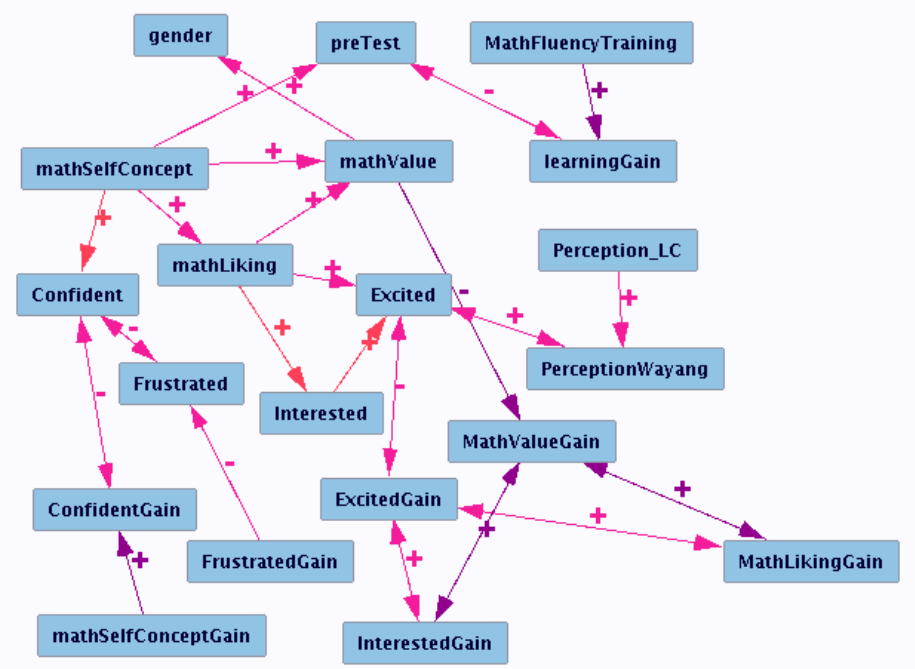


Figure 64 Causal model with Knowledge tiers; Wayang Outpost data

In Mathspring model in Figure 63, we regard the right cluster comprising SOF as performance oriented and left cluster comprising SHINT as enjoyment oriented. The students who used more tutor help features (SHINT) reported enjoying the system more (enjoyedSystem), finding the tutor more helpful (tutorHelpful) and being more excited (EXC) and interested (INTE). In the performance oriented cluster, the students who found math difficult (mathDifficult) solved less problems correctly in the first attempt (SOF), reported higher anxiety (AnxiPre) and higher frustration (frusPre).

5.3.6 What do these causal models say about SPP?

Interestingly, SPP usage is not related to any particular behavior pattern or cluster of variables. This comes somewhat as a surprise given that, in our personal observations of students using the SPP, students are generally found to be very receptive of it. The lack of empirical evidence of the SPP's impact could be due to multiple factors. One factor could be our experimental setup. Given that the data collected in this study comes from prompting the SPP to a group of students who report low affect, this could be causing selection bias. It is also possible that while some motivated students are using the SPP productively, some might be using it unproductively. We tried running a cluster analysis to tease different kinds of SPP usage patterns but the number of data cases was too small for this analysis to succeed. Though causal modeling did not reveal anything particular about SPP and SPP usage did not reveal anything noteworthy about students, the tutor itself nor learning with the tutor, we still see value of causal modeling as a research tool in our design of SPP. Causal modeling provides a tool to examine and validate the existing design of SPP and can act as a guide for updating the design of SPP.

Casual models informing re-design of SPP

An important goal of the Student Progress Page is to accurately represent a students' effort and knowledge through its visual representation, which is why the SPP divides a student's state in two major categories: knowledge and effort. Each student's knowledge and performance are represented by a mastery bar and by the growth of a pepper plant. A

student's effort is represented by flowers, fruits, growth and wilting of the pepper plant.

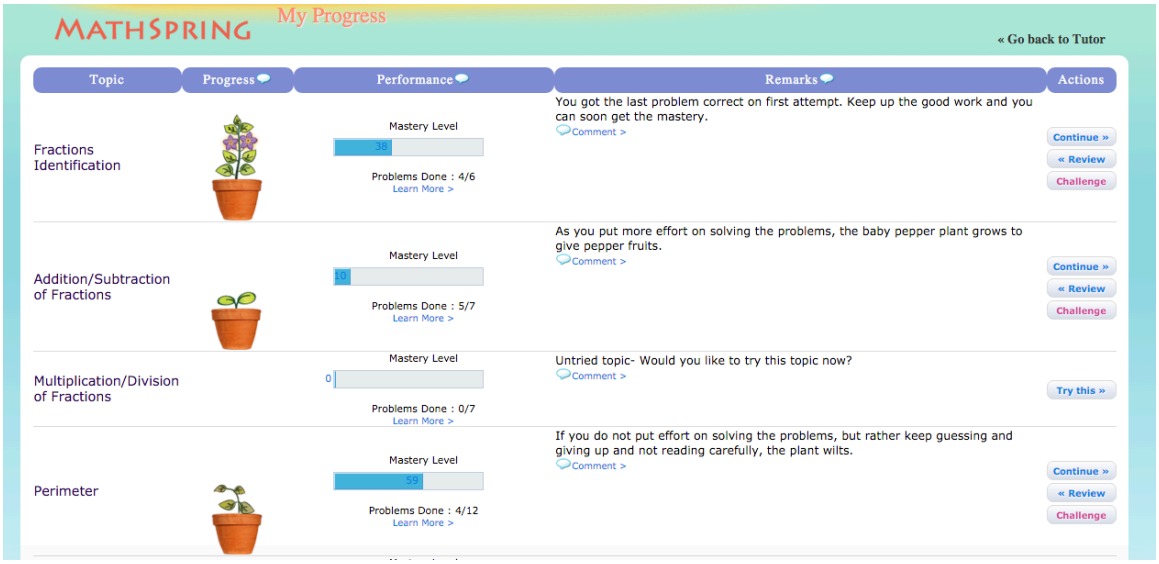


Figure 65 screenshot of Mathspring SPP

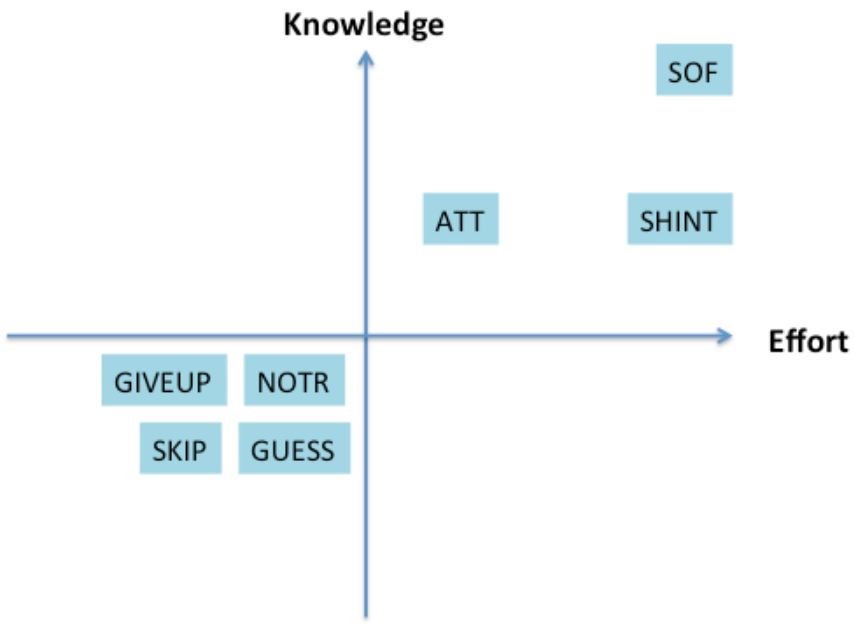


Figure 66 Student state variables in knowledge-effort quadrants

In the SPP, we make inferences of knowledge and effort based on student state variables. Figure 66 shows a knowledge-effort quadrant where student state variables are placed. SOF is indicator of high knowledge. SHINT is a reflection of both knowledge and effort. NOTR, GUESS, SKIP, GIVEUP are indicators of low effort.

In our causal models, we can see a pair of two clusters across the four quadrants. In Figure 60, we can see that NOTR and GUESS conform a highly inter-related cluster that shows disengaged behavior, while SOF and SHINT make a cluster reflecting engaged behavior.

On the other hand, Figure 63 showed that SOF is tightly associated to performance variables, whereas SHINT is further associated to enjoyment variables.

From the models, we can make the following generalizations with regards to student state variables.

- SOF is a proper measure of knowledge and performance
- SHINT seems be most effective towards inferences of student engagement
- NOTR and GUESS indicate disengagement
- GIVEUP and SKIP are not as effective as NOTR and GUESS towards inferences of disengagement
- ATT seems to be more indicative of disengaged behavior than engaged behavior

If we compare this finding with our current inference algorithm of what the SPP externalizes to the student, we can say that we have done a good job of relying on SOF and SHINT to demonstrate positive learning behaviors. We had given emphasis to SHINT as an indicator of engagement. However, we did not find evidence for ATT as an indicator of positive learning behavior. It is possible that while some ATT behaviors could represent real effort at finding the right solution, some ATT could be guessing. Currently, we are grouping the four variables, NOTR, GUESS, GIVEUP and SKIP, in one general category that indicates disengagement and low effort. It might be more effective if we give more weight to NOTR and GUESS while making an inference of disengaged behavior. I am not suggesting that we need to modify our SPP algorithm after each study, but instead that the causal models from our studies enable us to reflect on how accurate our algorithms and designs are, towards the final goal of accurately

reflecting effort and performance to the student. They can guide us in updating our assumptions and algorithms. It is not until we have a system that updates the design automatically, that we will need to manually update our design with continuous empirical evaluation.

5.4 Causal modeling with Mosaic

We ran a randomized controlled study with the Mosaic game within Mathspring, to analyze the hypothesis that Mosaic could be a game-like intervention to improve student affect. One hundred and eighty six students from urban schools in the Northeast USA participated in the study. The students used Mathspring over a single class period. There were two Mosaic mini-games: London and Manhattan. The students were randomly assigned to one of the three experimental conditions:

a. No-Mosaic: Students do not get Mosaic mini-games during the whole Mathspring sessions (N=60).

b. Mosaic Prompt: While using Mathspring, students will be asked randomly whether they want to try Mosaic mini-game. Students are free to accept the offer and play mini-game or reject and continue with Mathspring (N=62)

c. Force Mosaic: Students in this group are taken to Mosaic directly while working on Mathspring (N=64)

We used almost the same survey that we used in our Mathspring SPP experiment. However, since we ran the whole experiment within a single class session, we had to narrow down the number of survey questions. Similarly to the SPP study, we created variables from the survey responses and log records.

Note that student math performance and learning was not assessed as part of this study, as this was not part of the hypothesis, which stated that Mosaic is a game-like element that acts as an affective tool, thus helping students to feel better, and possibly engage more with the system due to this fact.

Mosaic Pre Survey

Similar to our previous experiment with SPP described in section 5.3, we are using Wigfield and Eccles (2000) Expectancy–Value Theory of Achievement Motivation to measure student attitude. We are using Carol Dweck's (1999) theory of motivation for learning orientation and Pekrun's theory of achievement emotions (2007) for baseline affect. The variables included in the pretest and posttest surveys are described next.

Attitude towards Math

likeMath*: Do you like your math class?*

mathDifficult*: Do you worry that math class is much too difficult for you?*

Learning Orientation

Pre_LOR*: When you are doing math exercises, is your goal to learn as much as you can? Is it your goal to demonstrate that you are better at math than your classmates?*

goodLearner*: Do you prefer learning about things that make you curious even if that means you have to work harder?*

Competitive*: Do you work hard in school so that you can beat everyone?*

HelpSeekPre*: Do you prefer to learn on your own, without being offered help?*

Baseline affect

BoredPre*: Does it make you feel bored to just think of your math homework assignments?*

GiveupPre*: When solving math problems, do you prefer to give up?*

IntePre**:** *In general, do you feel interested when solving math problems?*

ExciPre*: Do you feel that solving math problems is so exciting?*

EnjoyPre*: Do you enjoy your math class so much that you are strongly motivated to participate?*

FrusPre*: Does solving math problems make you feel frustrated?*

AnxiPre*: Do you get anxious while solving math problems?*

Self-discipline

We are still using Brief Self-Control Scale (BSCS) developed by Tangney et. al. (2004) for measuring self-discipline. Self-discipline survey was strongly predictive of student performance in our study with Assistment. However, self-discipline survey was less predictive in our study with Mathspring SPP. We had only two questions in SPP survey compared to the whole 13-item survey in Assistment. In Mosaic study again, we could only pick two items. We picked the two questions that we assumed were more straight-forward assessment of self-discipline.

hardWorker*: Are you a hard worker?*

setbacksDiscourage: *Do setbacks discourage you?*

Mosaic Post Survey

After the students completed the experiment, we asked them survey questions on their experience and perception of the tutor.

performedWell*: Do you think you have performed well in the math problems in Mathspring ?*

learnedLot*: Do you think that you have learned a lot on math topics using Mathspring ?*

enjoyMathspring*: Did you enjoy using Mathspring?*

hintsHelpful*: Do you think hints and other help features were helpful?*

SPPHelpful*: Was Progress Page useful?*

Within Tutor Variables

Students were asked about to self-report their affective state, in particular how they felt in terms of Interest and Frustration, while working within the tutor. We averaged their responses on those questions and created the two affect variables.

INTE: "How interested are you feeling?"

FRUS: "How frustrated are you feeling"

Some students skipped these affect surveys. For each student, we counted the number of times the survey has been skipped and created the following variables.

INTE_skip: number of times survey "How interested are you feeling" is skipped

FRUS_skip: number of times survey "How frustrated are you feeling" is skipped

We also calculated the number of times a student accessed Mosaic and accessed the Student Progress Page (SPP).

Mosaic: number of times Mosaic mini-game accessed

SPP: number of times SPP accessed

We did not conduct pretest and posttest on math skills in this study. We had only one class session for the whole experiment which limited us in time. Also, the min-games are regarded as affective boost and we did not expect the gain in student affect to translate to higher learning gain within one class session.

5.4.1 Causal modeling with pre-survey variables

A causal model was created from the pre-survey variables as in Figure 67. There are 186 rows of data, where each row corresponds to a student. Similar to the causal model from the SPP study (Figure 50), we can also see two clusters here. Variables such as likeMath , confPre, intPre, hardworker and goodLearner comprise of one cluster showing positive learning behavior. On the other hand, variables such as frusPre, anxiPre, giveupPre and mathdifficult comprise of another cluster of negative learning behavior. Gender was related to mathDifficult, implying that female students reported worrying about math class being too difficult.

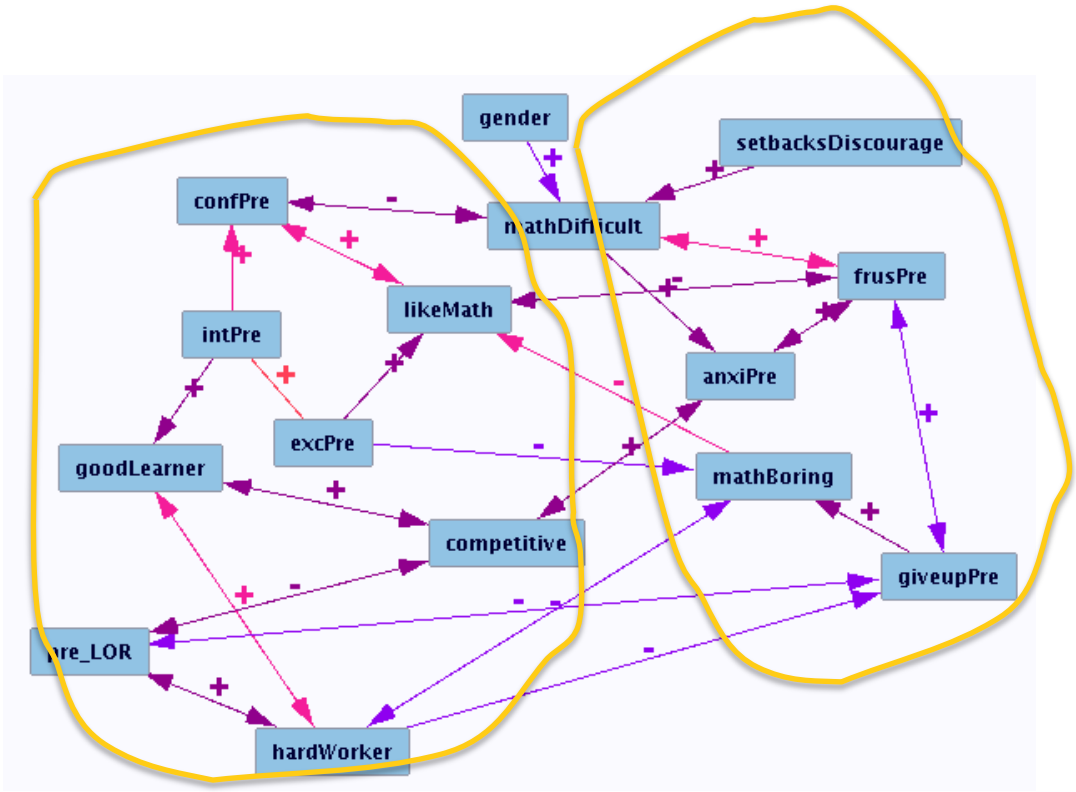


Figure 67 Causal modeling, pre-Survey variables, Mosaic Data

5.4.2 Causal modeling with within-tutor variables

We ran the causal modeling software including our student state variables of MathSpring usage and obtained a graph as shown in Figure 68. We also calculated average values of these variables in Table 21, across students. Since the experiment was only for one class session, and exposure was limited, and thus there are overall fewer instances of student state variables than in previous experiments. In particular, there are very few instances of SHINT. A possible reason is that students did not get to watch the Mathspring tutorial that taught them how to use the help features as part of this study.

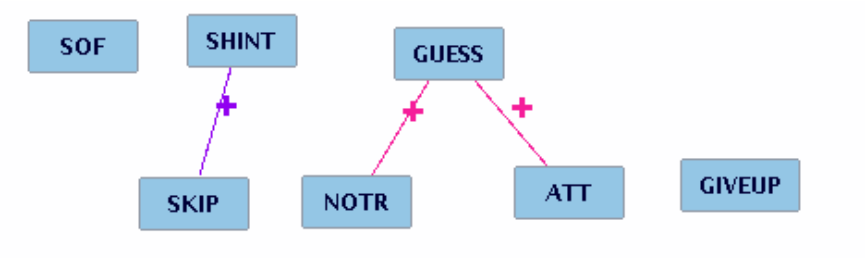


Figure 68 Student State variables, Mosaic data

Table 21 Student state variables, Mean and SD

SOF	SHINT	ATT	GUESS	SKIP	NOTR	GIVEUP
7.1 (4.9)	0.6 (1.3)	2.1 (1.9)	4.3 (3.8)	2.5 (2.7)	1.1(2.2)	0.5 (1.8)

When I added the affect self-report variables of interest and frustration within MathSpring (note that excitement and confidence were not assessed in this study) to the pool of data of student state variables, the graph in Figure 69 was obtained.

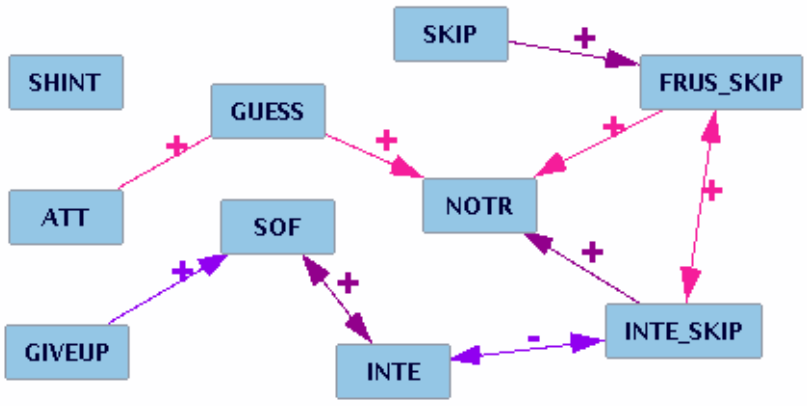


Figure 69 Student state and affective state variables within the tutoring session, Mosaic data

One interesting finding is that the students who skipped the affective survey were found to show more disengaged behaviors (NOTR and SKIP) . This makes me think that the experiment design should be reconsidered in future studies, as students who are already disengaged are going to skip the survey questions as well, and consequently the assessment of the emotion via self-reports may be biased. Consequently, when trying to add interventions based on affective surveys, researchers need to consider this scenario.

5.4.3 Causal modeling with pre-survey and within-tutor variables

We used the knowledge tiers in Figure 70 to create a causal graph of pre-survey and within-tutor variables, so that incoming variables are not allowed to be considered as consequences but causes, if a dependency is seen that involves these variables.

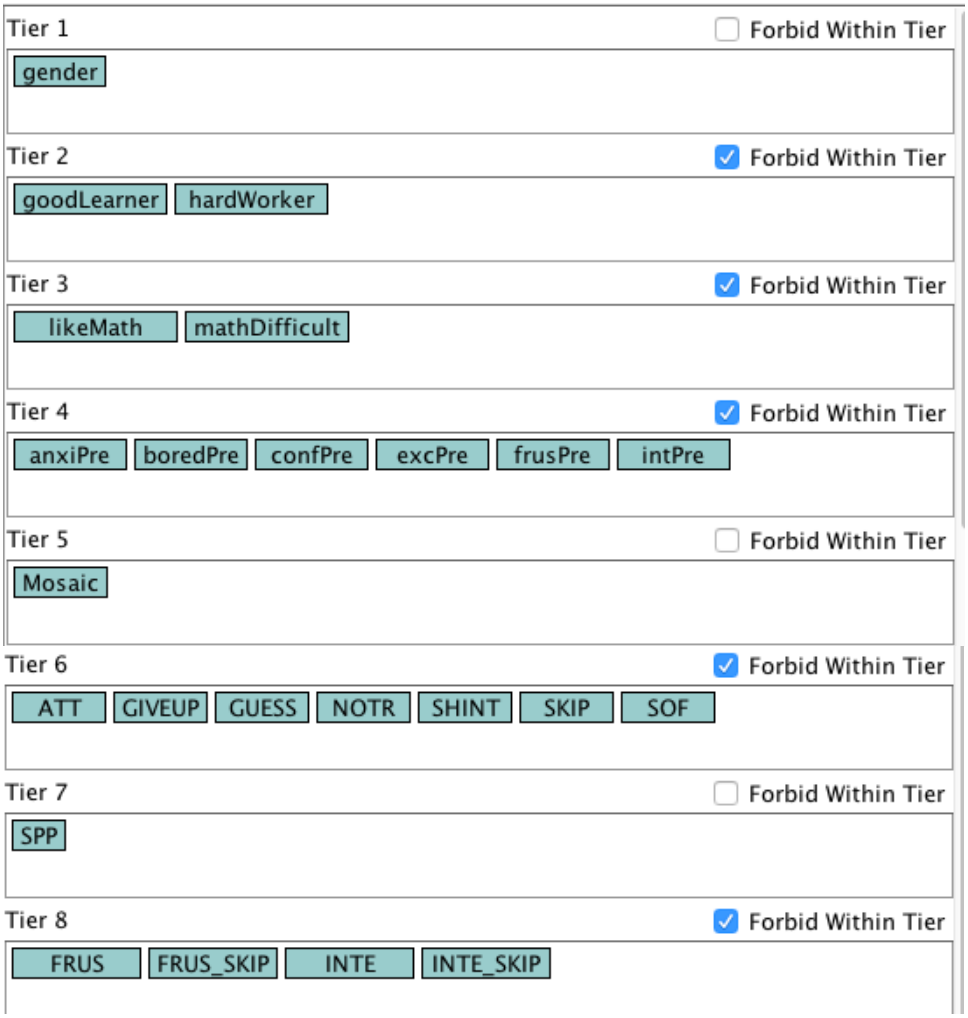


Figure 70 Knowledge tiers, pre-survey and within-tutor variables, Mosaic data

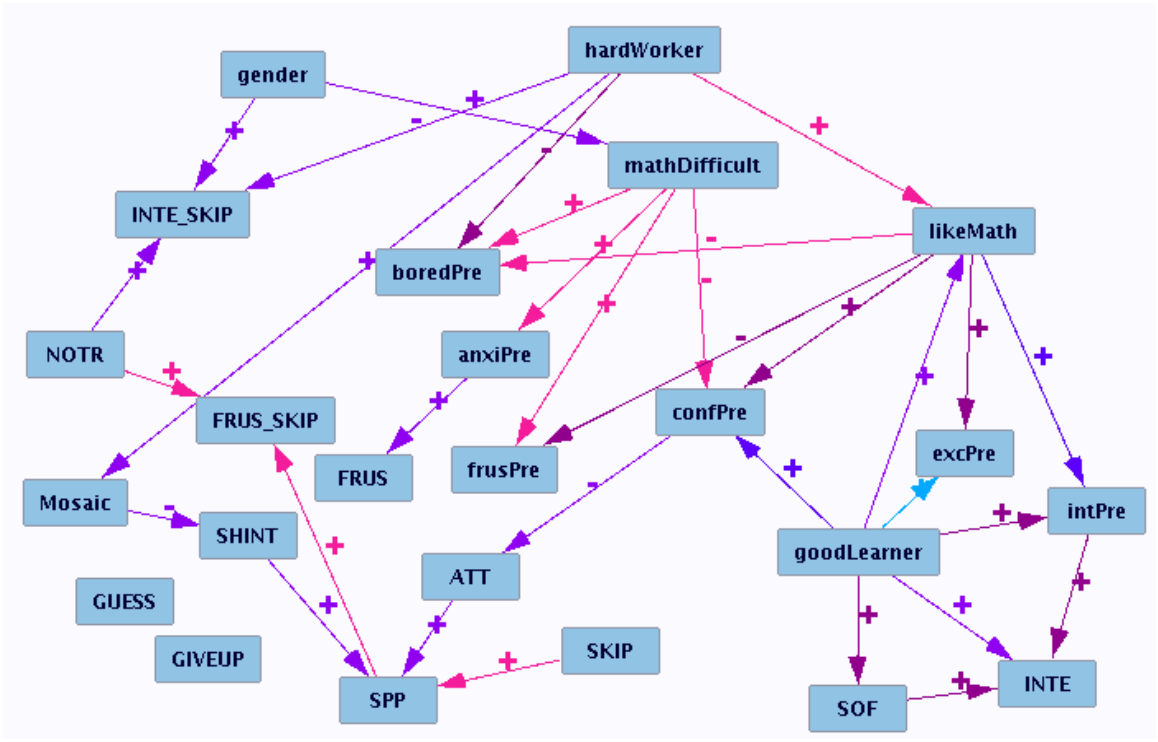


Figure 71 Causal modeling with knowledge tiers, pre-survey and within-tutor variables, Mosaic data

The students who reported that they prefer learning about things that make them curious even if that means they have to work harder (goodLearner) also solved more problems correctly (SOF), had higher baseline interest in math (intPre), higher baseline confidence (confPre) and higher baseline excitement (excPre), as well as higher interest level reported while working on the tutor (INTE). This is another instance that a student's learning trait and orientation is shown to impact her experience and performance within the tutor. The students who worried that math is too hard (mathDifficult) reported higher level of baseline anxiety (anxiPre) which led to higher frustration within the tutor

(FRUS). They also reported more baseline boredom (boredPre) and less confidence (confPre).

We see that the variable hardWorker was positively associated to Mosaic. A possible explanation is that the students who are more diligent may have progressed faster within the tutor which increases likelihood of getting Mosaic. However, we do not see the hardWorker variable associated to student state variables; this makes that explanation unlikely. We can also see a negative link between SHINT and Mosaic, as if Mosaic made students not see hints. A possible explanation is that the students who played Mosaic had less time for overall tutor activity and therefore had less SHINT instances.

5.4.4 Causal modeling with within-tutor and post-survey variables

We used knowledge tiers in Figure 72 to create a graph Figure 73 with within-tutor and post-survey variables. There were 116 rows of data, one row per student. We were not able to collect post-survey data from two classes with 23 and 25 students in each. The classes were terminated just before they tool the survey due to emergency fire alarms.

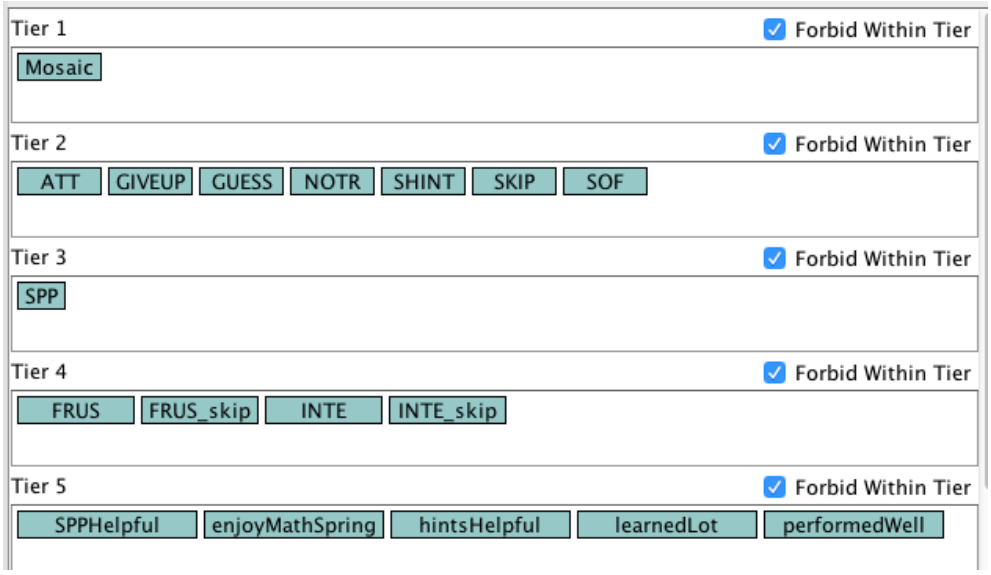


Figure 72 Knowledge tiers, within-tutor and post-survey variables, Mosaic data

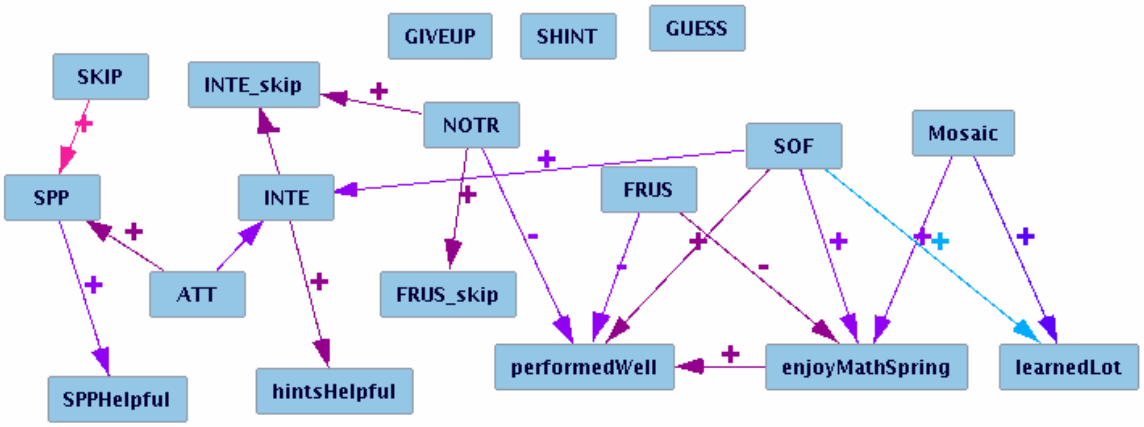


Figure 73 Graphical model with knowledge tiers, within-tutor and post-survey variables, Mosaic data

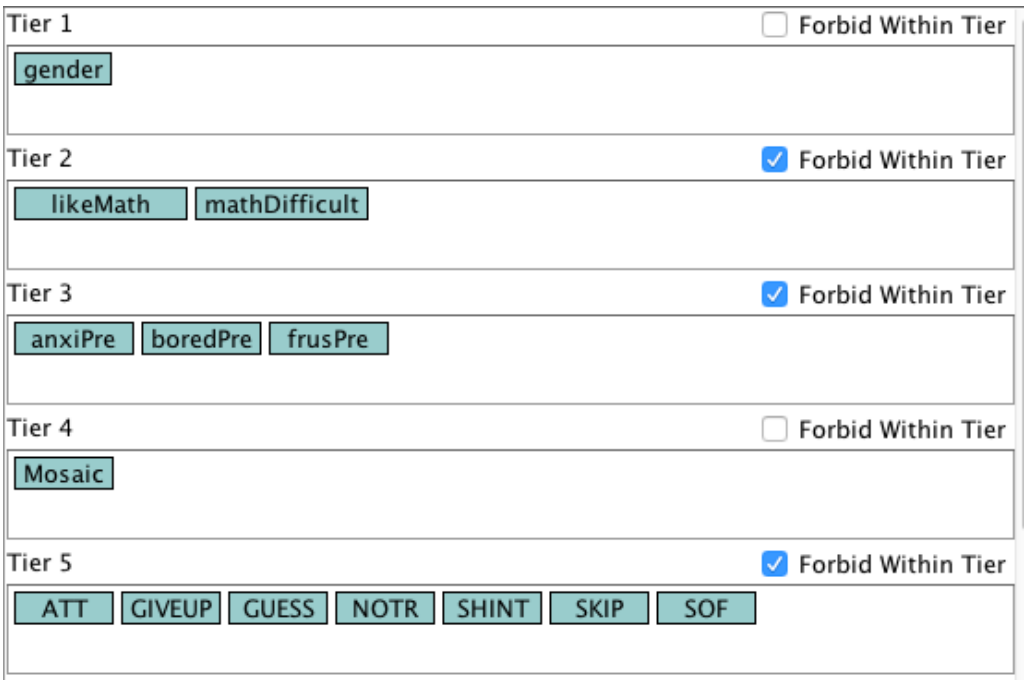
Students with higher number of instances of SOF reported more frequently that they performed well at posttest time (performedWell), that they enjoyed Mathspring (enjoyMathspring) and that they learned a lot from using Mathspring (learnedLot). The students who had more instances of not reading (NOTR) reported performing less well.

Both SHINT, GIVEUP and GUESS are not associated to any other variables.

We can see the variable Mosaic (how many times students used Mosaic mini-game) is positively linked to students reporting to enjoying Mathspring and learning a lot from Mathspring. Interestingly, Mosaic has no influence on students' perception that they performed well. This aligns with our initial hypothesis that Mosaic is an affective intervention designed to increase students' affective state. However, Mosaic is not associated to student's affective survey variables INTE and FRUS. This could be related to the fact that we did not sample INTE and FRUS sufficiently enough. We have described this in detail in section 4.3.

5.4.5 Causal modeling with Pre-survey , within-tutor and post-survey variables

We used the knowledge tiers in Figure 74 to create a graph using pre-survey, within-tutor and post-survey variables. Since the graph becomes too crowded once we have too many variables, we had to choose which variables we want to include and which we do not want. We first ran causal modeling with all variables and then excluded the pre-survey variables that had only indirect effects on within-tutor and post-survey variables.


Tier 1
Forbid Within Tier
gender
Tier 2
Forbid Within Tier
likeMath
mathDifficult
Tier 3
Forbid Within Tier
anxiPre
boredPre
frusPre
Tier 4
Forbid Within Tier
Mosaic
Tier 5
Forbid Within Tier
ATT
GIVEUP
GUESS
NOTR
SHINT
SKIP
SOF

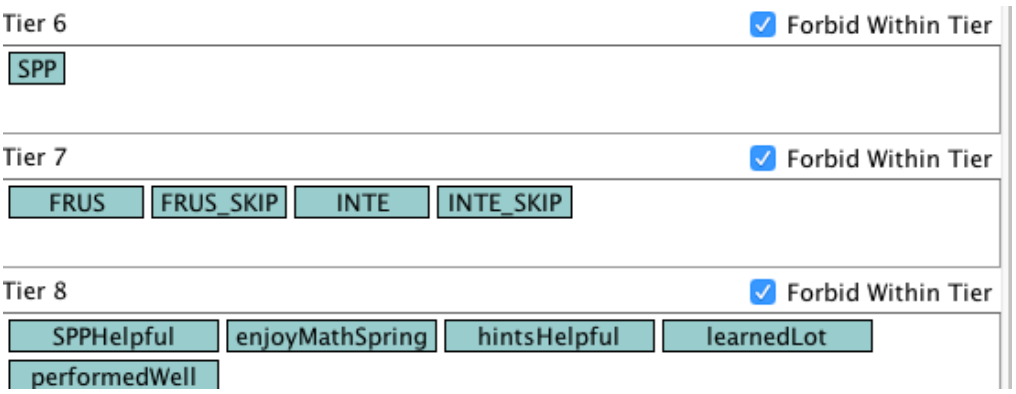


Figure 74 Knowledge tiers, pre-survey, within tutor and post-survey variables, Mosaic data

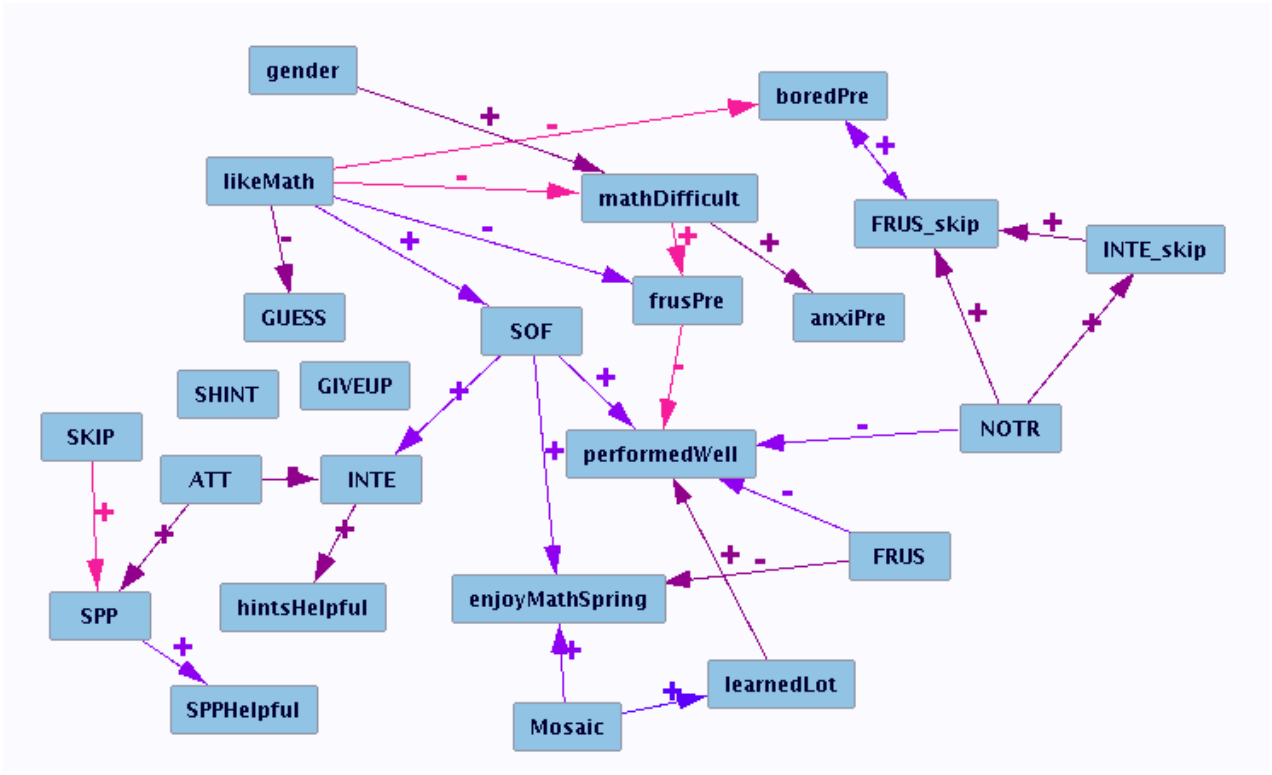


Figure 75 Causal modeling with Knowledge tiers, pre-survey, within tutor and post-survey variables, Mosaic data

This graph in Figure 75 , similar to the graph from SPP (Figure 63) shows that students' attitude to math, their learning orientation, baseline affect, performance within tutor, affect within tutor and eventual experience and perception of the tutor are all interrelated.

The students who like math more before starting will have better performance while using MathSpring (SOF), which leads to better affect self-reports in the tutor (INTE) and report having had a better experience with the tutor at posttest time (enjoyMathspring).

In the causal model in the previous experiment, the main intervention (the student Progress Page) had no significant influence on any of the outcome variables. However, in this causal model over a new set of students, the main intervention (Mosaic) did have a positive influence on students' post-tutor outcome variables, at least regarding enjoyment and their perception of learning from the tutor. Still, Mosaic did fail to influence students' affective states within the tutor, and their math performance inside of the tutoring system (at least indicators of good performance such as SOF). Our actual goal is to design and implement mini-games in such a way that playing the mini-games will give a boost to student's affect and which will then result in better performance. To explore whether this goal is achievable, we need to run this experiment over multiple sessions.

6 Conclusions and Implications

Educational games and intelligent tutoring systems have their own strengths and limitations in terms of offering meaningful learning experiences. While educational games can provide rich and engaging learning experiences, intelligent tutors can deliver measurable learning gains due personalized tutoring and practice. This dissertation has attempted to find practical ways to integrate those two approaches, identifying three distinct ways to add game features into intelligent tutors. I created three game-like systems that targeted cognitive, metacognitive and affective aspects of the student as their primary mode of intervention. Monkey's Revenge is a game-like math tutor that offered tutoring within a game-like environment. The Learning Dashboard is a game-like metacognitive support tool for students using the Mathspring tutoring system, which helps students reflect on their performance. Mosaic comprises of a series of mini-math games that pop-up within the Mathspring tutor to enhance students' affective states. We ran randomized controlled studies to understand how each of these three interventions may impact students' learning and affect.

Four versions of Monkey's Revenge were analyzed, to evaluate the impact of different degrees of game-like-ness on student learning and affect. One of the major concerns was that educational games might add extra cognitive load and take too much time away from learning; this, special effort was put into the creation of a minimalist interface and a

simple narrative. The intent was to strike an optimal balance so that the game-like tutor would be engaging enough but not overwhelming or distracting. Two randomized controlled studies were run with over 250 students in each study. Results indicated that students liked the game-like version of this math tutor significantly more than the basic tutor with no game-like elements. No conclusive results in learning gains were found, though students using the tutor with game-like elements performed marginally (but not statistically significantly) better than the students using basic tutor. Even though we do not have a proper measure of cognitive load, we use the math performance measure as an evidence that the game-like tutor is at least not penalizing students, not hindering student learning. In general, we did not manage to establish empirical evidence that game-like tutors can deliver superior learning gains than non-game-like tutors. One possibility is that the main reason for this failure is due to short exposure time. Monkey's Revenge should be expanded in the future so that it can be run for multiple sessions. This brings along the issue of resource constraint that I had identified as one of the major constraints in using educational games for the purpose of learning.

The Student Progress Page (SPP) is a component of the "learning dashboard" in Mathspring. The SPP presents information about a student's performance using game-like visualizations. In the study presented in this dissertation, students were randomly assigned to one of the four experimental conditions. The first group of students were taken to the SPP after they reported poor affect; the second group of students were prompted (offered) to go to SPP after they reported poor affect; the third group of students did not have access to the SPP at all; and the fourth group of students had access

to the SPP (through a "My Progress" button) but were not prompted nor forced to use the SPP. There were 209 participants who used the tutor over three sessions. We found that students who used the SPP reported having a significantly higher level of interest than the students who did not use the SPP. When looking at the transition between affective states across problems, for the whole population of students, we noted that having access to the SPP in a variety of ways promoted higher likelihood of excitement in future math problems, as well as decreased likelihood of affective paths leading to boredom.

We created two Mosaic mini-games that were integrated into Mathspring. We ran a study with 186 students, where they were randomly assigned to one of the following three conditions: students were *given* mini-games after they reported low affect, students were *offered* mini-games after they reported low affect, students were *not offered* mini-games at all. Students using mini-games reported significantly higher appraisal of their experience in tutor. We had expected that the students who got to play mini-games will report higher affective state. But we could not find any statistical difference between the affect level reported by the students who used and did not use Mosaic. Again, our intervention was only for one class session and this short duration was not enough to generate substantial impact on the students.

The second part of this dissertation consisted of understanding the mechanisms in which these game-like interventions impacted student's interaction, affect, perception and learning. We used causal modeling to unravel the interdependencies among different variables and constructs. After analyzing the affordances, limitations and pitfalls of

causal modeling as a methodology for data analysis, so that we can use this tool with caution, I created causal models from the data of the studies above, regarding the three game-like interventions. Findings suggest that students' prior (incoming) attitude and preference and personality traits are a major influence of how they interact with the tutoring systems and the interventions themselves, as well as how they perceive the interventions. Students' affect and their engagement behaviors and performance are highly related, creating a chain of cascading effects that suggest that students who appreciate mathematics (and their math ability) more tend to feel more positively, engage more, and make the most out of the software. The casual models in general allowed to see all of the following: a) associations that validated my prior assumptions (student's attitude and preferences prior to intervention affect how they interact and perceive the intervention); b) associations that provided new insights (among the four affective variables, confidence and frustration are more tightly linked with performance and ability whereas interest and excitement are more related to attitude and appreciation for math and the tutor); and c) associations that made me think about possible confounds (female students are solving more problems. But they neither do have higher prior knowledge nor are reporting higher self-discipline. Are they under-reporting their self-discipline or is solving more problems not a reflection of higher self-discipline?). In general, even when causal models did not manage to provide confirmatory causal claims, they provided intuitions about the learning process and guided us for new explorations (student variables can be divided into performance oriented and enjoyment oriented clusters).

Overall, this dissertation was an attempt to understand the mechanics of the student

learning process, while trying to find optimal solutions that generate both enjoyment and learning. I consider this dissertation does not provide solid empirical evidence to claim that the game-like solutions have effectively worked at improving learning and enjoyment. However, this dissertation work has provided sufficient data and models to suggest that these game-like interventions are on the right direction. Besides the fact that the results suggest some positive effects, these game-like interventions created from a very cautious and minimalist standpoint they were well received by the students, from a face validity perspective, providing them purpose to continue engaging with the learning software. If we could gather resources to expand these interventions and studies and observe students for longer periods of time, I consider that measurable learning gains should be achieved, which would manage to capture the benefits of game-like elements to math tutoring systems, and the effects of positive engagement and positive response we have seen in full classes of students.

6.1 Limitations

I started this research with some broad theoretical assumptions: three game-like interventions were identified as possible optimal methods to combine educational games and intelligent tutors. I hypothesized that these interventions would generate both enjoyment and learning gains. In retrospect, these statements seem quite ambitious, as we struggle to get the experimental results to back those assumptions. I will describe the limitations of this research in the following paragraphs.

Limitations regarding theoretical framework

I had hypothesized that a minimalist approach is the optimal solution for creating game-like systems. I had two reasons to make this assumption: development costs and cognitive overload. However, development costs are decreasing (lower than several years ago, when this dissertation work started) and it may no longer be a real constraint in future. Game engines and development tools are increasingly available which reduces the cost of game development. Unity (Creighton, 2010) is an example of a cross-platform game engine that is being used by millions of independent game developers. Similarly, there are new emergent technologies that automate the process of content generation. Procedural content generation (PCG) is the programmatic generation of game content using a random or pseudo-random process that results in an unpredictable range of possible game play spaces. PCG in games helps reduce cost by automating, or aiding in, game content generation (Hendrikx et al, 2013). With the advent of these new technologies and platforms, game development is more accessible and affordable to a wider population, giving rise to further inclusion and innovation in game design and development.

The second assumption that minimalist systems are better than rich environments to lower cognitive overload may be challenged. One possibility is that, as students use the system over time, they will be familiar with the environment and they will not be confused and distracted in subsequent sessions. A second possibility is that, as

educational software systems are becoming more immersive and engaged with the real environment and life-like, students may not need to focus on abstract symbolic associations as much, as their experience is situated; it is possible that transitions and associations between the real and the symbolic may become smoother. Immersive educational games is an emerging field where new research and innovations are happening; rich immersive educational games could be the seat to robust learning.

Limitations regarding intervention design and development

The concern that development cost would be a major limitation in the development of educational games has actually become applicable in this case. It was me, the researcher who designed and developed the interventions. While developing Monkey's Revenge, the conclusion was that it was too short to have a measurable impact on students. But due to resource constraints, it was not possible to extend it. I also realized that I was not able to focus on robust content design enough when producing Monkey's Revenge. I was consumed by design issues such as how to create an interface that was engaging but not distracting, how to create a narrative that was appealing to both genders, and how to integrate learning content within the narrative game-like context. In contrast, I did not face such design constraints while designing the learning dashboard and mini-games, as MathSpring was developed by a different team of people. Still, resource constraints was one of the issues that limited us in creating richer interventions across all interventions.

Limitations regarding study design and empirical analysis

I consider that study design is the major limitation of this research work. I used surveys along with log records as measurement tools. However, when using surveys, there is a trade-off between the number of constructs we can measure and how robustly we can measure them. Our studies were exploratory and we wanted to see interaction of multitude of constructs. This limited us in measuring those variables more robustly. In addition, such subjective measures that rely on students' reporting via surveys are not always reliable. We could have used more objective measures such as eye tracking and emotion sensors, as well as information from secondary sources such as student grades and homework completion rates.

Another major limitation was that we did not conduct enough usability studies and iterative-based design. We come from a research culture based on empirical analysis, with randomized controlled studies as a major research tool. If we had conducted more extensive usability studies, that would not have only helped us enhance our interventions but also given us better insights on how to design experiments, what constructs to focus on and which associations to explore further.

6.2 Future work

As discussed earlier, I believe that study design is the major limitation of this work and the most immediate future work. We need to extend experimental sessions so that students are exposed to the interventions for sufficient lengths of time to generate measurable impact. We also need to find ways to measure learning constructs in a more robust and reliable manner. In the future, I would like to focus on two constructs that I think are instrumental in educational game research: cognitive overload and persistence, as described next.

Measuring cognitive overload

Ferdig (2009) has stated that direct studies of cognitive load effects in game-based learning environments are extremely rare and mostly limited to the role of instructional guidance as an important factor in reducing high-load situations. Brunken et al. (2003) have listed various methods of assessing cognitive load such as self-reported data, objective observations of behavior, physiological conditions, or performance and dual-task measurement. Arroyo et al. (2009) had conducted successful study using emotion sensors tracking physical activities with Mathspring students. While we see the value in those sophisticated sensors, due to logistic concerns, we would like to prioritize non-intrusive methods as much as possible. We would also like to make distinction between the methods we may use for our small scale usability studies and for the bigger scale

randomized controlled studies. Think aloud and audio/video tracking would be suitable during our usability studies. Eye tracking and self reports would be more reasonable choice for larger studies.

Measuring persistence

I have tried to make inferences of students' engagement and effort. We looked at students' log records and associated asking for hints and solving correctly as evidence of effort whereas guessing and not reading was considered evidence of disengagement. We should not only look at momentary effort but should also observe persistence over time. Persistence is predictive of many academic outcomes. Games have an edge to enhance persistence and games that encourage higher persistence can be the path to superior learning. In the learning dashboard, we accumulate the performance over a math topic and visually represent performance as a pepper plant. The pepper plant gives special flowers for the math topics where students have demonstrated exceptional help usage. We especially consider the SHINT-SOF sequence, which we interpret as the student using help to solve a problem and then learning from that experience to solve the next problem correctly. Though this behavior can be one reflection of persistence, persistence is a more complex more robust construct.

Dicerbo (2014, 2016) has explored how we can measure persistence in educational games. She has looked at using a combination of data mining and factor analytic techniques. Future work would need to further the research on measuring persistence and devising a technique that works for our tutoring systems.

Games as assessment tool

When we analyze and compare the three game-like interventions that we devised, we find unique opportunities and limitations. Monkey's Revenge has the potential to generate learning gains but is resource intensive to create and scale. It is also challenging to design, as we need to maintain the delicate balance that prevents students from becoming confused, distracted, or overwhelmed. Alternatively, Mosaic mini-games are easy to design and integrate but are too simplistic and do not significantly contribute to the learning process. The Learning Dashboard has the potential to enhance both students' affect and learning, but requires that students be able to access and use it, understand and process its contents, and reflect on their progress in order to benefit.

In our pursuit of practical solutions to combining games and tutors, we are interested in using games as assessment tools. Game-based assessment is an emergent field (Shute, 2011; Kim & Shute, 2015, Halverson et al., 2015). DiCerbo & Behrens (2012) state that the promise of game-based assessment is being able to use data from in-game actions to make inferences about players' knowledge, skills and attributes, allowing us to make use of the ocean of data produced by daily interactions with software. Researchers such as Shute and DeCerbo view games as a platform of evidence-centered design (Kim et. al, 2015; Dicerbo, 2016). Our interest in games as assessment tools is primarily based on the fact that games can be a great platform for learners to play with their knowledge and skills. We envision an arrangement where intelligent tutoring systems teach students

knowledge and skills that they then apply to different game scenarios. We would be assessing the students as they play with their knowledge and use that assessment to give individualized, customized tutoring. For example, students would first learn about coordinate geometry in tutors such as Mathspring and ASSISTment; afterwards, they would use those skills to solve puzzles and build structures within a game. Since students enter the game scenarios with prior knowledge of the content, cognitive load is not as serious a concern when designing such games. This allows us to focus on creating rich learning experiences within an educational game while intelligent tutoring systems take care of providing robust learning.

Our dissertation is based on the belief that learning can and should be fun, but not always. Making learning both fun and effective is desirable but poses a complex challenge. Resource limitations and cognitive load are major constraints. We view our dissertation as an effort to find optimal solutions within these constraints. But as the education and technology community progresses, new tools and techniques are developed and we learn more about the nature of learning itself. With this progress, the landscape of constraints changes and new opportunities arise.

REFERENCES

O'Neil, H., Wainess, R., Baker, E. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal*, 16(4), 455-474.

Aleven, V. Myers, E. Easterday M., Ogan A. (2010). Toward a framework for the analysis and design of educational games, *Third IEEE International Conference on Digital Game and Intelligent Toy Enhanced Learning*, 69-76.

Gee, J. P. (2003). *What Video Games Have to Teach Us About Learning and Literacy*. New York: Palgrave/Macmillan.

Hays, R. T. (2005). The effectiveness of instructional games: A literature review and discussion. *Naval Air Warfare Center Training Systems Division*, Orlando, FL.

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, S. C., Jamie L. Estock, Orvis, K. L., Conkey, C. (2008). Relationships between game attributes and learning outcomes: Review and research proposals. *Simulation & Gaming,* 40 (2), 217-266.

Garris, R., Ahlers, R. & Driskell, J. E. Games. (2002). Motivation and learning: A research and practice model. *Simulation & Gaming*, 33(4), 441-467.

Malone, T. W., & Lepper, M. R. (1987). Making learning fun: A taxonomy of intrinsic motivations for learning. *Aptitude, learning and instruction*, 3, 223-253.

Shaffer, D. W., Resnick, M. (1999). "Thick" Authenticity: New Media and Authentic Learning. *Journal of Interactive Learning Research*, 10(2), 195-215.

Sweller, J. (1994). Cognitive load theory, learning difficulty and instructional design. *Learning and Instruction*, 4, 295–312.

Ferdig, R. E. (2009). *Handbook of research on effective electronic gaming in education*. Hershey, PA: Information Science Reference.

Brunken, R., Plass, J. L., & Leutner, D. (2003). Direct Measurement of Cognitive Load in Multimedia Learning. *Educational Psychologist*, 38(1), 53-61

Pellegrini, A. D., Dupuis, D., Smith, P. K. (2007). Play in evolution and development. *Developmental Review* 27. 261–276

Hegarty M., Mayer, R. E., & Monk C. A. (1995). Comprehension of Arithmetic Word Problems: A Comparison of Successful and Unsuccessful Problem Solvers. *Journal of Educational Psychology*, 87, 18-32.

Goldstone, R.L., Son, J.Y. (2005). The Transfer of Scientific Principles Using Concrete and Idealized Simulations. *The journal of the learning sciences*, 14(1), 69–110.

Klopfer, E., Osterweil. S., & Salen, K. (2009). Moving learning games forward, obstacles, opportunities sand openness, an educational arcade paper.

Szulborski, D. (2005). This is not a game: a guide to alternate reality gaming. DPRGRM.

Juul, J. (2009). *A Casual Revolution: Reinventing Video Games and Their Players*, The MIT press.

Glymour,C., Scheines,R. (2004). Causal modeling with the TETRAD program. *Synthese*. 37-64

Wittgenstein, L. (1953). *Philosophical Investigations*. Prentice Hall.

Boyer, K. E., Phillips, R., Wallis, M., Vouk, M., & Lester, J. (2008). Balancing Cognitive and Motivational Scaffolding in Tutorial Dialogue. *Intelligent Tutoring Systems*, 5091, 239-249.

Koedinger, K. R., & Corbett, A. (2006). Cognitive tutors: Technology bringing learning science to the classroom. *The cambridge handbook of the learning sciences,* 61-78.

Jackson, G.T., & McNamara, D.S. (2011). Motivational impacts of a game-based intelligent tutoring system. In R. C. Murray & P. M. McCarthy (Eds.), *Proceedings of the 24th International Florida Artificial Intelligence Research Society (FLAIRS) Conference,* 519-524.

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2011). Using tutors to improve educational games. In G. Biswas, S. Bull, J. Kay, & A. Mitrovic (Eds.), *Artificial intelligence in education: Lecture notes in artificial intelligence* 6738, 63-72.

Beck, J. E., Chang, K., Mostow, J., Corbett, A.T. (2008). Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Intelligent Tutoring Systems*: 383-394.

Mayer, R. (2009). *Multimedia Learning, Second Edition*. NY: Cambridge University Press

Clark, R. E. (2011). *Games for Instruction?* Presentation at the American Educational Research Association, New Orleans, LA.

Crocker, L. & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Wadsworth Publishing.

Ericsson, K. Anders; Krampe, Ralf T.; Tesch-Römer, Clemens. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*. Vol 100(3)

Salen, K., Zimmerman (2003). *Rules of Play: Game design fundamentals*. The MIT press

Prensky, M. (2007). *Digital game based learning*. Paragon House

Habgood, M.P.H. (2005). Zombie Division: Intrinsic Integration in Digital Learning Games. *Human Centred Technology Workshop*.

Fullerton, T., Swain, C., & Hoffman, S. (2008). *Game design workshop: A playcentric approach to creating innovative games*. Amsterdam: Elsevier Morgan Kaufmann.

Squire, K. (2004). Replaying history: Learning world history through playing Civilization III. *ProQuest Dissertations*, Indiana University.

Garris, R., Ahlers, R., Driskell, J.E. (2002). Games, Motivation, and Learning: *A Research and Practice Model*. *Simulation & Gaming* 33(4), 441–467

Virvou, M., Katsionis, G., Manos, K. (2005). Combining Software Games with Education: Evaluation of its Educational Effectiveness. *Education Technology & Society* v8 n2 p54-65

Defreitas, S., Oliver, M. (2006). How can exploratory learning with games and simulations within the curriculum be most effectively evaluated? *Computers & Education* 46, 249–264

Siang, A.C., Rao, R.K. (2003).Theories of learning: a computer game perspective. *Multimedia Software Engineering*, pp. 239–245

Thomas, P., Labat, J., Muratet, M., & Yessad, A. (2012). How to Evaluate Competencies in Game-Based Learning Systems Automatically? *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 168-173.

Amory, A., Naicker, K., Vincent, J., & Adams, C. (1999). The use of computer games as an educational tool: Identification of appropriate game types and game elements. *British Journal of Educational Technology*, 30(4), 311-321.

O, Neil, H. F., Wainess, R., & Baker, E. L. (2005). Classification of learning outcomes: Evidence from the computer games literature. The Curriculum Journal, 16(5), 455-474.

Ziemek, T. R. (2006). Two-D or not Two-D. *Proceedings of the 2006 Symposium on Interactive 3D Graphics and Games* - SI3D '06.

Aliya, S. K. (2002). The role of computer games in the development of theoretical analysis, flexibility and reflective thinking in children: A longitudinal study. *International Journal of Psychophysiology*, 45, 149.

Conati, C., & Zhou, X. (2002). Modeling Students' Emotions from Cognitive Appraisal in Educational Games. *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 944-954. doi:10.1007/3-540-47987-2_94

Malone, T. & Lepper (1987). Making Learning Fun: A Taxonomy of Intrinsic Motivations for Learning. In Snow, R. & Farr, M. J. (Ed), Aptitude, Learning, and Instruction Volume 3: Conative and Affective Process Analyses. Hillsdale, NJ

Barab, S. A., Gresalfi, M., & Ingram-Goble, A. (2010). Transformational Play: Using Games to Position Person, Content, and Context. *Educational Researcher*, 39(7), 525-536.

Bayraktar, S. (2001). A Meta-analysis of the Effectiveness of Computer-Assisted Instruction in Science Education. *Journal of Research on Technology in Education*, 34(2), 173-188.

Connolly, T. M., Boyle, E. A., Macarthur, E., Hainey, T., & Boyle, J. M. (2012). A systematic literature review of empirical evidence on computer games and serious games. Computers & Education, 59(2), 661-686.

Habgood, M. P., & Ainsworth, S. E. (2011). Motivating Children to Learn Effectively: Exploring the Value of Intrinsic Integration in Educational Games. *Journal of the Learning Sciences*, 20(2), 169-206.

Jonassen, D. H., & Hernandez-Serrano, J. (2002). Case-based reasoning and instructional design: Using stories to support problem solving. *Educational Technology Research and Development*, 50(2), 65-77.

Klopfer, E., Osterweil, S., & Salen. K. (2009). Moving learning games forward: Obstacles, opportunities, & openness. *MIT, The Education Arcade*

Marks, H. M. (2000). Student Engagement in Instructional Activity: Patterns in the Elementary, Middle, and High School Years. *American Educational Research Journal*, 37(1), 153-184.

Papastergiou, M. (2009). Digital Game-Based Learning in high school Computer Science education: Impact on educational effectiveness and student motivation. *Computers & Education*, 52(1), 1-12.

Sitzmann, T. (2011). A Meta-Analytic Examination Of The Instructional Effectiveness Of Computer-Based Simulation Games. *Personnel Psychology*, 64(2), 489-528.

Jackson, G.T., Boonthum, C., McNamara, D.S. (2009). iSTART-ME: Situating extended learning within a game-based environment. *Proceedings of the workshop on intelligent educational games with International Conference on AI and education*. AIED'2009, Brighton

Dondlinger, M. J. (2007). Educational Video Game Design: A Review of the Literature. *Journal of Applied Educational Technology* .Volume 4, Number 1 S

Lai, B., Slota, S. & Medin, D. (2012). Our Princess Is in Another Castle. A Review of Trends in Serious Gaming for Education. *Review of Educational Research*, 82(296), 295-299.

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, 38(1), 43-52.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2), 167-207.

Fullerton, T., Swain, C., & Hoffman, S. (2004). *Game design workshop*. San Francisco, CA: CMP books.

Habgood, J., & Overmars, M. H. (2006). *The game maker's apprentice: Game development for beginners*. Berkeley, CA: Apress.

Schell, J. (2008). *The art of game design: A book of lenses*. Morgan Kaufmann.

Csikszentmihalyi, M. (1990). *Flow: The psychology of optimal experience*. New York: Harper & Row

Everett, A. & Watkins, S. C. (2007). The Power of Play: The Portrayal and Performance of Race in Video Games. *The Ecology of Games: Connecting Youth, Games, and Learning*. Cambridge, MA: The MIT Press,

Deterding, S. (2011). From Game Design Elements to Gamefulness: Defining "Gamification". *Proceedings of the 15th International Academic MindTrek Conference*

Gee, J. P. (2007). *Good video games and good learning: Collected essays on video games, learning and literacy*. New York: Peter Lang

O'Neil, H., Wainess, R. and Baker, E. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal*, 16(4), 455-474

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2011). Using tutors to improve educational games. *Lecture notes in artificial intelligence* 6738. (pp. 63-72). Berlin: Springer

Clark, R. E. (2011). Games for Instruction? *Presentation at the American Educational Research Association*, New Orleans, LA

Wilson, K. A., Bedwell, W. L., Lazzara, E. H., Salas, E., Burke, S. C., Jamie L. Estock, Orvis, K. L., Conkey, C. (2008). Relationships between game attributes and learning outcomes: *Review and research proposals: Simulation & Gaming* 40 (2) p. 217-266

Rai, D., & Beck, J. E. (2012). Math Learning Environment with Game-Like Elements: An Incremental Approach for Enhancing Student Engagement and Learning Effectiveness. *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 90-100.

O'Neil, H., Wainess, R. and Baker, E. (2005). Classification of learning outcomes: Evidence from the computer games literature. *The Curriculum Journal*, 16(4) 455-474

Easterday, M. W., Aleven, V., Scheines, R., & Carver, S. M. (2011): Using tutors to improve educational games. *Lecture notes in artificial intelligence* 6738. (pp. 63-72). Berlin: Springer.

Rai, D., & Beck, J. E. (2011). Causal Modeling of User Data from a Math Learning Environment with Game-Like Elements. *Lecture Notes in Computer Science Artificial Intelligence in Education*, 528-530.

Rai D., & Beck, J. E. (2011). Exploring user data from a game-like math tutor: a case study in causal modeling. *Proceesings of 4th International Conference on Educational Data Mining*. Eindhoven, Netherlands

Efklides, A. (2006). Metacognition and affect: What can metacognitive experiences tell us about the learning process? *Educational Research Review*, 1(1), 3-14.

B. du Boulay, K. Avramides, R. Luckin, E. Martinesz-Miron, G. Rebolledo Mendez, A. (2010). Towards systems that care: a conceptual framework based on motivation, metacognition and affect. *International Journal of Artificial Intelligence in Education*, 20 (3)

Plass, J.L. Homer, B. D., Kinzer, C. , Frye, J. M., Perlin, K. (2011). Learning Mechanics and Assessment Mechanics for Games for Learning. White Paper # 01/2011 Version 0.1

Baker, R. S., D'mello, S. K., Rodrigo, M. T., & Graesser, A. C. (2010). Better to be frustrated than bored: The incidence, persistence, and impact of learners' cognitive–affective states during interactions with three different computer-based learning environments. *International Journal of Human-Computer Studies*, 68(4), 223-241

D'Mello, S. K., Person, N. K., & Lehman, B. (2009). Antecedent-Consequent Relationships and Cyclical Patterns between Affective States and Problem Solving Outcomes. Proceedings of *Artificial Intelligence in Education*. 57-64

Metcalfe, J., & Shimamura, A. P. (1994). *Metacognition: knowing about knowing*. Cambridge, MA: MIT Press.

Mitmansgruber, H., Beck, T. N., Höfer, S., & Schüßler, G. (2009). When you don't like what you feel: Experiential avoidance, mindfulness and meta-emotion in emotion regulation. *Personality and Individual Differences*, 46(4), 448-453

Clark, D. B., Tanner-Smith, E. E., & Killingsworth, S. S. (2015). Digital Games, Design, and Learning: A Systematic Review and Meta-Analysis. *Review of Educational Research*, 86(1), 79-122.

Wouters, P., Nimwegen, C. V., Oostendorp, H. V., & Spek, E. D. (2013). A meta-analysis of the cognitive and motivational effects of serious games. *Journal of Educational Psychology*, 105(2), 249-265

Nicholson, S. (2012). A User-Centered Theoretical Framework for Meaningful Gamification. *Games+Learning+Society* 8.0, Madison, WI

Millstone, J. (2012). Teachers attitudes about digital games in learning. *Joan Ganz Cooney center at Sesame Workshop*.

Miller, C. H. (2008). *Digital storytelling: A creator's guide to interactive entertainment*. *Amsterdam*: Focal Press/Elsevier.

Halverson, R., Berland, M., & Owen, V. W. (2015). Game-Based Assessment. *SAGE Encyclopedia of Educational Technology*

Robinson, K., & Aronica, L. (2015). *Creative schools: The grassroots revolution that's transforming education*.

O'Rourke, E., Haimovitz, K., Ballwebber, C., Dweck, C. K. , & Popovic, Z. (2014). Brain Points: A Growth Mindset Incentive Structure Boosts Persistence in an Educational Game. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. Pages 3339-3348

Duckworth, A. L., & Seligman, M. E. (2005). Self-Discipline Outdoes IQ in Predicting Academic Performance of Adolescents. *Psychological Science*, 16(12), 939-944.

Duckworth, A. L., & Seligman, M. E. (2006). Self-discipline gives girls the edge: Gender in self-discipline, grades, and achievement test scores. *Journal of Educational Psychology*, 98(1), 198-208

Tangney, J. P., Baumeister, R. F., & Boone, A. L. (2004). High Self-Control Predicts Good Adjustment, Less Pathology, Better Grades, and Interpersonal Success. J *Personality Journal of Personality*, 72(2), 271-324.

Dweck, C. S. (1986). Motivational processes affecting learning. *American Psychologist*, 41(10), 1040-1048

Dweck, C.S., (1999) *Self-Theories: Their role in motivation, personality and development*. Philadelphia: Psychology Press

Wigfield, A., & Eccles, J. S. (2000). Expectancy–Value Theory of Achievement Motivation. *Contemporary Educational Psychology*, Volume 25, Issue 1, Pages 68-81

Pekrun, R., Frenzel, A., Goetz, T., Perry, R. (2007). The Control-Value Theory of Achievement Emotions: An Integrative Approach to Emotions in Education. Chapter 2. *Emotion in education*. Academic Press: San Diego, CA.

Pekrun, R., Goetz, T., Frenzel, A.C. (2005). Academic Emotions Questionnaire—Mathematics (AEQ-M): *User's manual*. University of Munich, Department of Psychology.

Hendrikx, M., Meijer, S.,Velden, J. V. D., & Iosup, A. (2013). Procedural content generation for games: A survey. *ACM Transactions on Multimedia Computing, Communications, and Applications*. Appl. 9, 1, Article 1

Creighton, R. H. (2010). *Unity 3D game development by example: Beginner's guide: A seat-of-your-pants manual for building fun, groovy little games quickly*. Birmingham, UK

Arroyo, I., Cooper, D. G., Burleson, W., Woolf, B. P., Muldner, K., & Christopherson, R. (2009). Emotion Sensors Go To School. *In Proceedings of the 2009 conference on Artificial Intelligence in Education: Building Learning Systems that Care*. IOS Press, Amsterdam, The Netherlands, The Netherlands, 17-24.

Arroyo, I., & Woolf, B. P. (2005). Inferring learning and attitudes from a Bayesian Network of log file data, *12th international conference on artificial intelligence in education*

Corbett, A. T., & Anderson, J. R. (1995). Knowledge tracing: Modeling the acquisition of procedural knowledge. *User Modelling and User-Adapted Interaction*, 4(4), 253-278

Beck, J. E., Chang, K., Mostow, J., & Corbett, A. (2008). Does Help Help? Introducing the Bayesian Evaluation and Assessment Methodology. *Intelligent Tutoring Systems Lecture Notes in Computer Science*, 383-394

Baker, R., Walonoski, J., Heffernan, N., Roll, I., Corbett, A., & Koedinger, K. (2008). Why students engage in" gaming the system" behavior in interactive learning environments. *Journal of Interactive Learning Research*, 19(2), 185.

De Vicente, A., & Pain, H. (2002, June). Informing the detection of the students' motivational state: an empirical study. In Intelligent tutoring systems (pp. 933-943). Springer Berlin Heidelberg.

Beck, J. E., & Chang, K. M. (2007). Identifiability: A fundamental problem of student modeling. *In User Modeling* (pp. 137-146). Springer Berlin Heidelberg.

Beck, J. E., & Mostow, J. (2008). How who should practice: Using learning decomposition to evaluate the efficacy of different types of practice for different types of students. *9th International Conference on Intelligent Tutoring Systems*, 353-362

Martin, B., Mitrovic, A., Koedinger, K.R. and Mathan, S. (2010) *Evaluating and Improving Adaptive Educational Systems with Learning Curves*, Journal of User Modeling and User Adapted Interaction, Vol 21 number 3, Springer, pp. 249-283

San Pedro, M.O.C., Baker, R., Rodrigo, M.M. (2011) Detecting Carelessness through Contextual Estimation of Slip Probabilities among Students Using an Intelligent Tutor for Mathematics. *Proceedings of 15th International Conference on Artificial Intelligence in Education*, 304-31

Muldner, K., Wixon, M., Rai, D., Burleson, W., Woolf, B., & Arroyo, I. (2015). Exploring the Impact of a Learning Dashboard on Student Affect. *Lecture Notes in Computer Science Artificial Intelligence in Education*, 307-317

Young, M. F., Slota, S., Cutter, A. B., Jalette, G., Mullin, G., Lai, B.,Yukhymenko, M. (2012). Our Princess Is in Another Castle: A Review of Trends in Serious Gaming for Education. *Review of Educational Research*, 82(1), 61-89.

Egenfeldt-Nielsen, S., Smith, J. H., & Tosca, S. P. (2008). *Understanding video games: The essential introduction*. New York: Routledge.

Squire, K. D. (2013). Video Game-Based Learning: An Emerging Paradigm for Instruction. *Performance Improvement Quarterly*. 26(1)

Hunicke, R., LeBlanc, M., & Zubek, R. (2004). MDA: A formal approach to game design and game research. *In Proceedings of the AAAI Workshop on Challenges in Game AI* (Vol. 4, p. 1)

Rieber, L. P. (1996). Seriously considering play: Designing interactive learning environments based on the blending of microworlds, simulations, and games. *Educational Technology Research & Development*, 44(2), 43-58

Arroyo, I., Royer, J.M., Woolf, B.P. (2011). Using an Intelligent Tutor and Math Fluency Training to Improve Math Performance. *International Journal of Artificial Intelligence in Education*. 21 (2), pp. 135-152.

Wixon, M., Arroyo, I., Muldner, K., Burleson, W., Lozano, C., Woolf, B. (2014). The opportunities and limitations of scaling up sensor-free affect detection. *Proceedings of Educational Data Mining*, pp. 145–152

Baker, R.S.J.d., Gowda, S.M., Wixon, M., Kalka, J., Wagner, A.Z., Salvi, A., Aleven, V.,

Kusbit, G., Ocumpaugh, J., Rossi, L. (2012). Sensor-free automated detection of affect in a cognitive tutor for algebra. *Proceedings of Educational Data Mining*, pp. 126–133

Arroyo, I., Ferguson, K., et al. (2007). Repairing disengagement with non-invasive intervention. *Proceedings of Artificial Intelligence in Education*, pp. 195–202

Ryan, R., Deci, E. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *American psychologist* 55(1), 68–78

Shute, V. J. (2011). Stealth assessment in computer-based games to support learning. In S. Tobias & J. D. Fletcher (Eds.), *Computer games and instruction* (pp. 503-524). Charlotte, NC: Information Age Publishers.

Shute, V. J., Ventura, M., & Ke, F. (2015). The power of play: The effects of Portal 2 and Lumosity on cognitive and noncognitive skills. *Computers & Education*, 80, 58-67.

Kim, Y. J. & Shute, V. J. (2015). Opportunities and challenges in assessing and supporting creativity in video games. In G. Green & J. Kaufman (Eds.), *Research frontiers in creativity* (pp. 100-121). San Diego, CA: Academic Press.

Shute, V. J., Moore, G. R., & Wang, L. (2015). Measuring problem solving skills in Plants vs. Zombies 2. *Proceedings of the 8th International Conference on Educational Data Mining (EDM 2015)*. Madrid, Spain.

Kim, Y. J., & Shute, V. J. (2015). The interplay of game elements with psychometric qualities, learning, and enjoyment in gamebased assessment. *Computers & Education, 87*, 340-356.

Kim, Y. J., Almond, R. G., & Shute, V. J. (2015). Applying Evidence-Centered Design for the Development of Game-Based Assessments in Physics Playground. *International Journal of Testing*, 16(2), 142-163

DiCerbo, K. E. (2014). Game-Based Assessment of Persistence. *Educational Technology & Society*, 17 (1), 17–28

Dicerbo, K. E. (2016). Assessment of Task Persistence. *Handbook of Research on Technology Tools for Real-World Skill Development*, 778-804.